JPP JOURNAL OF PROTEINS AND PROTEOMICS
3(1), January-June 2012, pp. 39-45

TOOLS OF PROTEOMICS AND ITS APPLICATION IN BIOMARKER DISCOVERY

Kamala Vanarsa

Division of Rheumatic Diseases, Department of Internal Medicine, University of Texas Southwestern Medical School Dallas, TX 75390, USA

Abstract: In the recent years, the field of proteomics is progressing in leaps and bounds. In simple terms, the proteome is the complement of proteins expressed by the genome of an organism, tissue, or cell at any given time under any given condition. The study of the proteome has gained importance due to its ability to provide information about the post-translational modifications in proteins as well as protein-protein interactions, which play a crucial role in the pathophysiology of cancer and inflammatory disorders. This article will serve as a primer to help understand the basic tools and approaches used in the field of proteomics and compare and contrast the different approaches with each other. Additionally, it will also familiarize the reader with the clinical applications of proteomics in the field of cancer and other important diseases in which modified proteins are involved.

Keywords: proteomics; protein/peptide separation; mass spectrometry; biomarker

Introduction

Each cell contains a complete set of genetic information within its chromosomes. This genetic template encodes the information necessary for the cells to function, survive, and replicate. Within an organism, there are a number of diverse cell types, each with their own function and purpose. While genes essential for basic functions, such as metabolism, are expressed in every cell, those necessary for specific functions are confined to distinct cell types. Proteins are the products of expressed genes, and carry out necessary cellular functions. Because of this, protein expression patterns vary widely among the many cell types present in an organism. Even within one type of cell, differing subsets and amounts of proteins are required under different circumstances. Thus, even though the complete genome is present in every cell, gene expression patterns change depending upon the cell and the biological context. As such, there is one genome for an organism, but many expressed *proteomes*.

Corresponding Author: **Kamala Vanarsa**
E-mail: Kamala.Vanarsa@Utsouthwestern.edu

The terms *proteome,* and by extension *proteomics,* were coined by Marc Wilkins during the first Siena meeting in 1994 (Godovac-Zimmermann, 2008). By definition, the proteome is the complement of proteins expressed by the genome. Unlike this straightforward definition, the study of a proteome is much more complex. The main aim of proteomics is to understand the function of all proteins in a system rather than individual proteins. Therefore, in order to fully comprehend the proteome of a tissue, cell, or organelle, not only must the primary amino acid sequence be known but also the identity, post-translational modifications, and copy number of each protein. In addition, the ways in which these proteins interact with each other and other cellular components must be determined. In this way, proteomics yields a detailed picture of how multiple protein pathways in a cell interact with each other to carry out various physiological functions. The end product of this analysis is an identified network of protein system interactions known as the *interactome*. Proteomics allows the study of inherent differences in protein interactions between, for example, a diseased and

a healthy sample, two treatment protocols, or multiple disease processes. A significant change in protein expression in a diseased sample compared to the relevant control sample may allude to a potential role for one or more proteins in physiological or biological processes, and these proteins may constitute novel disease biomarkers, or even potential therapeutic targets.

As with any discipline, a successful proteomics study is the result of the proper application of the required tools. In proteomics, there are three basic tools that have to be considered for every study: protein/peptide separation, mass spectrometry, and database analysis. This primer will explore the basic concepts of these tools, and compare and contrast different approaches for each.

I. Tools of proteomics

After obtaining a complex protein mixture from a cell or tissue sample, individual proteins are initially separated, then subjected to enzymatic digestion to acquire the requisite peptides. These peptides are in turn separated by liquid chromatography and then loaded into a mass spectrometer. Proteins are subsequently identified using software to compare the obtained mass spectrometry results to a known database (Figure 1).

Protein/Peptide separation

Separation can be done either at the protein level or at the peptide level by using gel-based or non-gel-based techniques, respectively. While the end results are essentially the same, there are several different procedures for performing each separation. Proteins can be separated using sodium dodecyl sulfate polyacrylamide gel electrophoresis (Reinders *et al.*, 2006) (SDS-PAGE, based on molecular weight), isoelectric focusing (IEF, based on pI), or 2D gel electrophoresis (based on both molecular weight and pI). Indeed, 2D gel electrophoresis has been one of the most prevalently used techniques over the past decade, largely due to the fact that it is easy to use, relatively inexpensive, and a powerful means of separating complex protein mixtures. On the downside, 2D gel electrophoresis is not easily reproducible, large hydrophobic proteins tend to

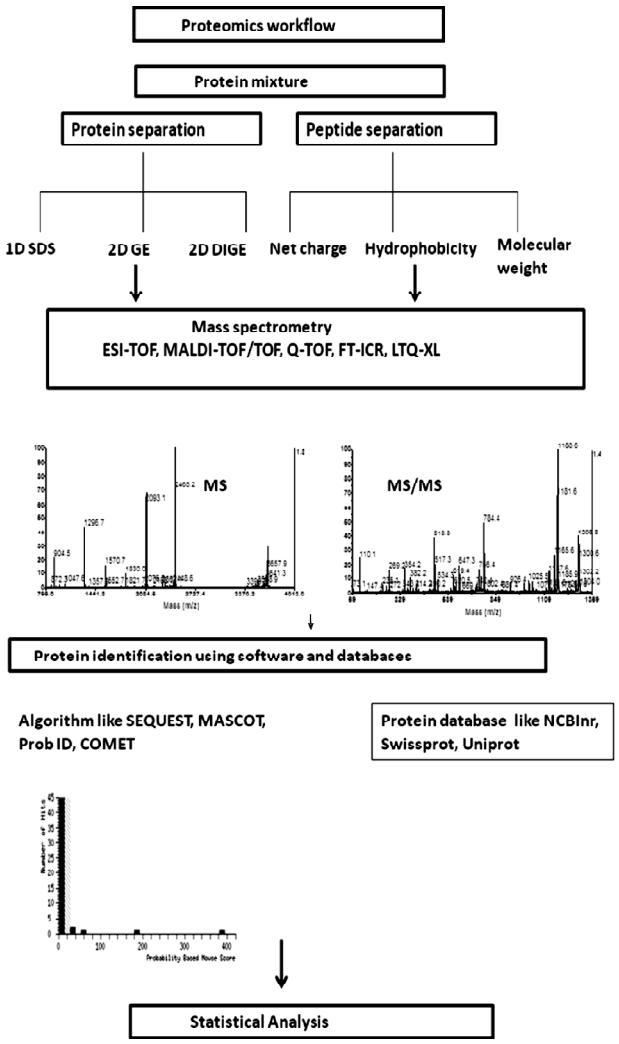

Figure 1: **An illustration of the main steps involved in proteomics analysis, including protein separation, digestion, mass spectrometric analysis and protein identification.**

smear under IEF, and staining is not efficient for proteins of low abundance. To address these problems, many researchers have turned to 2D differential in-gel electrophoresis (2D-DIGE), in which samples are differentially labeled with fluorescent dyes and then separated based on pH and molecular weight (Marouga *et al.*, 2005). The power of this method lies in the ability to label two unique samples with different fluorescent dyes and run them on the same gel, allowing for direct comparisons between samples (healthy controls and diseased samples, for instance). This not only reduces the need for multiple gels, but mitigates the inter-gel variability intrinsic to the traditional 2D electrophoresis method. The shift to fluorescent dyes has also resulted in increased

sensitivity, as background staining is dramatically reduced in comparison to traditional silver staining; this makes detection of low abundance proteins possible. While 2D-DIGE provides several improvements over 2D electrophoresis, the dynamic range of the technique is limited as it is not reliable for identifying proteins with extreme pI values or molecular weights (Monteoliva and Albar, 2004).

Alternately, peptide separation can be done using high performance liquid chromatography (HPLC, based on hydrophobicity), strong cation exchange chromatography (based on charge), or more recently, multidimensional protein identification technology (MUDPIT, based on hydrophobicity and charge) (Monteoliva and Albar, 2004; Washburn *et al.*, 2001; Wolters *et al.*, 2001). These methods offer greater separation of peptides, as the peptide is directly interfaced with the ion source, thereby maximizing sensitivity.

Mass Spectrometry

Following sample separation, mass spectrometry (MS) is used to generate and classify ionic fragments derived from peptides based on their mass-to-charge ratio, which is also referred to as the m/z value. There are many types of mass spectrometers and protocols in current use, but it is of paramount importance to consider the following three points in order to derive good MS data: (1) high sensitivity; which is normally expressed as the amount of analyte detected at a given signal-to-noise ratio with desired values in the femto (10^{-15}) to zepto (10^{-21}) molar range; (2) high resolution; defined as the ability to distinguish two ions that have nearly identical m/z values (Figure 2); (3) mass accuracy; defined as the ratio of the m/z measurement error to the true m/z with desired values approaching 1 ppm.

While there are a variety of mass spectrometers, each possesses three basic parts: (a) an ionizer- where the peptide sample is hit by a beam of electrons, thereby producing ionic fragments; (b) a mass analyzer- which resolves the ionic fragments based on m/z values; and (c) a detector- which qualifies and quantifies the ionic fragments (Figure 3). Various options for each component are compared and contrasted below:

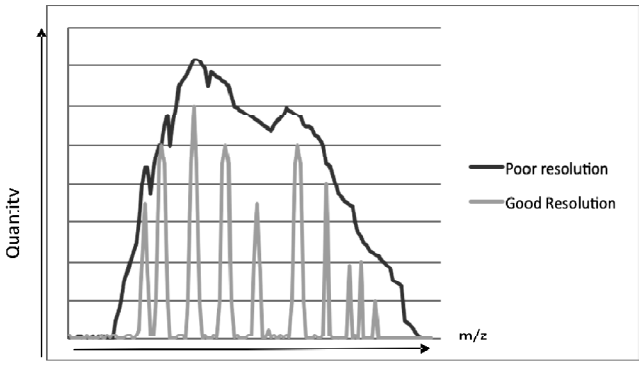

Figure 2: **A mass spectrometer with good mass resolution can determine the accurate mass of each molecule/ion/peptide while those with poor mass resolution determine only the average mass.**

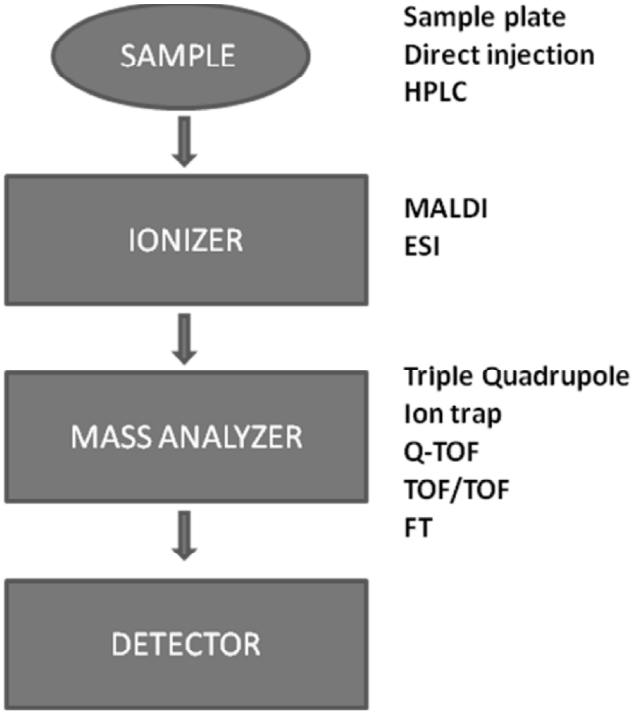

Figure 3: **Illustration of the basic steps in a mass spectrometric analysis.**

(a) Ionizer: The two most commonly used ionizers are MALDI (Matrix-assisted laser desorption ionization) and ESI (Electrospray ionization) (Bodnar *et al.*, 2003). In MALDI, the peptide sample is mixed with a matrix (composed of organic solvents plus a chromophore) and upon interaction with a laser beam the chromophore in the matrix absorbs photons and becomes electronically excited. This energy is picked up by the peptides, which are then released as positive ions that travel through the time-of-flight analyzer and strike the detector. In

ESI, peptides in an acidic solution are expelled from a narrow needle held under a strong electric field. Positively charged peptide ions then pass through a mass analyzer.

(b) Tandem Mass analyzers: Several versions of mass analyzers have been used. Table 1 summarizes the different characteristics of these analyzers.

Table 1
Comparison of the Mass Accuracy and Mass Resolution of the Leading Mass Analyzers

Mass Analyzer	*Mass Resolution*	*Mass Accuracy*
Time of Flight (TOF)	10,000 to 20,000	10-100 ppm
Quadrupole (Q)	20,000	100-1000 ppm
Quadrupole ion trap (LCQ or GCQ)	20,000	100-1000 ppm
Linear Quadrupole trap (LTQ)	20,000	100-1000 ppm
Fourier Transform (FT)	upto 1M	0.1-1ppm
Orbitrap	30,000 to 60,000	0.1-1ppm

(***i) Triple Quadrupole Mass analyzer***: This is the oldest version of the tandem MS. The name *triple quadrupole* refers to the presence of three sets of four metal rods. These *quadrupoles* act as mass filters with a collision cell in-between. A magnetic field is created by the application of direct current and radio frequency voltages across the metal rods. Depending upon the voltage, ions of specific m/z values pass through the rods in a corkscrew fashion, while the rest of the ions deviate and fly outwards.

(ii) Ion Trap mass analyzer: An Ion Trap mass analyzer consists of four electrodes: one top, one bottom, and two ring electrodes. Ions ejected from an ion source enter the ion trap and are maintained in orbit due to the applied direct current and radio frequency voltage. Depending upon the voltage applied, ions of specific m/z value are selected, and by increasing the voltage the ion is subjected to a collision with helium gas. The released fragments are then scanned according to their m/z value.

(***iii) Q-TOF (Quadrupole Time-of-Flight) mass analyzer***: The Q-TOF is functionally similar to a Triple Quadrupole mass analyzer, except the second quadrupole (Q2) is replaced by Time-of-Flight (TOF) analyzer. This analyzer is capable of much higher resolution than the Triple Quadrupole mass analyzer.

(***iv) TOF/TOF mass analyzer***: In this mass analyzer, ions head through a TOF analyzer that measures the time of flight before ions strike the detector. Earlier MALDI ionizers were equipped only with a single TOF and a detector. More recent MALDI instruments have incorporated a reflectron and a second TOF analyzer. The reflectron serves as a lens to improve the resolution of TOF instruments and to focus ions of a given m/z value, allowing them to pass through the second TOF analyzer and reach the detector at same time. The major drawback of TOF/TOF is that it is highly dependent on the quality of the sample and it is not well suited for gathering sequence information.

(***v) Fourier Transform-ion cyclotron resonance (FT-ICR) Mass analyzer***: The FT-ICR is similar to an Ion Trap mass analyzer, but employs a much higher magnetic field (3-7 tesla). Instruments with this mass analyzer can obtain spectacular resolution (figure 4). Weighed against this advantage is the higher cost of this type of equipment.

(***c) Detector***: Mass spectrometers with best resolution are equipped with a pair of metal surfaces that acts as a detector. The other commonly used detectors are faraday cups, electron multiplier, ion to photon detectors.

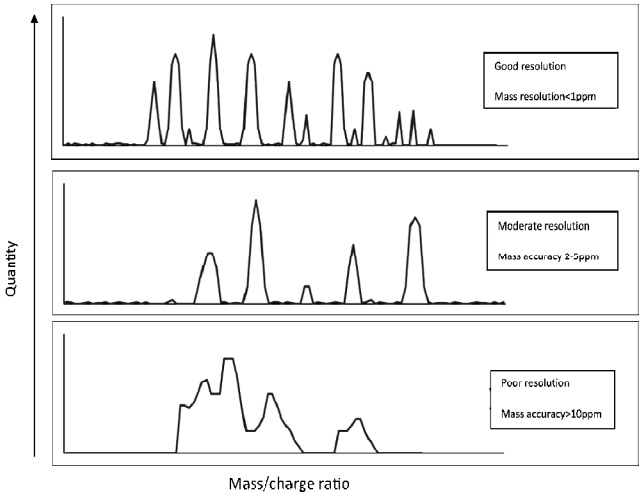

Figure 4: **Resolution versus mass accuracy. Mass measurement accuracy depends on mass resolution, with good resolution giving better mass accuracy.**

Software and databases

Mass spectrometers create output in a spectral readout form. The identification of proteins from the MS/MS spectrum can be done either by de novo interpretation or by using software algorithms like Mascot (by Matrix Science) and SEQUEST. De novo interpretation is done by BLAST search of an obtained peptide sequence against a published protein sequence database. This method is preferred when searching only a few proteins. When dealing with a complex mixture of proteins, it is preferable to use an algorithm to identify proteins by matching peptide MS/MS data to a computerized database. SEQUEST analyzes precursor ions with specific m/z values, then identifies peptides from a database that have the same mass as the precursor. In this way, virtual MS/MS spectra are produced. Correlation scores are calculated by comparing the virtual and original MS/MS spectra, and then proteins with the highest score are identified as the best matches. The Mascot program uses a probability-based mowse algorithm to identify proteins from databases. It can be used for peptide mass fingerprinting or MS/MS data analysis.

Quantitative protein profiling techniques

The above three steps are common to all mass spectrometry-based protein profiling or proteomic approaches. While all of the techniques offer qualitative information, they vary in their ability to yield quantitative insights. For example, although the proteomic procedure may be able to identify that a particular protein is increased in a diseased tissue, it may not indicate the precise fold change in expression. More recent protein profiling techniques have introduced important modifications that allow accurate quantitative comparisons to be executed.

Various quantitative proteomic techniques help researches identify and quantify proteins that vary significantly between different cell and tissue conditions. In contrast, protein profiling allows the comparison of protein ratios in disease states versus a reference control. Some commonly used quantitative protein profiling techniques include: 2D-DIGE, Cleavable isotope-coded affinity tag (cICAT), and Isobaric tag for relative and absolute quantitation (iTRAQ).

As discussed above, in 2D-DIGE two distinct protein samples are stained using two different dyes, and then co-loaded onto the same gel. Variations in these differentially-stained proteins are measured by densitometry analysis. In cICAT and iTRAQ, isotopic and isobaric tags are used for protein and peptide quantification, respectively. cICAT is a cysteine-specific labeling procedure, whereas iTRAQ is a stable isotopic labeling method used for multiplexed analysis.

II. Applications of proteomics techniques

In this review, the current state of proteomics technology has briefly been covered. However, without applying to a biological context, these techniques are merely theoretical. Many groups have utilized proteomics to identify potential biomarkers that have prognostic and/or diagnostic applications in disease physiology, progress, and treatment. Sampling of biomarkers that have been identified in disease conditions with specific proteomics approaches are listed below, but these are by no means comprehensive.

(i) Proteomics in Cancer: Using surface-enhanced laser desorption/ionization time-of-flight mass spectrometry (SELDI-TOF MS) four protein biomarkers were detected in urine of cancer patients: tryptophan (TRP), 5-hydroxytryptophan (5-HTP), 5-hydroxytryptamine (5-HT), and 5-hydroxyindole acetic acid (5-HIAA) (Kuo *et al.*, 2011). Using MALDI-TOF-MS, two potential biomarkers, element of component 3 and eukaryotic peptide chain release factor GTP-binding subunit ERF, were discovered in the serum of non-small cell lung carcinoma (NSCLC) patients and were described as surrogate markers having potential significance in detection and classification of NSCLC(Du *et al.*, 2011). Using 2D-DIGE and LC-tandem mass spectrometry, 12 specific proteins have been identified in breast carcinoma cells that had metastasized to the brain (MB231-Br) (Li *et al.*, 2011). These few examples are listed as successful prototypes in the application of proteomics to clinical oncology, but represent only a fraction of reported studies.

(***ii) Proteomics in Rheumatology***: Using the isobaric Tagging for Relative and Absolute protein Quantification (iTRAQ) proteomic

technique, 62 novel biomarkers were identified in the peripheral blood mononuclear cells (PBMC) of systemic lupus erythematosus (SLE) patients (Wang *et al*., 2012). MALDI-TOF MS was used to identify elevated levels of neutrophil gelatinase-associated lipocalin (NGAL) in synovial fluid of rheumatoid arthritis patients (Vanarsa and Mohan, 2010). Once again, these serve simply to illustrate the application of proteomics in clinical medicine.

(***iii) Proteomics in Coronary Diseases***: Using LC-ESI-MS/MS, Vaisar et al. documented that the composition of HDL proteins in patients with atherosclerosis is different from that of healthy subjects (Vaisar *et al*., 2007). In addition, use of 2D-DIGE and MS indicated that cathepsin D may mediate autophagy in chronically ischemic myocardium (Yan *et al*., 2005). Thus, proteomics is also helping define the underlying pathogenic mechanisms in various diseases.

(***iv) Proteomics in Alzheimer's disease***: Using immune precipitation and MALDI-TOF mass spectrometry, it was been determined that Aβ1-42 peptides are expressed at lower levels in the cerebrospinal fluid of Alzheimer's disease patients than healthy controls (Gelfanova *et al*., 2007). In addition, ESI and FT-ICR mass spectrometry were applied during extensive studies that examined amyloid fibril formation in Alzheimer's disease patients (Jarrett and Lansbury, 1992). Readers are encouraged to review these illustrative reports to understand how proteomics is reshaping medical research today.

Discussion

While much has been accomplished in the field of proteomics thus far, it is a technology that is still in its infancy. The number of publications in the proteomics field has increased tremendously in recent years, indicating an increased application of proteomic approaches and techniques in studies of biomarker discovery, biomolecule characterization, forensic analysis, pharmacokinetics, and other fields. Although there are many different techniques and sets of instrumentation that can be used to generate proteomics data, in general a clean sample separated extensively in all possible dimensions, then analyzed by an appropriate mass spectrometer with good resolution and accuracy will result in good data. The insights gained from this powerful technology can dramatically impact future science and medicine.

Abbreviations

SDS-PAGE, sodium dodecyl sulfate polyacrylamide gel electrophoresis; IEF, isoelectric focusing; 2D-DIGE, two dimensional differential in-gel electrophoresis; MUDPIT, multidimensional protein identification technology; MALDI, Matrix-assisted laser desorption ionization; ESI, Electrospray ionization; Q-TOF, Quadrupole Time-of-Flight; FT-ICR, Fourier Transform-ion cyclotron resonance; cICAT, Cleavable isotope-coded affinity tag; iTRAQ, Isobaric tag for relative and absolute quantitation; SELDI-TOF MS, Surface-enhanced laser desorption/ionization time-of-flight mass spectrometry; TRP, tryptophan; 5-HTP, 5-hydroxytryptophan; 5-HT, 5-hydroxytryptamine; 5-HIAA, 5-hydroxyindole acetic acid; NSCLC, non-small cell lung carcinoma; PBMC, peripheral blood mononuclear cells; SLE, systemic lupus erythematosus; NGAL, neutrophil gelatinase-associated lipocalin.

References

Bodnar, W. M., Blackburn, R. K., Krise, J. M., and Moseley, M. A. (2003). Exploiting the complementary nature of LC/MALDI/MS/MS and LC/ESI/MS/MS for increased proteome coverage. *J Am Soc Mass Spectrom 14*, 971-979.

Du, J., Yang, S.Y., Lin, X.L., Shang, W.L., Zhang, W., Huo, S.F., Bu, L.N., Zhou, B., Nan, Y.D., Zheng, H.D., *et al.* (2011). Biomarker discovery and identification from non-small cell lung cancer sera. *Front Biosci (Elite Ed) 3*, 1-10.

Gelfanova, V., Higgs, R. E., Dean, R. A., Holtzman, D. M., Farlow, M. R., Siemers, E. R., Boodhoo, A., Qian, Y. W., He, X., Jin, Z., *et al.* (2007). Quantitative analysis of amyloid-beta peptides in cerebrospinal fluid using immunoprecipitation and MALDI-Tof mass spectrometry. *Brief Funct Genomic Proteomic 6*, 149-158.

Godovac-Zimmermann, J. (2008). 8th Siena meeting. From genome to proteome: integration and proteome completion. *Expert Rev Proteomics 5*, 769-773.

Jarrett, J. T., and Lansbury, P. T., Jr. (1992). Amyloid fibril formation requires a chemically discriminating nucleation event: studies of an amyloidogenic sequence from the bacterial protein OsmB. *Biochemistry 31*, 12345-12352.

Kuo, T. R., Chen, J. S., Chiu, Y. C., Tsai, C. Y., Hu, C. C., and Chen, C. C. (2011). Quantitative analysis of multiple urinary biomarkers of carcinoid tumors through gold-nanoparticle-assisted laser desorption/ ionization time-of-flight mass spectrometry. *Anal Chim Acta 699*, 81-86.

Li, F., Glinskii, O.V., Zhou, J., Wilson, L.S., Barnes, S., Anthony, D.C., and Glinsky, V.V. (2011). Identification and analysis of signaling networks potentially involved in breast carcinoma metastasis to the brain. *PLoS One 6*, e21977.

Marouga, R., David, S., and Hawkins, E. (2005). The development of the DIGE system: 2D fluorescence difference gel analysis technology. *Anal Bioanal Chem 382*, 669-678.

Monteoliva, L., and Albar, J.P. (2004). Differential proteomics: an overview of gel and non-gel based approaches. *Brief Funct Genomic Proteomic 3*, 220-239.

Reinders, J., Zahedi, R.P., Pfanner, N., Meisinger, C., and Sickmann, A. (2006). Toward the complete yeast mitochondrial proteome: multidimensional separation techniques for mitochondrial proteomics. *J Proteome Res 5*, 1543-1554.

Vaisar, T., Pennathur, S., Green, P.S., Gharib, S.A., Hoofnagle, A.N., Cheung, M.C., Byun, J., Vuletic, S., Kassim, S., Singh, P., *et al*. (2007). Shotgun proteomics implicates protease inhibition and complement activation in the antiinflammatory properties of HDL. *J Clin Invest 117*, 746-756.

Vanarsa, K., and Mohan, C. (2010). Proteomics in rheumatology: the dawn of a new era. *F1000 Med Rep 2*, 87.

Wang, L., Dai, Y., Qi, S., Sun, B., Wen, J., Zhang, L., and Tu, Z. (2012). Comparative proteome analysis of peripheral blood mononuclear cells in systemic lupus erythematosus with iTRAQ quantitative proteomics. *Rheumatol Int 32*, 585-593.

Washburn, M.P., Wolters, D., and Yates, J.R., 3rd (2001). Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol 19*, 242-247.

Wolters, D. A., Washburn, M. P., and Yates, J. R., 3rd (2001). An automated multidimensional protein identification technology for shotgun proteomics. *Anal Chem 73*, 5683-5690.

Yan, L., Vatner, D.E., Kim, S.J., Ge, H., Masurekar, M., Massover, W.H., Yang, G., Matsui, Y., Sadoshima, J., and Vatner, S.F. (2005). Autophagy in chronically ischemic myocardium. *Proc Natl Acad Sci U S A 102*, 13807-13812.