1

Protein Analysis by Shotgun Proteomics[1]

Yu Gao[1] and John R. Yates III[2]

[1] *College of Pharmacy, University of Illinois at Chicago, Chicago, IL, USA*
[2] *Department of Molecular Medicine, Scripps Research, La Jolla, CA, USA*

1.1 Introduction

1.1.1 Terminology

In mass-spectrometry-based protein analysis, there are two major strategies, the top-down method and the bottom-up method [1, 2]. The terms "top" and "bottom" refer to the complexity of the analyte, namely the more complex "protein" and less complex "peptide." In top-down protein analysis, the intact protein is directly analyzed by mass spectrometer. Mass information and fragment ions are generated from the intact protein ions and are then used for direct protein identification and characterization. In comparison, bottom-up method starts with digesting the protein into peptides by either chemical or enzymatic digestion. The peptide product is then analyzed by a tandem mass spectrometer, and the peptide molecular weight and fragmentation information is matched back to the original protein or protein mixture. When a mixture of proteins is analyzed by a bottom-up method, it is also called shotgun proteomics, owing to the similarity to shotgun genomic sequencing.

1.1.2 Power of Shotgun Proteomics

In a typical shotgun proteomics experiment performed on a modern instrument, one should expect to identify anywhere from 1000 to 10 000 proteins

1 This chapter will discuss some of the most commonly used techniques and strategies in shotgun/bottom-up proteomics and their related applications.

Mass Spectrometry-Based Chemical Proteomics, First Edition.
Edited by W. Andy Tao and Ying Zhang.

from a mammalian cell lysate [3, 4]. In comparison, a typical top-down experiment is able to identify hundreds or a thousand proteins with a similar sample, but it require extensive fractionation to simplify the protein mixtures entering the mass spectrometer [5–7]. In top-down proteomics, intact protein is highly complex in terms of molecular weight, charge state, hydrophobicity, molecular structure (shape), and so on, therefore, it is hard to find optimal conditions for ideal separation, fragmentation, and detection of all proteins presented in the sample.

1.1.3 Advantage of Shotgun Proteomics

Comparing to intact protein, a peptide is a much more unified class of analyte, with a narrow range of molecular weight and charge state. Because most digested peptides are denatured, peptides also have a more unified shape [8, 9]. Therefore, starting with peptides instead of the intact protein presents advantages over the top-down method, including more robust liquid-chromatography (LC) separation for peptides, more uniform electrospray ionization, more complete fragmentation in tandem mass spectrometry (MS/MS), and easier interpretation of the simplified fragmentation patterns. Due to these advantages, bottom-up/shotgun proteomics method has become the easier strategy for protein analysis over the past two decades. However, these advantages also come with some nontrivial challenges in sample preparation, peptide separation, data acquisition, and informatics [10–12]. This chapter will discuss typical procedures of shotgun proteomics experiment and some recent advances regarding existing challenges.

1.2 Overview of Shotgun Proteomics

A typical shotgun proteomics experiment consists of three main steps: (i) sample preparation, (ii) mass spectrometry data acquisition, and (iii) data processing. The sample preparation step transforms the biological sample to a peptide mixture. The data acquisition step obtains MS/MS data from the peptide mixture. The final data processing step performs statistical and mathematical analyses to elucidate the identity and quantity of peptide and protein (Figure 1.1).

In the sample preparation step, a protein mixture is first obtained by separating protein and nonprotein contents from a biological sample such as cell lysate or serum. The separated protein mixture is then chemically modified (reduced and then alkylated) to break all Cys–Cys disulfide bonds in order to linearize protein. Protease, for example, trypsin, is then added to the modified protein mixture to digest protein into peptides. After digestion, the peptide mixture is

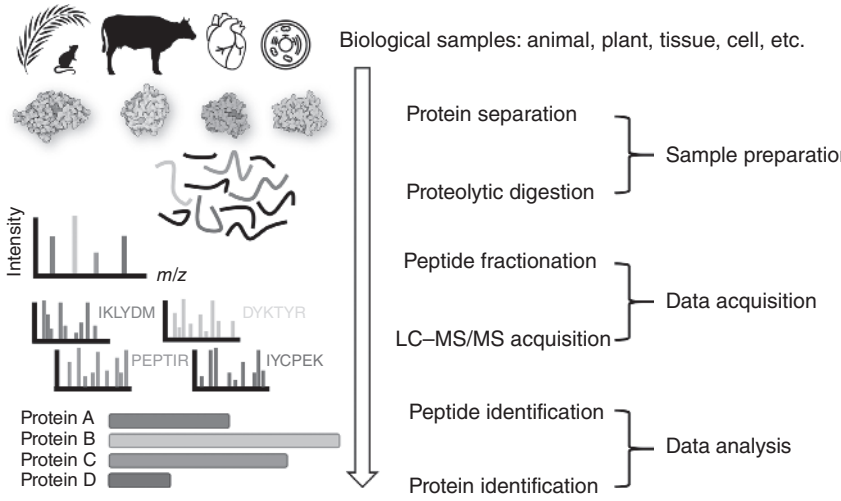


Figure 1.1 Typical workflow of a bottom-up proteomics experiment. Proteins are first separated from biological samples, then digested into peptides. An LC–MS/MS system is typically used to fractionate and fragment peptides. The acquired mass spectra are then matched to existing peptide sequence using a database search algorithm and then inferred back to proteins.

often loaded onto a C18 column and then washed to remove nonpeptide contents (salts, buffers, chaotropes, etc.).

Once the sample is digested and cleaned, an LC–MS system is used to fractionate peptides to increase the amount of MS/MS data obtained from the peptide mixture. As digested protein mixtures can create very complicated and complex peptide mixtures, to better resolve peptide mixture, various types of separation columns have been used either alone or in combinations, including reversed phase (RP), strong-cation exchange (SCX), size exclusion (SEC), hydrophilic interaction liquid chromatography (HILIC), and affinity purification. In general, the final separation method prior to introduction of peptides into electrospray ionization is reversed-phase as this method removes salts and other small-molecule interferants. The separated peptides are then ionized and injected into the mass spectrometer for analysis. In this step, the peptide mixture is first temporally separated by LC, then spatially separated by the electrical fields. This separation cascade provides enough resolution to separate hundreds of thousands of peptide species within hours.

In the final data processing step, the data obtained for each detected peptide species, including MS (whole mass) and tandem MS/MS (fragmentation masses) data, is analyzed by algorithms that search sequence databases to match spectra to the original protein sequence. If desired, the data can also be further analyzed for quantitation by either "labeled" or "label-free" methods.

1.3 Sample Preparation

1.3.1 Protein Separation

1.3.1.1 Overview

To analyze proteins from a complex biological sample, protein often needs to be separated from interfering small molecules and nucleotides. This is often done by nonspecific protein extraction such as protein precipitation or centrifugation [13–16]. Some of the most commonly used reagents/solvents/systems for protein precipitation include trichloroacetic acid (TCA)/water, chloroform/methanol, acetone, phenol/ammonium acetate/methanol, and so on. These methods can effectively separate the protein from other molecules such as salts, lipids, detergents (often introduced during lysis), DNA/RNA, and even the aqueous buffer. Therefore, the proteins are purified and concentrated for further processing. Centrifugation method such as sucrose gradient is also very useful for this purpose, but due to its lower throughput and efficiency, it is often used in combination of protein precipitation method to isolate proteins from specific cell organelles.

1.3.1.2 2D-Gel Approach

Two-dimensional polyacrylamide gel electrophoresis (2D PAGE) is a robust, orthogonal approach, popularly applied for the simultaneous separation and fractionation of complex protein mixtures that have been recovered from biological samples for proteomic analysis [17, 18]. The method allows separation of several thousand proteins, on the basis of their molecular mass and the isoelectric point in a single gel. It used to be one of the most widely used methods for protein separation, and it has been used in studies related to proteins and protein complexes [19]. Once the separation is achieved via 2D PAGE, a protein spot or band can be visualized and then extracted. Coomassie brilliant blue or silver staining is commonly used for protein visualization. Coomassie brilliant blue is generally preferred over silver staining as it is a reversible stain and compatible with MS analysis [20]. Despite greater sensitivity, due to its limited compatibility and nonlinearity with the signal, silver staining may give disappointing results [21]. After protein visualization, the gel spots are digested with trypsin and identified by either protein fingerprinting using matrix-assisted laser desorption/ionization–mass spectrometry (MALDI–MS) or via peptide sequencing using LC–MS. Although 2D electrophoresis is associated with the start of proteomics and is still widely used for various purposes, large-scale proteomics is now associated with advanced separation and mass spectrometry technologies for protein identification. Technologies like LC–MS/MS, which offer superior separations, have taken over from 2D gel-based methods.

1.3.1.3 Separation of Membrane Protein

Membrane proteins are integral parts of the nucleus and cell membranes. They are permanently anchored to the outer surface of the membrane or embedded into the lipid bilayer and are actively involved in many crucial cell functions, including transportation of ions and molecules, cellular communication via active cell signaling and cell interactions [22]. It is estimated that 20–30% of genes in most genomes are related to membrane proteins and hence are responsible for pathological induction of many deadly diseases, such as cancer, neurodegenerative disorders, diabetes, and so on [23]. As a result, these proteins have become the major targets of modern drugs. However, separating these proteins is a very challenging task, due to their high hydrophobicity and low abundance [24, 25]. Development of new technologies to identify and characterize membrane proteins was an important issue in proteomics.

Traditionally, membrane proteins are separated by sucrose gradient and similar methodologies that isolate subcellular membrane proteins directly from cell and tissues [26]. For tissue samples, Smolders et al. [27] have reported the use of biotinylation tagging on small tissue samples to overcome various problems that are commonly being encountered, such as poor extraction efficiency, weaker enrichment, sample contamination, and sample exhaustion. For cell samples, by using carefully optimized condition and specific neutral detergent (CA-630), Pankow et al. showed that it is possible to preserve membrane proteins in their native state during cell lysis [28, 29]. Membrane proteins together with their interactors could be co-immunoprecipitated in such condition and subsequently characterized by LC–MS.

1.3.1.4 Subcellular Fractionation

Separating membrane proteins with high specificity is not easy, more generally, probing the proteome in any subcellular location is a challenging task [30, 31]. This often involves the separation of specific organelle or compartments, such as nucleus and mitochondria, by biochemical methods or centrifugation from the whole cell extracts [32]. The specificity and the efficiency of the separation are often the major limiting factors of subcellular proteome measurements. Organelles with lipid bilayer membranes, such as nuclei and mitochondria, can be efficiently separated by either centrifugation or flow cytometry with preserved integrity [33]. After separation, further analysis could be carried out on separated intact organelles, such as mitochondria from the murine heart and skeletal muscles, respectively, to study posttranslational modifications (PTMs) like phosphorylation and carboxylation, and so on [34–37]. Other cellular apparatus such as endoplasmic reticulum, Golgi, and lysosome have been studied much less frequently due to difficulties in separation [38–42]. Separation of these organelles is often complicated and lacks specificity and

efficiency [30, 31]. Moreover, the protein contents of these organelles are also believed to be more transitory and dynamic [38, 41]. Overall, from the data that is being accumulated so far from all the subcellular proteomic analysis, it has been quite evident that the purity of subcellular organelle is of utmost importance to produce quality data.

1.3.1.5 Protein Enrichment

In many cases, a subset of proteins, such as low-abundance proteins or proteins with PTM, needs to be specifically enriched from the cell lysate to achieve better quantification and identification. This is largely due to the mismatch of a very high dynamic range of protein expression and a limited dynamic range provided by the mass spectrometer. Therefore, protein enrichment techniques such as affinity purification, by either antibody or other affinity methods, are often performed during sample preparation for the detection of low-abundance proteins from complex samples.

1.3.1.6 Phosphoprotein

Phosphorylated proteins are often of low abundance, but they play a vital role in cell signaling [43]. Enrichment by affinity purification, using either immobilized metal affinity chromatography (IMAC) or metal oxide affinity chromatography (MOAC), can effectively improve phosphoprotein identification by several orders of magnitude [44, 45]. The method is based upon the high affinity of phosphate groups to cations such as Zn^{2+}, Fe^{3+}, Ti^{4+}, and so on. The approach has been successfully used for both off-line and on-line separation of phosphoproteins and phosphopeptides. Metal oxide such as titanium dioxide (TiO_2) can be used as a very robust chelating agent and thus provide specific phosphopeptide enrichment [46]. To overcome nonspecific chelation, esterification of acidic residues before IMAC enrichment may significantly improve the specificity of the enrichment [46, 47]. Immunoprecipitation is another technique that has been used for enrichment of phosphorylated proteins [48, 49]. By using highly specific antiphosphotyrosine antibodies, phosphoproteins containing phosphotyrosine can be enriched with high specificity. However, the high specificity of antibodies is also associated with enrichment bias toward a certain type of phosphorylation or certain peptide sequences [50, 51]. Therefore, two or more antibodies can be used together to target different phosphorylation sites.

1.3.1.7 Glycoprotein

Another very common PTM is mono- or oligosaccharide glycosylation of serine, threonine, and asparagine residues of proteins [52]. Similar to phosphorylation, glycoproteins benefit from enrichment before mass spectrometry analysis [53, 54]. Some of the most common glycoprotein enrichment techniques include HILIC, ion exchange chromatography, lectin-based affinity

purification, antibody-based affinity purification, and the formation of covalent interaction such as hydrazide and boronic acid chemistry [55–60]. Among all existing techniques, ZIC-HILIC (zwitterionic hydrophilic interaction liquid chromatography) based glycopeptide enrichment has been shown to be highly efficient and specific in the separation of *N*-glycopeptides when compared with several other techniques [61, 62]. However, due to the diverse nature of complex glycans, different types of techniques are often quite complimentary to each other. Combination of multiple techniques could significantly improve glycoprotein identification and quantitation results.

1.3.1.8 AP–MS and Interactome

In the past decade, affinity purification under nondenaturing conditions coupled to mass spectrometry (AP–MS) has become a popular technique for the identification of target proteins and interactor proteins [63–67]. When a protein is captured under nondenaturing conditions by affinity purification, its interactors are often captured together. By carefully evaluating background proteins and nonspecific interactors by using appropriate controls, specific interactors can often be differentiated. Iterative AP–MS experiments of multiple members of the same protein complex can also be used to cross-validate true interactions. Recently, two large protein interactome maps were published, both using large-scale AP–MS. [63–65] Together, the two interactome studies cover more than 6000 bait proteins with more than 12 000 interactor proteins from human cells (HeLa and HEK293T). This information provides invaluable knowledge about protein–protein interaction, protein dynamics, and cellular behaviors including the function of multienzyme complexes, the cross-talk between cells and tissues, and the function of enzymes.

Sample protocol for simple protein separation from cultured cell:

1. Harvest cell, wash the cell with PBS buffer a few times to remove extracellular fluid and centrifuge down the cell pellet.
2. Add urea lysis buffer (30 mM Tris, 8 M urea, 2 M thiourea, 4% CHAPS, add 1 tablet of cOmplete™ Mini EDTA-free protease inhibitor per 10 ml solution, add benzonase to digest DNA) to the cell pellet in a volume ratio of approximately 3 : 1 to 5 : 1 (buffer to pellet). Pippette up and down to resuspend pellet in buffer.
3. Sonicate on ice for 30 seconds, cool down for 60 seconds, repeat three times.
4. Centrifuge down pellet debris on max spin speed, take supernatant, and use Bradford assay or UV to determine protein concentration in the supernatant.
5. To a supernatant of 100 µl, add 400 µl methanol, vortex well, then add 100 µl chloroform. Vortex well again and add 300 µl water. Sample should look cloudy at this point. Vortex well and centrifuge at 14 000*g* for two minutes.
6. Pipette off the top aqueous layer. Protein exists between layers and may be visible as a thin wafer.

7. Add 400 µl methanol, vortex well, centrifuge at 14 000*g* for three minutes, then pipette out methanol as much as possible without disturbing the protein pellet (at the bottom).
8. Speed-vac to remove the remaining organic solvent. Avoid drying for too long or the pellet may be harder to resolubilize.

1.3.2 Protein Modification

1.3.2.1 Overview

One of the main purposes of protein modification in the sample preparation stage is to linearize the protein and thus facilitate downstream protein digestion and later protein inference. This mainly involves reduction of Cys–Cys disulfide bond by DTT or TCEP and then alkylation by either chloro- or iodo-acetamide to prevent re-formation. However, in some special cases, Cys–Cys disulfide bonds can be preserved for structure elucidation [68]. Modification of the peptide N-terminus and the E-amine of lysine are often used to add "tags" for "labeled-quantitation" (will be discussed later in Section 1.5.2) such as dimethyl labeling, isobaric tags for relative and absolute quantitation (iTRAQ), and tandem mass tag (TMT) labeling [69–72]. Another important reason for protein modification is to preserve protein–protein interaction information before digestion. This generally involves the chemical modification of adjacent proteins. The classic technique is chemical crosslinking, which uses chemical reactions to convert noncovalent, transient protein interaction to covalent, permanent chemical bonds [73]. A recent development is the proximity labeling methods using a fused, promiscuous biotin ligase or a peroxidase to label proteins in close proximity [74, 75].

1.3.2.2 Reduction of Disulfide Bond and Alkylation

Typically, after protein separation, the disulfide linkage is first reduced by DTT or TCEP. Breaking the Cys–Cys bond linearizes the protein and therefore prevents the formation of branched peptides that contain a disulfide bond after digestion. However, in some special cases, where the disulfide bonds are located is needed to elucidate either protein structure or protein interaction, these disulfide bonds can be either preserved or partially reduced for downstream analysis. Alkylation of the free Cys by either chloro- or iodo-acetamide prevents the re-formation of the disulfide bond. It is worth noting that iodoacetamide can quickly alkylate other amino acids as well as the N-terminus due to its higher reactivity [76]. Therefore, chloroacetamide is often used as an alternative to prevent off-target alkylation.

1.3.2.3 Chemical Crosslinking

Another important application of protein modification prior to digestion is to preserve information regarding protein interaction and nearest neighbors.

This may generally be associated with chemical modification, such as crosslinking by converting noncovalent, transient protein interactions to a permanent, covalent linkage. In a typical protein crosslinking experiment, the protein complex is chemically crosslinked in their simplest form, to be suitable for further digestion. This is generally achieved with the incubation of purified protein complex along with the crosslinking reagent; which replaces noncovalent interactions of the surface exposed to amino acid residues, with the covalent one [77–80]. Then the protein sample is then digested with the help of a suitable protease and is further separated and analyzed by LC–MS/MS. Moreover, it has been shown that with the combination of novel chemical crosslinkers and advanced data analysis platforms, it is possible to obtain structural information of protein complexes and protein–RNA complexes from crosslinked proteomics analysis [81]. This ability to undertake a large-scale analysis of crosslinked peptides from complex mixtures of protein has been one of the major developments in the field.

1.3.2.4 Proximity Labeling

Proximity analysis provides a way to investigate proteins in the vicinity of a protein that may or may not be interacting by introducing covalent tags to neighboring proteins of the bait. Typically, a promiscuous biotin ligase is fused with the bait protein and then expressed together with the bait in the cell. Upon the addition of biotin, the biotin ligase will label adjacent proteins with biotins. The cell is then lysed and enriched for biotinylated proteins by streptavidin beads. By carefully comparing the biotinylated proteins with control, proteins in the proximity of the bait protein can be identified. The approach is very interesting because it can be used in living cells and allowing direct investigation of physiologically relevant interactions. The two prerequisites for the method to be successfully accomplished are appropriate matching with the fusion protein and isolation of specifically labeled protein. Three different types of enzymes have been used extensively for proximity labeling; the BirA biotin ligase to introduce a biotin into proteins in proximity to BirA (BioID), horseradish peroxidase (HRP) to introduce hydroxyl groups in adjacent proteins (APEX), and an engineered ascorbate peroxidase for faster labeling [82–88]. Various versions of these enzymes have been developed, providing faster and cleaner labeling. It is worth noting that a variation of this method can also be used on fixed tissue and fixed cell samples using specific antibodies and an HRP-conjugated secondary antibody [89].

1.3.3 Protein Digestion

The key step of bottom-up proteomics is the protein digestion, which converts vastly different proteins into peptides that are more uniform in size, shape, and charge. Digestion is often allowed to occur at different levels and also with

different combinations of proteases. Commonly used proteases for digestion are trypsin, chymotrypsin, elastase, and endoproteases such as Lys-C, Lys-N, and Arg-C. The most commonly used enzyme is trypsin. Trypsin is a highly specific serine protease, active at an optimum pH of 8 and at 37 °C. It cuts at the C-terminal side of lysine (K) and arginine (R), except when proline (P) is on the carboxyl side of Arg or Lys. Both specificity and speed of hydrolysis are reduced when acidic residues are present at either side of the cleavage site. The specificity of trypsin ensures that most trypsin-digested peptides have at least two positively charged residues (two ends being R–R, R–K, K–R, or K–K), which is helpful for the downstream peptide identification by LC–MS/MS. Several other strategies can be implemented on a routine basis to enhance the quality of digestion and improve protein identification. Addition of MS-compatible surfactants helps to better solubilize and unfold proteins. Various commercially available surfactants such as ProteaseMAX, Invitrosol, Rapigest, and so on, can be added to reduce protein digestion time and to digest proteins that are difficult to digest otherwise.

Typically, when a single enzyme is used for digestion, the sequence coverage for proteins identified from a shotgun proteomics experiment is far less than 50%, that is, most of the amino acid sequence of that protein has not been detected by the mass spectrometer. The reason is that many of the peptides from the digested proteins are simply too long, too short, or hard to ionize, making them difficult to detect. To improve sequence coverage, multiple enzymes, including highly specific and nonspecific enzymes, can be used in combinations. Therefore, the same sequence is digested differently and produces various digested peptides, which improve the chance of detection by mass spectrometer analysis.

Sample protocol for peptide modification and digestion from protein:

1. Dissolve protein pellet in 8 M urea solution (for 2.4 g urea, add 1 ml 500 mM Tris pH 8.5, 2.2 ml water). For every 50–100 µg protein, add 60 µl of the above solution.
2. For 60 µl protein solution, add 0.3 µl 1 M TCEP to make a final concentration of 5 mM. Incubate at room temperature for 20 minutes with mild shaking.
3. For 60 µl protein solution, add 6.6 µl of 500 mM 2-chloro-acetamide and incubate at room temperature for 15 minutes, keep in the dark.
4. For 60 µl protein solution, dilute sample with 180 µl 100 mM Tris pH 8.5 buffer to 240 µl total. Add 2.4 µl 100 mM $CaCl_2$ to a final concentration of 1 mM. Add sequence-grade trypsin solution (0.5 µg/µl) at 1 : 20 to 1 : 100 weight ratio (trypsin:protein).
5. Incubate at 37 °C in the dark for four hours to overnight.
6. Add 13.5 µl 90% formic acid to a 5% final concentration.
7. Centrifuge at max speed for 15 minutes, transfer the supernatant to a new tube, freeze at −80 °C, or directly send for LC–MS/MS analysis.

1.4 Peptide Separation and Data Acquisition

1.4.1 Peptide Separation

One of the most significant advantages by using bottom-up proteomics is easier LC separation of peptides, comparing to intact proteins. After enzymatic digestion, digested peptides are much more uniform in shape, size, and charge than proteins. Using chromatographic techniques such as ion exchange (IXC), RP, or combinations of IXC and RP such as Multidimensional Protein Identification Technology (MudPIT), peptides can be efficiently separated by both their surface charge and hydrophobicity.

1.4.1.1 Reversed Phase (RP)

Today, the most commonly used technique for peptide separation is nanoelectrospray along with reversed-phase nanoflow LC. The method involves direct loading of peptide fragments onto a nanoflow capillary column, wherein they are separated on the basis of differential hydrophobicity and are processed further. Once the separation is achieved, separated protein fragments are directly electrosprayed from capillary tip into the mass spectrometer. The efficient separation by high performance liquid chromatography (HPLC) or ultra performance liquid chromatography (UPLC), when combined with the advanced mass spectrometer, is sufficient to identify more than 1000 proteins within an hour. With a longer column and separation time, more than 5000 protein identification can also be achieved in certain cases by a single reversed-phase LC–MS/MS system.

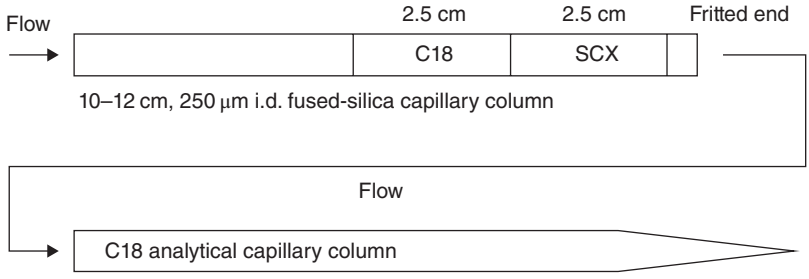
1.4.1.2 HILIC

HILIC separates peptides based on their hydrophilic interactions with an ionic resin and has found most application in peptide fractionation and PTM analysis. Electrostatic repulsion-hydrophilic interaction chromatography (ERLIC) is a specific form of HILIC, using a weak anion exchange (WAX) resin. Unlike reversed phase liquid chromatography (RPLC), peptides are retained under two separation modes. Early in the organic to aqueous gradient, hydrophilic interactions dominate, as in HILIC and inversely to RPLC. However, as the aqueous content of the elution buffer is increased, basic peptides electrostatically repel the WAX resin while acidic peptides are retained until their hydrophilic interaction with the WAX resin is disrupted late in the gradient. These superimposed separation mechanisms with ERLIC distribute peptides over the gradient better than RPLC and outperform it based on peptide and protein identifications by higher confidence spectral matching of larger peptides.

1.4.1.3 MudPIT

After enzymatic digestion, the number of peptide species and the huge differences in quantity among protein species result in a highly complex peptide

mixture. A combination of orthogonal separation methods, such as SCX and RP, helps to better separate different peptide species and therefore achieve better proteome coverage. In a typical MudPIT separation, peptide mixture is first loaded onto a short C18 + SCX capillary column (2.5 cm of C18 followed by 2.5 cm of SCX resin, as shown in Figure 1.2). The C18 + SCX column is then connected with an analytical C18 capillary column with a needle end for electrospray. In a typical 12-step MudPIT experiment, the first step uses a gradient of buffer B to elute all peptides from the short 2.5 cm C18 column to the SCX resin. In all the subsequent steps, buffer C is first used to elute a portion of peptides with different surface charges from SCX resin to the analytical column. The eluted peptides are then analyzed by C18 analytical column with a gradient of buffer B.



Example of a 12-step MudPIT gradient table

Step	0–3 min	3–5 min	5–15 min	15–25	25–110 min	110–120 min
1	100% A		0–100% B			100% A
2	100% A	10% C	100% A	0–15% B	15–45% B	100% A
3	100% A	15% C	100% A	0–15% B	15–45% B	100% A
4	100% A	20% C	100% A	0–15% B	15–45% B	100% A
5	100% A	25% C	100% A	0–15% B	15–45% B	100% A
6	100% A	30% C	100% A	0–15% B	15–45% B	100% A
7	100% A	35% C	100% A	0–15% B	15–45% B	100% A
8	100% A	40% C	100% A	0–15% B	15–45% B	100% A
9	100% A	45% C	100% A	0–15% B	15–45% B	100% A
10	100% A	50% C	100% A	0–15% B	15–45% B	100% A
11	100% A	60% C	100% A	0–15% B	15–45% B	100% A
12	100% A	100% C		0–15% B	15–70% B	100% A

Figure 1.2 Example of MudPIT setup and 12-step MudPIT gradient table. Buffer A: 5% acetonitrile, 95% water, 0.1% formic acid. Buffer B: 95% acetonitrile, 5% water, 0.1% formic acid. Buffer C: 500 mM ammonium acetate in water, 5% acetonitrile, 0.1% formic acid.

1.4.1.4 Capillary Electrophoresis

Capillary electrophoresis has also re-emerged as a complementary, more sensitive, and viable option in shotgun proteomics, largely due to improvements in electrospray interfaces. Fractionation of peptides prior to nLC-ESI to improve comprehensiveness was initially performed online with SCX resin minimizing sample losses from transfers intrinsic to offline fractionation and autosamplers.

1.4.2 Peptide Ionization

After separation, peptides are then ionized by various ionization methods to the gas phase and enter the mass spectrometer. Shotgun proteomics allows implementation of two primary ionization methods for ionic charging and transfer peptide fragments into the gas phase, noted as nanoelectrospray (nESI) and MALDI. The technique of nanoelectrospray is widely used in analytical mass spectrometry of oligosaccharides, glycosides, and glycoproteins due to its ease of use and remarkable sensitivity. In bottom-up proteomics, nESI provides excellent sensitivity with only a minimum amount of sample. In contrast to nanoelectrospray, MALDI offers both nondestructive vaporization as well as ionization of many large and small molecules. Although nanoelectrospray is the most commonly used method in shotgun proteomics, MALDI has been used more and more often for mass-spectroscopy-based imaging as it can provide spatial information together with the mass information.

1.4.3 Mass Analyzer

Over the last two decades, important advances in mass spectrometers, development of front-end automated methodologies, and completion of human genome project; applications for further peptide analysis, such as peptide identification, characterization, and so on, have greatly increased. In this regard, multiple automated instruments have been developed with hybrid technology, involving simultaneous separation, quantification, and data analysis. In this regard, some of the common mass analyzers that have proven to be adept in analysis of complex peptide mixtures can be noted as linear ion trap (LIT), Orbitrap, Fourier transform ion cyclotron resonance (FT-ICR), quadrupole, and time of flight (TOF). All these mass analyzers allow easy isolation and accurate data measurement of peptide masses at different interfaces, using different mechanisms; by maintaining proportionate balance between speed and sensitivity. Out of these analyzers mentioned herewith, most advanced version of mass spectrometers exploited widely in the field of proteomics is LIT. Furthermore, it should be noted that ion trap mass spectrometry is performing a leading role in modern instrumental world, for being capable of identifying and quantifying high- and low-molecular-weight pure peptides, with the same

sensitivity and specificity. Thus, linear-ion-trapped mass analyzer essentially serves the role of all, that is, ion selection, ion trapping, ion fragmentation as well as low-resolution mass analysis. Identification of peptides within the sample is accessed with the help of data-dependent acquisition on the basis of initially unbiased sampling. An upgraded version of LIT involves four elongated planer electrodes, mounted in parallel to maximize the potential of ion trapping in both radial and axial directions. Once the sample is ionized, peptide ions are being trapped within the LIT. A radiofrequency voltage applied within the trap is increased, thus initiating ion ejection from the ion trap to detectors outside of the quadrupoles. With the initial precursor ion scan, it is possible to identify abundant peptide precursor ion m/z values and then subsequently an identified peptide precursor ion is selected and then isolated by scanning out all other ions. Trapped ions are translationally excited causing collisions with the helium bath gas to vibrationally excite the ions, through conversion of translational energy into vibrational energy. As vibrational energy increases, covalent bonds begin to fragment and when this happens, the resulting fragment ions are no longer excited. Accordingly, the fragment ions are scanned out of the ion trap. Some scan strategies are implemented to create unbiased sampling of peptide ions by using a data-independent acquisition with consecutive small (10–25 m/z) ion isolation windows. Additionally, sampling speed in ion traps improved from the 3D ion trap by the invention of the 2D LIT and then creation of the segmented 2D LIT to separate ion trapping and fragmentation from mass analysis. The segmented trap allows the use of different gas pressures

The 2D LIT has also been a useful technology to create hybrid mass spectrometers to combine ion trapping and MS/MS capability to mass analyzers where these steps are difficult or impossible to perform in the mass analyzer. For example, the 2D LIT was interfaced with a FT-ICR mass analyzer to add a high-resolution and high-accuracy mass analyzer to the capabilities of the LIT. The LIT was also added to the Orbitrap mass analyzer to create a powerful hybrid mass spectrometer.

The Orbitrap mass analyzer detects the frequency of ion current produced by peptide ions, which oscillate along a central electrode with a frequency proportional to $(m/z)^{-1/2}$. The frequency-based signal can be measured repetitively without losing the peptide ion and therefore enhance the accuracy [117]. Fourier transformation is then used to convert the frequency signal to highly accurate m/z values. The introduction of Orbitrap mass analyzers significantly improved the analysis for PTMs and quantification with isotopic labeling.

FT-ICR instruments, analyzers specifically working on the principle of mass to charge ratio (m/z) of ions, are still capable of a higher mass accuracy as compared to Orbitrap. However, the price and size of FT-ICR instrument is often inferior when compared to Orbitrap instruments, which limit the use of FT-ICR in many applications.

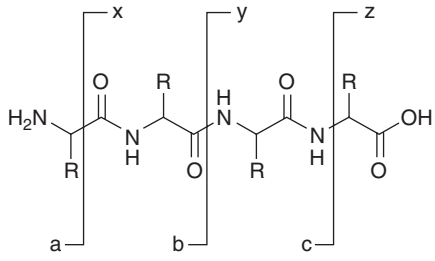
1.4.4 Peptide Fragmentation Method

Shotgun proteomics has advanced with improvements in mass spectrometry technology providing better data accuracy and quantification to achieve wider proteome coverage. Further improvements on proteome coverage could be achieved by advances in gas phase fragmentation methods for peptides. Although development is still going on, at present, various modern methods are available to allow ion fragmentation, providing different information about the structure and composition of given molecule. The most commonly used methods are still collision-induced dissociation (CID) and collisionally activated dissociation (CAD).

1.4.4.1 CID/HCD

CID, a mass spectrometry technique, has been recognized as the most trusted choice for fragmenting gaseous molecular ions, due to its high efficiency, predictable fragmentation, and ease of use. The ions that are being generated through CID are exploited for several purposes.

This type of excitation is primarily associated with LIT and beam-type collision activation, preferable to mass spectrometers such as triple quadruple. In ion trap instruments, the motion of precursor ions is increased by resonance excitation to create more forceful collisions of ions with neutral molecules. Resonance excitation is associated with 3D and 2D ion trap instruments. In general, commonly produced ions are b- and y-type, leaving positive as well as negative charges on the N- and C-terminus (Figure 1.3). As resonance excitation is based on the motion of ions in the trap and that motion is based on the m/z value of the ion as soon as a fragmentation event occurs, the



Method	Fragmentation
CID	b and y ions
HCD	Mainly b and y ions, a ions also possible
ECD/ETD	c and z ions
UVPD	a, b, c, x, y, z ions

Figure 1.3 Common fragmentation methods used in bottom-up proteomics and the typical fragmentation ions generated from peptide.

resulting fragment ions fall off the excitation frequency and they are no longer excited.

Ion activation in a beam-type instrument such as a triple quadrupole or a quadrupole TOF occurs when ions are passed through a quadrupole containing a high gas pressure (a few millitorr) and ions collide with the gas until sufficient vibrational energy is reached for fragmentation to occur. Unlike ion trapping instruments when fragment ions are produced, they continue to undergo energetic collisions as they pass through the quadrupole collision cell. A variation on the collision cell used in traditional beam instruments was developed for Orbitrap hybrids. In this device, the precursor ion is accelerated into the collision cell and then returned to the injection site effectively passing up and back in the cell. An advantage to fragmentation in the collision versus fragmentation in an ion trap is that a collision does not have a lower m/z cutoff and thus immonium ions can be detected and reporter ions from TMT-like experiments can be observed.

1.4.4.2 ETD/ECD

Collision-based fragmentation methods are driven by the input of vibrational energy into ions. This energy gets randomized throughout the bonds of the ion and then weakest bonds fragment. This is an ergodic process. Electron capture methods such as ECD and ETD result in fast fragmentation of ions probably at the site of electron capture or transfer and thus is considered to be nonergodic. The electrons that are being employed in this method are either thermal electrons or electrons transferred from negatively charged fluoranthene, generating mostly c- and z-types of ions (Figure 1.3). ETD/ECD can generally provide a "softer" fragmentation method in the sense that labile PTMs such as phosphorylation and glycosylation are preserved. The localization of labile y-CO_2 and SO_3 modifications can be identified by ECD fragmentation. Studies have shown that CID is efficient for phosphopeptide identification and ECD is better for phosphorylation site localization [90, 91]. When combined, complementary information and higher confidence are offered for identified phosphopeptides. ETD is also suggested to be more advantageous over CID method for the detection of phosphorylation and glycosylation sites due to retention of labile modification moieties.

1.4.4.3 IRMPD/UVPD

High-energy light has also been used to fragment ions. There are two main types of photodissociation methods for bottom-up proteomics, the infrared multiphoton dissociation (IRMPD) and ultraviolet photodissociation (UVPD). IRMPD is typically coupled with FT-ICR mass spectrometry and has been used for a large variety of molecules. IR lasers produce low-energy radiation that mainly excites the O–H and N–H stretching frequencies in ions. As a method to fragment ions, the goal is to pump enough energy into ions that it

then gets distributed throughout the ion and causes broader fragmentation of the ion. However, due to lower activation efficiency of peptide cations in ion trap instruments, IRMPD is not a common fragmentation method for bottom-up proteomics. A UV laser produces much higher-energy photons, which can be used to effectively fragment peptides and produce all six types (a-, b-, c-, x-, y-, z-) of fragmentation ions (Figure 1.3). Because of its unique fragmentation ability, UVPD can significantly improve the confidence of peptide identification and sequence coverage, especially for acidic peptides, which would preferentially ionize in negative mode rather than the commonly used positive mode.

1.4.5 Acquisition Mode

There are two main strategies for data of acquisition in mass spectrometry experiments to scan and fragment peptide ions: the data-dependent analysis (DDA) mode, and the data-independent analysis (DIA) mode. In the DIA mode, mass spectrometer acquires data based on the sequential isolation and fragmentation of specific precursor windows, for example, 4–25 Da. All peptide species within this m/z window are fragmented together and acquired on the same MS/MS scan. Because the mass spectrometer does not select for any ion in the DIA mode, data are being acquired at a steady pace and in a systematic way. Because of the lack of ion selection, MS/MS acquired in DIA mode normally contains multiple peptide ions, which makes peptide identification in DIA a challenging task. In contrast, under DDA mode, the mass spectrometer scans parent ions and selects the most abundant ions for fragmentation with a window of about 2–3 Da. Each MS/MS acquired under DDA mode normally contains only one peptide, which simplifies peptide identification. Moreover, dynamic exclusion is used to prevent the mass spectrometer from repeating collection of MS/MS from a particular peptide ion m/z value over set time period (3060 seconds). This technique allows the mass spectrometer to scan for unique peptide species and can improve the range of detection. The major difference between the two scans modes is that one is stochastic in the collection of MS/MS (DDA) and the other is systematic (DIA), but the performance of both methods improves from increases in scan speed.

1.5 Informatics

Proteomics exists because of the ability to use computer algorithms to analyze and interpret the mass spectrometry data. After protein separation, digestion and LC–MS/MS data acquisition, the acquired data, both MS and MS/MS spectra need to be further processed to identify and quantify peptides and then proteins. The data analysis process can greatly influence the overall biological

conclusion of the whole experiment and therefore needs to be conducted with high precision and accuracy. On the other hand, a modern LC–MS/MS system is able to generate 20 000–30 000 spectra and gigabytes of data per hour. This enormous amount of data, combined with the need for precision and accuracy, pose significant computational challenge for data analysis. In the past 25 years, numerous algorithms and tools have been developed to facilitate data analysis for bottom-up proteomics, mostly focused on three aspects of proteomics, peptide identification, quantitation, and protein interaction.

1.5.1 Peptide Identification

Peptide identification is the foundation of bottom-up proteomics, as the peptide is the analyte being directly measured by the mass spectrometer. It was a tedious and time-consuming process of protein biochemistry with manual interpretation of MS/MS before the invention of SEQUEST, a computer program that can automatically match MS/MS to peptide sequences derived from a reference sequence database. Automatic peptide identification enables researchers to identify thousands of peptides and therefore greatly improves the efficiency and scale of proteomics research. Since its introduction in 1994, the field of bottom-up proteomics has flourished, and bottom-up proteomics has become the method of choice for probing and interrogating the proteome. Currently, there are three major computational methods for peptide identification, namely the database search, spectral library search, and *de novo* sequencing (Figure 1.4).

1.5.1.1 Database Search

The first peptide identification algorithm is SEQUEST, a database search algorithm for automated peptide identification. The software computes on the basis of two important calculations, confirming whether experimental peptide sequence is a perfect match of a fragmentation spectrum: XCorr and DeltaCN. The first one can be referred to as a statistical calculation of correlation of theoretical and experimental data; whereas the later one indicates the difference between two possibilities: the best spectrum match and the second-best peptide-spectrum match (PSM). Although first introduced over two decades ago, it is still the benchmark standard for peptide identification algorithms today. More generally, a database search algorithm first compares each MS/MS spectrum to all the theoretical spectra of all possible candidate peptides for a molecular weight, to generate PSMs. Candidate peptides are computationally generated from a predefined protein reference database using predefined enzyme specificity and precursor mass. The best matches (PSMs) between experimental and theoretical spectra will be calculated and ranked, then filtered using a specific false discovery rate (FDR) level to provide the final peptide identification results. The protein reference database must contain the

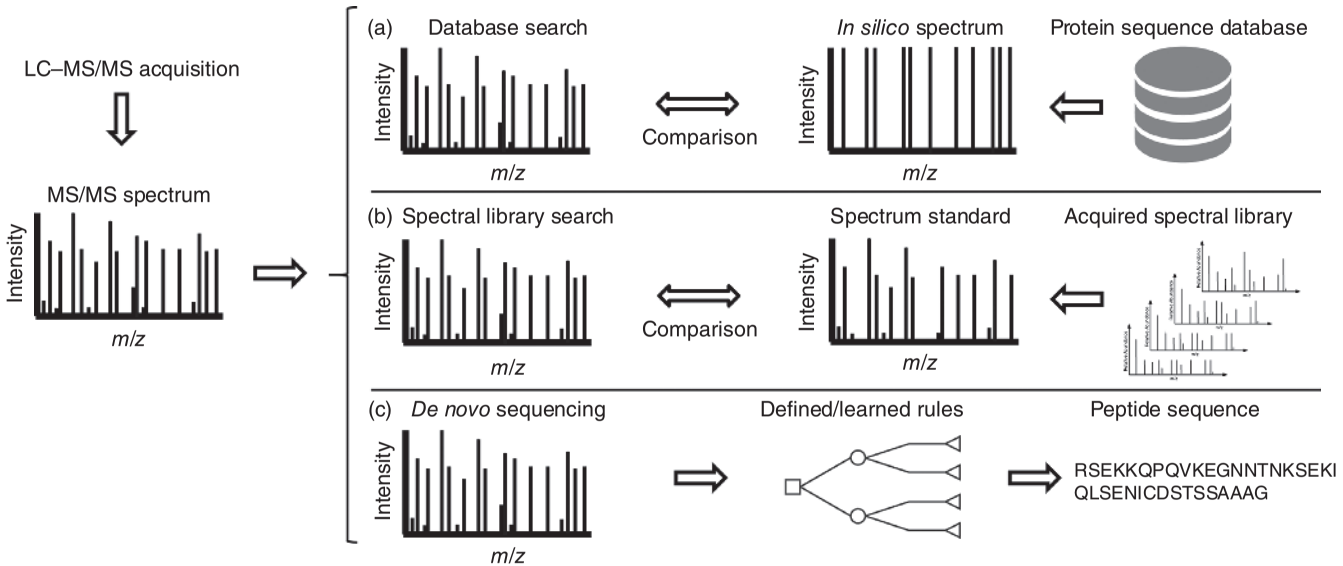


Figure 1.4 Comparison of three most commonly used peptide identification strategies: (a) Database search compares theoretical spectrum generated *in silico* from a protein sequence database to the actual MS/MS spectrum. (b) Spectral library search compares the current MS/MS spectrum to previously acquired standard spectrum from a spectral library. (c) *De novo* sequencing uses a set of predefined or previously learned rules to directly derive peptide sequence from MS/MS spectrum.

sequences of all proteins that are expected from the sample, similar to the reference genome used for genomic sequencing experiment. In most cases, reversed or random protein sequences are also appended to the reference database as true negatives (decoy sequence). After peptide identification, these true negatives are then used as "internal standards" to estimate and adjust the FDR of the peptide identification process. Over the past two decades, various database search programs have been developed for bottom-up proteomics, some of the most popular ones include SEQUEST, MASCOT, Andromeda, MS-GF+, X!Tandem, OMSSA, Comet, Crux, ProLuCID, PeaksDB, pFind, MSFragger, and so on. All these programs use different scoring scheme and algorithms, but they all require a protein reference database. If a specific protein or peptide does not exist in the reference database, it cannot be identified by any of the abovementioned programs. Fortunately, the development of next-generation genomic sequencing techniques now provides us with an unprecedented capability to sequence almost any genome with acceptable cost. These genomes can then be used to predict and build protein reference databases. The most popular source of protein reference database is Uniprot.

Sample protocol for peptide identification with database search:

Software used in this protocol:
RawConverter, **ProLuCID-GUI**, **ProLuCID**, **DTASelec2** can all be found at fields.scripps.edu (Yates laboratory webpage), under Resources, Download page.
Proteowizard can be downloaded at proteowizard.sourceforge.net

1. Download **ProLuCID-GUI** and check if Java is properly installed following the instruction within the downloaded package.
2. Convert the vendor raw data files obtained from the LC–MS/MS system to ms2 format. **RawConverter** or **Proteowizard** can be used to convert most common formats to ms2. **ProLuCID-GUI** also comes with an example.ms2 file for practice.
3. Download the protein reference database according to the sample species, for example, human, mice, zebrafish, and so on, in FASTA format from Uniprot. (www.uniprot.org/proteomes). Store the database FASTA file together with the ms2 files obtained from step 1. Use **ProLuCID-GUI** to add the reverse sequence as decoy. **ProLuCID-GUI** also comes with a human reference database in FASTA format.
4. Load the downloaded FASTA file and all ms2 files with **ProLuCID-GUI**, set precursor tolerance to 50 ppm if data is acquired from a high resolution (e.g. orbitrap) instrument. Precursor range is normally set from 600 to 6000 Da, number of isotopes is set to 3 to cover M – 1, M, M + 1 peaks. Fill in the name of the protease and the cleavage specificity. Here we use "trypsin" as the protease name, and "KR" as

specificity. Set enzyme specificity to 0 for nonspecific, 1 for one end (semi-tryptic), or 2 for both ends (fully-tryptic). When set to 2, both ends of the cleaved peptides are required to be cleaved specifically (after KR or being termini) by the enzyme.

5. Specify all the chemical modification in "static modification." For a typical experiment using chloro- or iodoacetamide for alkylation, "57.021 46 C" is used to specify the alkylation adduct on cysteine.
6. Specify all dynamic (differential) modifications, such as phosphorylation, using the same syntax. For standard phosphorylation on Serine/Threonine/Tyrosine, we use "79.966 331 STY." If any N-terminus or C-terminus static modification is present, add them to the corresponding textboxes. For dynamic modifications at terminus, add them together with other differential modifications, for example, "79.966 331 STY; 15.994 915 N-term."
7. After peptide identification, all PSMs, including those from decoy sequence, will be filtered by **DTASelect2** to remove false positive as much as possible, till the FDR reaches the desired level. For FDR estimation and peptide filtering, fill in the minimum number of peptides required per protein (e.g. 2) and the minimum number of tryptic end to consider for each peptide (e.g. 2). Then select the target FDR level: protein-, peptide-, or spectrum-level FDR. Finally set the target FDR. For 5% FDR, input 0.05.
8. An output folder needs to be specified to store results, do not use the same folder that stores the ms2 files. Then click "Run ProLuCID and DTASelect" to run the program, or click "Re-run DTASelect only" to only rerun DTASelect for a different target FDR filtering.
9. After successful peptide identification and filtering, the filtered result files will be stored at the output folder specified at step 8.

1.5.1.2 Spectral Library Search

The idea of spectral library search is simple: compare the experimental spectra with existing standard spectra library to find all the matches. This concept is a long-standing method used in mass spectrometry since the 1960s. Mass spectral libraries were based on electron ionization spectra and when MS/MS techniques were developed, it was not clear the spectra produced had sufficient reproducibility to create libraries. Yates et al. demonstrated the ability to match MS/MS of peptides to a library [92]. Software algorithms for library searching include SpectraST [93], NIST MS Search [94], M-SPLIT, X!Hunter, BiblioSearch, Pepitome, Bonanza, Tremolo, and pMatch [95]. Because most of spectral library search algorithms use a dot product calculation, which utilizes not only the m/z value but also the intensity of the fragmentation ions, spectral library is believed to be more sensitive and more reproducible than other peptide identification methods. However, the biggest challenge within this process

is to create an appropriate spectra library, which contains high-quality spectra of most of the expected peptides, for the comparison. The human proteome is estimated to contain more than 20 000 proteins. When trypsin is used as the protease, without considering any miss-cleavage or any PTM, the number of all possible tryptic peptides is estimated to be around 500 000–1 000 000. Synthesizing peptide standards and acquiring mass spectra for most of these peptides is an enormous, but not impossible effort [96, 97]. If two or more PTMs are considered, the number will immediately surpass our current synthetic capability. Moreover, as explained before, the type of ionization, type of mass analyzer, and the fragmentation mode greatly influence the contents of acquired spectra. Standard spectra library acquired by ETD fragmentation mode contains few b- and y-ions, and therefore, they simply cannot be used to match experimental data acquired from CID fragmentation. To circumvent this problem, high confident peptides identified by DDA and database search are commonly used to construct spectral library for future peptide identification in the experimental systems of interest. This approach has become popular in recent years for DIA data analysis, as conventional database search algorithms were not designed to analyze multiplexed MS/MS.

1.5.1.3 *De novo* Sequencing

Both database search and spectral library search are limited by either known sequence or known spectral library. These methods limit the search space for peptide "sequencing" to the known "universe" of sequences or spectra, which in turn improves sensitivity, specificity, and accuracy. For unknown protein or peptide samples, such as soil extracts, it is often hard to perform peptide identification by either method, as the sequence space is large and not covered by available genome sequences. In contrast, *de novo* sequencing must search a virtually infinite sequence space with no *a priori* sequence information to use as a guide. Many advanced statistical models and machine learning methods have been introduced for the development of *de novo* sequencing algorithms. The key is to learn the peptide sequence from the MS/MS fragmentation pattern. This is a daunting task as even the physical rules of peptide fragmentation are known, it is not guaranteed that each spectrum has a singular solution because ambiguities can exist in the spectrum due to missing fragment ions. In the past two decades, many *de novo* sequencing tools have been developed, such as PepNovo, PEAKS, NovoHMM, MSNovo, pNovo, UniNovo, and Novor. These methods have greatly improved the speed and sensitivity of data analysis, as well as the confidence in the peptide sequence derived from *de novo* sequencing. More recently, the incorporation of deep learning algorithms such as convolutional neural network (CNN) has shown great improvement for processing noisy MS/MS spectra. However, it should be noted that despite constant advancements in the methodology, *de novo* sequencing still cannot be directly compared with the output obtained from database searching and spectral library analysis in both accuracy and specificity.

1.5.1.4 Peptide-Centric Analysis

DIA has become increasingly popular. In DIA, multiple peptides are fragmented together, creating a multiplexed MS/MS spectrum. The multiplexed MS/MS cannot be easily processed by conventional identification programs since it contains an unknown number of peptide species, that is, one spectrum may contain 1, 2, 3 or more peptides. To solve this problem, peptide-centric analysis starts by asking a different question. Instead of asking "what is the peptide" in a particular spectrum, peptide-centric analysis focuses on testing "if a particular peptide exists" within a series of spectra [98]. By testing each peptide against all MS/MS, a probability function can be derived for each peptide, and thus can be used to calculate the probability of the match and to quantitate each peptide [99, 100]. Although very attractive, peptide-centric analysis is still a relatively new method and requires more time to prove its usefulness.

1.5.2 Peptide/Protein Quantitation

Mass spectrometry is the method of choice for large-scale quantitation of proteomes. As most methods use peptides as surrogates for quantitation, care must be taken when trying to infer abundance changes at the peptide level to the protein level. Abundance changes are a direct reflection of the cellular response to the environment or treatments. A quantitative proteomics experiment is able to provide quantitative information for thousands of proteins simultaneously, provides useful insights of key proteins and PTMs involved in certain cellular process or disease. Initial MS-based quantification approaches were dependent upon chemical labeling by the addition of isotope-coded reagents to reactive groups, via peptide terminus or through side chain of the amino acids. Over the years, numerous techniques have been developed to provide quantitative information. In general, these quantitation methods can be divided by two main categories, labeled quantitation methods and label-free quantitation methods.

1.5.2.1 Labeled Quantitation

In labeled quantitation, different samples are modified by isotope containing molecules and then combined for further preparation and LC–MS/MS analysis. There are generally two types of labeling approach, metabolic labeling and chemical labeling. Metabolic labeling uses isotope-labeled amino acids to incorporate heavy isotope atoms into newly synthesized proteins during cell or animal growth. Cells or animals receiving different treatments are fed and labeled with amino acids with different weight isotope compositions. An advantage to metabolic labeling is that labels are introduced into actively growing cells and thus are incorporated into intact proteins. Proteins can then be mixed after cell lysis. In chemical labeling, isotope-labeled chemical probes

are used to label proteins or peptides after cell lysis or protein digestion, respectively. Chemical labeling is usually performed on either the terminus or reactive side chains such as thiol and amino groups of the peptides. After labeling with either approach, multiple samples are then combined and analyzed by LC–MS/MS. During data analysis, the origin of a particular peptide can be traced by the composition of the incorporated isotope atoms. When two or more isotopic weights of the same peptide are detected, these are directly compared to estimate the relative abundance of the two peptides based on ion signal between among samples. For both methods, heavy and light samples are combined together for analysis to reduce errors during sample preparation and LC–MS/MS. A clever chemical labeling approach uses tags that have isobaric weights that reveal reporter ions of different m/z value upon fragmentation in the mass spectrometer. These multiplexed labeling methods such as TMTs and iTRAQ can multiplex up to 10 and 8 experiments, respectively. An advantage to these methods is that multiple samples can be analyzed with a single LC–MS/MS experiment instead of 10 or 8 and thus greatly reduces mass spectrometry time and the associated cost. A disadvantage to the isobaric tagging method is the issue of precursor ion contamination. Multiple peptide ions can be present in the isolation window of an MS/MS acquisition and as the ions are fragmented, the reporter ions will all contribute to the values in the low m/z end of the spectrum even though the other fragment ions are different. An issue for labeling methods in general is the completeness of labeling for both metabolic labeling and chemical labeling approaches. Incomplete labeling can significantly complicate the data analysis and jeopardize quantitation accuracy.

SILAC and SILAM Stable isotope labeling with amino acids in cell culture (SILAC) [101] and stable isotope labeling in mammals (SILAM) [102] are metabolic labeling methods, which introduce isotopically labeled amino acids during the growth of a cell or an animal. In a typical SILAC experiment, two populations, for example, control vs. treatment, of cells are cultured; one is fed with a simple growth medium containing normal amino acids (Arg-0 or Lys-0), and the second is fed with an enriched medium containing heavy stable-isotope-labeled arginine and lysine (Arg-6, Lys-6) instead of light Arg and Lys (Figure 1.5). In the growing state, cells will gradually incorporate stable isotopes into the entire proteome. For HEK293 cells, near complete labeling can be achieved after five to eight cell divisions. Metabolic labeling is generally easier for fast dividing organisms such as yeast, bacterial, or cultured cells. For larger organisms, such as mice, it is more challenging to label proteins with heavy stable isotopes, therefore SILAM was developed to address this problem. In SILAM, mice are fed with ^{15}N-labeled spirulina as the only protein source mixed with other nonprotein material for a balanced diet. The time required for labeling any specific tissues depends upon local protein turnover rates [102].

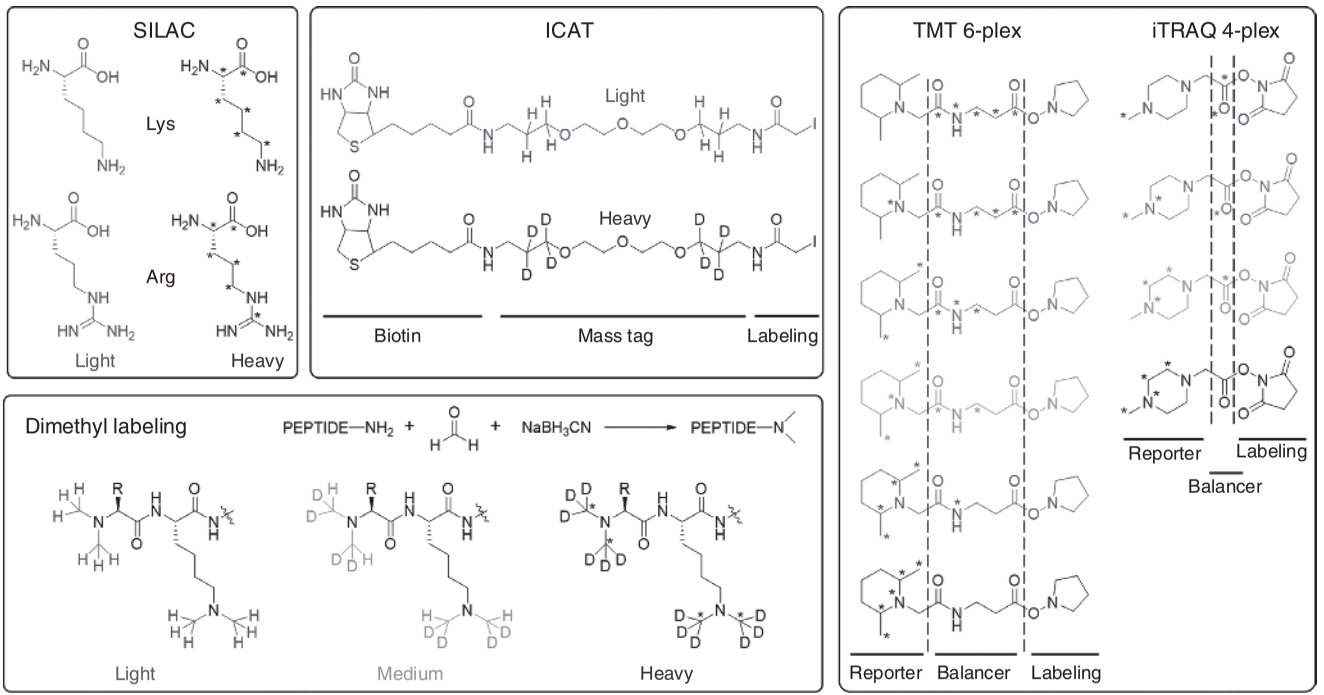


Figure 1.5 Examples of different labeling strategy for quantitation, where * denotes the position of heavy isotope atoms.

In both SILAC and SILAM, a high percentage of isotope incorporation into the protein is desired for successful quantitation, incomplete metabolic labeling may generate overlapping heavy and light isotope envelopes, which jeopardize the accuracy of the quantitation. Labeling cells or tissues with heavy isotopes may result in changes to the proteome, known as the isotopic effect [103]. To circumvent the problem, heavy-isotope-labeled samples can also be used as an internal standard to bridge the quantitation of different samples under various treatment conditions. For example, if Samples A and B need to be compared, a heavy-isotope-labeled Sample C can be used as the internal standard. Sample A is first mixed with Sample C and analyzed by LC–MS/MS, the ratio of protein abundance [A:C] is then calculated for each protein. Similarly, Sample B is then mixed with Sample C and analyzed the same way to calculate the protein ratio [B:C]. The ratio between A and B can then be calculated as [A:C]/[B:C]. Samples A and B cannot be directly mixed and analyzed together since they are both "light" (no heavy isotope labeling), and therefore all peptides are indistinguishable.

Most metabolic labeling approaches are expensive and time-consuming since heavy-isotope-labeled amino acids or food needs to be supplied during a long period of time to support cell or animal growth and reproduction. However, metabolic labeling does offer highly accurate protein quantitation results when properly implemented.

ICAT Isotope-coded affinity tagging (ICAT) is one of the early chemical labeling methods developed for mass-spectrometry-based quantification. It utilizes the reaction between thiol side chain of the cysteine residue and isotope-coded tags. (Figure 1.5) In principle, ICAT comprises three different elements tagging, a biotin affinity tag, a thiol-specific reactive group, and a linker with specific light or heavy isotopes. In a typical ICAT experiment, cysteine residues are labeled with ICAT reagents during the protein alkylation step, with either eight ^{1}H or eight ^{2}H atoms (or ^{12}C and ^{13}C). A mixture of separated proteins is then subjected to avidin affinity chromatography for further purification of ICAT tagged proteins. Tagged proteins are digested with trypsin, and the labeled peptides are eluted from the avidin column and then analyzed by LC–MS/MS to obtain a detailed spectrum. Because ICAT tags only cysteine residues (observed amino acid frequency of 3.3% in vertebrates), the method may miss those proteins that don't contain Cys residues. Additionally, some proteins may only contain a single Cys residue, which makes calculating statistics for quantitation difficult.

Dimethyl Labeling Dimethyl labeling is a chemical labeling approach that uses reductive amination to alkylate the N-terminus and the amino group on lysine side chain with a dimethyl tag [69]. As most peptides contain at least an N-terminus, dimethyl labeling can be used to cover a large portion of the

proteome. After protein separation and digestion, formaldehyde (CH_2O) and cyanoborohydride ($NaBH_3CN$) are added for reductive amination. Different combination of isotopically labeled formaldehyde (CH_2O, CD_2O, $^{13}CD_2O$) and cyanoborohydride ($NaBH_3CN$, $NaBD_3CN$) can be used to perform duplex (e.g. $CH_2O + NaBH_3CN$ and $CD_2O + NaBH_3CN$) to quadruplex (e.g. $CH_2O + NaBH_3CN$, $CH_2O + NaBD_3CN$, $CD_2O + NaBH_3CN$, and $CD_2O + NaBD_3CN$) experiments (Figure 1.5). If the mass resolution of the instrument is high enough, an isotopologue strategy can also be implemented (e.g. $CH_2O + NaBD_3CN$ and $^{13}CH_2O + NaBH_3CN$). The reductive amination of primary amines with formaldehyde is generally a fast and clean reaction with high yield. Therefore, most of the peptides can be efficiently labeled using this method, giving a good coverage of the proteome. The simplicity and the effectiveness of dimethyl labeling make it a very attractive chemical labeling approach to compare two to four samples in the same experiment.

TMT/iTRAQ TMTs and iTRAQ are both isobaric mass tags for chemical labeling. Accordingly, both methods share several functional elements, like a unique mass reporter, a linker that can further initiate cleavage and balance the mass, and an amino reactive group. (Figure 1.5) Commercially available TMT reagents are available in 2-, 6-, 10-, and 11-plex, while iTRAQ reagents are available in 4- and 8-plex formats. When multiple samples are labeled and then combined for LC–MS/MS analysis, the same peptide species from different samples have the same total mass weight (isobaric mass tagging) allowing coelution of peptides during LC and simultaneous isolation of fragments, during MS/MS. During peptide fragmentation, the linker is fragmented to release the mass reporter, wherein the relative abundance of the original peptide is estimated on the basis of relative intensity. When contaminations in the reporter mass region or interfering ions from coeluting peptides are found, MS^3 (MS/MS/MS) can be performed to improve quantitation accuracy, often referred to as MultiNotch quantitation [104]. Isobaric labeling has widespread applications as its multiplexing ability provides a highly efficient way to quantify a large number of samples.

1.5.2.2 Label-Free Quantitation

Although accurate quantification of peptide fragments can be achieved through tagging with stable isotopes, these methods often suffer from increased processing steps, increased cost of labeling reagents, inefficient labeling, and difficulty in the analysis of low-abundance proteins. Alternatively, label-free protein quantification has been used to perform fast, efficient, and cost-effective protein quantitation. In a label-free analysis, each sample is analyzed independently without any labeling and then compared across samples for quantitation. Generally, label-free quantitation methods can be categorized into two different categories: (i) spectral count and (ii) ion intensity or area

under curve (AUC). Spectral counting uses the count of total PSM related to a particular protein as a measure of relative abundance. Ion intensity or AUC quantitation uses the sum of the signal intensity of peptide peaks, belonging to a particular protein.

Spectral Count Spectral counting is a label-free quantitation method that uses the number of MS/MS generated by a particular protein to estimate protein abundance [105]. Spectral counting is based on the correlation of the frequency of PSMs with the amount of protein in the sample, that is, assume that a higher amount of protein generates more PSMs. It is a method that benefits from good fractionation of complex peptide mixtures by LC. If a complex peptide mixture is separated using a short LC gradient, then the spectral counts will be compressed and not very accurate. The method can be applied from low to moderate mass resolution (0.1–1 Da) LC–MS/MS data, and it is observed to provide good results for proteins with highly redundant PSMs [105]. The simplest comparison is to compare the number of PSMs for each protein. However, such comparison is only quantitative when all proteins in comparison are similar, for example, protein isoforms. To improve the quantitation accuracy by spectral counting, several scoring models have been developed over the years, including spectral index (SI) [106], protein abundance index (PAI) [107], exponentially modified protein abundance index (emPAI) [108], normalized spectral abundance factor (NSAF) [109], and distributed normalized spectral abundance factor (dNSAF) [110]. Different scoring models take different parameters into account for protein quantitation. For example, SI corrects for ion intensity and both NSAF and dNSAF consider the length of the protein as a parameter for scoring. To date, the most widely used scores are the emPAI and NSAF scores, due to their simplicity and superior results [108, 111].

emPAI score for protein x is calculated as:

$$\mathrm{emPAI}_x = \frac{10^{\frac{P_x^{\mathrm{observed}}}{P_x^{\mathrm{observable}}}} - 1}{\sum_{i=1}^{n}\left(10^{\frac{P_i^{\mathrm{observed}}}{P_i^{\mathrm{observable}}}} - 1\right)}$$

where n is the total number of detected proteins. The emPAI score was proposed initially based on an observation that the PAI, that is, observed peptides divided by the number of observable peptides per protein, is linearly correlated to the logarithm of protein concentration in LC–MS/MS experiments [108].

NSAF score for protein x is calculated as:

$$\mathrm{NSAF}_x = \frac{\mathrm{number\_of\_PSMs\_to\_}x\,/\,\mathrm{length\_of\_}x}{\sum_{i=1}^{n}\left(\mathrm{number\_of\_PSMs\_to\_}i\,/\,\mathrm{length\_of\_}i\right)}$$

where n is the total number of detected proteins. One of the most significant features of NSAF score is that it is adjusted for protein length. Therefore, larger proteins that are able to produce more fragments than smaller proteins are normalized and therefore receive comparable scores. NSAF scores can be easily calculated and implemented for almost all bottom-up proteomics experiments, thus widely used in spectral counting-based label-free quantitation. It has been shown to provide significantly better linearity than most other scores in certain tests [111].

Ion Intensity/AUC Although simple and powerful, spectral counting-based label-free quantitation does not always provide satisfactory results. This is due to a number of factors, including the dynamic exclusion of precursor ions in DDA, which sometimes creates bias during detection. Ion intensity or area-under-curve (AUC) quantitation uses the MS spectra to integrate the volume of the chromatographic peak in the time, m/z, and ion intensity to estimate peptide abundance. The concentration of a peptide is typically correlated from the area under the chromatographic peak, in the range of 10 fmol to 100 pmol. The overall protein quantification is based upon measuring ion abundance of peptides for each protein within the detection limit of an instrument. In spite of being a direct detection system, measurements are generally being influenced by various factors such as the accuracy of the results and reproducibility. Some of the major challenges in ion intensity quantitation include chromatographic drift and disagreement among different peptides from the same protein. Dynamic warping is a very popular technique to align peaks among samples, which greatly reduces retention time drift. When a large discrepancy is observed among observed ions for the same protein or peptide, the most abundant top N ions can be used for quantitation, excluding all the low abundant and low confident ions. Different data analysis platforms have slightly different implementations to calculate the area or volume under the curve. Some of the proprietary implementations of algorithms, that is, algorithm not published, such as Progenesis QI were shown to outperform other implementations such as MaxQuant, Proteome Discoverer, and Scaffold Q+ [112].

1.5.3 Protein Inference

Shotgun proteomics is a technique commonly used for identification of different peptide mixtures in a given sample. The final results are analyzed on the basis of correct and false PSMs. However, for most proteomic experiments, the final goal is to identify which proteins are present in the biological sample. Peptide identification only serves as an intermediate step to infer the existence of proteins. The process that assembles peptides back to proteins is protein inference.

In order to infer protein with high confidence, the first step is to assess the quality of the identified peptides. As discussed earlier (Section 1.5.1.1), a decoy protein database can be used to assess the FDR and differentiate high confident PSMs from low confident ones. A decoy protein database is generally created by a sequential reversal of normal protein database and is then appended to the original database. After protein identification by either SEQUEST or other algorithms, programs such as DTASelect2 and PeptideProphet can be used to filter the identification results to the desired FDR level [113–115]. For a typical whole lysate proteomics analysis, a protein level FDR of 1–5% is typically used.

After filtering, all the high confident PSMs are then further assembled into proteins. Currently, there are more than a dozen of programs supporting protein assembly from peptide identification. Some of the most common ones include DTASelect2, ProteinProphet, Barista, DBParser, PANORAMICS, PeptideClassifier, and MSNet [116]. These programs employ various mathematical models, including optimistic model, parametric model, nonparametric model, parsimonious model, and graph-based model to solve mainly two problems: the shared peptide problem and the singleton problem. When an identified peptide sequence is shared among multiple proteins, it is hard to determine whether it should be assembled into one or multiple proteins. In a typical proteomics experiment, multiple proteins can be found without any unique PSM. It is crucial to carefully assess all the evidence from shared peptides in order to correctly infer the existence of those proteins. The singleton problem is defined as a protein identified by only one PSM. Theoretically, without other evidence, it is hard to infer proteins from a single PSM. However, the co-existence of other proteins and the known protein–protein interaction networks can sometimes be used to improve the confidence of such one-hit wonders.

Similar to DNA sequencing and genome assembly, protein sequence can also be assembled through the "overlap → layout → consensus" from digested protein samples. Multiple enzymes, both specific and nonspecific, can be used in combinations to generate peptide sequences with high coverage and large overlap, allowing the assembly of the full protein sequence. This is an especially useful technique for sequencing highly homogenous protein mixtures with large sequence overlap, such as antibodies.

Overall, existing solutions such as DTASelect2 and ProteinProphet can efficiently infer proteins from peptide identification results for most of the highly abundant or highly enriched proteins. On the other hand, a significant portion of proteins, especially the low abundant ones, are still difficult to be identified with confidence. Sophisticated statistical models and circumstantial evidence borrowed from existing genomic and protein interaction data can be further explored in the future to improve these results.

References

1 Wolters, D.A., Washburn, M.P., and Yates, J.R. III (2001). An automated multidimensional protein identification technology for shotgun proteomics. *Anal. Chem.* 73: 5683–5690.
2 Yates, J.R. III (1998). Mass spectrometry and the age of the proteome. *J. Mass Spectrom.* 33: 1–19.
3 Zhang, Y., Fonslow, B.R., Shan, B. et al. (2013). Protein analysis by shotgun/bottom-up proteomics. *Chem. Rev.* 113: 2343–2394.
4 Yates, J.R. III (2013). The revolution and evolution of shotgun proteomics for large-scale proteome analysis. *J. Am. Chem. Soc.* 135: 1629–1640.
5 Catherman, A.D., Skinner, O.S., and Kelleher, N.L. (2014). Top down proteomics: facts and perspectives. *Biochem. Biophys. Res. Commun.* 445: 683–693.
6 Toby, T.K., Fornelli, L., and Kelleher, N.L. (2016). Progress in top-down proteomics and the analysis of proteoforms. *Annu. Rev. Anal. Chem. (Palo Alto, Calif.)* 9: 499–519.
7 Tran, J.C., Zamdborg, L., Ahlf, D.R. et al. (2011). Mapping intact protein isoforms in discovery mode using top-down proteomics. *Nature* 480: 254–258.
8 Yoshida, T. (2004). Peptide separation by hydrophilic-interaction chromatography: a review. *J. Biochem. Bioph. Methods* 60: 265–280.
9 Di Palma, S., Hennrich, M.L., Heck, A.J., and Mohammed, S. (2012). Recent advances in peptide separation by multidimensional liquid chromatography for proteome analysis. *J. Proteomics* 75: 3791–3813.
10 Yates, J.R. III (2015). Pivotal role of computers and software in mass spectrometry – SEQUEST and 20 years of tandem MS database searching. *J. Am. Soc. Mass. Spectrom.* 26: 1804–1813.
11 Schubert, O.T., Rost, H.L., Collins, B.C. et al. (2017). Quantitative proteomics: challenges and opportunities in basic and applied research. *Nat. Protoc.* 12: 1289–1294.
12 Chandramouli, K. and Qian, P.Y. (2009). Proteomics: challenges, techniques and possibilities to overcome biological sample complexity. *Hum. Genomics Proteomics* 2009: 239204.
13 Wessel, D. and Flugge, U.I. (1984). A method for the quantitative recovery of protein in dilute solution in the presence of detergents and lipids. *Anal. Biochem.* 138: 141–143.
14 Simpson, D.M. and Beynon, R.J. (2010). Acetone precipitation of proteins and the modification of peptides. *J. Proteome Res.* 9: 444–450.
15 Simpson, R.J. (2006). Precipitation of proteins by organic solvents. *Cold Spring Harbor Protoc.* 2006.
16 Link, A.J. and LaBaer, J. (2011). Trichloroacetic acid (TCA) precipitation of proteins. *Cold Spring Harbor Protoc.* 2011: 993–994.

17 Klose, J., Willers, I., Singh, S., and Goedde, H.W. (1983). Two-dimensional electrophoresis of soluble and structure-bound proteins from cultured human fibroblasts and hair root cells: qualitative and quantitative variation. *Hum. Genet.* 63: 262–267.

18 Klose, J. and Kobalz, U. (1995). Two-dimensional electrophoresis of proteins: an updated protocol and implications for a functional analysis of the genome. *Electrophoresis* 16: 1034–1059.

19 Kimura, Y. and Hirano, H. (2004). Two dimensional gel electrophoresis-based proteomics. *Tanpakushitsu Kakusan Koso* 49: 1866–1870.

20 Chevalier, F., Rofidal, V., and Rossignol, M. (2007). Visible and fluorescent staining of two-dimensional gels. *Methods Mol. Biol.* 355: 145–156.

21 Rabilloud, T., Chevallet, M., Luche, S., and Lelong, C. (2008). Fully denaturing two-dimensional electrophoresis of membrane proteins: a critical update. *Proteomics* 8: 3965–3973.

22 Rabilloud, T. (2009). Membrane proteins and proteomics: love is possible, but so difficult. *Electrophoresis* 30 (Suppl. 1): S174–S180.

23 Li, B., Chang, J., Chu, Y. et al. (2012). Membrane proteomic analysis comparing squamous cell lung cancer tissue and tumour-adjacent normal tissue. *Cancer Lett.* 319: 118–124.

24 Santoni, V., Molloy, M., and Rabilloud, T. (2000). Membrane proteins and proteomics: un amour impossible? *Electrophoresis* 21: 1054–1070.

25 Rabilloud, T. (2003). Membrane proteins ride shotgun. *Nat. Biotechnol.* 21: 508–510.

26 Kashino, Y. (2003). Separation methods in the analysis of protein membrane complexes. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* 797: 191–216.

27 Smolders, K., Lombaert, N., Valkenborg, D. et al. (2015). An effective plasma membrane proteomics approach for small tissue samples. *Sci. Rep.* 5: 10917.

28 Pankow, S., Bamberger, C., Calzolari, D. et al. (2015). ΔF508 CFTR interactome remodelling promotes rescue of cystic fibrosis. *Nature* 528: 510–516.

29 Bamberger, C., Pankow, S., Park, S.K., and Yates, J.R. III (2014). Interference-free proteome quantification with MS/MS-based isobaric isotopologue detection. *J. Proteome Res.* 13: 1494–1501.

30 Lee, Y.H., Tan, H.T., and Chung, M.C. (2010). Subcellular fractionation methods and strategies for proteomics. *Proteomics* 10: 3935–3956.

31 Cox, B. and Emili, A. (2006). Tissue subcellular fractionation and protein extraction for use in mass-spectrometry-based proteomics. *Nat. Protoc.* 1: 1872–1878.

32 Larance, M. and Lamond, A.I. (2015). Multidimensional proteomics for cell biology. *Nat. Rev. Mol. Cell Biol.* 16: 269–280.

33 Baghirova, S., Hughes, B.G., Hendzel, M.J., and Schulz, R. (2015). Sequential fractionation and isolation of subcellular proteins from tissue or cultured cells. *MethodsX* 2: 440–445.

34 Zhang, J., Li, X., Mueller, M. et al. (2008). Systematic characterization of the murine mitochondrial proteome using functionally validated cardiac mitochondria. *Proteomics* 8: 1564–1575.

35 Lau, E., Cao, Q., Ng, D.C. et al. (2016). A large dataset of protein dynamics in the mammalian heart proteome. *Sci. Data* 3: 160015.

36 Bousette, N., Kislinger, T., Fong, V. et al. (2009). Large-scale characterization and analysis of the murine cardiac proteome. *J. Proteome Res.* 8: 1887–1901.

37 Chakravarti, B., Oseguera, M., Dalal, N. et al. (2008). Proteomic profiling of aging in the mouse heart: altered expression of mitochondrial proteins. *Arch. Biochem. Biophys.* 474: 22–31.

38 Gao, Y., Chen, Y., Zhan, S. et al. (2017). Comprehensive proteome analysis of lysosomes reveals the diverse function of macrophages in immune responses. *Oncotarget* 8: 7420–7440.

39 Gannon, J., Bergeron, J.J., and Nilsson, T. (2011). Golgi and related vesicle proteomics: simplify to identify. *Cold Spring Harbor Perspect. Biol.* 3: a005421.

40 Bell, A.W., Ward, M.A., Blackstock, W.P. et al. (2001). Proteomics characterization of abundant Golgi membrane proteins. *J. Biol. Chem.* 276: 5152–5165.

41 Smirle, J., Au, C.E., Jain, M. et al. (2013). Cell biology of the endoplasmic reticulum and the Golgi apparatus through proteomics. *Cold Spring Harbor Perspect. Biol.* 5: a015073.

42 Taylor, R.S., Wu, C.C., Hays, L.G. et al. (2000). Proteomics of rat liver Golgi complex: minor proteins are identified through sequential fractionation. *Electrophoresis* 21: 3441–3459.

43 Humphrey, S.J., James, D.E., and Mann, M. (2015). Protein phosphorylation: a major switch mechanism for metabolic regulation. *Trends Endocrinol. Metab.* 26: 676–687.

44 Andersson, L. and Porath, J. (1986). Isolation of phosphoproteins by immobilized metal (Fe^{3+}) affinity chromatography. *Anal. Biochem.* 154: 250–254.

45 Posewitz, M.C. and Tempst, P. (1999). Immobilized gallium(III) affinity chromatography of phosphopeptides. *Anal. Chem.* 71: 2883–2892.

46 Han, G., Ye, M., and Zou, H. (2008). Development of phosphopeptide enrichment techniques for phosphoproteome analysis. *Analyst* 133: 1128–1138.

47 Ficarro, S.B., McCleland, M.L., Stukenberg, P.T. et al. (2002). Phosphoproteome analysis by mass spectrometry and its application to *Saccharomyces cerevisiae*. *Nat. Biotechnol.* 20: 301–305.

48 Ballif, B.A., Carey, G.R., Sunyaev, S.R., and Gygi, S.P. (2008). Large-scale identification and evolution indexing of tyrosine phosphorylation sites from murine brain. *J. Proteome Res.* 7: 311–318.

49 Rush, J., Moritz, A., Lee, K.A. et al. (2005). Immunoaffinity profiling of tyrosine phosphorylation in cancer cells. *Nat. Biotechnol.* 23: 94–101.

50 Sugiyama, Y., Katayama, S., Kameshita, I. et al. (2015). Expression and phosphorylation state analysis of intracellular protein kinases using Multi-PK antibody and Phos-tag SDS-PAGE. *MethodsX* 2: 469–474.

51 Delom, F. and Chevet, E. (2006). Phosphoprotein analysis: from proteins to proteomes. *Proteome Sci.* 4: 15.

52 Bause, E. (1983). Structural requirements of N-glycosylation of proteins. Studies with proline peptides as conformational probes. *Biochem. J* 209: 331–336.

53 Pan, S., Chen, R., Aebersold, R., and Brentnall, T.A. (2011). Mass spectrometry based glycoproteomics—from a proteomics perspective. *Mol. Cell. Proteomics* 10: R110.003251.

54 Yang, Y., Franc, V., and Heck, A.J.R. (2017). Glycoproteomics: a balance between high-throughput and in-depth analysis. *Trends Biotechnol.* 35: 598–609.

55 Gonzalez-Begne, M., Lu, B., Liao, L. et al. (2011). Characterization of the human submandibular/sublingual saliva glycoproteome using lectin affinity chromatography coupled to multidimensional protein identification technology. *J. Proteome Res.* 10: 5031–5046.

56 Mechref, Y., Madera, M., and Novotny, M.V. (2008). Glycoprotein enrichment through lectin affinity techniques. *Methods Mol. Biol.* 424: 373–396.

57 Zhang, H., Li, X.J., Martin, D.B., and Aebersold, R. (2003). Identification and quantification of N-linked glycoproteins using hydrazide chemistry, stable isotope labeling and mass spectrometry. *Nat. Biotechnol.* 21: 660–666.

58 Hagglund, P., Bunkenborg, J., Elortza, F. et al. (2004). A new strategy for identification of N-glycosylated proteins and unambiguous assignment of their glycosylation sites using HILIC enrichment and partial deglycosylation. *J. Proteome Res.* 3: 556–566.

59 Yeh, C.H., Chen, S.H., Li, D.T. et al. (2012). Magnetic bead-based hydrophilic interaction liquid chromatography for glycopeptide enrichments. *J. Chromatogr. A* 1224: 70–78.

60 Monzo, A., Bonn, G.K., and Guttman, A. (2007). Boronic acid-lectin affinity chromatography. 1. Simultaneous glycoprotein binding with selective or combined elution. *Anal. Bioanal.Chem.* 389: 2097–2102.

61 Boersema, P.J., Divecha, N., Heck, A.J., and Mohammed, S. (2007). Evaluation and optimization of ZIC-HILIC-RP as an alternative MudPIT strategy. *J. Proteome Res.* 6: 937–946.

62 Di Palma, S., Boersema, P.J., Heck, A.J., and Mohammed, S. (2011). Zwitterionic hydrophilic interaction liquid chromatography (ZIC-HILIC and ZIC-cHILIC) provide high resolution separation and increase sensitivity in proteome analysis. *Anal. Chem.* 83: 3440–3447.

63 Huttlin, E.L., Bruckner, R.J., Paulo, J.A. et al. (2017). Architecture of the human interactome defines protein communities and disease networks. *Nature* 545: 505–509.

64 Huttlin, E.L., Ting, L., Bruckner, R.J. et al. (2015). The BioPlex network: a systematic exploration of the human interactome. *Cell* 162: 425–440.

65 Hein, M.Y., Hubner, N.C., Poser, I. et al. (2015). A human interactome in three quantitative dimensions organized by stoichiometries and abundances. *Cell* 163: 712–723.

66 Morris, J.H., Knudsen, G.M., Verschueren, E. et al. (2014). Affinity purification-mass spectrometry and network analysis to understand protein-protein interactions. *Nat. Protoc.* 9: 2539–2554.

67 Mehta, V. and Trinkle-Mulcahy, L. (2016). Recent advances in large-scale protein interactome mapping. *F1000Research* 5: 782.

68 Lu, S., Fan, S.B., Yang, B. et al. (2015). Mapping native disulfide bonds at a proteome scale. *Nat. Methods* 12: 329–331.

69 Hsu, J.L., Huang, S.Y., Chow, N.H., and Chen, S.H. (2003). Stable-isotope dimethyl labeling for quantitative proteomics. *Anal. Chem.* 75: 6843–6852.

70 Boersema, P.J., Raijmakers, R., Lemeer, S. et al. (2009). Multiplex peptide stable isotope dimethyl labeling for quantitative proteomics. *Nat. Protoc.* 4: 484–494.

71 Ross, P.L., Huang, Y.N., Marchese, J.N. et al. (2004). Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. *Mol. Cell. Proteomics* 3: 1154–1169.

72 Thompson, A., Schafer, J., Kuhn, K. et al. (2003). Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal. Chem.* 75: 1895–1904.

73 Yang, B., Wu, Y.J., Zhu, M. et al. (2012). Identification of cross-linked peptides from complex samples. *Nat. Methods* 9: 904–906.

74 Chen, C.L. and Perrimon, N. (2017). Proximity-dependent labeling methods for proteomic profiling in living cells. *Wiley Interdiscip. Rev.: Dev. Biol.* 6: e272.

75 Kim, D.I. and Roux, K.J. (2016). Filling the void: proximity-based labeling of proteins in living cells. *Trends Cell Biol.* 26: 804–817.

76 Boja, E.S. and Fales, H.M. (2001). Overalkylation of a protein digest with iodoacetamide. *Anal. Chem.* 73: 3576–3582.

77 Liu, F., Rijkers, D.T., Post, H., and Heck, A.J. (2015). Proteome-wide profiling of protein assemblies by cross-linking mass spectrometry. *Nat. Methods* 12: 1179–1184.

78 Rappsilber, J. (2011). The beginning of a beautiful friendship: cross-linking/mass spectrometry and modelling of proteins and multi-protein complexes. *J. Struct. Biol.* 173: 530–540.

79 Trakselis, M.A., Alley, S.C., and Ishmael, F.T. (2005). Identification and mapping of protein–protein interactions by a combination of cross-linking, cleavage, and proteomics. *Bioconjugate Chem.* 16: 741–750.

80 Petrotchenko, E.V. and Borchers, C.H. (2010). Crosslinking combined with mass spectrometry for structural proteomics. *Mass Spectrom. Rev.* 29: 862–876.

81 Dorn, G., Leitner, A., Boudet, J. et al. (2017). Structural modeling of protein–RNA complexes using crosslinking of segmentally isotope-labeled RNA and MS/MS. *Nat. Methods* 14: 487–490.

82 Rhee, H.W., Zou, P., Udeshi, N.D. et al. (2013). Proteomic mapping of mitochondria in living cells via spatially restricted enzymatic tagging. *Science* 339: 1328–1331.

83 Martell, J.D., Deerinck, T.J., Sancak, Y. et al. (2012). Engineered ascorbate peroxidase as a genetically encoded reporter for electron microscopy. *Nat. Biotechnol.* 30: 1143–1148.

84 Lam, S.S., Martell, J.D., Kamer, K.J. et al. (2015). Directed evolution of APEX2 for electron microscopy and proximity labeling. *Nat. Methods* 12: 51–54.

85 Choi-Rhee, E., Schulman, H., and Cronan, J.E. (2004). Promiscuous protein biotinylation by *Escherichia coli* biotin protein ligase. *Protein Sci.* 13: 3043–3050.

86 Cronan, J.E. (2005). Targeted and proximity-dependent promiscuous protein biotinylation by a mutant *Escherichia coli* biotin protein ligase. *J. Nutr. Biochem.* 16: 416–418.

87 Kim, D.I., Jensen, S.C., Noble, K.A. et al. (2016). An improved smaller biotin ligase for BioID proximity labeling. *Mol. Biol. Cell* 27: 1188–1196.

88 Roux, K.J., Kim, D.I., Raida, M., and Burke, B. (2012). A promiscuous biotin ligase fusion protein identifies proximal and interacting proteins in mammalian cells. *J. Cell Biol.* 196: 801–810.

89 Bar, D.Z., Atkatsh, K., Tavarez, U. et al. (2018). Biotinylation by antibody recognition-a method for proximity labeling. *Nat. Methods* 15: 127–133.

90 Sweet, S.M., Bailey, C.M., Cunningham, D.L. et al. (2009). Large scale localization of protein phosphorylation by use of electron capture dissociation mass spectrometry. *Mol. Cell. Proteomics* 8: 904–912.

91 Boersema, P.J., Mohammed, S., and Heck, A.J. (2009). Phosphopeptide fragmentation and analysis by mass spectrometry. *J. Mass Spectrom.* 44: 861–878.

92 Yates, J.R. III, Morgan, S.F., Gatlin, C.L. et al. (1998). Method to compare collision-induced dissociation spectra of peptides: potential for library searching and subtractive analysis. *Anal. Chem.* 70: 3557–3565.

93 Lam, H., Deutsch, E.W., Eddes, J.S. et al. (2007). Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* 7: 655–667.

94 Stein, S.E. and Scott, D.R. (1994). Optimization and testing of mass spectral library search algorithms for compound identification. *J. Am. Soc. Mass. Spectrom.* 5: 859–866.

95 Griss, J. (2016). Spectral library searching in proteomics. *Proteomics* 16: 729–740.

96 Marx, H., Lemeer, S., Schliep, J.E. et al. (2013). A large synthetic peptide and phosphopeptide reference library for mass spectrometry-based proteomics. *Nat. Biotechnol.* 31: 557–564.

97 Zolg, D.P., Wilhelm, M., Schnatbaum, K. et al. (2017). Building ProteomeTools based on a complete synthetic human proteome. *Nat. Methods* 14: 259–262.

98 Ting, Y.S., Egertson, J.D., Payne, S.H. et al. (2015). Peptide-centric proteome analysis: an alternative strategy for the analysis of tandem mass spectrometry data. *Mol. Cell. Proteomics* 14: 2301–2307.

99 Ning, Z., Zhang, X., Mayne, J., and Figeys, D. (2016). Peptide-centric approaches provide an alternative perspective to re-examine quantitative proteomic data. *Anal. Chem.* 88: 1973–1978.

100 Ting, Y.S., Egertson, J.D., Bollinger, J.G. et al. (2017). PECAN: library-free peptide detection for data-independent acquisition tandem mass spectrometry data. *Nat. Methods* 14: 903–908.

101 Ong, S.E., Blagoev, B., Kratchmarova, I. et al. (2002). Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics* 1: 376–386.

102 McClatchy, D.B., Dong, M.Q., Wu, C.C. et al. (2007). 15N metabolic labeling of mammalian tissue with slow protein turnover. *J. Proteome Res.* 6: 2005–2010.

103 Filiou, M.D., Varadarajulu, J., Teplytska, L. et al. (2012). The 15N isotope effect in *Escherichia coli*: a neutron can make the difference. *Proteomics* 12: 3121–3128.

104 McAlister, G.C., Nusinow, D.P., Jedrychowski, M.P. et al. (2014). MultiNotch MS3 enables accurate, sensitive, and multiplexed detection of differential expression across cancer cell line proteomes. *Anal. Chem.* 86: 7150–7158.

105 Liu, H., Sadygov, R.G., and Yates, J.R. III (2004). A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.* 76: 4193–4201.

106 Griffin, N.M., Yu, J., Long, F. et al. (2010). Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis. *Nat. Biotechnol.* 28: 83–89.

107 Rappsilber, J., Ryder, U., Lamond, A.I., and Mann, M. (2002). Large-scale proteomic analysis of the human spliceosome. *Genome Res.* 12: 1231–1245.

108 Ishihama, Y., Oda, Y., Tabata, T. et al. (2005). Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Mol. Cell. Proteomics* 4: 1265–1272.

109 Paoletti, A.C., Parmely, T.J., Tomomori-Sato, C. et al. (2006). Quantitative proteomic analysis of distinct mammalian Mediator complexes using normalized spectral abundance factors. *Proc. Natl. Acad. Sci. U.S.A.* 103: 18928–18933.

110 Zhang, Y., Wen, Z., Washburn, M.P., and Florens, L. (2010). Refinements to label free proteome quantitation: how to deal with peptides shared by multiple proteins. *Anal. Chem.* 82: 2272–2281.

111 McIlwain, S., Mathews, M., Bereman, M.S. et al. (2012). Estimating relative abundances of proteins from shotgun proteomics data. *BMC Bioinf.* 13: 308.
112 Al Shweiki, M.R., Monchgesang, S., Majovsky, P. et al. (2017). Assessment of label-free quantification in discovery proteomics and impact of technological factors and natural variability of protein abundance. *J. Proteome Res.* 16: 1410–1424.
113 Ma, K., Vitek, O., and Nesvizhskii, A.I. (2012). A statistical model-building perspective to identification of MS/MS spectra with PeptideProphet. *BMC Bioinf.* 13 (Suppl. 16): S1.
114 Tabb, D.L., McDonald, W.H., and Yates, J.R. III (2002). DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *J. Proteome Res.* 1: 21–26.
115 Park, G.W., Hwang, H., Kim, K.H. et al. (2016). Integrated proteomic pipeline using multiple search engines for a proteogenomic study with a controlled protein false discovery rate. *J. Proteome Res.* 15: 4082–4090.
116 Huang, T., Wang, J., Yu, W., and He, Z. (2012). Protein inference: a review. *Briefings Bioinf.* 13: 586–614.
117 Makarov A. (2000), Electrostatic Axially Harmonic Orbital Trapping: A High-Performance Technique of Mass Analysis. *Anal. Chem.* 72(6), 1156–1162.