Mass Spectrometry-based Methods of Proteome Analysis

Boris L. Zybailov and Michael P. Washburn
Stowers Institute for Medical Research, Kansas City, MO, USA

Encyclopedia of Molecular Cell Biology and Molecular Medicine, 2nd Edition. Volume 8
Edited by Robert A. Meyers.

ISBN: 3-527-30550-5

Keywords

2D Page
Two-dimensional polyacrylamide gel electrophoresis; a technology that separates intact proteins according to their pI in the first dimension and molecular weight in the second.

MS/MS
A mass spectrometry technique in which precursor ions are selectively fragmented, thus allowing detailed structural information to be obtained.

MudPIT
A multidimensional protein identification technology; a proteomic method based on liquid chromatography of peptide mixtures with subsequent identification by tandem mass spectrometry.

Quantitative Proteomics
A group of methods that allow large-scale quantitative assessment of protein expression and identification.

Proteomics aims to identify, characterize, and map gene functions at the protein level for whole cells or organisms. A typical experimental scheme for a large-scale proteomics inquiry involves fractionation of a complex protein mixture by electrophoretic or chromatographic means followed by subsequent identification of the components in the individual fractions by mass spectrometry. Owing to

continuous and rapid improvement in instrument sensitivity, throughput capacity, software versatility, and techniques of statistical validation of the data, mass spectrometry–based approaches are becoming mainstream methods in a proteome analysis.

1 Principles and Instrumentation

1.1 Proteome and Proteomics

The term *proteome* was first introduced in mid-1990s to name the functional complement of a genome. By analogy with *genomics*, the term *proteomics* refers to studies of a gene's function at the protein level. Both these terms have a large-scale flavor to them. Indeed, studies that fall under the category ''proteomics'' frequently deal with large-scale analyses of proteins on the level of the whole organism, tissue, cell, or subcellular compartments.

Examples of typical biological questions addressed by modern proteomics experiments include but are not limited to establishing ''news of difference'' between healthy and pathological states, monitoring global gene expression during growth and development, establishing cellular localizations of a particular subset of an expressed genome, and identifying networks of protein–protein interactions. In fact, a recent review classifies proteomics studies according to the type of the addressed biological questions into ''profiling proteomics,'' ''functional proteomics,'' and ''structural proteomics.'' Profiling proteomics amounts to large-scale identification of the proteins in a cell or tissue present at a certain physiological conditions. Functional proteomics refers to the studies of functional characteristics of proteins – posttranslational modifications, protein–protein interactions, and cellular localizations. Structural proteomics includes studies of protein tertiary structure, typically made by a combination of X-ray, NMR, and computational techniques. Adopting this classification, this chapter primarily focuses on the profiling and functional proteomics.

1.2 Need for Large-scale Analyses of Gene Products at the Protein Level

Rapid success of the various genome-sequencing programs is one of the major factors that led to the development of large-scale proteomics methods. In fact, a search of the genome-derived sequence databases is usually an intrinsic part of mass spectrometry–based proteome analysis. While of tremendous value, genomic information is, in principle, insufficient for understanding the complex processes of cellular function. Indeed, in addition to the projection of the information from genes into proteins, cells of the living organisms must metabolically extract useful information from the environment. Therefore, the total informational content necessarily increases in a proteome compared to the corresponding genome.

An immediate consequence of this increase in the informational content is that all of the following – protein isoforms, protein posttranslational modifications, protein conformational states, protein abundance levels, as well as dynamical

changes of these properties – have their own important functional implications in the living cells. In certain cases, genomic sequence still can be used to assess some of the functional properties of the corresponding proteins. For example, the codon adaptation index (CAI) and codon bias can be used to predict the expression level of genes. Another important case is a global sequence-based prediction of protein–protein interactions, an example of which is discussed in Sect. 3 of this chapter.

Also, to some degree, the functional properties of proteins can be assessed by analysis of the mRNA transcripts. Specifically, mRNA expression profiles are frequently used to estimate the corresponding protein levels. Most common methods that are used to measure global mRNA expression are cDNA microarray and serial analysis of gene expression (SAGE). However, in many cases, the correlation between mRNA and protein is insufficient to quantitatively predict protein expression. Potential reasons for these discrepancies in mRNA and protein levels include the differences in half-lives of proteins and mRNAs and posttranslational mechanisms that control the rates of translation. Thus, neither DNA nor mRNA sequence information is sufficient for understanding of the cellular functions. This fact provides motivation to improve the old and develop new technologies for the large-scale analyses of gene products at the protein level.

1.3 General Problems in Proteome Analyses

In a given organism, there are several times more different proteins than there are genes. Estimates for humans give such numbers as about 30 000 different genes and more than 100 000 different proteins. Factors such as posttranslational modifications, isoforms, and expression levels increase the complexity and interconnectedness of a proteome compared to a genome. Problems associated with modern proteomics methods include (1) difficulties in the detection of low-abundant proteins; (2) difficulties in obtaining quantitative information; (3) biases in a proteome coverage; (4) difficulties in characterization of posttranslational modifications; (5) difficulties in data analysis and interpretation in the large-scale experiments; (6) poor reproducibility.

1. Protein levels as low as several dozens copies per cell can be of functional significance. This is true for certain receptors, signaling proteins, and regulatory proteins. However, detection of the low-abundance proteins is often an issue in proteomics experiments. The ability of a proteomics method to identify low-abundant proteins in the presence of high-abundant ones is characterized by the dynamic range of a method. The dynamic range of a proteomics method is defined as the ratio of concentration of the most abundant to the concentration of the least abundant proteins identified by a method. Proteomics methods based on two-dimensional polyacrylamide gel-electrophoresis (2D PAGE) separations typically have a lower dynamic range than the methods based on affinity separation or multidimensional liquid chromatography.
2. Quantitative information on the protein abundance in different physiological contexts is the goal of comparative proteomics studies. 2D PAGE methods are good at solving these types of problems, when combined with scintillation

counting or fluorescence imaging spectroscopy. In the MS-based methods, quantitation is achieved by differential labeling (discussed in detail in Sect. 2).

3. Biases in proteome coverage can be inherent to a particular method, or be related to the design of an experimental scheme. For example, 2D PAGE methods are biased against highly hydrophobic proteins. At the same time, there is no inherent bias against hydrophobic proteins in MS identification. However, sample enrichment or fractionation steps necessarily preceding MS analysis are often biased against one or the other class of proteins.
4. Information on types and degrees of posttranslational modifications and protein isoforms is a necessary component of comprehensive functional description of a proteome. Unfortunately, it is difficult to design a proteomics method that would simultaneously give both good proteome coverage and identify all posttranslational modifications. Methods, such as 2D PAGE, which separate intact proteins, are good at the detection of different posttranslational modifications and isoforms present at the same time. However, owing to large diversity in protein masses and shapes, biases in proteome coverage are inherent in the intact protein methods.
5. During a large-scale proteome analysis, thousands of data points are generated. Hence, there are inherent problems related to the data reduction and extraction of the useful information. Additionally, the obtained information needs to be presented in an accessible format to the scientific community. These issues often require significant computational support and software development.
6. Most of the high-throughput, large-scale methods suffer from poor reproducibility. This is especially true in the case of 2D PAGE and MS-based schemes with several fractionation and separation steps involved. Possible ways to improve reproducibility include reduction in the number of separation and fractionation steps, adhering to standard protocols, and use of automation whenever possible.

Most of these challenges, however, are merely technical limitations. Modern technologies continuously progress toward higher sensitivity, higher range of proteome coverage, and higher reproducibility. Also, even if it may be impossible to have 100% proteome coverage with a single method, proper design of a proteomics program that takes advantage of several different methods can achieve impressive results.

1.4 MS Instrumentation

Mass spectrometry (MS) is a platform technology for all proteomics. It is the mass spectrometer that is used to identify the protein present in samples. Furthermore, there are a variety of mass spectrometry approaches that can be used to achieve this end. Figure 1 illustrates the general concept of MS analysis in proteomics. Generally, MS analysis involves creation of gaseous ions from an analyte followed by separation of the produced ions according to their mass-to-charge ratio (m/z). Instruments that perform analysis of this type are called *mass spectrometers*. Principle components of a mass spectrometer are an ion source, where ionization takes place; a mass analyzer, where the m/z is measured; and a detector, where amount

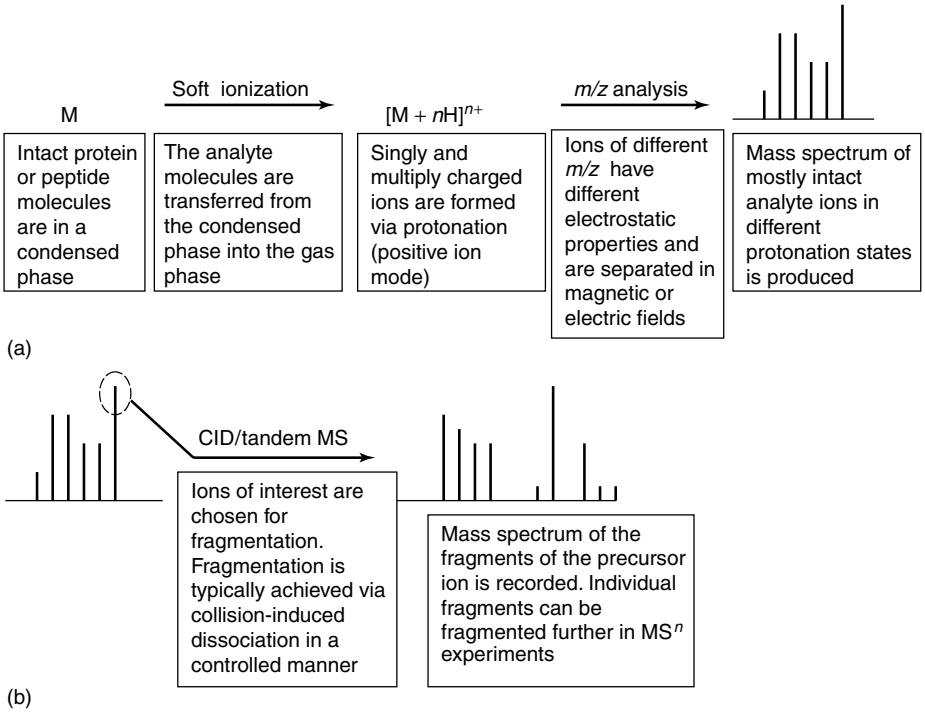


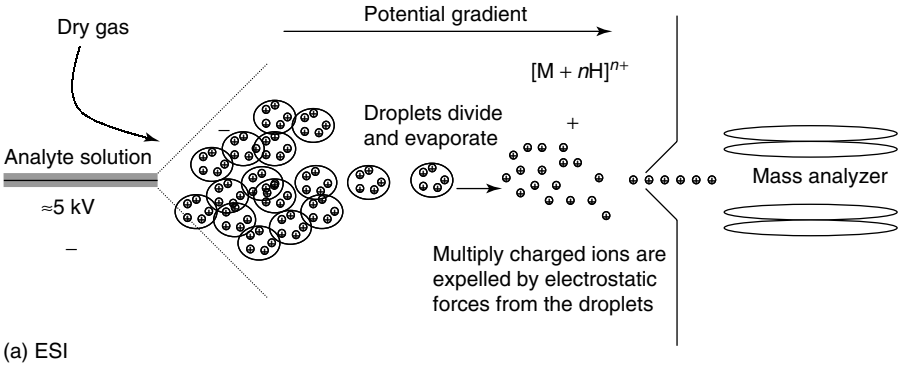
Fig. 1 Principle scheme of single MS (a) and MS/MS (b) analyses. (a) Intact proteins or peptides are transferred from the condensed phase into the gas phase and are ionized through capture (positive mode) or loss (negative mode) of protons. In high-throughput protein identification experiments, positive mode of ionization is used. Negative mode of ionization can be used for analysis of carbohydrates and nucleic acids. Ions of different m/z are discriminated by magnetic or electric fields. (b) Ions whose structure needs to be determined are chosen for fragmentation via CID reactions in the MS/MS experiment. The obtained fragmentation patterns can be searched against protein databases for identification.

of the ions corresponding to a particular m/z is recorded. Large macromolecules, such as proteins, are nonvolatile, and owing to a lack of proper ionization techniques for a long time, MS was limited to the analysis of smaller molecules. Finally, in the 1980s, protein ionization methods were developed, thus making MS-proteomics feasible.

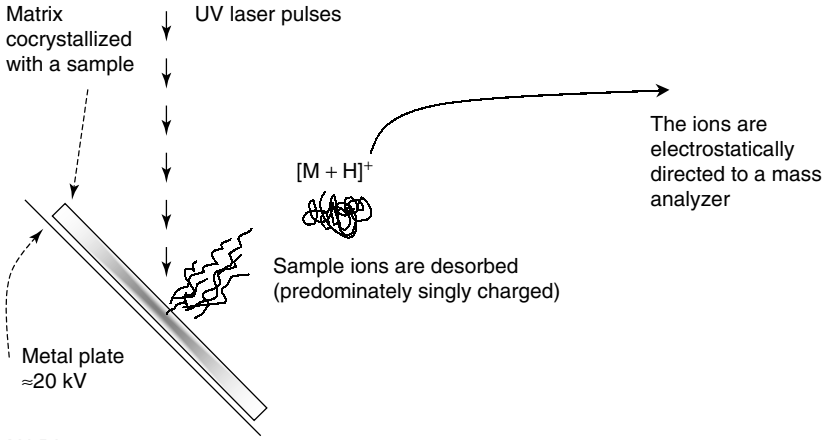
1.4.1 Ionization Methods

The two most commonly used methods of protein ionization are electrospray ionization (ESI) and matrix-assisted laser desorption ionization (MALDI). In both of these techniques, ionization occurs through uptake (positive mode) or loss (negative mode) of one or several protons. In large-scale proteomics studies, the positive mode is typically used. The negative mode of ionization finds its uses in analyses of carbohydrates and nucleic acids. In any of the available ionization methods, there is no definite relationship between the amount of ions formed and the amount of the analyte. This fact makes mass spectrometers inherently nonquantitative devices.

ESI produces mostly multiply charged ions by creating a fine spray of charged droplets in a strong electric field. With the application of a dry gas, the droplets evaporate, and the electrostatic repulsion causes transfer of the analyte ions into the gas phase (Fig. 2a). The multiple charging allows analysis of very large molecules with analyzers that have relatively small m/z range. Another advantage of the multiple charging is that more accurate molecular weight can be obtained from analysis of the distribution of multiply charged peaks. ESI is used in a wide range of proteomics applications but is limited by susceptibility to high salt concentrations and to contaminants in the sample.



(a) ESI



(b) MALDI

Fig. 2 Conceptual schemes of (a) Electrospray ionization (ESI) and (b) Matrix-assisted laser desorption ionization (MALDI). (a) During ESI, a fine spray of charged droplets is created. The droplets evaporate and the multiply-charged ions are expelled by electrostatic forces. ESI is frequently used for analysis of complex peptide mixtures via coupling to multidimensional liquid chromatography. (b) During MALDI, the analyte molecules are ejected by laser pulses from the sample cocrystallized with a matrix. During MALDI, mostly singly charged ions are produced. Because the analyte molecules are ejected in bundles, MALDI method is ideally suited for coupling to TOF mass analyzers. MALDI-TOF instruments are frequently used to analyze individual spots on 2D PAGE.

MALDI produces predominately singly charged ions by aiming laser pulses at a sample cocrystallized with a molecular matrix on a metal plate under high voltage (~20 kV). Typical matrices used in MALDI-assisted protein analysis include α-Cyano-4-hydroxycinnamic acid and 2-(4-Hydroxyphenylazo)-benzoic acid. The laser is tuned to the absorption maximum of the matrix. The sample ions are preformed in the condensed phase in sufficient quantities. The matrix absorbs some of the laser pulse energy, thus minimizing sample damage. The sample ions and matrix molecules gain enough kinetic energy and are ejected into a gas phase (Fig. 2b). Table 1 outlines and compares principle characteristics of these ionization methods.

Importantly, in both the MALDI and ESI sources, ion fragmentation occurs rarely and the analyte molecules remain largely intact. This nondestructiveness is what makes these methods so attractive for characterization of biomolecules. However, in certain applications (e.g. peptide sequencing), it may be necessary to fragment molecular ions (preferably in a predictable manner) in order to extract additional information. This is typically achieved via collision-induced dissociation (CID). Such types of MS analysis are called *tandem MS* (MS/MS) and often are denoted by MS^n, where n is the number of generation of fragment ions analyzed (Fig.1).

1.4.2 Mass Analyzers

Mass analyzers separate ions according to their m/z in electric or magnetic fields. Most common types of mass analyzers used in proteomics research are quadrupole (QD), ion trap (IT), time-of-flight (TOF), and Fourier transform ion cyclotron resonance (FTICR). Often in modern instruments, several analyzers of the same or different types are combined together to achieve maximum performance.

Tab. 1 Comparison of ESI and MALDI sources.

	ESI	*MALDI*
Principle of action	Uses electric field to produce sprays of fine droplets; as the droplets evaporate, ions are formed	Uses laser pulses to desorb and ionize analyte molecules cocrystallized with a matrix on a metal surface
Ions formed	Multiply charged (The larger the analyte molecule, the more likely it acquires multiple charges)	Singly charged
Mass range	>100 kDa	>100 kDa
Resolution	~2500 (with IT/QD mass analyzers)	~10 000 (with TOF mass analyzers and ion reflectors)
Typical application	10^{-15} mole LC/MS of peptide mixtures; tandem MS; protein identification by comparing experimental MS/MS spectra with theoretical MS/MS spectra produced from protein databases	10^{-15} mole Analysis of spots on 2D PAGE; determination of molecular weight; protein identification by ''peptide mass fingerprinting''

QD mass analyzers are frequently used in conjunction with ESI source. They offer moderate resolution (up to 2500 mass units) and moderate sensitivity. QD consists of four parallel rods with a hyperbolic cross section. Diagonally opposite rods are connected to radio frequency and direct-current voltage sources thus establishing the quadrupole field. Ions that are produced at an ion source are electrostatically accelerated into this quadrupole field. Mass-selection is achieved by proportional changes of amplitudes of radio frequency and direct current in a way that for any given pair of these amplitudes there, only the ions of specific m/z reach the detector. QDs are known for their tolerance of high pressures (up to 10^{-4} torr), which makes them attractive to use with ESI for liquid chromatography (LC)-coupled applications.

IT mass analyzers work by confining ions to a small volume via radio-frequently oscillating electric fields. ITs offer moderate sensitivity at a relatively low monetary cost and are good for MS/MS applications. In the latter case, ITs are often used cojointly with QDs. Mass accuracy of regular ITs is rather low, since only limited amount of ions can be accumulated in a small volume. Mass accuracy, as well as resolution can be significantly improved by using two-dimensional ITs (sometimes also called *linear*), which increase the ion storage volume.

TOF mass analyzers measure the time traveled by an ion from a source to a detector. The longer the flight path, the better the resolving power. However, longer flight paths also increase the scan time. In commercial TOF analyzers, compromise is achieved with the length of the flight path on the order of several meters. The resolving power of a simple TOF analyzer is poor – several to ten times less than that of QD. Significant improvement in the resolving power of TOF is offered by additions of one or several ion reflectors. Upon reflection, the velocity distribution of ions at particular m/z narrows, thus increasing resolution and sensitivity. As a result, modern TOF and double-TOF spectrometers can achieve resolving power of 10 000 and more. Measurement of the ion's TOF can be achieved only if the analyte ions are presented to TOF analyzer in discrete bundles. Because of this need for discrete ion bundles, TOF analyzers are particularly suited to the pulsed nature of MALDI. In fact, MALDI-TOF is one of the most common types of mass spectrometers used in proteomics research.

FTICR mass analyzers are based on the resonance absorption of energy by ions that precess in a magnetic field. The recorded array of the precession time-curves is Fourier-transformed to obtain the component frequencies of the different ions. Next, the component frequencies are related to the ion's m/z. Because frequency is a parameter that is easy to measure with high precision and accuracy, FTICR has the highest resolving power amongst MS analyzers – up to 10^6 and more. With FTICR, it is also particularly easy to do MS^n experiments. Unfortunately, current FTICR instruments are cumbersome, expensive, and not readily available.

1.4.3 MS/MS

In some applications, it is necessary to fragment molecular ions produced by ESI or MALDI further to obtain additional structural information. Figure 3 illustrates a possible way to do this with a triple–QD ESI mass spectrometer. To obtain a tandem spectrum, the first quadrupole scans across a set m/z range and selects ions of interest. In the second quadrupole, CID

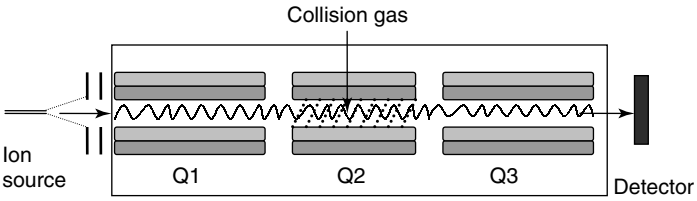


Fig. 3 A triple quadrupole mass spectrometer with MS/MS capability. Q1 selects ion to be fragmented and allows the selected ion to pass into Q2. The ions in Q2 are fragmented via collisions with inert gas (typically argon). Q3 analyzes the fragments generated in Q2. MS^n ($n \leq 4$) experiments can be performed with this setup as well. To achieve this, fragmentation must be increased at the level of ionization. In modern instruments, quadrupole-ion trap mass analyzers are used for MS^n ($n > 8$) experiments.

reaction takes place – ions that were selected by the first quadrupole undergo collisions with argon gas and fragment. Finally, the third quadrupole analyzes the resulted fragments. Importantly, the fragmentation via CID occurs in a predictable manner – protein and peptide ions break mostly at peptide bonds – which makes the large-scale sequencing easier.

1.5 Methods of Sample Fractionation

The large complexity of protein mixtures from biological samples poses additional challenges for proteomics experiments. Depending on the scope of a particular proteomics task, it is desirable to introduce sample enrichment and fractionation steps prior to MS identification. The scope of a particular task can range either from an analysis of a subset of a proteome – for example, analysis of proteins specific to a particular organelle and proteins modified in a certain way – to a simultaneous analysis of all the proteins in the proteome. In both of these cases, a proper choice of separation strategy determines the overall throughput and sensitivity of the proteomics experiment. While it may be desirable to fractionate protein mixture down to individual proteins (aim of 2D PAGE separation), significant gain in the throughput can be achieved when mixtures of proteins are introduced to the mass spectrometer.

MS analysis of intact proteins is impractical in most of the high-throughput tasks with the current instrumentation because of the large range in protein masses. To simplify the measurements, in most of the MS analyses, proteins are enzymatically digested (either individually purified or in mixtures) to produce peptide fragments. The resulted peptide mixtures are often fractionated further. Fractionation of the peptide mixtures is usually achieved via in-liquid chromatography methods.

Isolation of a particular subcellular compartment or an organelle is usually achieved by a combination of centrifugation and solubilization steps. Proteins modified posttranslationally in a specific way (e.g. phosphoproteins) can be isolated by chemical- or immunoaffinity methods (see examples in Sect. 3). Large-scale isolation of protein complexes can be done by the epitope tagging and subsequent immunoaffinity purification (see examples in Sect. 3). Another promising affinity

method is that of protein chips – a method that uses large arrays of antibodies, or other binding factors to isolate proteins of interests.

1.5.1 2D-electrophoretic Separation of Complex Protein Mixtures

2D PAGE separation methods were introduced in mid-1970s and were extensively used for analysis of complex protein mixtures. The principle of 2D PAGE separation is illustrated in Fig. 4. 2D PAGE separates proteins in the first dimension by their pI, and in the second, by their molecular weight. Prior to the development of MS-identification techniques suitable for macromolecules, identification of individual proteins in the gel spots was difficult. Typically, the identification of individual proteins in the gel spots was done by immunostaining methods or by N-terminal degradation sequencing. Nowadays, in most laboratories, analysis of 2D PAGE spots is performed with MS. Typically, the individual spots are excised, digested, and analyzed by MALDI-MS. Protein identification is achieved by peptide mapping – comparison of the observed peptide peak patterns with the predicted digest fragments of proteins in a database. However, if some of the spots on the 2D PAGE are overlapping, the method of peptide mapping can give incorrect results.

As a separation technique, 2D PAGE offers high resolution, and is able to distinguish between different protein isoforms and also between different posttranslational modification

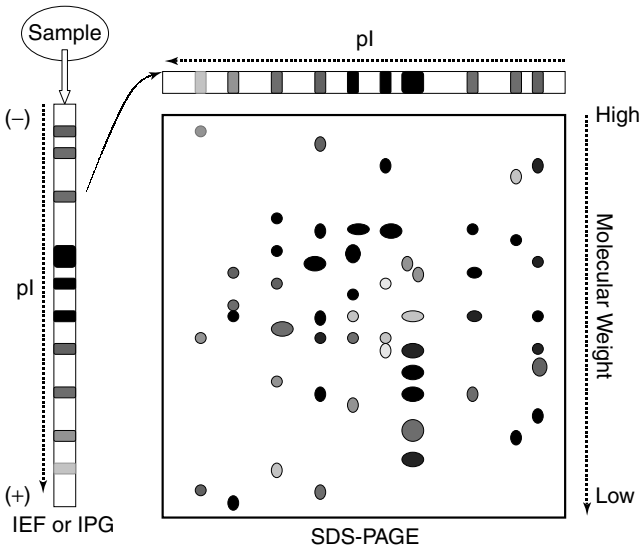


Fig. 4 Principle of the two-dimensional electrophoretic separation of intact proteins. The sample is loaded onto IEF to IPG strip, where proteins are separated according to their isoelectric points (pI). Next, the strip is loaded on SDS-PAGE, where the proteins are separated according to their molecular weights (MW). Visualization of protein spots is achieved via chemical staining methods. Individual spots can be further excised and analyzed by mass spectrometry.

states. However, 2D PAGE is a denaturing technique; hence, it is not suitable for direct analysis of protein complexes and protein–protein interactions. Also, 2D PAGE is known to discriminate against low-abundance and membrane proteins. Other limitations of 2D PAGE include biases against proteins with pIs outside the 2–10 range and biases against heavy proteins (>100 kDa). The dynamic range of 2D PAGE is also limited by resolution of the spot visualization methods. Recently, a number of ways to overcome some of these limitations of 2D PAGE have been developed.

Significant improvement to 2D PAGE was the development of immobilized pH gradients (IPGs), which increased both reproducibility and the load capacity of 2D PAGE. Sample prefractionation into discrete isoelectrical fractions and subsequent analysis by several narrow-range IPGs is yet a further improvement to this technology, which both increases the method's dynamic range and somewhat reduces the biases in proteome coverage. Despite these technological improvements, 2D PAGE remains a time-consuming and labor-intensive technique. Alternative methods of protein separations employing LC exist and can potentially overcome some of the 2D PAGE's limitations in the large-scale proteome analysis.

1.5.2 **Liquid-phase Separation Methods**

An ESI source allows MS characterization of proteins and peptides that are present in solutions and therefore is ideally suited for online coupling with liquid-phase separations. Liquid-phase separation can be achieved by using high performance liquid chromatography (HPLC), capillary isoelectric focusing (CIEF), capillary electrophoresis (CE), or by other methods. When two or more distinct liquid phases are used in a separation, with each of the phases relying on a unique independent physical property of the analyte, this liquid-phase separation is called *multi-dimensional*. Independent physical properties correspond to ''dimensions'' and these can be size, charge, hydrophobicity, or affinity to a particular substrate. The use of several independent dimensions significantly increases resolving power of a separation.

In-liquid separation methods can be used both for the intact proteins and for the peptide mixtures. One promising technique for characterization of global extracts of intact proteins is CIEF-FTICR mass spectrometry. In this method, the CIEF separates intact proteins by pI and subsequent analysis by FTCIR produces a two-dimensional display similar to the one obtained in 2D PAGE. Jensen et al. reported resolution of 400 to 1000 proteins in the mass range of 2–100 kDa from global protein extracts of *Escherichia coli* and *Deinococcus radiodurans*. While of good resolving power and sensitivity, this technique needs further improvement, perhaps by addition of protein fragmentation so that the resolved proteins could be identified.

Wall et al. developed a method that can potentially obtain both intact protein molecular weights as well as sequence information for a large portion of a proteome. In this method, proteins were separated by pI using isoelectric focusing (IEF) in the first dimension and by hydrophobicity using nonporous reversed-phase HPLC in the second dimension (IEF-NP RP HPLC). Next, the fractions that were eluted from HPLC were analyzed in two parallel experiments by (1) MALDI-TOF-MS and (2) ESI-TOF-MS. Analysis of human erythroleukemia cell lysate by this method resulted in resolution and identification of

several hundreds of unique proteins in a mass range from 5 to 85 kDa. Having both intact protein molecular weights as well as sequence information alleviates characterization of protein posttranslational modifications and isoforms. Thus, Wall et al. reported detection of posttranslational modifications of cytosolic actin, heat shock 90 beta, HINT and α-enolase.

The intact protein approaches are usually limited to proteins of smaller sizes, mostly because the resolving power of mass analyzers drops with increase in the ions masses. Also, biases toward one or the other types of intact proteins are unavoidable with any chromatographic separation method. A conceptually different approach, in which purification of intact proteins is avoided, involves proteolytic digestion of protein mixtures. Proteolytic digestion transforms protein mixtures into more uniform (in terms of mass and chromatographic properties) mixtures of peptides. Because of the increased uniformity, analysis of peptide mixtures essentially eliminates biases in proteome coverage. Unfortunately, the absence of information on intact proteins makes characterization of posttranslational modifications more difficult.

Chemical or enzymatic cleavage of proteins into peptides followed by multidimensional liquid-phase separation is typically used in a global proteome survey type of experiments. Generally, proteins from whole cells or tissue homogenates are first enzymatically digested and then the resulting peptide mixtures are separated by electrophoresis or liquid chromatography in a microcapillary format. A microcapillary column is attached directly to the ionization source so that the eluting fractions are ionized and introduced to a mass analyzer. While fragmentation of the peptide ions is optional with MALDI-MS/peptide mapping analysis of 2D PAGE spots, it is required with the liquid-phase/MS analysis of complex peptide mixtures. The sequence information obtained by MS/MS for individual peptide ions is used to match peptides to the corresponding proteins in a database.

Multidimensional protein identification technology (MudPIT) is a good example of a liquid-phase/MS method. In MudPIT, a biphasic SCX/RP microcapillary column is used to fractionate a complex peptide mixture (Fig. 5). The peptide fractions are eluted from the column by series of HPLC gradients into the tandem mass spectrometer. The obtained MS/MS spectra are used to search a sequence database via SEQUEST algorithm. MudPIT was used to detect and identify ~1500 proteins from *Saccaromyces cerevisiae* proteome, ~2500 proteins from the *Oryza sativa* proteome and ~2500 proteins from the *Plasmodium falciparum* proteome.

1.5.3 Affinity Methods (Epitope Tagging)

Affinity-purification methods are frequently employed in those proteomics programs that analyze specific subsets of a proteome, such as phosphorylated or glycosylated proteins. Also, affinity methods are often used in large-scale studies of protein–protein interactions and protein complexes. Affinity methods are based on targeted interactions between proteins and antibodies or between proteins and other protein-specific molecules, immobilized in a column or in the form of an array. For example, to isolate phosphorylated proteins from the rest of the proteome, one can use an immunoaffinity column prepared with antibodies specific to the phosphorylated amino acids.

Epitope tagging is used in large-scale analyses of protein–protein interactions and protein complexes (Fig. 6). In this

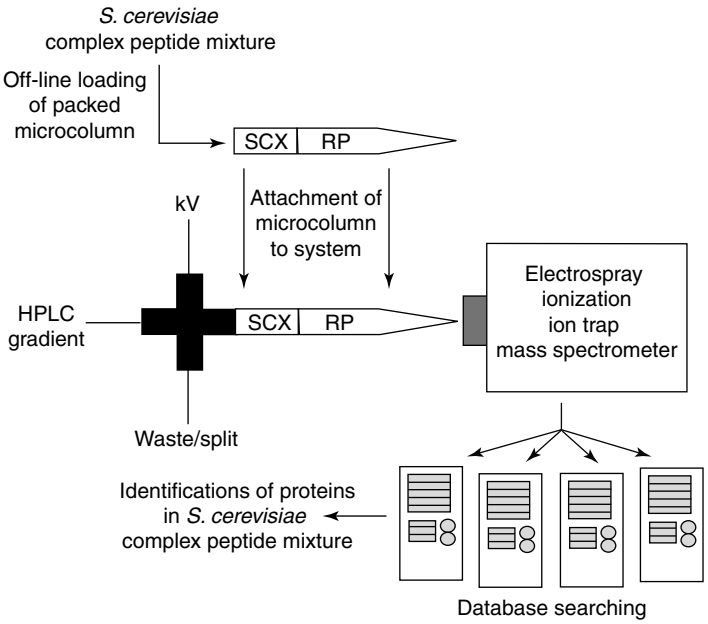


Fig. 5 Multidimensional protein identification technology (MudPIT). Complex peptide mixtures are loaded onto a biphasic microcapillary column packed with strong cation exchange (SCX) and reverse-phase (RP) materials. Peptides directly elute into the tandem mass spectrometer because a voltage (kV) supply is directly interfaced with the microcapillary column. Peptides are first displaced from the SCX to the RP by a salt gradient and are eluted off the RP into the MS/MS. The tandem mass spectra generated are correlated to theoretical mass spectra generated from protein or DNA databases by the SEQUEST algorithm. Figure is reproduced from Washburn, M.P., Wolters, D., Yates, J.R. III (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology, *Nat. Biotechnol.* **19**, 242–247 by permission of Nature Publishing Group.

strategy, proteins to be isolated are fused with a motif recognizable by a specific antibody. This fusion is typically done by incorporating the epitope sequence into C-terminal of genes that encode proteins of interest. Then the fused proteins are expressed and purified along with their interaction partners by the immunoaffinity chromatography using antibodies specific to this particular epitope. If a particular protein is of low abundance, it can be overexpressed to increase the overall recovery.

1.6 Protein Identification by MS

Accurate identification of proteins is the essential requirement of proteomics studies. High complexity of protein mixtures derived from tissues, whole cells, or subcellular compartments adds an additional requirement of high throughput in the identification of proteins. In this section, we review common strategies that match MS data with proteins present in a biological sample.

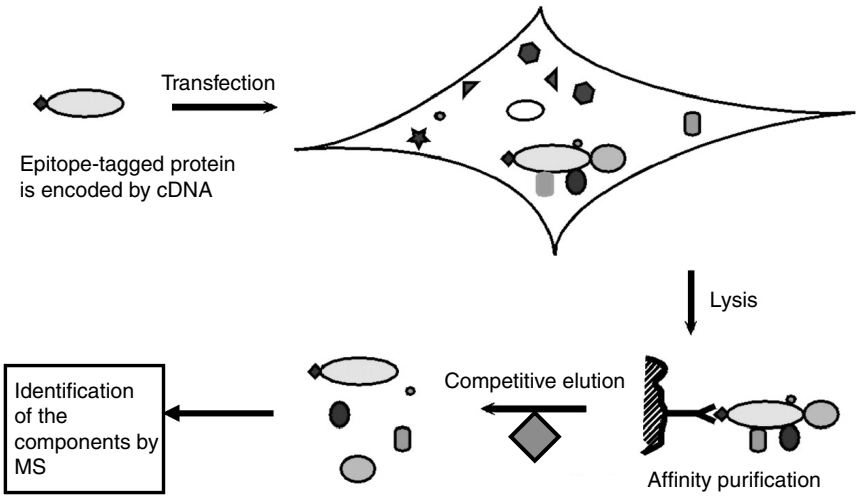


Fig. 6 Principle of affinity purification with epitope tagging. Bait proteins are expressed as fusion with a motif recognizable by a certain antibody (epitope). The cells are lysed, and proteins are purified by immunoaffinity chromatography. Next, the bound fraction is competitively eluted from the column, and the pulled proteins are analyzed by LC/MS or 1D PAGE/MS.

Typically, protein identification amounts to the deduction of the sequence-specific information from MS data followed by searches of the sequence databases. Databases that MS data can be matched against are protein, expressed sequence tag (EST), and genomic databases. Because currently there are no suitable ways to fragment intact proteins inside a mass spectrometer, high-throughput sequencing by MS is possible only if intact proteins are fragmented into peptides first. Depending on a particular type of proteomics scheme and MS instrumentation used, protein sequence information can be obtained by (1) peptide mass fingerprinting; (2) accurate mass tags (AMTs); (3) peptide fragmentation in MS/MS; (4) sequence tags; or by combinations of (1) through (4).

Peptide mass fingerprinting method is usually adopted in those cases in which individual proteins are separated during purification steps, such as in 2D PAGE. In the case of 2D PAGE, individual spots are picked, digested, and masses of the peptides are recorded. Next, the experimentally obtained peptide masses are matched against theoretical peptide libraries generated from protein sequence databases. MS/MS data can also be obtained along with mass fingerprinting, and nowadays this is a common practice.

With MALDI-TOF instrumentation – which is typically used with the peptide mass fingerprinting type of analysis – several peptide masses are needed to unambiguously identify a protein. Using more accurate instrumentation, such as FTICR-MS, it is possible to identify proteins based on the mass of a single peptide, without MS/MS data. In this case, a peptide that uniquely corresponds to a protein is called *accurate mass tag* (AMT). Conrads et al. evaluated utility of AMTs for identification of proteins from *S. cerevisiae* and *Caenorhabditis elegans*. The

authors demonstrated that up to 85% of the predicted tryptic peptides from these two organisms could be used as AMTs at mass accuracies typical of FTICR-MS instruments (~1 ppm). The authors also discussed utility of AMTs with highly accurate mass measurements in detection of phosphorylated proteins. They argued that because mass defect of P is larger than that of H, C, and O, the average mass of phosphopeptides is slightly lower than the mass of unmodified peptides of the same nominal weight; thus enabling the identification of phosphorylated peptides if the mass measurement accuracy is sufficient.

Whenever it is possible to match a single peptide to a protein, it is no longer necessary to purify samples down to individual proteins. Instead, fractions containing mixtures of proteins can be digested with trypsin followed by analysis of the resulted peptide fragments. For example, the above-discussed AMT method can be used for analysis of complex protein mixtures, because it matches single peptide to a unique protein. However, the requirement of high mass accuracy makes uses of AMT limited. Additionally, FTICR instrumentation is cumbersome and is not a widely distributed technology today. Far more superior in terms of proteome coverage per monetary cost are methods that employ peptide sequence analysis by MS/MS. The key fact that enables high-throughput peptide sequencing by MS/MS is predictability of the peptide precursor ion fragmentation in the CID reactions.

Figure 7 shows the adopted nomenclature for the fragment ions series. Ions that form through dissociation of the peptide bonds are the most abundant ones if moderate collision energies are used (30–50 V). The ions that retain their charge on the N-terminal part after fragmentation of the precursor are called *y* ions, and the

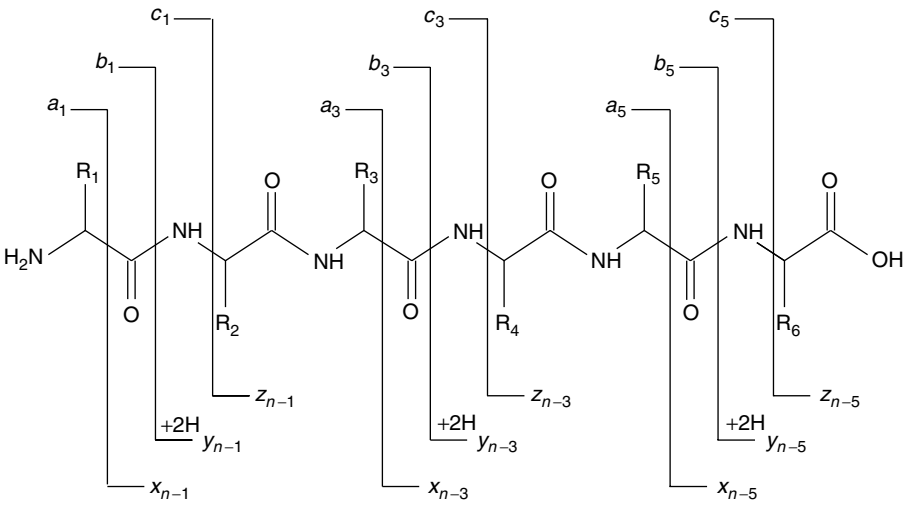


Fig. 7 Nomenclature of the ions formed during peptide fragmentation. In a typical MS/MS experiment via CID reactions, *y* and *b* ions are predominant. These two types are used for computer-assisted peptide identification. Through the loss of CO fragment, *a* ions can be produced from *b* ions. While not important for identification, *a* ion series can be used for independent result validation. Other types of ions are not produced in CID reactions.

ions that retain their charge on the C-terminal part after fragmentation of the precursor are called *b* ions. The subscript to the right of *b* and *y* symbols equals the number of amino acid residues in the corresponding fragment. The difference in m/z values between consecutive ions within a given series corresponds to the difference in the sequences of the two fragments. Because the consecutive ions within a series represent peptide fragments that differ in exactly one amino acid, and each amino acid residue has a unique nominal weight (except I and L), the pattern of m/z values of *y* and *b* ions corresponds to the amino acid sequence of the precursor peptide. Unfortunately, some expected peaks could be missing from the MS/MS. Additionally, experimental spectra can be complicated by unwanted fragmentation. Also, it is not always possible to unambiguously determine from which series a particular ion fragment comes. For these reasons, manual interpretation of MS/MS spectra can be tedious and ineffective. As a consequence, *de novo* sequencing is rarely done in high-throughput proteomics experiments. Instead of direct interpretation of the MS/MS spectra, computer programs that find the best matches to the spectra from a database are used.

Typical example of an algorithm that finds best matches to MS spectra is SEQUEST software package developed by Eng et al. Analysis of MS data by SEQUEST starts with reduction of tandem spectra complexity. Only a certain number of most abundant ions are considered and the rest are discarded. Also, the unfragmented precursor ion is removed from the spectra in order to prevent its misidentification as a fragment. Next, SEQUEST selects sequences from a database. First, SEQUEST creates list of peptides that have masses at or near the mass of the precursor ion. Second, SEQUEST generates virtual MS/MS spectrum for each of the candidate and compares them to the observed spectrum. As the result of this comparison, a cross correlation score is produced (Xcorr). Another score, DeltaCN, reflects the difference in correlation of the second ranked match from the first one. The higher this score is, the more likely it is that the first match is the correct one.

SEQUEST also has the ability to detect posttranslational modifications. In this case, SEQUEST looks for increased masses of amino acids. For example, cysteines are typically carboxymethylated by iodoacetomide prior to enzymatic digestion. In this case, masses of all cysteine residues are considered to increase by 57 Da. To look for amino acids that can be either modified or not modified, the database size is increased by allowing different weights to represent the same amino acid. Unfortunately, only known posttranslational modification can be detected by SEQUEST – the types of modifications that are being looked at need to be explicitly specified in input parameters.

Another useful technique for peptide identification is sequence tag. This technique is implemented in algorithms such as GutenTag. In GutenTag, a partial sequence is inferred directly from the tandem spectrum, and then the database is searched for matches that include this partial sequence and that match the masses on C- and N-sides of the fragment. Importantly, GutenTag method is error-tolerant, and allows detection of unknown posttranslational modifications. A recent variation of this method, MultiTag also allows protein identification from organisms with unsequenced genomes.

2 Quantitative Methods of Proteome Analysis Using MS

Quantitative proteome analysis aims at large-scale identification of differences in protein expression. Two-dimensional difference gel electrophoresis (2D-DIGE) is one of the techniques that can achieve that aim. Recent example of application of this technique is the report by Friedman et al. The authors used 2D-DIGE in combination with MALDI-TOF to analyze the proteome of human colon cancer. However, while acknowledging the importance of the gel-based methods in modern quantitative proteomics, we believe that methods based on multidimensional chromatography have greater potential, and therefore, we limit our further discussion to quantitation within the LC/LC/MS/MS paradigm.

Even though absolute quantification (i.e. clear relationship between the peak intensity and amount of the analyte) is a challenge for modern mass spectrometers, methods of finding relative abundances exist. These methods are based on labeling through mass modifications of the whole proteome or some of its subsets. Labeling through mass modification is particularly suited for quantitative analysis in differential expression profiling, where two states under study are differentially labeled via ''light'' or ''heavy'' mass modifications. Figure 8 schematically depicts a quantitative proteomics approach. Mass modification can be introduced at different steps in the sample preparation – during growth, after growth, during digestion, or after digestion. Labeling at the early stages of sample preparation minimizes losses of the analytes. Instrumentation for quantitative studies is essentially the same as instrumentation for qualitative studies. Nevertheless, quantitative analysis is more laborious and typically achieves lower proteome coverage. Also, quantitative methods in proteomics are still at the stage of development and are not as broadly applied as qualitative methods.

2.1 Metabolic Labeling

Widely used in structural biology to prepare samples for analysis by nuclear magnetic resonance (NMR) analysis, metabolic labeling methods have become useful quantitative proteomics tools. Metabolic labeling is a mass modification method that is done very early in the experiment – at the stage of cell growth. Figure 8 illustrates the principle behind this strategy. The goal is to compare protein abundances in cells grown at different conditions. To do this, the growth media from one of the conditions is enriched in stable low-abundant isotopes. The enrichment can be done either by labeling all amino acids by ^{15}N, or by supplementation with a single labeled amino acid. The processed and digested cell extracts from the two different conditions are combined in a one-to-one ratio and analyzed by liquid-phase/MS/MS methods. In the resultant MS spectra, peaks corresponding to the same peptide from different conditions are offset according to the degree of labeling. The ratio between the two peaks corresponds to the difference in abundances.

The first demonstrations of metabolic labeling in quantitative proteomics used gel electrophoresis and spot excision as the protein isolation method. For example, Oda et al. identified proteins that were altered in expression between two strains of *S. cerevisiae* grown in ^{14}N or ^{15}N media, and determined the phosphorylation levels of a specific protein in the same

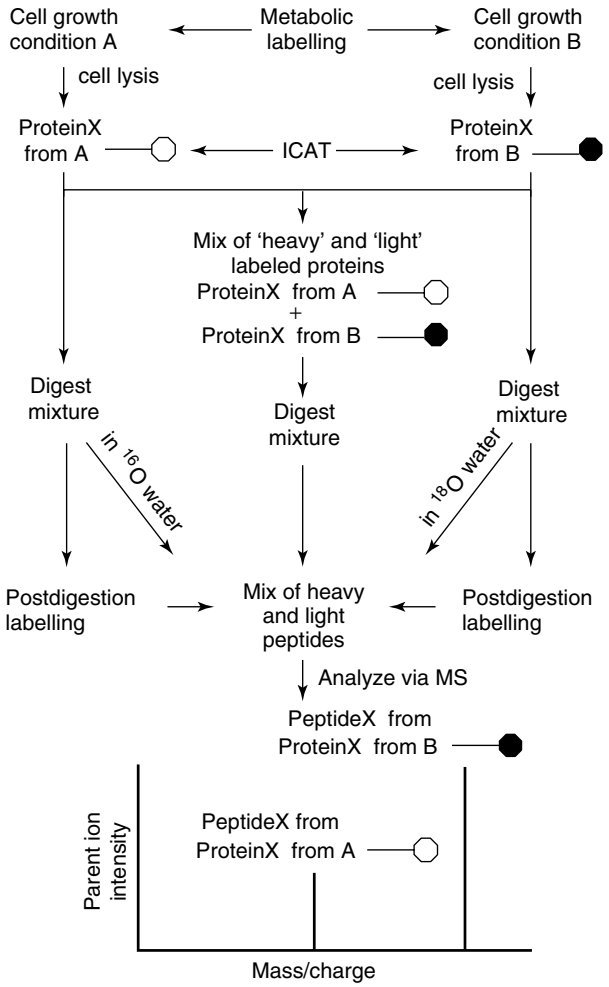


Fig. 8 Quantitative proteomics approach. When carrying out a quantitative proteomic analysis, the key is for the same peptide from two unique growth conditions to have unique masses when being analyzed by a mass spectrometer. ''Heavy'' and ''light'' peptides may be generated at many points in a sample preparation pathway. Metabolic labeling introduces a label during the growth of the organism and is therefore the earliest point of introduction of ''heavy'' and ''light'' labels. Metabolic labeling is followed by ICAT, digestion in ^{16}O and ^{18}O water, and lastly postdigestion labeling. Only after a label has been introduced can the samples be mixed and further processed. Figure and figure legend are reproduced from Washburn, M.P., Ulaszek, R., Deciu, C., Schieltz, D.M., Yates, J.R. III (2002) Analysis of quantitative proteomic data generated via multidimensional protein identification technology, *Anal. Chem.* **74**, 1650–1657 by permission of Analytical Chemistry.

sample. Complete metabolic labeling has been demonstrated using chromatography based proteomics methods on *D. radiodurans*, mouse B16 cells, and on *S. cerevisiae*.

In addition, isotopically enriched single amino acids may be used for the selective metabolic labeling of a cell type for a quantitative proteomic analysis. In *S. cerevisiae*, Jiang et al. have described the single amino acid isotopic enrichment of *S. cerevisiae* with D_{10}-Leu, and Berger et al. described the comparative analysis of *S. cerevisiae* cultured in media containing either ^{13}C-Lys or unlabeled lysine. Stable isotope labeling by amino acids in cell culture (SILAC) was explored as an alternative to ^{15}N labeling. Ong et al. studied mammalian cells grown either with deuterated or nondeuterated leucine. Importantly, parameters such as cell morphology, doubling time, and ability to differentiate did not change in deuterated sample compared to nondeuterated one. The authors used this technique to study changes in protein expression induced by muscle cell differentiation. The authors reported that glyceraldehyde-3-phosphate dehydrogenase, fibronectin, and pyruvate kinase M2 were upregulated. Another approach to metabolic labeling is the rare isotope depletion of the growth media from one of the conditions under study. If the rare isotopes are removed, then one expects m/z distributions to shift to the lighter values. With conventional instrumentation, this could work for large proteins. However, high-resolution instrumentation like FTCIR must be used to analyze peptide mixtures.

2.2 Isotope Coded Affinity Tags

Metabolic labeling with stable isotopes while analytically advantageous to other methods is available only in cases when studied cells or organisms are cultivable. For this reason, metabolic labeling is not suitable for diagnostic and clinical applications. When cultivation or controlling growth is difficult, different strategies need to be used for quantitative proteome analysis. One of the possible approaches is the method that uses cysteine-specific reagents – isotope coded affinity tags (ICATs) – to differentially label proteomics samples. ICATs have three functional elements: cysteine-specific reactive group, isotopically labeled linker, and affinity group (Fig. 9). For the comparative analysis, the cysteine residues in samples are separately labeled with either labeled or unlabeled ICAT. The derivatized samples are then combined in a one-to-one ratio and digested with a protease resulting in both labeled and unlabeled peptide fragments. The labeled peptide fragments are then purified by affinity chromatography, fractionated by reversed-phase chromatography and analyzed by tandem mass spectrometry, which provides both qualitative analysis and the relative abundance of the peptide isoforms in the samples. The MS analysis is analogous to metabolic labeling case – ratio of ''heavy'' to ''light'' peptide ions correlates with their relative abundance. Also, the ICAT approach has the obvious conceptual limitation – only peptides that contain cysteine can be detected. It has been shown that in yeast only ~10% of all tryptic peptides contain cysteine. Therefore, full sequence coverage may not be possible even for the most abundant proteins.

Significant improvement to the ICAT method is the cleavable ICAT (cICAT), which is presently supplied by Applied Biosystems in a kit format. The cICAT reagent has four essential structural elements. The first is a protein reactive group

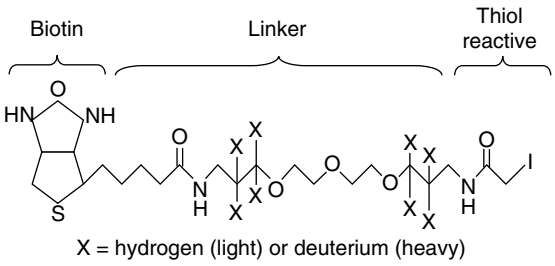


Fig. 9 Structure of isotope coded affinity tags. The ICAT reagent consists of a biotin group linked to a cysteine reactive group. The linker may be deuterated eight times or protonated at each site allowing for the generation of D_0- or D_8-ICAT. The differential masses of the linker group allow for the use of ICAT in a quantitative proteomic scheme. The figure and figure legend have been reproduced from Hunter, T.C., Washburn, M.P. (2003) The integration of chromatography and peptide mass modification for quantitative proteomics, *J. Liq Chromat. Rel. Technol.* **26**, 2285–2301 by permission of Marcel Dekker, Inc.

(iodoacetamide) that covalently links the isotope-coded affinity tag to the protein through alkylation of cysteines. The second structural element is the biotin affinity tag that allows enrichment of the tagged peptides. The third is an isotopically labeled linker ($C_{10}H_{17}N_3O_3$). Nine carbon atoms of the linker can be either ^{12}C or ^{13}C giving light and heavy version of the tag respectively. The light and heavy molecules have the same chromatographic properties, but differ in mass (9 Da). Once the sample is subjected to mass spectrometric analysis, the ratio of intensities between heavy and light peptides provides a relative quantitation of the proteins in the original sample. The fourth structural element of cICAT is an acid cleavage site that allows removal of part of the tag prior to MS analysis. After avidin-affinity purification of the cICAT-labeled peptides, biotin portion of the label and part of the linker can be removed by adding trifluoroacetic acid. This reduces the overall mass of the tag on the peptides and improves the overall peptide fragmentation efficiency.

The ICAT methodology has been successfully applied to a variety of biological questions including studies of several cell types, organelles, and different classes of proteins. Quantitative proteomic analysis via ICAT has been coupled with cDNA array analysis to investigate the galactose utilization pathway in *S. cerevisiae* and to investigate the mRNA and protein expression changes brought about by culturing *S. cerevisiae* in either galactose or ethanol. In addition, ICAT detected changes in protein expression of peripheral and integral membrane proteins by analyzing the effect on 12-phorbol 13-myristate acetate on the microsomes of HL-60 cells. Of all non-gel-based quantitative proteomic strategies, ICAT is the most mature as demonstrated by the successful use of ICAT in biologically driven analyses.

2.3 ^{18}O Labeling

The global modification of all proteolytic peptides in a mixture may be carried out via the labeling of carboxyl groups that occurs through incorporation of ^{18}O from $H^{18}O$ during proteolytic hydrolysis (Fig. 10). In order to introduce a 4-Da mass shift into the C-terminus of a peptide, proteins may be digested in the presence of $H_2{}^{18}O$. Proteases like trypsin will carry out this

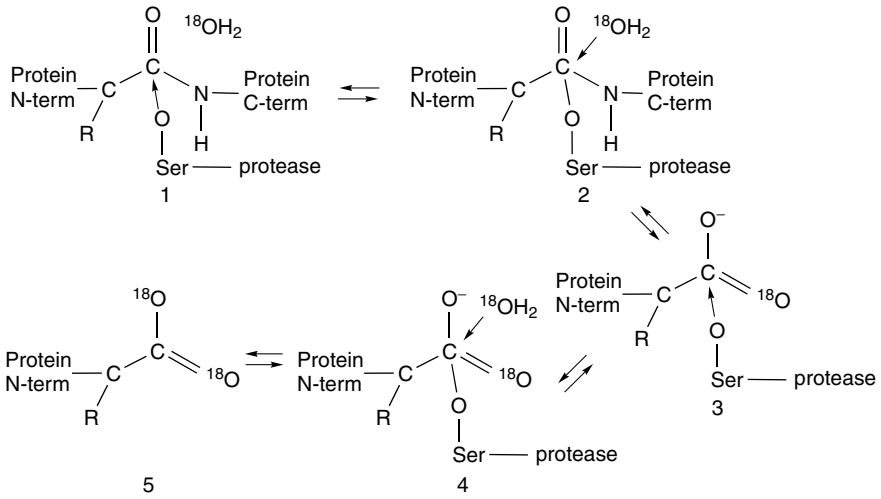


Fig. 10 C-terminal digestion modification with ^{18}O. ^{18}O may be incorporated into the C-terminus of a peptide during digestion with enzymes such as trypsin, endoproteinase Lys-C, and endoproteinase Glu-C. A simplified version of this reaction scheme is shown. (1) To begin, the peptide needs to be digested in $^{18}OH_2$ in order to then incorporate ^{18}O. The serine in the active site of the proteases listed attacks the carbonyl carbon in a peptide bond. (2) Next, $^{18}OH_2$ attacks the protein–protease intermediate also at the carbonyl carbon displacing the NH group on the peptide bond. (3) As a result, a peptide with a single ^{18}O has been generated. (4) A repeat of steps (1) and (2) is needed to drive the reaction to completion as shown in (5) where two atoms of ^{18}O have now been incorporated into the peptide C-terminus. Labeling of one sample with ^{18}O by digesting in $^{18}OH_2$ and mixing this with the other sample digested in ^{18}O depleted water allows for the determination of the relative abundance of peptides from a mixture. Figure and figure legend is reproduced from Hunter and Washburn (2003) by permission of Marcel Dekker, Inc.

reaction during the process of enzymatic cleavage. By mixing a sample with proteins digested in the presence of ^{18}O and the absence of ^{18}O, a pairwise comparison may be made to determine the relative abundance of peptides in a sample.

2.4 Postdigestion Labeling

There are several alternatives to the residue-specific modification of cysteine, which include methods for differential modification of lysine and O-phosphorylated serine residues. The phosphoprotein isotope coded affinity tag (PhIAT) method has been shown to be capable of enriching and identifying mixtures of low-abundance phosphopeptides. The PhIAT method uses a chemical modification of phosphorylated serine and threonine residues to cysteine before introduction of a standard ICAT reagent. The mass-coded abundance tag (MCAT) approach uses a residue-specific modification lysine residues by *O*-methylisourea to introduce a differential tag. In addition, 2-methoxy-4,5-dihydro-1H-imidizole has also been used to modify lysine residues for the purpose of introducing a differential mass tag.

The N/C termini of peptides after digestion may also be labeled through a variety of means. A C-terminal modification

method is methyl esterification of carboxyl groups using either methanolic HCl or the deuterated analog. In this case, multiple sites in a peptide may be modified with labels introduced at aspartic acids, glutamic acids, and C-termini. The N-terminal labeling of tryptic peptides with *N*-hydroxysuccinimide or 1-Nicotinoyloxysuccinimide esters and their stable isotope analogs is another approach that can be potentially used for quantitative proteomic analyses. In fact, coupling ^{18}O labeling and N-terminal labeling methods for protein expression profiling produced more comprehensive results than when either method was used alone.

2.5 Global mRNA and Protein Expression Analyses

An emerging application of quantitative proteomics approaches includes the large-scale analysis of protein expression correlated to large-scale mRNA expression analyses. In three independent comparisons of mRNA and protein levels in *Saccharomyces cerevisiae*, overall partial positive Spearman rank correlation coefficients ranging between 0.21, 0.45, and 0.57 were obtained. These studies employed ICAT, chromatography, and mass spectrometry, ^{15}N labeling and MudPIT or chemiluminescence of SDS-PAGE approaches to determine protein expression levels in cells grown under different conditions in each study. In all likelihood, this pattern of partial positive correlation between mRNA and protein expression levels could be expected to persist under a variety of conditions in *S. cerevisiae*. When these approaches begin to be globally applied to other organisms, it will be interesting to see if this trend persists.

3 Specific Examples of Applications

3.1 Global Proteome Sampling

The goal of global proteome sampling is the simultaneous identification of proteins in a cell or tissue at a given condition. Data obtained in such experiments can be further used to answer more specific biological questions, such as difference between healthy and pathological states. Typically, the proteins are identified by mass spectrometry and are grouped into functional categories.

3.1.1 Global Proteome Sampling Based on 2D Page

Global proteome analysis by 2D PAGE method is difficult because each spot needs to be picked and identified individually, thus increasing time and cost of the analysis. Also, as we discussed earlier, 2D PAGE separation suffers from biases – certain classes of proteins, such as hydrophobic or those of high molecular weight are difficult to detect. Apart from these limitations, 2D PAGE approach has an important advantage over other methods – it easily resolves protein isoforms.

In their analysis of *Haemophilus influenzae* proteome, Langen et al. used several techniques to maximize the 2D PAGE performance. To increase the proteome coverage, they used immobilized pH gradient strips covering several pH regions. Also, to visualize low-copy-number proteins, the authors performed a series of protein extractions, such as heparin chromatography, chromatofocusing, and hydrophobic interaction chromatography. In order to detect cell-envelope-bound proteins, the authors used immobilized pH gradient strips in

combination with a two-detergent system with a cationic detergent in the first and an anionic detergent in the second dimensions. The isolated proteins were identified by MALDI/MS and peptide fingerprinting. As a result, 502 unique proteins were identified (about 30% of all ORFs)

Analysis of the mouse brain proteome performed by Klose et al. provides a good illustration of what 2D PAGE can do for the global proteome sampling type of experiments. By using 2D PAGE, the authors performed comparative analysis of the two distantly related mouse strains, *Mus musculus* C57Bl/6 (B6) and *Mus spretus* (SPR). About 8700 proteins from the cytosolic fraction of brain proteome were compared between the two species. By analyzing 2D PAGE of B6 and SPR strains, as well as of F_1 (B6 × SPR) and B_1 (F_1 × SPR) hybrids, the authors detected 1324 species-specific polymorphisms. Among these, 466 proteins were identified by MALDI-TOF/MS using peptide mass fingerprinting. To detect the polymorphisms, the authors considered variations in electrophoretic mobility, spot intensity, and the number of different isoforms corresponding to one protein. Additionally, through the analysis of F_1 and B_1 generations, the authors established which polymorphisms were genetically dominant. The key feature that enabled this comprehensive study was the high quality of the 2D PAGE. To analyze the mouse brain proteome, the authors used the large-gel 2D PAGE, which employs IEF gel incubation, and large (46 × 30 cm) format. Implemented in this way, the 2D PAGE gives both high resolution and high sensitivity – more than 10 000 protein spots from mouse tissues can be visualized simultaneously.

3.1.2 Global Proteome Sampling Based on Multidimensional LC

When it is necessary to catalog proteins present in a cell or an organism in a given environmental context, the multidimensional LC separation of peptide mixtures followed by MS/MS is the most convenient method to use. While it is not as good at determining protein isoforms and posttranslational modifications as 2D PAGE, the biases in proteome coverage are greatly reduced.

Florens et al. performed proteomics studies of the life cycle of the human pathogen *Plasmodium falciparum* (malaria) life cycle. The authors identified 2415 proteins and assigned them to functional groups at four stages of the cycles (sporozoites, merozoites, trophozoites, and gametocytes). The sporozoite is the form in which *P. falciparum* is injected by a mosquito. The merozoite is the form that invades erythrocytes. The trophozoite is the form that multiplies in the erythrocytes. The gametocyte is the sexual stage of malaria parasite life cycle. The analysis was performed by MudPIT. The authors found that about 50% of sporozoite proteins were unique to that stage. In sporozoites, about 25% were shared with any other stage. Trophozoites, merozoites, and gametocytes had 20 to 30% unique proteins and they had 40 to 60% of their proteins shared. Only 6% of all identified proteins were shared between all four stages, which were mainly histones, ribosomal proteins, and transcription factors. Out of the 2415 identified proteins, 51% were previously annotated as hypothetical.

Koller et al. used both 2D PAGE and MudPIT to analyze *Oryza sativa* (rice) proteome. The analyses were performed on the protein extracts from leaf, root, and seed tissue. The goal of this study was to determine tissue-specific expression of

proteins. 2D PAGE separation followed by MS/MS yielded 556 unique protein identifications, comprising 348 proteins from leaf, 199 from root, and 152 from seed. MudPIT analysis resulted in significantly larger coverage: 2363 total proteins, with 867 from leaf, 1292 from root, and 822 from seed. A total of 165 proteins were uniquely detected by 2D PAGE, whereas 1972 proteins were uniquely detected by MudPIT. Next, the authors searched the nonredundant protein database by BLAST and grouped the identified proteins into functional categories. The largest category (32.8%) included proteins that had no homology to the predicted proteins. Proteins classified as involved in metabolism comprised 20.8% of all identified proteins. Out of the 2528 detected proteins, 189 were shared among all three tissues. These included housekeeping proteins that are involved in transcription, mRNA biosynthesis, translation, and protein degradation. However, most of the proteins had tissue-specific expression: 622 specific to leaf, 862 specific to root, and 512 specific to seeds.

To characterize the proteome of *S. cerevisiae* mitochondria, Sickmann et al. combined four separation methods: IEF-incubated 2D PAGE; digestion with four different proteases, followed by multidimensional LC/MS/MS; SDS/PAGE combined with multidimensional LC/MS/MS; treatment of mitochondria with trypsin, followed by SDS/PAGE or HPLC and MS/MS. The authors identified a total of 750 mitochondrial proteins. When classified into functional categories, 24.9% of all the identified proteins were of unknown function, 24.9% were involved in genome maintenance and gene expression, and 14.1% were involved in energy metabolism. The rest of the identified proteins were involved in metabolism, transport, and cell rescue.

3.1.3 MS-assisted Disease Diagnosis from Serum Samples

Proteomics technologies recently emerged as a useful tool in clinical disease diagnosis. For example, Petricoin et al. used surface-enhanced laser desorption ionization time-of-flight (SELDI-TOF) and artificial-intelligence-based informatics algorithms to discriminate between control group and ovarian cancer patients. First, the authors generated a preliminary training set of mass spectra derived from 50 unaffected women and 50 women with ovarian cancer. Next, they used an iterative searching algorithm to find the best discriminatory pattern amongst these MS data. As a result, the algorithm correctly identified all cancer cases in the masked set. Additionally, out of 66 cases of malignant disease, only 3 were recognized as cancer (false-positives). Thus, the study by Petricoin et al. demonstrated good sensitivity and predictive power of MS-based proteomics when applied to clinical disease diagnosis. For further information on this subject, we refer the interested reader to the comprehensive review by Rosenblatt et al.

3.2 Analysis of Protein Modifications by Mass Spectrometry

In living organisms, protein activity is regulated mainly by covalent modifications, which occur either co- or post-translationally. Identification of types of modifications and their locations is often a necessary requirement for an understanding of the regulation and function of a given protein. There are hundreds of known protein modifications. Among

these, phosphorylation is, perhaps, the most important and widespread – about one-third of all proteins from mammalian genomes are thought to be phosphorylated. Another functionally important modification is glycosylation – glycosylated proteins are ubiquitous components of cellular surfaces where their oligosaccharide groups participate in a wide range of cell–cell recognition events. Comprehensive analysis of glycosylated proteins is more challenging than analysis of other modifications, mainly because the structure of oligosaccharide varies.

Other commonly occurring modifications that are involved in protein regulation and function are disulfide bonds, acetylations, and ubiquitinations. Some of these and other modifications are listed in Table 2. Changes in a protein length, either as a result of alternative splicing or protein truncations, also may be considered as protein modifications. Generally, it is difficult to identify protein truncations by methods that deal with protein/peptide mixtures, and often purification down to individual proteins is required in such cases (e.g. by 2D PAGE).

Currently available methods of large-scale analysis of modified proteins can be grouped into two major classes: those that use sample enrichment or chemical treatment prior to MS, and those that rely on MS data alone. Enrichment methods include affinity purification, chemical tagging followed by affinity purification, and immunoprecipitation. MS methods of detection and identification of modified peptides include neutral loss scan, precursor ion scan, postsource decay, and others.

Sometimes it is of special interest to obtain information on several types of modifications at once. If that is the case, computer programs such as GutenTag (see Sect. 2) can be used to analyze MS/MS data. Additionally, if mass changes introduced by modifications are known, the search for modified proteins can be done by SEQUEST with input parameters modified in accordance with the mass changes. However, generally, it is difficult to analyze modified peptides in the background of nonmodified peptides. Therefore, when it is clear what type of modification needs to be analyzed, the fractionation steps that enrich that particular modification need to be introduced into the experimental scheme.

Tab. 2 Common protein modification.

Modification	*Monoisotopic/average mass change*
Phosphorylation[a]	+79.9663/79.9799
Acetylation	+42.0106/42.0373
N-acetylglucosamine (GlcNAc)[a]	+203.0794/203.1950
Disulfide bond	−2.01565/2.0159
Methylation	+14.0157/14.0269
Hydroxylation	+15.9949/15.9994
Oxidation of methionine	+15.9949/15.9994
Ubiquitination of lysines	+114.0429/114.1040[b]

[a] Occurs on tyrosine, serine, threonine. Widespread throughout the proteome. Functions include protein regulation, signal transduction.
[b] Mass change is due to Gly–Gly residue, which is left on ubiquitinated lysines after trypsin digestion.

Below we discuss several illustrative examples of analysis of phosphorylated, glycosylated, and ubiquitinated proteins from recent literature. In addition, tools and techniques used in analysis of phosphorylation are generally applicable to analysis of many other modifications such as methylation and acetylation, while analysis of glycosylation poses additional analytical and instrumental challenges due to variability in structures of oligosaccharide groups. Ubiquitination, while not as frequent as phospho- and glyco modifications, is important for protein degradation in proteasomes.

3.2.1 Phosphorylated Proteins

Main tools in large-scale identification of phosphoproteins are enrichment by immobilized metal affinity chromatography (IMAC), chemical tagging, and immunoprecipitation by phosphor-specific antibodies. IMAC technology is based on methods developed by Andersson et al. and relies on interaction of phosphate group with immobilized Fe^{3+} ions. Ficarro et al. used IMAC combined with LC/MS/MS to characterize the phosphorylated portion of the yeast proteome. The authors showed that carboxylic acid interfered with IMAC purification, and needed to be protected. The protection was achieved by esterification with methanol in the presence of HCl. Phosphorylated tryptic peptides were identified via SEQUEST. From the whole cell lysate, Ficarro et al. detected more than 1000 phosphopeptides. From these, 383 sites of phosphorylation were determined. A potential improvement to this analysis would be the use of other proteinases in parallel with trypsin to increase the sequence coverage.

An enrichment technique that could also complement IMAC is immunoprecipitation with antibodies that bind to any protein that contain phosphorylated residues. While antibodies exist for phosphorylated serine, threonine, and tyrosine, only the anti-phosphotyrosine antibody binds strongly enough to allow enrichment. Pandey et al. used phosphotyrosine immunoprecipitation to study phosphorylation in HeLa cells in response to epidermal growth factor (EGF). The phosphopeptides were immunoprecipitated from untreated and EGF-treated cell lysates and resolved by electrophoresis. Individual gel bands were excised and studied by MALDI-MS and ESI-MS/MS. As a result, the authors identified Vav-2 as a substrate of EGF-receptor.

A report by Salomon et al. also gives a nice demonstration of analysis of phosphorylation in human cells. The authors used phosphotyrosine immunoprecipitation along with methyl esterification and IMAC combined with multidimensional LC/MS to assess tyrosine phosphorylation that occurs over time in myelogenous leukemia cells in response to treatment. The authors reported identification of 64 unique tyrosine phosphorylation sites in 32 proteins.

In another report by Ficarro et al., the authors used anti-phosphotyrosine immunoblots to study capacitation of human sperm. Capacitation is a cAMP-dependent process that is necessary for fertilization. The authors performed a comparative analysis of capacitated versus noncapacitated sperm. First, they separated sperm proteins by 2D PAGE followed by western blotting with anti-phosphotyrosine antibodies. In the next step, they excised and digested spots that exhibited phosphorylation, followed by IMAC to enrich for phosphopeptides with subsequent MS/MS analysis. As a result, the authors pinpointed several proteins that undergo phosphorylation upon capacitation

of the sperm. Additionally, the authors used differential isotopic labeling to quantify the phosphorylation. The labeling was achieved at the stage of protecting the carboxy groups prior to IMAC, by treatment of the peptide mixtures from capacitated and noncapacitated digests with CD_3OD/DCl and CH_3OH/HCl respectively. These two peptide mixtures were further combined in one-to-one ratio and analyzed by IMAC/LC/MS/MS. As a result of this quantitation, the authors found 20 unique peptides that exhibited different phosphorylation levels between capacitated and noncapacitated sperm.

Metabolic labeling strategy was first described to quantitate changes in phosphorylation. In this approach, cells from two batches that have potentially different levels of phosphorylated proteins are metabolically labeled with N^{14} and N^{15}. Next, the cells are lysed; the target proteins are purified, digested, and analyzed by MS. Changes in peak intensities that correspond to modified and unmodified peptides from the two conditions provide quantitation of phosphorylation.

Another way to quantitate phosphorylation levels is to use modified ICAT strategy. In this approach, the phosphate groups in phosphopeptides derived from two different conditions are chemically replaced with either labeled or unlabeled tags. The tagging involves the following steps: (1) beta-elimination of the phosphate groups; (2) addition of 1,2-ethanedithiol containing either four hydrogens (EDT-D_0) or four deuteriums (EDT-D_4); (3) biotinylation of the EDT group using (+)-biotinyl-iodoacetamidyl-3,6-dioxaoctanediamine. The tagged peptides are further affinity purified by avidin column and analyzed by LC/MS.

3.2.2 Glycosylated Proteins

The importance of protein glycosylation, especially during cell–cell communication in multicellular organisms, is often acknowledged by using the terms *glycobiology* and *glycomics*. Owing to a wide range of possible polysaccharide structures, analysis of glycosylation is not as straightforward as analysis of other posttranslational modification. Currently, there are no satisfactory methods for global, proteome-wide analysis of all glycoprotein forms. It is possible, however, to characterize glycoproteins with glycogroups of constant structure. It is also possible to globally map glycosylation sites.

As an example, consider a broad research question such as to identify and characterize glycosylated proteins in a given biological system. In such a case, hypothetical analysis could include the following steps: (1) proteolytic digestion; (2) enrichment for glycopeptides; (3) identification of glycopeptides by MS and MS/MS; (4) structure determination of polysaccharide groups by MS^n. Alternatively, one could also separate proteins by 2D PAGE, and use glycospecific staining methods to identify spots of interest. Also, during the MS part of analyses, it could be useful to separate constant glyco structures from the variable ones, as well as N-linked from O-linked ones. One of the problems in analysis of glycopeptides is that glycogroups are very labile. Because of this lability, peptide fragmentation in CID reactions is reduced, thus making sequencing by MS/MS more difficult. An alternative way is to chemically (e.g. with beta-elimination of O-linked oligosaccharides) or enzymatically (e.g. with *N*-glycosidase F) remove glycogroups prior to analysis and to use chemical tags.

In eukaryotes, the most widespread constant type of glycosylation is O-linked *N*-acetylglucosamine (*O*-GlcNAc), which is found on many nuclear and cytoplasmic proteins. Glycosylation of serine and threonine residues by *O*-GlcNAc is believed to compete with and complement phosphorylation in mediating protein–protein interactions. The proteins that are glycosylated by *O*-GlcNAc include RNA polymerase II, transcription factors, chromatin-associated proteins, nuclear pore proteins, protooncogenes, tumor suppressors, and proteins involved in translation.

Because *O*-GlcNAc and phospho groups modify essentially the same amino acids, it is of special interest to establish methods that can characterize these modifications simultaneously. In this pursuit, Wells et al. developed a method based on beta-elimination followed by Michael addition (BEMAD) of dithiothreitol (DTT) or biotine pentylamine (BAP). This method relies on the fact that *O*-GlcNAc groups are more prone to elimination than phosphate groups. With the right conditions of elimination, *O*-GlcNAc can be tagged selectively. The DTT and BAP tags also allow enrichment by affinity chromatography and are stable in MS/MS fragmentation, thus allowing identification of the modified sites. First, the authors tested their method on synthetic peptides, and then they performed analysis of several biological samples: Synapsin I from rat brain and nuclear pore complex (NPC). In Synapsin I, three novel *O*-GlcNAc sites, as well as three previously known sites were mapped, thus validating the method. In the nuclear pore complex, BEMAD also mapped novel *O*-GlcNAc sites in Lamin B receptor and Nup155. Using BEMAD along with modification-specific antibodies and enzymes, the authors were able to distinguish between *O*-GlcNAc- and phosphopeptides.

During CID reactions, oligosaccharide moieties fragment mainly at glycosidic bonds. This fact potentially allows discerning of a primary structure of an oligosaccharide in MS^n experiments. While technology for a large-scale structural analysis of glycoforms is not developed yet, oligosaccharide structure determination is certainly possible for individually isolated glycopeptides. Notable examples include characterization of lipooligosaccharides from *Haemophilus influenzae* and *Neisseria gonorrhoeae*.

3.2.3 Ubiquitinated Proteins

Degradation of proteins in living organisms is a complex, highly regulated process, which plays important roles in many cellular pathways. The first step in protein degradation is the attachment of ubiquitin moieties to lysines of the substrate. The second step is proteolysis of the tagged protein by proteasome. Given the importance of this posttranslational modification, it is surprising that the report by Peng et al. is perhaps the only paper in the literature that addresses large-scale analysis of protein ubiquitination.

Peng et al. described a systematic approach based on LC/LC/MS/MS to analyze protein ubiquitination in yeast. In this pursuit, Peng et al. expressed His-tagged ubiquitin in *S. cerevisiae cells* followed by purification over Ni-chelating resin. Denaturing conditions were used at the enrichment step in order to minimize copurification of proteins that are not ubiquitinated, but form complexes with ubiquitin. The enriched ubiquitinated proteins were digested with trypsin and analyzed by SCX/RP/MS/MS followed by identification by SEQUEST. The tryptic

digestion of ubiquitinated peptides results in glycine–glycine fragments at the sites of modification. The corresponding increase in mass by 114 Da of the modified lysine residues allows localization of the sites of ubiquitination. As a result, the authors identified 110 ubiquitination sites present in 72 ubiquitinated proteins. However, some of the known ubiquitinations were not detected. This is probably due to the fact that the method is biased toward more abundant species. Indeed, precise localization of a posttranslational modification via SEQUEST requires a high sequence coverage. As a consequence, in their analysis, Peng et al. identified 1075 proteins totally after Ni-resin purification, but were able to confirm ubiquitination in only 10% of them.

3.3 Analysis of Protein–Protein Interactions by Mass Spectrometry

In a cell, proteins exert their functions through interactions with other proteins. Proteomics methods developed in the past few years can be applied directly to the analysis of the protein–protein interactions and protein complexes. Protein–protein interactions can be studied either on the level of individual protein complexes, or on the level of the whole proteome. Examples of different types of analyses include the possibility of predicting interactions from amino acid sequence, focused mass spectrometric analyses of individual multiprotein complexes: yeast ribosome, SAGA-like complex (SLIK), Pol II preinitiation complex (PIC), nuclear pore complex (NPC), and proteome-wide analyses of protein–protein interactions. In addition, quantitative proteomic methods may be used to analyze the dynamics of protein complexes.

3.3.1 Computational Methods of Protein–Protein Interaction Prediction

To some degree, protein–protein interactions can be deduced indirectly from the amino acid sequences. In this pursuit, Bock et al. created a learning algorithm that was trained to recognize and predict protein–protein interactions. The training was achieved on the experimentally known interactions from a variety of organisms. As an outcome of the training, the decision function was constructed, which was statistically evaluated using unseen test data. As a result, on average, about 80% of the interactions were predicted accurately from the unseen datasets. Obviously, computational methods alone do not give exhaustive description and the obtained results need to be validated experimentally. Nevertheless, the interaction datasets obtained *in silico* provide useful reference points for experimental types of analyses.

3.3.2 Yeast 2-hybrid Arrays

One of the most common approaches to analyze protein–protein interactions is the yeast-2-hybrid (Y2H) screen, which is a selection method that detects protein–protein interactions in the yeast nucleus. This is a well-developed technique that can be easily optimized for a high-throughput analysis. The Y2H screen was successfully applied to mapping of a large-scale protein interaction network in the yeast *S. cerevisiae*. The Y2H has a good resolution and can detect weak and transient interactions. However, the Y2H method detects only binary interactions between proteins and may not be used to study transcriptional activators. Another drawback of the Y2H is that interactions occur in the nucleus, so that interactions for many proteins take place out of their native environment.

3.3.3 Direct Analysis of Large Protein Complexes; Composition of the Yeast Ribosome

Link et al. employed multidimensional liquid chromatography coupled with MS/MS to study composition of the yeast ribosome. The authors named their method *direct analysis of large protein complexes*, and demonstrated its wide applicability. In this study, the ribosomes were purified by sucrose gradients and then denatured. Next, the ribosomal RNA was removed, the ribosomes were enzymatically digested, and the digests were loaded on a 2D separation column, consisting of the SCX and the RP dimensions. After the separation, peptides were eluted directly into the mass spectrometer for the MS/MS analysis. Finally, SEQUEST algorithm was used to search nucleotide databases and to match the peptide fragmentation patterns. The authors demonstrated the high-throughput capacity of this approach – more than 100 proteins could be identified in a single experiment. As a result of this study, new protein components of yeast and human 40 S subunit were discovered.

3.3.4 Analysis of Multisubunit Protein Complexes Involved in Ubiquitin-dependent Proteolysis by Mass Spectrometry

In a number of reports, Deshaies et al. (and references therein) used MS-based strategies to characterize the composition of various protein complexes involved in proteolysis. The authors used sequential epitope tagging, affinity purification, and mass spectrometry (SEAM) to study regulation and function of SCF ubiquitin ligases. In the application of this method, SCF subunits Skp1 and Cul1 were C-terminally tagged with Myc epitope. Next, the cells expressing the tagged proteins were lysed, and the soluble fractions were affinity purified, digested, and analyzed by mass spectrometry. As a result, a total of 16 Skp1- and Cul1-interacting proteins were detected. Several of these proteins were not previously known, including Hrt1, Rav1, and Rav2. These new proteins were further subjected to SEAM and this led to identification of the new complex Rav1/Rav2/Skp1. Subsequently, it was found that Rav1/Rav2/Skp1 complex interacts with V_1 component of V-ATPase, the vacuolar membrane ATPase. In different studies by biochemical methods, the authors further determined that Rav1/Rav2/Skp1 regulates the assembly of V-ATPase from V_1 and V_0 domains.

The same group used multidimensional protein identification technology (MudPIT) to identify proteins that interact with the 26 S proteasome. As it was introduced in the previous sections of this chapter, MudPIT allows analysis of immunoprecipitated fractions without a gel separation step. Instead, the immunoprecipitated fractions are digested, and the peptide mixtures are separated in two dimensions (SCX and RP) followed by MS analysis. Using MudPIT, the authors identified every known subunit within affinity-purified 26 S proteasome, as well as one subunit that was not previously known. Additionally, a set of proteins potentially interacting with the proteasome (PIPs) was found. By immunoblotting methods, six of these PIPs were further confirmed to associate with the proteasome.

3.3.5 Proteomics of the Nuclear Pore Complex

MS-based protein identification is a useful tool that can efficiently determine the composition of a given multiprotein complex. However, this method by itself does not necessarily provide information on the complex's spatial architecture. Nor does it directly answer questions about the

multiprotein complex function. Therefore, when such questions arise, the mass spectrometric tools need to be complemented with other techniques, for example, immunoblotting, immunofluorescence, electron microscopy and so on. Several studies of the nuclear pore complexes (NPCs) that are discussed here give a good illustration of these types of integrative strategies.

In the early 1990s, studies by three-dimensional cryoelectron microscopy revealed the basic shape and architecture of NPCs in *Xenopus* nucleus. It was determined that NPCs are proteinaceous structures situated in the double membrane of the nuclear envelope. Estimated sizes of NPCs vary from ~125 MDa (*Xenopus*) to 66 MDa (*Saccharomyces*). NPCs have an eightfold rotational symmetry with the rotational axis normal to the nuclear envelope membrane and a twofold mirror symmetry with the symmetry plane parallel to the nuclear envelope membrane. Current research efforts are aimed at understanding the mechanism of the biological function of the NPCs – nucleotransport. To understand the NPC's function, it is useful to catalog all the protein components of NPCs in different organisms. Ideally, this should lead to testable hypotheses on how these components contribute to the overall structure and function of NPCs. Biological problems of this kind can be addressed by proteomics methods as was elegantly demonstrated for yeast and mammalian NPCs.

In the yeast study, Rout et al. prepared highly enriched fraction of NPC proteins, followed by separation on ceramic hydroxyapatite HPLC, which gave efficient recovery of the loaded proteins. Reverse-phase TFA-HPLC was used in parallel for resolution of the low molecular mass proteins from the NPC fraction. The next step in the separation involved SDS-PAGE with visualization of protein bands by copper staining. Subsequently, peptide mixtures were prepared from bands of interest via in-gel trypsin digestion followed by analysis by MALDI-MS. The peptide mass matching method was used to search nonredundant protein sequence database. Previously known nucleoporins were genomically tagged with a protein A epitope (ProtA), which allowed further immunolocalization by fluorescence microscopy. As a result of this study, the authors identified 29 nucleoporins and 11 transport factors and NPC-associated proteins. The authors also determined stoichiometry and position of each of the nucleoporins found within the NPC by quantitative immunoblotting and by immunoelectron microscopy. In the immunoelectron microscopy analysis, the ProtA-tagged proteins were labeled using gold-conjugated antibody, which aided visualization. On the basis of the deciphered architecture of the NPC, Rout et al. proposed a model of nucleotransport called a *Brownian affinity gating model*. The core idea of the proposed model is that translocation through the NPC occurs via diffusion: diffusive movements of the filamentous nucleoporins on the cytoplasmic face of the NPC exclude macromolecules that do not bind to them, but when the binding does occur (and that happens when a cargo molecule is associated with its transport factor), the residence time of the cargo at the NPC gate increases, which in turn facilitates the diffusion of the cargo into the nucleus.

In the mammalian study, Cronshaw et al. enriched NPCs fractions from rat liver nuclei by sequential solubilization. At each step of the enrichment, the authors used electron microscopy, SDS-PAGE, and immunoblotting to confirm that NPCs remained intact. After the enrichment, NPCs were treated with detergent, which

produced a solution of monomeric nucleoporins. The individual proteins in this nucleoporin mixture were separated by C4 reverse-phase chromatography followed by SDS-PAGE. Protein identification was performed using single-MS and MS/MS. In addition to previously known 23 nucleoporins, Cronshaw et al. identified six novel nucleoporins, and also four proteins containing WD repeats. One of these four WD-containing proteins was ALADIN, the gene mutated in Allgrove syndrome.

Spatial organization of a protein complex can be assessed by the cross-linking method. In this method, a protein complex is affinity-purified and then treated with a cross-linking agent, a chemical that introduces new covalent bonds between neighbor proteins. New bands, that appear on SDS-PAGE because of this treatment can be excised and identified by mass spectrometry. As a result, if a pair of the proteins gets cross-linked, this usually means that the two proteins are located close to each other within the complex. A good example of this strategy is the study by Rappsilber et al. in which the authors used cross-linking/MS method to deduce the spatial composition of the six-member subcomplex Nup84p of the yeast NPC. The authors emphasized generic applicability of this approach. One of the significant challenges in application of this method, however, is the choice of the proper cross-linking reaction condition – usually a number of different cross-linkers has to be screened, before the right degree of cross-linking is obtained. Additionally, interaction between subunits that are hidden deep inside a complex may be inaccessible to cross-linkers. Because of these and other difficulties, the cross-linking method is limited to complexes of small sizes and is difficult to apply on the broader scale. Nevertheless, the cross-linking method may prove to be useful in the studies of conformational or compositional changes of individual protein complexes, when the specific structural states can be ''frozen'' through interaction with the cross-linking agents.

3.3.6 High-throughput Analyses of Protein–Protein Interactions

In most cases, MS-based analytical schemes similar to those employed in the studies of individual complexes can be redesigned for use at the proteome-wide scale. A report by Gavin et al. is one of the first examples of the MS-based proteome-wide analyses of the protein complexes. For the large-scale isolation of the protein complexes from yeast Gavin et al. used tandem affinity purification (TAP), as first introduced by Rigaut et al. In the TAP method, a gene-specific fusion cassette – which contains calmodulin-binding domain, a specific protease cleavage site, and ProtA domain – is introduced at the C- or N-terminal of yeast's ORFs of interest. Then, assuming that expression of fusion proteins is maintained close to the natural level, the first affinity purification is performed. In this step, fusion proteins along with their interaction partners (so-called protein assemblies) are isolated from cell extract by affinity selection on IgG matrix. Next, the bound proteins are released by addition of the protease. Finally, the second affinity purification is done, which involves incubation with calmodulin beads in the presence of calcium. The advantage of tandem affinity purification when compared to standard epitope tagging approaches is that it removes most of the nonspecific interactions. Out of 1548 yeast strains generated by Gavin et al., 1167 expressed the fusion proteins at detectable levels. After the purification of the protein complexes by TAP, the authors

subjected the complexes to electrophoretic separation followed by trypsin digestion, and subsequent analysis by MALDI-MS. Overall, by MS analysis, Gavin et al. identified 1440 gene products (~25% of the genome) from various organelles. However, identification of membrane proteins in this study has proven to be difficult – only 40 membrane proteins were purified successfully out of total 293 membrane proteins detected. The authors then proceeded with grouping of the identified proteins into complexes. This was done by the analysis of overlaps in composition of the pulled-down assemblies from 589 different bait proteins. The authors reported a total of 245 purifications that corresponded to 98 known complexes from the yeast protein database (YPD). Another 242 purifications out of the 589 were assembled into 134 new complexes. The authors were able to identify proteins as low-abundant as 15 copies per cell, thus showing high sensitivity of the TAP method. However, reproducibility was rather poor – the authors estimated that probability of finding the same protein from the same bait in two purifications is about 70%. Another weakness of the TAP method comes from possibility of interference of the TAP tag with complexes assembly and protein function. In fact, Gavin et al. found that when the essential genes were TAP-tagged; in about 20% of these cases, nonviable strains were obtained. Also, the authors reported significant bias against proteins with molecular weight below 15 KDa.

Another notable report of a high-throughput protein complex identification is the study by Ho et al. In this case, the bait proteins contained Flag epitope tag and were overexpressed from *GAL1* or *tet* promoters. Next, the protein assemblies were isolated in one-step immunoaffinity purification followed by resolution on SDS-PAGE, digestion and MS, and MS/MS analyses. The authors called this method "high-throughput mass spectrometric protein complex identification" (HMS-PCI). The immunoaffinity purification of complexes assembled around overexpressed baits should, in theory, generate more false-positives than TAP method would generate, because of the nonphysiological concentrations of the baits. On the other hand, weak and transient interactions that would not be detected in the TAP method could be captured by HMS-PCI. In fact, Ho et al. were able to assess certain regulatory and signaling pathways, by studying complexes pulled-down with phosphatases and kinases used as baits.

As of today none of the methods of mapping of protein–protein interactions within a proteome is comprehensive enough to provide full coverage. Hence, it is useful to compare datasets obtained with different approaches. In their article, Christian von Mering et al. evaluated all available interaction datasets obtained in yeast. These included the data from MS-based studies discussed above, as well as data from Y2H, correlated mRNA expression, synthetic lethality, and *in silico* predictions. The evaluation was done through comparisons with the reference dataset (MIPS and YPD). As a result, percentage coverage (fraction of the reference covered) and accuracy (fraction of data confirmed by the reference) were estimated for every method. According to the authors' analysis, TAP method provides both higher coverage and higher accuracy than either Y2H or HMS-PCI. Also, the analysis shows that HMS-PCI is the least accurate method amongst the three.

In a series of reports, Bader and Hogue developed algorithmic approaches for finding molecular complexes from datasets

obtained in different interaction studies. By analyzing combined data from TAP and HMS-PCI studies, they found a novel nucleolar complex of 148 proteins that included 39 proteins with unknown function. Further, they described a graph-theoretic clustering algorithm *molecular complex detection* (MCODE) that allows detection of dense regions (potential complexes) within the interaction networks. Importantly, the authors showed that MCODE algorithm is not affected by a high rate of false-positives in datasets from the high-throughput experiments.

To summarize, none of the current high-throughput experimental schemes provide sufficient coverage and accuracy. Therefore, integrative approaches that take advantage of different methods are necessary. Additionally, all of the discussed cases dealt with cells in a certain growth conditions. It is of special interest, however, to study dynamics within protein interaction networks in response to environmental stimuli, in progression through the cell cycle, or in pathological states. Some of these questions can be addressed by quantitative proteomics techniques.

3.3.7 Quantitative Proteomics Methods in the Studies of Protein Complexes

Methods of quantitative proteomics can be used to study dynamical changes in abundance, composition, and activities of multiprotein complexes. A good example of such a study is the work by Ranish et al. in which composition of a large RNA polymerase preinitiation complex (PIC) was assessed by the isotopically coded affinity tags (ICATs) method. This and other differential labeling methods are discussed in detail in the previous section of this chapter. The ICAT approach as employed by Ranish et al., consists of the four major steps summarized below:

1. The samples from two different conditions are labeled with either ''light'' or ''heavy'' tags.
2. The labeled samples are combined together and enzymatically digested.
3. The mixture is fractionated by SCX chromatography, then the labeled peptides are isolated by avidin-affinity chromatography followed by separation by the reversed-phase microcapillary chromatography.
4. The labeled peptides that are eluting from the reversed phase column are analyzed by ESI-MS/MS.

In MS analysis, the peak intensity ratios of the differentially labeled peptides on the ion chromatogram are related to the relative abundances of the corresponding proteins in the two different environments. The peptide identities are established in the MS/MS spectra and the corresponding proteins are identified by SEQUEST. With this quantitative ICAT method, the authors were able to distinguish between components of the PIC and the copurifying background proteins, some of which had higher abundances. Thus, the authors demonstrated the high analytical power of this approach. In their analysis, Ranish et al. identified a total of 326 proteins, 42% of which participate in the Pol II-mediated transcription. Also, the authors used the ICAT method to monitor changes in the PIC composition in the presence or absence of TBP. TBP is a transcription factor that binds to the TATA element and is required for the functional PIC assembly. According to Ranish et al., most of the Pol II components are increased in abundance by a factor of at least 1.9 upon addition of TBP, and several Pol II components showed no increase in abundance. In addition, potentially new component of the PIC was discovered. A

limitation of this approach, as was noted by the authors, is that using only the cysteine specific tags leaves out tryptic peptides that do not contain cysteines. In this respect, strategies that use metabolic labeling or N-terminal labeling may be more promising.

As an example, in their recent paper, Blagoev et al. used stable isotopic amino acids in cell culture (SILAC) to study EGF signaling. The control and EGF-stimulated HeLa cell populations were labeled with ^{12}C- arginine and ^{13}C-arginine respectively via metabolic incorporation. Combined cell lysates from these two conditions were affinity-purified with SH2 domain of GST-SH2 fusion protein used as bait. SH2 domain specifically binds phosphorylated EGF receptor. Protein complexes obtained in this purification were digested with trypsin and the peptide mixtures were analyzed by MS. As a result, the authors identified 228 proteins, 28 of which were enriched upon EGF simulation.

Bibliography

Books and Reviews

Aebersold, R., Mann, M. (2003) Mass spectrometry-based proteomics, *Nature* **422**, 198–207.

Bakhtiar, R., Nelson, R.W. (2000) Electrospray ionization and matrix-assisted laser desorption ionization mass spectrometry. Emerging technologies in biomedical sciences, *Biochem. Pharmacol.* **59**, 891–905.

Chalmers, M.J., Gaskell, S.J. (2000) Advances in mass spectrometry for proteome analysis, *Curr. Opin. Biotechnol.* **11**, 384–390.

Choudhary, J.S., Blackstock, W.P., Creasy, D.M., Cottrell, J.S. (2001) Matching peptide mass spectra to EST and genomic DNA databases, *Trends Biotechnol.* **19**, S17–S22.

Figeys, D. (2003) Proteomics in 2002: a year of technical development and wide-ranging applications, *Anal. Chem.* **75**, 2891–2905.

Koonin, E.V. (2001) Computational genomics, *Curr. Biol.* **11**, R155–R158.

McLachlin, D.T., Chait, B.T. (2001) Analysis of phosphorylated proteins and peptides by mass spectrometry, *Curr. Opin. Chem. Biol.* **5**, 591–602.

Noordewier, M.O., Warren, P.V. (2001) Gene expression microarrays and the integration of biological knowledge, *Trends Biotechnol.* **19**, 412–415.

Siuzdak, G. (1994) The emergence of mass spectrometry in biochemical research, *Proc. Natl. Acad. Sci. U.S.A.* **91**, 11290–11297.

Watson, J.T. (1997) *Introduction to Mass Spectrometry*, Lippincott-Raven, Philadelphia, NY.

Wilkins, M.R., Appel, K.D., Hochstrasser, D.F. (Eds.) (1997) *Proteome Research: New Frontiers in Functional Genomics (Principles and Practice)*, Springer-Verlag, New York.

Primary Literature

Adam, S.A. (2001) The nuclear pore complex, *Genome Biol.* **2**, REVIEWS0007.1–0007.6.

Aebersold, R. (2003) Quantitative proteome analysis: methods and applications, *J. Infect. Dis.* **187**(()Suppl. 2), S315–S320.

Aebersold, R., Mann, M. (2003) Mass spectrometry-based proteomics, *Nature* **422**, 198–207.

Akey, C., Radermacher, M. (1993) Architecture of the Xenopus nuclear pore complex revealed by three-dimensional cryo-electron microscopy, *J. Cell. Biol.* **122**, 1–19.

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

Ardekani, A.M., Liotta, L.A., Petricoin, E.F. III (2002) Clinical potential of proteomics in the diagnosis of ovarian cancer, *Expert Rev. Mol. Diagn.* **2**, 312–320.

Bacher, G., Korner, R., Atrih, A., Foster, S.J., Roepstorff, P., Allmaier, G. (2001) Negative and positive ion matrix-assisted laser desorption/ionization time-of-flight mass spectrometry and positive ion nano-electrospray ionization quadrupole ion trap mass spectrometry of peptidoglycan fragments isolated from various Bacillus species, *J. Mass Spectrom.* **36**, 124–139.

Bader, G.D., Hogue, C.W. (2002) Analyzing yeast protein-protein interaction data obtained from different sources, *Nat. Biotechnol.* **20**, 991–997.

Bader, G.D., Hogue, C.W. (2003) An automated method for finding molecular complexes

in large protein interaction networks, *BMC Bioinform.* **4**, 2.

Bakhtiar, R., Nelson, R.W. (2000) Electrospray ionization and matrix-assisted laser desorption ionization mass spectrometry. Emerging technologies in biomedical sciences, *Biochem. Pharmacol.* **59**, 891–905.

Berger, S.J., Lee, S.W., Anderson, G.A., Pasa-Tolic, L., Tolic, N., Shen, Y., Zhao, R., Smith, R.D. (2002) High-throughput global peptide proteomic analysis by combining stable isotope amino acid labeling and data-dependent multiplexed-MS/MS, *Anal. Chem.* **74**, 4994–5000.

Bock, J.R., Gough, D.A. (2001) Predicting protein–protein interactions from primary structure, *Bioinformatics* **17**, 455–460.

Bock, J.R., Gough, D.A. (2003) Whole-proteome interaction mining, *Bioinformatics* **19**, 125–134.

Cagney, G., Emili, A. (2002) De novo peptide sequencing and quantitative profiling of complex protein mixtures using mass-coded abundance tagging, *Nat. Biotechnol.* **20**, 163–170.

Cagney, G., Uetz, P., Fields, S. (2000) High-throughput screening for protein-protein interactions using two-hybrid assay, *Methods Enzymol.* **328**, 3–14.

Carr, S.A., Huddleston, M.J., Bean, M.F. (1993) Selective identification and differentiation of N- and O-linked oligosaccharides in glycoproteins by liquid chromatography-mass spectrometry, *Protein Sci.* **2**, 183–196.

Chalmers, M.J., Gaskell, S.J. (2000) Advances in mass spectrometry for proteome analysis, *Curr. Opin. Biotechnol.* **11**, 384–390.

Choudhary, J.S., Blackstock, W.P., Creasy, D.M., Cottrell, J.S. (2001) Matching peptide mass spectra to EST and genomic DNA databases, *Trends Biotechnol.* **19**, S17–S22.

Comer, F.I., Hart, G.W. (1999) O-GlcNAc and the control of gene expression, *Biochim. Biophys. Acta* **1473**, 161–171.

Comisarow, M.B., Marshall, A.G. (1996) The early development of Fourier transform ion cyclotron resonance (FT-ICR) spectroscopy, *J. Mass Spectrom.* **31**, 581–585.

Conrads, T.P., Alving, K., Veenstra, T.D., Belov, M.E., Anderson, G.A., Anderson, D.J., Lipton, M.S., Pasa-Tolic, L., Udseth, H.R., Chrisler, W.B., Thrall, B.D., Smith, R.D. (2001) Quantitative analysis of bacterial and mammalian proteomes using a combination of cysteine affinity tags and 15N-metabolic labeling, *Anal. Chem.* **73**, 2132–2139.

Conrads, T.P., Anderson, G.A., Veenstra, T.D., Pasa-Tolic, L., Smith, R.D. (2000) Utility of accurate mass tags for proteome-wide protein identification, *Anal. Chem.* **72**, 3349–3354.

Conrads, T.P., Issaq, H.J., Veenstra, T.D. (2002) New tools for quantitative phosphoproteome analysis, *Biochem. Biophys. Res. Commun.* **290**, 885–890.

Corthals, G.L., Wasinger, V.C., Hochstrasser, D.F., Sanchez, J.C. (2000) The dynamic range of protein expression: a challenge for proteomic research, *Electrophoresis* **21**, 1104–1115.

Cronshaw, J.M., Krutchinsky, A.N., Zhang, W., Chait, B.T., Matunis, M.J. (2002) Proteomic analysis of the mammalian nuclear pore complex, *J. Cell. Biol.* **158**, 915–927.

Deshaies, R.J., Seol, J.H., McDonald, W.H., Cope, G., Lyapina, S., Shevchenko, A., Verma, R., Yates, J.R. III (2002) Charting the protein complexome in yeast by mass spectrometry, *Mol. Cell. Proteomics* **1**, 3–10.

Dohmen, R.J., Madura, K., Bartel, B., Varshavsky, A. (1991) The N-end rule is mediated by the UBC2(RAD6) ubiquitin-conjugating enzyme, *Proc. Natl. Acad. Sci. U.S.A.* **88**, 7351–7355.

Dotsch, V., Wagner, G. (1998) New approaches to structure determination by NMR spectroscopy, *Curr. Opin. Struct. Biol.* **8**, 619–623.

Dreger, M. (2003) Subcellular proteomics, *Mass Spectrom. Rev.* **22**, 27–56.

Eng, J., McCormack, A., Yates, J.R. (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequence in a protein database, *J. Am. Soc. Mass Spectrom.* **5**, 976–989.

Fenn, J.B., Mann, M., Meng, C.K., Wong, S.F., Whitehouse, C.M. (1989) Electrospray ionization for mass spectrometry of large biomolecules, *Science* **246**, 64–71.

Fey, S.J., Larsen, P.M. (2001) 2D or not 2D. Two-dimensional gel electrophoresis, *Curr. Opin. Chem. Biol.* **5**, 26–33.

Ficarro, S., Chertihin, O., Westbrook, V.A., White, F., Jayes, F., Kalab, P., Marto, J.A., Shabanowitz, J., Herr, J.C., Hunt, D.F., Visconti, P.E. (2003) Phosphoproteome analysis of capacitated human sperm. Evidence of tyrosine phosphorylation of a kinase-anchoring protein 3 and valosin-containing

protein/p97 during capacitation, *J. Biol. Chem.* **278**, 11579–11589.

Fields, S., Song, O. (1989) A novel genetic system to detect protein-protein interactions, *Nature* **340**, 245–246.

Figeys, D. (2003) Proteomics in 2002: a year of technical development and wide-ranging applications, *Anal. Chem.* **75**, 2891–2905.

Florens, L., Washburn, M.P., Raine, J.D., Anthony, R.M., Grainger, M., Haynes, J.D., Moch, J.K., Muster, N., Sacci, J.B., Tabb, D.L., Witney, A.A., Wolters, D., Wu, Y., Gardner, M.J., Holder, A.A., Sinden, R.E., Yates, J.R., Carucci, D.J. (2002) A proteomic view of the plasmodium falciparum life cycle, *Nature* **419**, 520–526.

Fountoulakis, M., Takacs, M.F., Berndt, P., Langen, H., Takacs, B. (1999) Enrichment of low abundance proteins of escherichia coli by hydroxyapatite chromatography, *Electrophoresis* **20**, 2181–2195.

Fountoulakis, M., Takacs, M.F., Takacs, B. (1999) Enrichment of low-copy-number gene products by hydrophobic interaction chromatography, *J. Chromatogr., A* **833**, 157–168.

Friedman, D.B., Hill, S., Keller, J.W., Merchant, N.B., Levy, S.E., Coffey, R.J., Caprioli, R.M. (2004) Proteome analysis of human colon cancer by two-dimensional difference gel electrophoresis and mass spectrometry, *Proteomics* **4**, 793–811

Futcher, B., Latter, G.I., Monardo, P., McLaughlin, C.S., Garrels, J.I. (1999) A sampling of the yeast proteome, *Mol. Cell. Biol.* **19**, 7357–7368.

Gaucher, S.P., Cancilla, M.T., Phillips, N.J., Gibson, B.W., Leary, J.A. (2000) Mass spectral characterization of lipooligosaccharides from haemophilus influenzae 2019, *Biochemistry* **39**, 12406–12414.

Ghaemmaghami, S., Huh, W.K., Bower, K., Howson, R.W., Belle, A., Dephoure, N., O'Shea, E.K., Weissman, J.S. (2003) Global analysis of protein expression in yeast, *Nature* **425**, 737–741.

Glickman, M.H., Ciechanover, A. (2002) The ubiquitin-proteasome proteolytic pathway: destruction for the sake of construction, *Physiol. Rev.* **82**, 373–428.

Goff, S.A., Ricke, D., Lan, T.H., Presting, G., Wang, R., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H., Hadley, D., Hutchison, D., Martin, C., Katagiri, F., Lange, B.M., Moughamer, T., Xia, Y., Budworth, P., Zhong, J., Miguel, T., Paszkowski, U., Zhang, S., Colbert, M., Sun, W.L., Chen, L., Cooper, B., Park, S., Wood, T.C., Mao, L., Quail, P., Wing, R., Dean, R., Yu, Y., Zharkikh, A., Shen, R., Sahasrabudhe, S., Thomas, A., Cannings, R., Gutin, A., Pruss, D., Reid, J., Tavtigian, S., Mitchell, J., Eldredge, G., Scholl, T., Miller, R.M., Bhatnagar, S., Adey, N., Rubano, T., Tusneem, N., Robinson, R., Feldhaus, J., Macalma, T., Oliphant, A., Briggs, S. (2002) A draft sequence of the rice genome (Oryza sativa L. ssp. japonica), *Science* **296**, 92–100.

Goodlett, D.R., Keller, A., Watts, J.D., Newitt, R., Yi, E.C., Purvine, S., Eng, J.K., von Haller, P., Aebersold, R., Kolker, E. (2001) Differential stable isotope labeling of peptides for quantitation and de novo sequence derivation, *Rapid. Commun. Mass Spectrom.* **15**, 1214–1221.

Gorg, A., Obermaier, C., Boguth, G., Harder, A., Scheibe, B., Wildgruber, R., Weiss, W. (2000) The current state of two-dimensional electrophoresis with immobilized pH gradients, *Electrophoresis* **21**, 1037–1053.

Goshe, M.B., Conrads, T.P., Panisko, E.A., Angell, N.H., Veenstra, T.D., Smith, R.D. (2001) Phosphoprotein isotope-coded affinity tag approach for isolating and quantitating phosphopeptides in proteome-wide analyses, *Anal. Chem.* **73**, 2578–2586.

Goshe, M.B., Veenstra, T.D., Panisko, E.A., Conrads, T.P., Angell, N.H., Smith, R.D. (2002) Phosphoprotein isotope-coded affinity tags: application to the enrichment and identification of low-abundance phosphoproteins, *Anal. Chem.* **74**, 607–616.

Goto, N.K., Kay, L.E. (2000) New developments in isotope labeling strategies for protein solution NMR spectroscopy, *Curr. Opin. Struct. Biol.* **10**, 585–592.

Greenbaum, D., Colangelo, C., Williams, K., Gerstein, M. (2003) Comparing protein abundance and mRNA expression levels on a genomic scale, *Genome Biol.* **4**, 117.

Greenbaum, D., Jansen, R., Gerstein, M. (2002) Analysis of mRNA expression and protein abundance data: an approach for the comparison of the enrichment of features in the cellular population of proteins and transcripts, *Bioinformatics* **18**, 585–596.

Greis, K.D., Hayes, B.K., Comer, F.I., Kirk, M., Barnes, S., Lowary, T.L., Hart, G.W. (1996) Selective detection and site-analysis of

O-GlcNAc-modified glycopeptides by beta-elimination and tandem electrospray mass spectrometry, *Anal. Biochem.* **234**, 38–49.

Griffin, T.J., Gygi, S.P., Ideker, T., Rist, B., Eng, J., Hood, L., Aebersold, R. (2002) Complementary profiling of gene expression at the transcriptome and proteome levels in saccharomyces cerevisiae, *Mol. Cell. Proteomics* **1**, 323–333.

Gronborg, M., Kristiansen, T.Z., Stensballe, A., Andersen, J.S., Ohara, O., Mann, M., Jensen, O.N., Pandey, A. (2002) A mass spectrometry-based proteomic approach for identification of serine/threonine-phosphorylated proteins by enrichment with phospho-specific antibodies: identification of a novel protein, frigg, as a protein kinase a substrate, *Mol. Cell. Proteomics* **1**, 517–527.

Gygi, S.P., Corthals, G.L., Zhang, Y., Rochon, Y., Aebersold, R. (2000) Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 9390–9395.

Gygi, S.P., Rist, B., Gerber, S.A., Turecek, F., Gelb, M.H., Aebersold, R. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags, *Nat. Biotechnol.* **17**, 994–999.

Gygi, S.P., Rochon, Y., Franza, B.R., Aebersold, R. (1999) Correlation between protein and mRNA abundance in yeast, *Mol. Cell. Biol.* **19**, 1720–1730.

Ha, K.S. (2001) Principle and biological applications of protein chip system, *Exp. Mol. Med.* **33**, 127–132.

Haab, B.B., Dunham, M.J., Brown, P.O. (2001) Protein microarrays for highly parallel detection and quantitation of specific proteins and antibodies in complex solutions, *Genome Biol.* **2**, RESEARCH0004.1–0004.13.

Han, D.K., Eng, J., Zhou, H., Aebersold, R. (2001) Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry, *Nat. Biotechnol.* **19**, 946–951.

Hansen, A.P., Petros, A.M., Mazar, A.P., Pederson, T.M., Rueter, A., Fesik, S.W. (1992) A practical method for uniform isotopic labeling of recombinant proteins in mammalian cells, *Biochemistry* **31**, 12713–12718.

Hansen, K.C., Schmitt-Ulms, G., Chalkley, R.J., Hirsch, J., Baldwin, M.A., Burlingame, A.L. (2003) Mass spectrometric analysis of protein mixtures at low levels using cleavable 13C-isotope-coded affinity tag and multidimensional chromatography, *Mol. Cell. Proteomics* **2**, 299–314.

Hart, G.W., Haltiwanger, R.S., Holt, G.D., Kelly, W.G. (1989) Nucleoplasmic and cytoplasmic glycoproteins, *Ciba. Found. Symp.* **145**, 102–112, discussion 112–108.

Herbert, B. (1999) Advances in protein solubilization for two-dimensional electrophoresis, *Electrophoresis* **20**, 660–663.

Herbert, B., Righetti, P.G. (2000) A turning point in proteome analysis: sample prefractionation via multicompartment electrolyzers with isoelectric membranes, *Electrophoresis* **21**, 3639–3648.

Ideker, T., Thorsson, V., Ranish, J.A., Christmas, R., Buhler, J., Eng, J.K., Bumgarner, R., Goodlett, D.R., Aebersold, R., Hood, L. (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network, *Science* **292**, 929–934.

Imam-Sghiouar, N., Laude-Lemaire, I., Labas, V., Pflieger, D., Le Caer, J.P., Caron, M., Nabias, D.K., Joubert-Caron, R. (2002) Subproteomics analysis of phosphorylated proteins: application to the study of B-lymphoblasts from a patient with Scott syndrome, *Proteomics* **2**, 828–838.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., Sakaki, Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 4569–4574.

Jansen, R., Bussemaker, H.J., Gerstein, M. (2003) Revisiting the codon adaptation index from a whole-genome perspective: analyzing the relationship between gene expression and codon occurrence in yeast using a variety of models, *Nucleic Acids Res.* **31**, 2242–2251.

Jensen, P.K., Pasa-Tolic, L., Anderson, G.A., Horner, J.A., Lipton, M.S., Bruce, J.E., Smith, R.D. (1999) Probing proteomes using capillary isoelectric focusing-electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry, *Anal. Chem.* **71**, 2076–2084.

Jensen, P.K., Pasa-Tolic, L., Peden, K.K., Martinovic, S., Lipton, M.S., Anderson, G.A., Tolic, N., Wong, K.K., Smith, R.D. (2000) Mass spectrometric detection for capillary isoelectric focusing separations of complex protein mixtures, *Electrophoresis* **21**, 1372–1380.

Ji, J., Chakraborty, A., Geng, M., Zhang, X., Amini, A., Bina, M., Regnier, F. (2000)

Strategy for qualitative and quantitative analysis in proteomics based on signature peptides, *J. Chromatogr., B. Biomed. Sci. Appl.* **745**, 197–210.

Jiang, H., English, A.M. (2002) Quantitative analysis of the yeast proteome by incorporation of isitopically labeled leucine, *J. Proteome Res.* **1**, 345–350.

Karas, M., Hillenkamp, F. (1988) Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons, *Anal. Chem.* **60**, 2299–2301.

Klose, J. (1999) Large-gel 2-D electrophoresis, *Methods Mol. Biol.* **112**, 147–172.

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

Koller, A., Washburn, M.P., Lange, B.M., Andon, N.L., Deciu, C., Haynes, P.A., Hays, L., Schieltz, D., Ulaszek, R., Wei, J., Wolters, D., Yates, J.R., III (2002) Proteomic survey of metabolic pathways in rice, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 11969–11974.

Koonin, E.V. (2001) Computational genomics, *Curr. Biol.* **11**, R155–R158.

Krishna, R., Wold, F. (1993) Post-Translational Modification of Proteins, *Adv. Enzymol. Relat. Areas Mol. Biol.* **67**, 265–298.

Lahm, H.W., Langen, H. (2000) Mass spectrometry: a tool for the identification of proteins separated by gels, *Electrophoresis* **21**, 2105–2114.

Leavell, M.D., Leary, J.A., Yamasaki, R. (2002) Mass spectrometric strategy for the characterization of lipooligosaccharides from neisseria gonorrhoeae 302 using FTICR, *J. Am. Soc. Mass Spectrom.* **13**, 571–576.

Lee, Y., Mrksich, M. (2002) Protein chips: from concept to practice, *A TRENDS Guide to Proteomics II*. **20**, s14–s18.

Li, J., Steen, H., Gygi, S.P. (2003) Protein profiling with cleavable isotope-coded affinity tag (cICAT) reagents: the yeast salinity stress response, *Mol. Cell. Proteomics* **2**, 1198–1204.

Link, A. (2002) Multidimensional peptide separations in proteomics, *A TRENDS Guide to Proteomics II*. **20**, s8–s13.

Link, A.J., Eng, J., Schieltz, D.M., Carmack, E., Mize, G.J., Morris, D.R., Garvik, B.M., Yates, J.R. III (1999) Direct analysis of protein complexes using mass spectrometry, *Nat. Biotechnol.* **17**, 676–682.

Lipton, M.S., Pasa-Tolic, L., Anderson, G.A., Anderson, D.J., Auberry, D.L., Battista, J.R., Daly, M.J., Fredrickson, J., Hixson, K.K., Kostandarithes, H., Masselon, C., Markillie, L.M., Moore, R.J., Romine, M.F., Shen, Y., Stritmatter, E., Tolic, N., Udseth, H.R., Venkateswaran, A., Wong, K.K., Zhao, R., Smith, R.D. (2002) Global analysis of the deinococcus radiodurans proteome by using accurate mass tags, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 11049–11054.

Liska, A.J., Shevchenko, A. (2003) Expanding the organismal scope of proteomics: cross-species protein identification by mass spectrometry and its implications, *Proteomics* **3**, 19–28.

Liu, P., Regnier, F. (2002) An isotope coding strategy for proteomics involving both amine and carboxyl labeling, *J. Proteome Res.* **1**, 443–450.

Mamyrin, B.A., Karataev V.I., Shmikk D.V., Zagulin V.A. (1973) Reflectron for TOF-MS, *Sov. Phys. JETP* **37**, 45–48.

Mamyrin, B.A., Shmikk D.V. (1979) Linear reflection, *Sov. Phys. JETP* **49**, 762–764.

Mann, M., Wilm, M. (1994) Error-tolerant identification of peptides in sequence databases by peptide sequence tags, *Anal. Chem.* **66**, 4390–4399.

McLachlin, D.T., Chait, B.T. (2001) Analysis of phosphorylated proteins and peptides by mass spectrometry, *Curr. Opin. Chem. Biol.* **5**, 591–602.

von Mering, C., Kruse, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S., Bork, P. (2002) Comparative assessment of large scale data sets of rotein-protein interactions, *Nature* **417**, 399–403.

Mirgorodskaya, O.A., Kozmin, Y.P., Titov, M.I., Korner, R., Sonksen, C.P., Roepstorff, P. (2000) Quantitation of peptides and proteins by matrix-assisted laser desorption/ionization mass spectrometry using 18O-labeled internal standards, *Rapid. Commun. Mass Spectrom.* **14**, 1226–1232.

Molloy, M.P. (2000) Two-dimensional electrophoresis of membrane proteins using immobilized pH gradients, *Anal. Biochem.* **280**, 1–10.

Moskovets, E. (2000) Optimization of the mass reflector parameters for direct ion extraction, *Rapid. Commun. Mass Spectrom.* **14**, 150–155.

Moskovets, E., Karger, B.L. (2003) Mass calibration of a matrix-assisted laser desorption/ionization time-of-flight mass

spectrometer including the rise time of the delayed extraction pulse, *Rapid. Commun. Mass Spectrom.* **17**, 229–237.

Mrksich, M., Dike, L.E., Tien, J., Ingber, D.E., Whitesides, G.M. (1997) Using microcontact printing to pattern the attachment of mammalian cells to self-assembled monolayers of alkanethiolates on transparent films of gold and silver, *Exp. Cell. Res.* **235**, 305–313.

Munchbach, M., Quadroni, M., Miotto, G., James, P. (2000) Quantitation and facilitated de novo sequencing of proteins by isotopic N-terminal labeling of peptides with a fragmentation-directing moiety, *Anal. Chem.* **72**, 4047–4057.

Muszynska, G., Andersson, L., Porath, J. (1986) Selective adsorption of phosphoproteins on gel-immobilized ferric chelate, *Biochemistry* **25**, 6850–6853.

Noordewier, M.O., Warren, P.V. (2001) Gene expression microarrays and the integration of biological knowledge, *Trends Biotechnol.* **19**, 412–415.

Oda, Y., Huang, K., Cross, F.R., Cowburn, D., Chait, B.T. (1999) Accurate quantitation of protein expression and site-specific phosphorylation, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 6591–6596.

O'Farrell, P.H. (1975) High resolution two-dimensional electrophoresis of proteins, *J. Biol. Chem.* **250**, 4007–4021.

Oh-Ishi, M., Satoh, M., Maeda, T. (2000) Preparative two-dimensional gel electrophoresis with agarose gels in the first dimension for high molecular mass proteins, *Electrophoresis* **21**, 1653–1669.

Ong, S.E., Blagoev, B., Kratchmarova, I., Kristensen, D.B., Steen, H., Pandey, A., Mann, M. (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.

Pappin, D.J. (2003) Peptide mass fingerprinting using MALDI-TOF mass spectrometry, *Methods Mol. Biol.* **211**, 211–219.

Pasa-Tolic, L., Harkewicz, R., Anderson, G.A., Tolic, N., Shen, Y., Zhao, R., Thrall, B., Masselon, C., Smith, R.D. (2002) Increased proteome coverage for quantitative peptide abundance measurements based upon high performance separations and DREAMS FTICR mass spectrometry, *J. Am. Soc. Mass Spectrom.* **13**, 954–963.

Peters, E.C., Horn, D.M., Tully, D.C., Brock, A. (2001) A novel multifunctional labeling reagent for enhanced protein characterization with mass spectrometry, *Rapid. Commun. Mass Spectrom.* **15**, 2387–2392.

Petricoin, E.F., Ardekani, A.M., Hitt, B.A., Levine, P.J., Fusaro, V.A., Steinberg, S.M., Mills, G.B., Simone, C., Fishman, D.A., Kohn, E.C., Liotta, L.A. (2002) Use of proteomic patterns in serum to identify ovarian cancer, *Lancet* **359**, 572–577.

Puig, O., Caspary, F., Rigaut, G., Rutz, B., Bouveret, E., Bragado-Nilsson, E., Wilm, M., Seraphin, B. (2001) The tandem affinity purification (TAP) method: a general procedure of protein complex purification, *Methods* **24**, 218–229.

Rademacher, T.W., Parekh, R.B., Dwek, R.A. (1988) Glycobiology, *Annu. Rev. Biochem.* **57**, 785–838.

Reynolds, K.J., Yao, X., Fenselau, C. (2002) Proteolytic 18O labeling for comparative proteomics: evaluation of endoproteinase Glu-C as the catalytic agent, *J. Proteome Res.* **1**, 27–33.

Rosenblatt, K.P., Bryant-Greenwood, P., Killian, J.K., Mehta, A., Geho, D., Espina, V., Petricoin, E.F., Liotta, L.A. (2004) Serum proteomics in cancer diagnosis and management, *Annu. Rev. Med.* **55**, 97–112.

Rout, M.P., Aitchison, J.D., Suprapto, A., Hjertaas, K., Zhao, Y., Chait, B.T. (2000) The yeast nuclear pore complex: composition, architecture, and transport mechanism, *J. Cell. Biol.* **148**, 635–652.

Rudiger, H., Gabius, H.J. (2001) Plant lectins: occurrence, biochemistry, functions and applications, *Glycoconj. J.* **18**, 589–613.

Sanchez, J.C., Hochstrasser, D.F. (1999) High-resolution, IPG-based, mini two-dimensional gel electrophoresis, *Methods Mol. Biol.* **112**, 227–233.

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

Schwartz, J.C., Jardine, I. (1996) Quadrupole ion trap mass spectrometry, *Methods Enzymol.* **270**, 552–586.

Schwartz, J.C., Senko, M.W., Syka, J.E. (2002) A two-dimensional quadrupole ion trap mass spectrometer, *J. Am. Soc. Mass Spectrom.* **13**, 659–669.

Sheeley, D.M., Reinhold, V.N. (1998) Structural characterization of carbohydrate sequence, linkage, and branching in a quadrupole Ion trap mass spectrometer: neutral oligosaccharides and N-linked glycans, *Anal. Chem.* **70**, 3053–3059.

Siuzdak, G. (1994) The emergence of mass spectrometry in biochemical research, *Proc. Natl. Acad. Sci. U.S.A.* **91**, 11290–11297.

Smith, R.D., Anderson, G.A., Lipton, M.S., Pasa-Tolic, L., Shen, Y., Conrads, T.P., Veenstra, T.D., Udseth, H.R. (2002) An accurate mass tag strategy for quantitative and high-throughput proteome measurements, *Proteomics* **2**, 513–523.

Smith, R.D., Loo, J.A., Edmonds, C.G., Barinaga, C.J., Udseth, H.R. (1990) New developments in biochemical mass spectrometry: electrospray ionization, *Anal. Chem.* **62**, 882–899.

Snow, D.M., Hart, G.W. (1998) Nuclear and cytoplasmic glycosylation, *Int. Rev. Cytol.* **181**, 43–74.

Sunyaev, S., Liska, A.J., Golod, A., Shevchenko, A. (2003) MultiTag: multiple error-tolerant sequence tag search for the sequence-similarity identification of proteins by mass spectrometry, *Anal. Chem.* **75**, 1307–1315.

Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., Rothberg, J.M. (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae, *Nature* **403**, 623–627.

Unlu, M., Morgan, M.E., Minden, J.S. (1997) Difference gel electrophoresis: a single gel method for detecting changes in protein extracts, *Electrophoresis* **18**, 2071–2077.

Varshavsky, A. (1996) The N-end rule: functions, mysteries, uses, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 12142–12149.

Verma, R., Chen, S., Feldman, R., Schieltz, D., Yates, J., Dohmen, J., Deshaies, R.J. (2000) Proteasomal proteomics: identification of nucleotide-sensitive proteasome-interacting proteins by mass spectrometric analysis of affinity-purified proteasomes, *Mol. Biol. Cell.* **11**, 3425–3439.

Vosseller, K., Wells, L., Hart, G.W. (2001) Nucleocytoplasmic O-glycosylation: O-GlcNAc and functional proteomics, *Biochimie* **83**, 575–581.

Wall, D.B., Kachman, M.T., Gong, S., Hinderer, R., Parus, S., Misek, D.E., Hanash, S.M., Lubman, D.M. (2000) Isoelectric focusing nonporous RP HPLC: a two-dimensional liquid-phase separation method for mapping of cellular proteins with identification using MALDI-TOF mass spectrometry, *Anal. Chem.* **72**, 1099–1111.

Wall, D.B., Kachman, M.T., Gong, S.S., Parus, S.J., Long, M.W., Lubman, D.M. (2001) Isoelectric focusing nonporous silica reversed-phase high-performance liquid chromatography/electrospray ionization time-of-flight mass spectrometry: a three-dimensional liquid-phase protein separation method as applied to the human erythroleukemia cell-line, *Rapid. Commun. Mass Spectrom.* **15**, 1649–1661.

Washburn, M.P., Koller, A., Oshiro, G., Ulaszek, R.R., Plouffe, D., Deciu, C., Winzeler, E., Yates, J.R. III (2003) Protein pathway and complex clustering of correlated mRNA and protein expression analyses in Saccharomyces cerevisiae, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 3107–3112.

Washburn, M.P., Ulaszek, R., Deciu, C., Schieltz, D.M., Yates, J.R. III (2002) Analysis of quantitative proteomic data generated via multidimensional protein identification technology, *Anal. Chem.* **74**, 1650–1657.

Washburn, M.P., Wolters, D., Yates, J.R. III (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology, *Nat. Biotechnol.* **19**, 242–247.

Washburn, P.W., Yates, J.R. III (2000) New methods of proteome analysis: multidimensional chromatography and mass spectrometry, *Trends Biotechnol.* **18**, S27–S30.

Wasinger, V.C., Cordwell, S.J., Cerpa-Poljak, A., Yan, J.X., Gooley, A.A., Wilkins, M.R., Duncan, M.W., Harris, R., Williams, K.L., Humphery-Smith, I. (1995) Progress with gene-product mapping of the Mollicutes: Mycoplasma genitalium, *Electrophoresis* **16**, 1090–1094.

Watson, J.T. (1997) *Introduction to Mass Spectrometry*, Lippincott-Raven, Philadelphia, NY.

Wells, L., Whalen, S.A., Hart, G.W. (2003) O-GlcNAc: a regulatory post-translational modification, *Biochem. Biophys. Res. Commun.* **302**, 435–441.

Weng, S., Gu, K., Hammond, P.W., Lohse, P., Rise, C., Wagner, R.W., Wright, M.C., Kuimelis, R.G. (2002) Generating addressable protein microarrays with PROfusion covalent mRNA-protein fusion technology, *Proteomics* **2**, 48–57.

Wilkins, M.R., Sanchez, J.C., Gooley, A.A., Appel, R.D., Humphery-Smith, I., Hochstrasser, D.F., Williams, K.L. (1996) Progress with proteome projects: why all proteins expressed by a genome should be identified and how to do it, *Biotechnol. Genet. Eng. Rev.* **13**, 19–50.

Wilkins, M.R., Williams, K.L., Appel, R.D., Hochstrasser, D.F. (Eds.) (1997) *Proteome Research: New Frontiers in Functional Genomics (Principles and Practice)*, Springer-Verlag.

Wilson, N.L., Schulz, B.L., Karlsson, N.G., Packer, N.H. (2002) Sequential analysis of N- and O-linked glycosylation of 2D-PAGE separated glycoproteins, *J. Proteome Res.* **1**, 521–529.

Winslow, R.L., Boguski, M.S. (2003) Genome informatics: current status and future prospects, *Circ. Res.* **92**, 953–961.

Wolters, D.A., Washburn, M.P., Yates, J.R. III (2001) An automated multidimensional protein identification technology for shotgun proteomics, *Anal. Chem.* **73**, 5683–5690.

Yao, J., Burton, J.L., Saama, P., Sipkovsky, S., Coussens, P.M. (2001) Generation of EST and cDNA microarray resources for the study of bovine immunobiology, *Acta Vet. Scand.* **42**, 391–405.

Yates, J.R., Eng, J.K., McCormack, A.L., Schieltz, D. III (1995) Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database, *Anal. Chem.* **67**, 1426–1436.

Yost, R.A., Boyd, R.K. (1990) Tandem mass spectrometry: quadrupole and hybrid instruments, *Methods Enzymol.* **193**, 154–200.

Yu, J., Hu, S., Wang, J., Wong, G.K., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X., Cao, M., Liu, J., Sun, J., Tang, J., Chen, Y., Huang, X., Lin, W., Ye, C., Tong, W., Cong, L., Geng, J., Han, Y., Li, L., Li, W., Hu, G., Li, J., Liu, Z., Qi, Q., Li, T., Wang, X., Lu, H., Wu, T., Zhu, M., Ni, P., Han, H., Dong, W., Ren, X., Feng, X., Cui, P., Li, X., Wang, H., Xu, X., Zhai, W., Xu, Z., Zhang, J., He, S., Xu, J., Zhang, K., Zheng, X., Dong, J., Zeng, W., Tao, L., Ye, J., Tan, J., Chen, X., He, J., Liu, D., Tian, W., Tian, C., Xia, H., Bao, Q., Li, G., Gao, H., Cao, T., Zhao, W., Li, P., Chen, W., Zhang, Y., Hu, J., Liu, S., Yang, J., Zhang, G., Xiong, Y., Li, Z., Mao, L., Zhou, C., Zhu, Z., Chen, R., Hao, B., Zheng, W., Chen, S., Guo, W., Tao, M., Zhu, L., Yuan, L., Yang, H. (2002) A draft sequence of the rice genome (Oryza sativa L. ssp. indica), *Science* **296**, 79–92.

Zhu, H., Bilgin, M., Bangham, R., Hall, D., Casamayor, A., Bertone, P., Lan, N., Jansen, R., Bidlingmaier, S., Houfek, T., Mitchell, T., Miller, P., Dean, R.A., Gerstein, M., Snyder, M. (2001) Global analysis of protein activities using proteome chips, *Science* **293**, 2101–2105.