Journal of Proteomics & Bioinformatics - Open Access

JPB/Vol.2/October 2009

Review Article | OPEN ACCESS Freely available online | doi:10.4172/jpb.1000102

LC–MS Based Detection of Differential Protein Expression

Leepika Tuli and Habtom W. Ressom*

Georgetown University, Lombardi Comprehensive Cancer Center, Washington DC, USA

Abstract

While several techniques are available in proteomics, LC-MS based analysis of complex protein/peptide mixtures has turned out to be a mainstream analytical technique for quantitative proteomics. Significant technical advances at both sample preparation/separation and mass spectrometry levels have revolutionized comprehensive proteome analysis. Moreover, automation and robotics for sample handling process permit multiple sampling with high throughput.

For LC-MS based quantitative proteomics, sample preparation turns out to be critical step, as it can significantly influence sensitivity of downstream analysis. Several sample preparation strategies exist, including depletion of high abundant proteins or enrichment steps that facilitate protein quantification but with a compromise of focusing on a smaller subset of a proteome. While several experimental strategies have emerged, certain limitations such as physiochemical properties of a peptide/protein, protein turnover in a sample, analytical platform used for sample analysis and data processing, still imply challenges to quantitative proteomics. Other aspects that make analysis of a proteome a challenging task include dynamic nature of a proteome, need for efficient and fast analysis of protein due to its constant modifications inside a cell, concentration range of proteins that exceed dynamic range of a single analytical method, and absence of appropriate bioinformatics tools for analysis of large volume and high dimensional data.

This paper gives an overview of various LC-MS methods currently used in quantitative proteomics and their potential for detecting differential protein expression. Fundamental steps such as sample preparation, LC separation, mass spectrometry, quantitative assessment and protein identification are discussed.

For quantitative assessment of protein expression, both label and label free approaches are evaluated for their set of merits and demerits. While most of these methods edge on providing "relative abundance" information, absolute quantification is achieved with limitation as it caters to fewer proteins. Isotope labeling is extensively used for quantifying differentially expressed proteins, but is severely limited by successful incorporation of its heavy label. Lengthy labeling protocols restrict the number of samples that can be labeled and processed. Alternatively, label free approach appears promising as it can process many samples with any number of comparisons possible but entails reproducible experimental data for its application.

Keywords: Liquid chromatography-mass spectrometry (LC-MS); Quantitative proteomics; Labeling; Label-free; Tandem mass spectrometry (MS/MS)

Abbreviations: 2DE: Two-dimensional electrophoresis; ACO: Ant colony optimization; CAD: collision activated dissociation; CART: Classification and regression tree; CID: Collision induced dissociation; COW: Correlation optimized warping; CPM: Continuous profile model; Da: Dalton; DTW: Dynamic time warping; ECD: Electron capture dissociation; ESI: Electrospray ionization; ETD: Electron transfer dissociation; FT-ICR: Fourier transform-ion cyclotron resonance; GC: Gas chromatography; HIC: Hydrophobic interaction chromatography; HILIC: Hydrophilic interaction liquid chromatography; HPLC: High performance liquid chromatography; ICAT: Isotope coded affinity tags; ICPL: Isotope coded protein labeling; ICR: Ion cyclotron resonance; IEX: Ion exchange; IMAC: Immobilized metal affinity chromatography; IPC: Ion pair chromatography; iTRAQ: Isobaric tag for relative and absolute quantification; IT: Ion Trap; LC: Liquid chromatography; LDI: Laser desorption ionization; LIT: Linear ion trap; MALDI: Matrix-assisted laser desorption; MeCAT: Metal coded tags; MRM: Multiple reaction monitoring; MS: Mass spectrometry; MS/MS: Tandem mass spectrometry; MW: Molecular weight; NP: Normal phase; PCA: Principal component analysis; PF2D: Protein fractionation two dimensional; PLS: Partial least squares; PMF: Peptide mass fingerprint; PSAQ: Protein standard absolute quantification; PTM: Post-translational modification; Q: Quadrupole; QIT: Quadrupole ion trap; Q-TOF: Quadrupole-time of flight; RFE: Recursive feature elimination; RP: Reverse phase; RT: Retention time; SAX: Strong anion exchange; SCX: Strong cation exchange; SEC: Size exclusion chromatography; SELDI: Surface-enhanced laser desorption ionization; SILAC: Stable isotope labeling by amino acids in cell culture; SRM: Selected reaction monitoring; SVM: Support vector machine; TIC: Total ion current; TLC: Thin layer chromatography; TOF: Time of flight; UPLC: Ultra-performance liquid chromatography; WCX: Weak cation exchange; XIC: Extracted ion current

***Corresponding author:** Habtom W. Ressom, Georgetown University, Lombardi Comprehensive Cancer Center, Washington DC, USA, Tel: 1 (202) 687-2283; E-mail: hwr@georgetown.edu

Received August 12, 2009; **Accepted** October 01, 2009; **Published** October 02, 2009

Citation: Tuli L, Ressom HW (2009) LC–MS Based Detection of Differential Protein Expression. J Proteomics Bioinform 2: 416-438. doi:10.4172/jpb.1000102

Proteomics

Genomics era arrived with a promise of providing complete genome sequence needed for comprehensive analysis of an organism. However, it was discovered much later that a genome could be predominantly static and unaltered in response to extra- and intracellular influences (Souchelnytskyi, 2005). The obvious line of attack was to explore transcriptomics as both genome and proteome were dynamically linked to it. Transcriptomic studies were successful as they provided quantitative information on mRNA transcripts generated for certain point of time. But a lack of correlation observed, between mRNA and protein expression levels, eventually led investigators to focus directly on proteins, referred to as ultimate effectors of a cell.

Proteomics plays a central role in the discovery process due to its diverse applications – mechanism of disease process, drug targets, nutritional and environmental science, functional genomics etc. Proteomics focuses on identifying and quantifying proteins, characterizing them based on interaction, pre-translational and post-translational modifications, sub-cellular localization, and structure under physiological conditions. Based upon its underlying approach, proteomics is categorized as: expression, structural, and functional. Expression proteomics deals with quantitative comparison of proteins that differ by an experimental condition. Proteins are profiled on expression level changes or any modifications that may have occurred between groups that are being compared (Souchelnytskyi, 2005). Structural proteomics, on the other hand aims at mapping out structure of a protein complex or specific protein isolated from a system. Likewise, functional proteomics characterizes a selective group of proteins and assigns function derived from protein signaling and/or drug interaction mechanism. In this article, we focus on quantitative aspects of expression proteomics and its workflow. Due to its ability to reflect dynamic nature of all cellular processes and provide a global integrated view of all entities, quantitative proteomics has been extensively used for monitoring both physiological phenomenon and pathological conditions (Souchelnytskyi, 2005).

One such area that has greatly benefited from quantitative comparisons is biomarker discovery in cancer research. Biomarker discovery entails quantitative analysis and identification of proteins that can be mapped back to the cause of the condition. The basis for biomarker discovery is to develop diagnostic techniques that facilitate early detection and treatment options. Despite its enormous clinical importance, the overall process of biomarker finding is long with validation needed at several steps of discovery process. (Chambers et al., 2000). Irrespective of individual objectives of a proteome study, most researchers find themselves interested in identifying proteins of relevance or comparing protein abundances under different conditions (normal and diseased); with mass spectrometry (MS) as the enabling technology. Recently, several protein profiling technologies have been developed that allow identification of hundreds of proteins and facilitate quantitative comparison of analytes from cell, tissue, or human fluid samples (Graves and Haystead, 2002). Such progress has primarily resulted from initial and continuous development of instrumentation and analytical methods such as mass spectrometry and chromatographic and electrophoretic separation as well as from data analysis tools. Since protein identification and quantification are complimentary to each other most proteomics studies include the following: i) Extraction and isolation of proteins, ii) Separation of proteins/peptides (2D gel and non-gel) iii) Data acquisition of protein fragmentation pattern using mass spectrometry, and iv) Database search to reveal protein identification. While these steps are still being improved and developed to accommodate multi-dimensional data (Graves and Haystead, 2002), inherent factors that still make proteome investigations a difficult task are broad dynamic nature of a proteome and the in ability to capture constantly changing dynamics of a proteome (Graves and Haystead, 2002).

Quantitative Proteomics

Quantitative proteomics is an extension of expression proteomics as it provides quantitative information (relative or absolute) on existing proteins within a sample. Since biological processes are mainly controlled by proteins, it is desirable to study and compare proteins directly. Obtaining accurate information on protein is crucial, as any change in response to external influence, indicates toward proteins that control underlying biological mechanisms. Further to this, quantitative protein information can also be used towards modeling of biochemical networks. Absolute concentrations of proteins help build high definition models; whereas relative quantitative data can be used to compare protein expression levels among samples, provided the expression levels are normalized to a reference protein within sample (Souchelnytskyi, 2005).

Even though current quantitative proteomics is far from characterizing comprehensive proteome, several techniques exist that are successful in extracting quantitative information on proteins in their own limited way. These techniques include two-dimensional electrophoresis, protein microarrays, microfludics, and liquid chromatography coupled with mass spectrometry as outlined below.

Two-dimensional electrophoresis (2DE) method resolves proteins as spots on a gel each spot specified by its molecular weight (MW) and isoelectric point (pI). Widely used for complex proteins mixtures, this technique has been successfully employed as a tool for examining pathological processes such as Alzheimer's disease, schizophrenia (neuropathology), cardiac hypertrophy, and cardiomyopathy (cardiovascular diseases) progression (Souchelnytskyi, 2005). However, certain drawbacks exist: difficulty in accommodating hydrophobic proteins or extracting less soluble proteins (membrane proteins) and inability to achieve an entire representation of a proteome. To broaden the range of proteins covered and improve loading amounts, protocols can include sequential extraction of proteins by fractionation, however that leads to an extensive workflow which is a major bottleneck for 2DE methodology. Bottlenecks also occur at protein detection and quantitation levels, as intensive image analysis is needed for single or doubly stained gels. In view of current limitations of the 2DE platform, many gel free techniques were developed that can include insoluble proteins as well and are more global in set of proteins being analyzed.

Protein microarray is one such gel free technique that consists of a library of peptides, proteins or analyte of interest spotted on a solid support. Spotted protein samples are labeled with

a fluorescent tag binding to the individual targets for quantification and measurement purpose (Veenstra and Yates, 2006). In addition to fluorescence, chromogen, chemi-luminescent and radioisotopic labeling has also been utilized used for detection purposes. Even though protein array are flexible and have a great deal of potential to complement other prevalent proteomic technologies, its utilization and development has been severely limited due to several technical challenges (Hall et al., 2007). Some of these consist of: a) absence of a wider variety of affinity reagents (besides monoclonal antibodies, recombinant proteins) b) Improved surface chemistry to facilitate immobilization and capture of affinity reagents and c) need for self-assembling protein array platform(Veenstra and Yates, 2006). While protein microarray is a high throughput method to probe an entire collection of proteins, it is as good as the quality of proteins fixed on the chip.

Microfluidics is a miniaturized technique that has rapidly advanced with the aim of analyzing small volumes of proteins. Recent advances have been made towards implementing microfluidics for protein sample treatment, cell manipulation, sample cleanup, protein fractionation as well as on chip proteolytic digestion (Veenstra and Yates, 2006). With dramatic advances observed for microfabrication and microfluidic applications, several protein profiling strategies such as microfluidics based isoelectric focusing system, microdialysis of small volumes of proteins have emerged that have the potential for improved delivery to MS setup. Both elctrospary ionization (ESI-MS) and matrix assisted laser desorption ionization (MALDI) interfaces have been investigated for microfabricated microfluidic devices, to be successfully applied towards protein/peptide separation using chromatographic and/or electrokinetic-based principles. While ESI-MS emitters have been effective for an infusion analysis, the tips can contribute toward peak broadening not conducive for microfluidics based separation. Similarly for MALDI purposes, crystallized peptides have been presented along the edge of a disc for MS analysis. In such cases, the device has an increased surface-to-volume ratio which allows protein digestion by means of proteolytic enzymes immobilized on the chip surface (Veenstra and Yates, 2006). One such successful application is multi-dimensional separation of yeast cell lysate proteins, demonstrated with microfluidics interface (Veenstra and Yates, 2006).

Liquid chromatography coupled with mass spectrometry (LC-MS) is another attempt at achieving a fully integrated proteomic system for analyzing protein components and is our main topic of our discussion. LC is the most commonly used mechanism for separating peptides and proteins. Separated analytes are detected and identified by mass spectrometry. Inside a mass spectrometer, biomolecules are further separated before fragmentation, by their mass to charge ratio (*m/z*). Discussed in the following sections are the various LC-MS methods currently used for quantitative expression proteomics.

Sample Preparation

Given the large differences observed in concentration of thousands of proteins present in a sample, it is imperative to generate consistent and reliable data at all times. To achieve this, an optimal protocol is needed to minimize the impact of various factors that influence data quality. These include sample type (body fluid, cells, tissue, etc.), sample collection method, sample storage, physiochemical properties of analytes extracted, and/or solubilized and reagents used.

Human samples are the most studied species for bimolecular profiling, as they carry physiological information for onset of a disease. Constantly monitored for their diagnostic abilities, human specimens include serum, plasma, cerebrospinal fluid, bile, urine, milk, seminal fluid, hair, skin, saliva, etc. Careful selection of samples is necessary as certain groups of specimens are more conducive to proteomic investigations. For example, serum and plasma have high protein content, compared to saliva that is 99% water and 0.3% protein, making the former more suitable for proteomic studies. Conversely, urine samples mainly composed of metabolites or end products of blood are more appropriate for metabolomic studies. Similarly, hair is recognized as an attractive specimen for drug analysis. Hence, a compromise in terms of sample availability and study objective is needed, as all analytes may not show up in all body fluids (Rieux, 2006). Once samples are selected, sample integrity is maintained with optimal storage conditions, to minimize sample variability before analysis. Storage conditions vary according to its duration and kind of analytes needed to be preserved. Most body fluids collected can be stored at -80°C. Likewise, mammalian cell lines and tissue samples collected at the time of biopsy are usually stored for long durations in cryo-vial units. These materials are frozen in liquid nitrogen and maintained at low temperatures for years. Protein loss can occur due to interaction with surfaces through adsorption or aggregation. In such cases, samples sensitive to certain surfaces or light can be stored in special dark vials (polypropylene) under a regulated environment (Rieux, 2006).

Processing of fluid samples can get difficult as it may contain cells, proteins, peptides, nucleic acids, lipids, sugars, metabolites, and small molecules etc. An initial step is to separate cells and cellular debris from soluble components by means of low-speed centrifugation and sample clarification via filters. Following this would be steps that ensure proper processing, sampling, and storage of specimens. Since incorrect sampling can activate endogenous processes, leading to variation in analyte composition, it is essential to ensure optimal handling conditions after sample collection. One such example is serum preparation from plasma through coagulation. Since coagulation is a cascade of proteolytic events, it is a difficult to control process which ultimately affects composition of the resulting serum. Experimental variation due to erroneous sampling can be minimized by careful selection of sample population, reduced number of pretreatments (after collection), and by running appropriate number of samples.

In the absence of a universal sample preparation protocol, extraction of proteins from samples is accomplished by a combination of mechanisms such as cell lysis, density gradient centrifugation, fractionation, ultrafiltration, depletion/enrichment, and precipitation (Wells et al., 2003).

Cell lysis can be carried out in appropriate solubilization buffer by means of mechanical and chemical disruption methods. Gentle-lysis methods include osmotic, freeze thaw, and detergent based lysis; whereas vigorous methods include sonication, grinding, mechanical homogenization, and glass bead disruption (Wells et al., 2003).

Density gradient centrifugation can resolve complex cell lysates by separating intact organelles based on their molecular weight, size, and shape. This approach has been successful in isolating specific protein complexes and/or proteins from certain sub cellular compartments such as nuclei, mitochondria etc. Proteins from homogenate fractions are then further analysed for protein identification (Graves and Haystead, 2002). Due to their complexity, cell lysates need further fractionation to isolate low abundant proteins from the high abundant ones.

Fractionation reduces sample complexity by enriching for a specific subset of proteins, before separation and MS analysis (Wells et al., 2003). While fractionation can help detect more proteins via mass spectrometry, it is limited to soluble proteins only as they can be easily recovered (Graves and Haystead, 2002).

Ultrafiltration can also reduce sample complexity by removing high molecular weight (MW) proteins, based on a MW cut-off, thus increasing relative concentration of low MW proteins in a given sample. It cannot be used towards targeted protein profiling, however, when interfaced with other protein depletion/enrichment techniques it can enhance dynamic range of proteomic analysis (Graves and Haystead, 2002). Use of magnetic beads is one such method that is used in conjunction with ultrafiltration. This method not only enriches and desalts peptides from a mixture but subsequently leads to noise suppression with the ability to quantify low mass analytes. Magnetic beads are coated with functional groups, such as reverse phase C8, with recovery of peptides (from beads) being reproducible via an automated sample processing robot (Orvisky et al., 2006).

Besides ultrafiltration, high abundant protein depletion and/or low abundant protein enrichment are other methods that primarily aim towards low abundant proteins, which can be potential biomarkers. Removal of high abundant proteins is facilitated by using antibody- based or affinity dye-based resins. Different suppliers support different matrices that are targeted to a variety of proteins, however, one needs to evaluate various methodologies (MARS-Agilent, ProteoMiner-BioRad, Sigma-16 protein depletion) before deciding on a method that gives best results. For label free approach, it has been established that depletion improves linearity of protein intensity within the dynamic range of an MS instrument (Wang et al., 2006a).

Another technique commonly used for concentrating proteins in sample extracts is protein/peptide precipitation which removes any contaminating species present, such as nucleic acids, lipids, salts etc. (Graves and Haystead, 2002). Salting out, use of isoelectric point and organic solvents are some additional methods commonly used in protein sample preparation (Graves and Haystead, 2002). Use of organic solvent or denaturing conditions allows release of smaller proteins /peptides/ hormones bound to large carrier proteins which can otherwise be lost in sample preparation methods. Clean up kits can be used instead of precipitation to remove any insoluble components from sample before enzymatic digestion. Furthermore, for a sensitive MS analysis, salts, detergents and electrolytes need to be removed, as they lead to ion suppression (Graves and Haystead, 2002).

Once analytes of interest are extracted, peptides are generated by enzymatic or chemical cleavage of intact proteins and subjected to MS analysis for detection and identification purposes. Trypsin enzyme, commonly used for proteolytic digestion cleaves at the carboxyl side of lysine and arginine residues. Due to trypsin specificity it is possible to predict peptides which ensure reproducible and effective generation of peptides. These peptides are then cleaned/desalted using C18 cartridge and/or Ziptip prior to downstream MS analysis (Graves and Haystead, 2002).

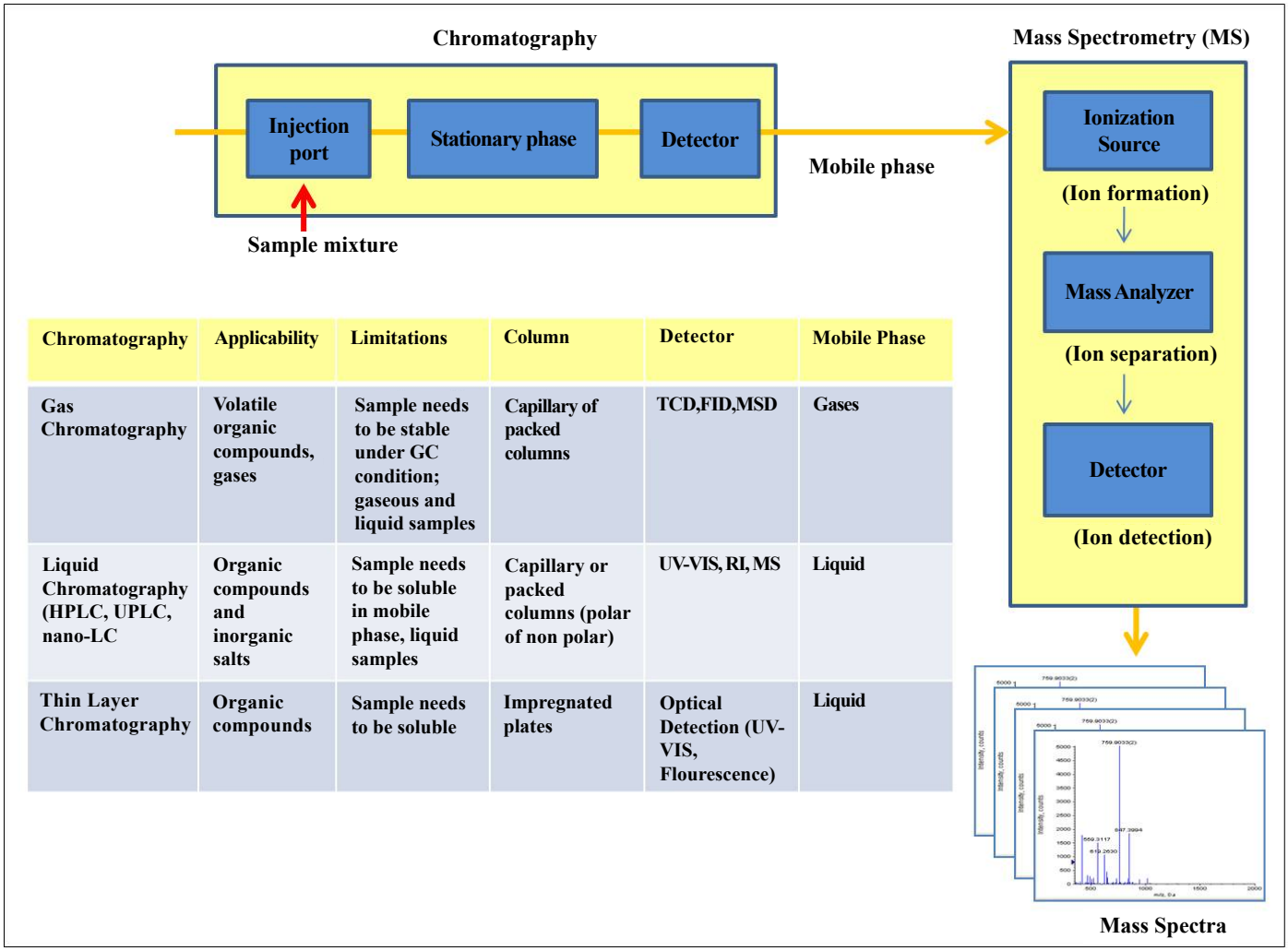
LC Separation Mechanisms

Since a proteome of an organism is too complex to be analyzed by a single separation step, multi-dimensional separation is needed to achieve greater selectivity and peak capacity. Chromatography is one such technique that can separate protein/peptide before downstream analysis. Based on its application, chromatography can be categorized as: gas chromatography (GC), LC, or thin layer chromatography (TLC) (Fig. 1).

Separation with liquid chromatography is achieved in two phases: mobile phase (liquid) and stationary phase. The mobile phase permeates through a stationary phase at high pressure while separating analytes, which are subsequently analyzed by mass spectrometry. Stationary phase is a column packed with irregular or spherically shaped particles or a porous monolithic layer. Separation is accomplished by standard liquid chromatography, high performance liquid chromatography (HPLC), ultra-performance liquid chromatography (UPLC). UPLC is a new technique similar to HPLC, except for the decreased run time and less use of solvent. Separation on UPLC is performed under high pressure using small particle packed columns (5 μm) that improves separation efficiency. Parameters affecting performance of LC include solvent strength, pH, organic modifier, ion pairing reagent, type of buffer and ionic strength. Based on the polarity of the mobile and stationary phases, LC is further divided into two sub-classes: reverse phase and normal phase. Reverse phase has water-methanol mixture as the mobile phase and C18 packing as the stationary phase; whereas normal phase has stationary phase as more polar and a non-polar mobile phase (toluene).

Discussed below are some of the most commonly employed LC-MS based separation methodologies (adsorption, partition, ion exchange, affinity, size exclusion) currently used in quantitative proteomics.

Reverse phase (RP) chromatography is a commonly used liquid-based separation that utilizes physiochemical properties of proteins/peptides for its chromatographic profile. Since column media is used to differentially retard migration of peptides, RP-LC is more suited for complex peptide mitures. Parameters such as stationary and mobile phase, retention mode, analyte charge, hydrophobicity, and analyte conformation are critical for RP-LC separation of analytes (Rieux, 2006). Typically, differences in hydrophobicity are the driving force for separation between analytes. Analytes elute in combination with mobile phase that differs in its aqueous composition with organic modifiers. Depending upon the kind of organic modifier used (mobile phase), different solute-solvent interactions occur that allow orthogonal selectivity in RP-RP (2D) LC setup (Rieux, 2006). For a multidimensional setup, RP is the last dimension as volatility of



Chromatography	Applicability	Limitations	Column	Detector	Mobile Phase
Gas Chromatography	Volatile organic compounds, gases	Sample needs to be stable under GC condition; gaseous and liquid samples	Capillary of packed columns	TCD,FID,MSD	Gases
Liquid Chromatography (HPLC, UPLC, nano-LC	Organic compounds and inorganic salts	Sample needs to be soluble in mobile phase, liquid samples	Capillary or packed columns (polar of non polar)	UV-VIS, RI, MS	Liquid
Thin Layer Chromatography	Organic compounds	Sample needs to be soluble	Impregnated plates	Optical Detection (UV-VIS, Flourescence)	Liquid

Figure 1: Basic LC-MS system and different kinds of chromatography (GC, LC, TLC) (TCD: thermal conductivity detector, FID: flame ionization detector, MSD: mass selective detector, UV-VIS: ultraviolet-visible, RI: refractive index).

organic mobile phase facilitates its coupling to MS interface. Advantages include fewer ion suppression effects during LC elution and efficient retention of polar compounds (Rieux, 2006).

Normal phase (NP) chromatography is another but rarely used method of separation that focuses upon compounds that are water/organic solvent insoluble and has been applied for isomer separation (McMaster, 2005). Normal phase chromatography uses stationary phase made of silica, amino, cyan, and diol packing, with mobile phase being mainly organic (Rieux, 2006). NP has good retention for polar compounds with fine separation capabilities, but suffers from solubility issues and poor reproducibility (Rieux, 2006).

Ion exchange (IEX) is another method commonly used for purifying proteins or pre-fractionated peptides. This method uses charge and surface distribution of analytes to determine their interaction with solid phase. Since pH affects charges on amino acids, pH control of mobile phase is critical, as it helps exploit different IEX that can be used. Ion exchange chromatography can be categorized as: weak anion exchange (WAX), weak cation exchange (WCX), strong anion exchange (SAX), and strong cation exchange (SCX). The weak ion exchanges occur over a relatively narrow pH whereas strong ion exchanges ionize over a wide range of pH. Elution is attained by varying pH or salt concentration of mobile phase which displaces peptides adhering to stationary phase of a column (Rieux, 2006). At low pH (< 3), N terminus amino acids as well as basic amino acids contribute towards net positive charge of an analyte, which can be separated by cation exchange chromatography. Anion exchange chromatography however, needs higher pH values. Besides using buffers with different pH, solvents with salt concentration increased in a stepwise manner can also be used to elute peptides. Use of salt displacer leads to adduct formation, which can be minimized by adding ammonium acetate or formate, as they are preferably chosen over Na+ and K+ salts. Several studies using ion exchange chromatography find SCX a preferred mode of fractionation as it allows buffer exchange before separating onto second dimension RP-LC. For all 2D LC separations evaluated, SCX-RP and RP-RP LC techniques were found to be comparable (Delmotte et al., 2007). Groups using SAX-RP reported similar results when switching over to RP-RP 2-D LC and SCX-RP 2-D LC model (Rieux, 2006). Advantages of IEX include high resolving power, fast separation, compatibility with downstream chromatographic separation or assay, high recovery and the ability to concentrate proteins from a dilute solution. However, certain constraints apply: application of

sample under low ionic strength and controlled pH conditions, protection of chromatographic instrumentation against salt-induced corrosion, and post-chromatographic concentration of recovered proteins resulting in high salt concentrations (>1 *M*) making it unsuitable for downstream applications (Stanton, 2003).

Affinity chromatography enriches for a set of proteins, with certain structural features such as phosphorylation, glycosylation, nitration or histidine, from a mixture of analytes. Selection is based on analyte interaction with immobilized molecules present on solid packing material. Depending upon the kind of immobilized molecules used, affinity chromatography is described as: immuno-affinity chromatography (antibodies), immobilized metal affinity chromatography (IMAC) or as ligands such as dye, lectin, hexapeptides, etc. Immuno-affinity chromatography uses antibody specificity for selecting analytes (Rieux, 2006). Amino acids on antibody's binding site engage in various non-covalent interactions with amino acids of peptides/proteins. The tridimensional structure of immobilized antibodies is preserved by using buffer with a certain composition and pH that resembles physiological condition. Acidic buffers are used to disrupt interactions that elute analytes of interest. For example protein A and G can form complexes with immunoglobins used to deplete serum/plasma of existing antibodies. (Rieux, 2006). Similarly, immobilized metal affinity chromatography relies upon complex formation between metal ions and specific amino acids and their functional groups, particularly histidine. For example, phospho-proteins and amino acids can bind to immobilized Fe3+ ions and metal oxides such as aluminium, zirconium, and titanium (Rieux, 2006) which is used towards retaining phosphoproteins/peptides from protein mixtures. Likewsie, glycosylated analytes can be selected using lectin packed columns that easily recognize glycosylation motifs present on peptides/proteins. Such columns use higher concentration NaCl with 0.2-0.8M sugar added to it (Rieux, 2006). Affinity chromatography is applicable to most organic and non volatile compounds and is flexible for a wide range of parameters (solid and liquid phase) that can be varied to accommodate better separation. However, affinity chromatography can be time consuming and it can be realized with certain analytical detectors only (Stanton, 2003). Also IMAC can be biased towards proteins and peptides with histidine residues (Rieux, 2006).

Size exclusion chromatography (SEC) uses matrices of different pore size to exclude analytes based on their size. Proteins/peptides that are small enough to penetrate through the pores elute toward the end of the run, thus separating them from high MW compounds and polymers. Restricted access media (RAM) is one kind of SEC extensively employed in pharmaceuticals to separate low-molecular weight analytes from their counterparts. Porous silica based packing is used to deplete samples of albumin, by means of size-exclusion and adsorption chromatography (McMaster, 2005). Advantages include less sample loss due to minimal interaction with stationary phase, low molecular weight cleanup of samples, ability to screen for small molecules in a sample mix. Since analytes are separated by size, a 10% difference between molecular masses of peptides/proteins is needed for better resolution. Also a limited number of bands can be accommodated in one run time (McMaster, 2005).

Hydrophilic interaction liquid chromatography (HILIC) and hydrophobic interaction chromatography (HIC) belong to a normal phase chromatography that uses miscible solvents on a polar stationary phase. Analytes absorbed on a stationary phase are partitioned into mobile phase. In HILIC, charged polar compounds undergo cation exchange with silanol groups of stationary phase (silica particles coated with hydrophilic moieties) (McMaster, 2005). HIC on the other hand is more closely related to RP except that it exploits hydrophobic properties of proteins in a more polar and less denaturing environment. The assumption being all proteins precipitate at high salt concentrations (neutral salts) and are released from adsorbing surfaces at lower salt concentrations Protein binding to HIC is accomplished by high concentrations of anti-chaotropic salts, whereas elution is carried out by decreasing salt concentration of adsorbent buffer (McMaster, 2005). Both HILIC and/or HIC have the following advantages: (i) they are complementary to RP-LC, (ii) they enhance ESI-MS as higher organic composition allows better sensitivity, and (iii) they allow sample preparation from liquid phase. Both techniques suffer from versatility issues as well as their inability to analyze non-polar compounds (McMaster, 2005).

Ion pair chromatography (IPC) differs from reverse phase chromatography as it focuses on polar or ionic biomolecular separation. Analyte selectivity is determined by mobile phase where organic eluent is supplemented with an ion pairing reagent that has a charge opposite to the analyte of interest. IPC (in elueant) forms an ion-pair with the counter ion retained on stationary phase through hydrophobic moiety (McMaster, 2005). Due to diverse interactions between analytes and IPC reagent, each ion pair is retained differently which facilitates sharper separation. One successful application of IPC is in separation of biogenic amines: adrenaline, tryptamine, dopamine that share similar retention times but are imparted different retention times by means of an ion-pair. Selection of IPC reagents is influenced by the presence of necessary counter ion and is further optimized by adjusting pH and IPC reagent concentration (McMaster, 2005). Disadvantages include short column life and poor reproducibility.

Protein fractionation two dimensional (PF2D) is an alternative technique to classical proteome approach where protein fractions are collected in the first dimension based on iso-electric point and second dimension by hydrophobicity (Ruelle et al., 2007). PF2D represents a two dimensional (similar to 2DE) liquid phase separation technique in which fractions collected can be analyzed by mass spectrometry. Samples separated on a first dimension column are put through second dimension reverse phase HPLC (Ruelle et al., 2007). Limitations of PF2D inlcude: the need for large volume of sample and repeated replacement of separation column. PF2D has been favorably applied towards characterizing immunogens from nonpathogenic bacteria *Bacillus subtilis*, in conjunction with polyclonal antibody and tandem mass spectrometry (Ruelle et al., 2007). This setup appropriately referred to as "i-F2D-MS/MS", successfully integrated analytical 2D LC (PF2D) with immuno-blotting and mass spectrometry (Ruelle et al., 2007).

Mass Spectrometry

While significant advances have been successfully accomplished at liquid chromatography level, mass spectrometry re-

mains an integral tool for protein identification and quantitation purposes. A mass spectrometer consists of (i) an ionization source on the front end that converts eluting peptides into gas phase ions; (ii) a mass analyzer that separates ions based on m/z ratios; and (iii) a detector that registers relative abundance of ions at discrete m/z. Two ionization methods that have revolutionized the use of mass spectrometers are: MALDI and ESI. Both methods are soft-ionization techniques that allow formation of intact gas-phase ions prior to molecular masses measurement by mass analyzer. There are various types of mass analyzer currently used in mass spectrometers inlcuding: ion trap (IT), time-of-flight (TOF), quadrupole (Q), iv) ion cyclotron resonance (ICR) and Orbitrap (Yates et al., 2009). Mass analyzer plays a critical role in mass spectrometry as it can store and separate ion based on mass to charge ratio. Uniqueness of a mass analyzer is assessed by its sensitivity, mass resolution, accuracy, analysis speed, ion transmission and dynamic range (Yates et al., 2009). A mass spectrometer can have various arrangements of ion source and analyzer, some including one mass analyzer or more than one analyzer, referred to as hybrid instruments. The most common hybrid instruments include ESI-Q-Q-Q, ESI-Q-TOF, Q-IT, IT-TOF, TOF-TOF, IT-FTICR, IT-Orbitrap, MALDI-TOF and MALDI-QIT-TOF (Veenstra and Yates, 2006; Panchaud et al., 2008).

Based on the kind of analysis pursued, mass analyzers are categorized as: scanning (TOF), ion beam (Q), or trapping (such as Orbitrap, IT, ICR) (Yates et al., 2009).The scanning analyzers are coupled to ionization techniques that proceed with pulsed analysis (MALDI) whereas ion beam and trapping analyzers are coupled to continuous ESI source (Yates et al., 2009). Instruments with ion trap analyzers feature high sensitivity, fast scan rates, high duty cycle, and multiple MS scans with high resolution and mass accuracy (100ppm). It is observed that mass spectrometers with one ion trap are more suited to bottom up approach mainly due to their high sensitivity and fast scan rates. Due to its ability to select, trap and manage ionic reactions, ion trap is mostly used at the front end of Orbitrap and Fourier transform-ion cyclotron resonance (FT-ICR) (Yates et al., 2009). Orbitrap and FT-ICR are recent additions to mass spectrometry instrumentation and discussed later in this section.

Once ions pass through mass analyzer they are detected and transformed into a signal by a detector. Three kinds of detector available are: i) electron multiplier, ii) photomultiplier conversion dynode, and iii) Faraday cup. Each of these amplify incident ion signal to output current which is then directly measured (Graves and Haystead, 2002).

In ionization, *MALDI* relates to laser desorption ionization (LDI) of analytes. While LDI encompasses airdrying of sample on a metal surface, MALDI uses a matrix compound that absorbs and transfers energy from the laser. A variety of matrices including aromatic acids can be used towards this objective. The aromatic group absorbs energy at the level of laser wavelength which results in proton transfer to the analyte (Veenstra and Yates, 2006). MALDI is fast and efficient in ionizing peptides and proteins, but the quality of its spectra greatly depends on matrix preparation. The sample co-crystallizes with an excess matrix solution which has led to experimentation with various methods of sample-matrix preparation. All methods aim towards achieving a homogenous layer of analyte crystals as non-uniform sample/matrix crystals give low resolution accompanied by low correlation between analyte concentration and its intensity. In general, MALDI ion source interfaced with TOF mass analyzer can scan tryptic digest of target proteins with an average of 30-40% protein sequence coverage. *ESI* on the other hand forms ions at atmospheric pressure followed by droplet evaporation. Peptide/ protein solution passed through a fine needle at high potential helps generate analyte ions. The electrical potential produces charged droplets which shrink by evaporation resulting in charge density. The ESI ion source has a tendency to produce multiply charged peptide ions depending upon the number of groups on a polypeptide chain that are available for ionization (Veenstra and Yates, 2006). Tandem MS of polypeptides is often done in positive ionization although negative ionization can also be applied towards identifying sulfated or phophorylated peptides. A common setup for ESI includes reverse phase-liquid chromatography (RP-LC) coupled to ESI-MS/MS. The flow rates of the solution can be adjusted depending upon the nano or micro bore RP columns. Typically the flow rates used for LC systems are 100-300nL/min and 1-100µL/min for the nano or micro LC, respectively (Veenstra and Yates, 2006).

Comparisons between MALDI and ESI reveal both strategies are complementary to each other, each having its own strength and weakness. MALDI mainly produces singly charged peptide ions which make mass spectra interpretation very straightforward. Being a sensitive technique, it is more tolerant of presence of buffers, salts or detergent than ESI. It works best with simpler protein mixtures as high ion yields of intact analyte can be achieved with high accuracy. However, factors that limit its application include: i) Inability of certain peptides to co-crystallize with matrix ii) Disparity in ionization affinity being observed, as all expected tryptic peptides do not show up iii) Need for homogenous sample-matrix crystals as good target (Veenstra and Yates, 2006). Additionally, LC-ESI can generate multiple charged ions, directly from sample solution. When coupled to LC, peptides separated continuously can be examined sequentially with high efficiency and increased throughput. However, ESI is less tolerant of interfering compounds in the sample matrix. With peptide elution (from chromatographic column) exceeding MS/MS scan, sometimes peptides present in samples can be severely under-sampled (Veenstra and Yates, 2006).

Fourier transform-ion cyclotron resonance (FT-ICR) is a novel mass analyzer with increased resolution, peak capacity and resolving power. In FT-ICR, ions in the cyclotron are irradiated with same frequency electromagnetic wave which leads to resonance absorption of the wave (Yates et al., 2009). Energy transferred to the ion increases its kinetic energy which further increases its trajectory radius. All ions in a cyclotron are then simultaneously excited by a rapid scan of large frequency range within 1 microsecond time span (Yates et al., 2009). Based on extensive calculations, Fourier transform mass spectrometer can achieve time spans of 1 sec per spectrum. Due to high sensitivity, the dynamic range of these instruments is limited to 10^6 with an ability to scan ions for a longer duration. Ability to select ions of a single mass based on resonance frequencies increases the resolution for these instruments from100,000 to 500,000 (Yates et al., 2009).

Orbitrap is the most recent addition to the pool of mass analyzers currently available, used in the form of LTQ-Orbitrap. This application traps moving ions in an electrostatic field that forces them to move in complex spiral patterns (Yates et al., 2009). Accurate reading on measuring m/z is achieved by means of oscillation frequencies of ions with different masses through use of a Fourier transform. The orbitrap mass analyzer presents a dynamic range greater than 10^3 with high resolution (150,000), high mass accuracy (2-5 ppm) and a m/z range of 6000 (Yates et al., 2009). Upon coupling to LTQ ion trap, the hybrid instrument can provide high resolution and mass accuracy along with faster scans and high sensitivity. Orbitrap has been successfully used for large-scale analysis of *Mycobacterium tuberculosis* proteome and applied to a virtual multiple reaction monitoring (MRM) approach (Yates et al., 2009). Due to its two complete mass analyzers capable of detecting and recording ions, orbitrap can operate for both top-down and bottom-up analyses (Scigelova and Makarov, 2006). Large scale bottom up proteomics is plagued by false identifications which can be minimized if all data acquired is with high mass accuracy. In such instances, linear ion trap of LIT-Orbitrap isolates and fragments ions selected for analysis, for full scan and subsequent MS fragmentation in orbitrap (Scigelova and Makarov, 2006) which considerably increase mass accuracy. While MS/MS in ion trap is similar to that of orbitrap, the most significant difference is in the resolution and mass accuracy observed for its peaks. Alternatively high resolving power of orbitrap can also facilitate analysis of intact proteins and help locate modifications on fragment sequences (Scigelova and Makarov, 2006). Moreover, orbitrap has been applied for extensive characterization of phosphopeptides by means of MS^3 and *de novo* sequencing using computer algorithms such as PEAKS (Scigelova and Makarov, 2006). Overall benefits of Orbitrap via high mass accuracy ability, includes quantification of low abundant peptides, profiling of complex samples, and identification of proteins from limited sequence proteomes (Yates et al., 2009).

Comparisons reveal Orbitrap offers comparable mass accuracy to FT-ICR instruments, but at a very low cost and less maintenance. While Orbitrap has been used in both bottom-up and top-down approaches, FT-ICR offers broader mass/charge range more suited to top-down protein analysis (Yates et al., 2009).

Protein/Peptide Identification

Protein identification by mass spectrometry is categorized as either top-down or bottom-up approach. In top-down proteomics, intact proteins or large protein fragments are introduced into the mass analyzer whereas in bottom-up peptides are introduced as ESI ions. Upon entry precursor ions receive multiple charge, which are then further fragmented to produce product ions, using electron capture dissociation (ECD) and/or electron transfer dissociation (ETD) mechanism. Interpreting top-down MS/MS spectra can be difficult as each multiply charged precursor can generate a set of multiply charged product ions. This limitation is circumvented by employing charge state manipulation or use of high mass accuracy instruments such as FT-ICR. While ion charge state manipulation is easy to accomplish by ion proton transfer, having access to a FT-ICR instrument is not easy. While top-down is still an upcoming field, major advantages include complete protein sequence along with characterization and location of post translational modifications (PTM), no time wasted on protein digestion due to use of intact proteins. However, it is mainly restricted to proteins smaller than 50kDa, requires use of high end mass spectrometers (FT-ICR, Orbitrap) and generates a composite spectrum that is more suited for simple protein mixtures (Mikesh et al., 2006).

Alternatively, bottom-up ionizes peptides separated over online-chromatography coupled to a mass spectrometer. Peptides from first scan form a peptide mass fingerprint (PMF) that is directly searched against a theoretical database for protein identification or subjected to tandem MS by collision induced fragmentation. Mass spectrometry instruments typically used for bottom-up approach include ions traps, hybrid Q-TOF and TOF-TOF mass spectrometers. Bottom-up is the most widely used approach successfully applied to identifying proteins from complex mixtures. With online reverse phase LC coupled to MS the whole setup can be automated minimizing experimental variation. However, certain limitations exist: partial sequence coverage as only a small fraction of peptides are identified, loss in PTM information, extended run times on multi-dimensional LC and loss of low abundant peptides masked by high abundant protein information.

Although both (top-down and bottom-up) approaches terminate with mass spectrometry identification, top down uses an offline separation, as coupling online chromatography to FT-ICR and other high mass accuracy spectrometers is difficult. Commonly used separation methods for top-down are pre-fractionation, protein fractionation, and purification. Also included are affinity capture and solution phase isoelectric focusing techniques that separate protein based on sequence and pI information. Ion exchange, size exclusion, and hydrophilic- or hydrophobic-interaction chromatography and capillary electrophoresis are other viable options for protein fractionation (Mikesh et al., 2006). Gel electrophoresis separation can also be used but with restraint as it is difficult to extract proteins and one encounters detergents that interfere with MS analysis. Likewise, bottom-up uses either gel electrophoresis (1D or 2-D GE) with peptides extracted from in gel digested proteins, separated over reverse phase LC or multidimensional LC of complex peptide mixture. Gel separation offers great advantages in terms of access to additional protein information such as mass, pI and PTM. However, drawbacks include labor intensive gel analysis, predominance of high abundant proteins, poor recovery of hydrophobic proteins, etc.

With common mass spectrometers, peptides introduced by ESI are sequenced for information by tandem mass spectrometry (MS/MS). By far, the most common method used is low energy collision activated dissociation (CAD) that cleaves amide bonds on peptide backbone to produce b and y ions. However, CAD is not conducive to detecting post-translational mondifications (PTM) due to presence of missed cleavages by trypsin, low charge peptides and size limitation imposed by CAD (Mikesh et al., 2006). While modifications on cellular proteins are widespread, any protein alteration is an indication towards its role in biological phenomenon. PTM's provide insights into the function of a protein; therefore analyzing them is critical for disease investigations. To facilitate better peptide sequence identification with labile PTM's being retained on different peptides; two alternative methods of dissociation have been developed (ETD and ECD). ETD utilizes ion/ion chemistry to fragment peptides,

by transferring an electron to multiply charged positive precursor ions. ECD however, relies on peptide cations in magnetic field of FT-ICR to capture floating low energy electrons, themselves. These reactions result in peptide cations containing an odd electron that undergoes subsequent dissociation. ECD is indifferent to peptide sequence and length, therefore results in random breakage of peptide backbone while retaining labile modifications. Although both techniques elicit c and z ions in the end by cleaving Cα-N bond, they differ in the kind of instruments and method of peptide dissociation (Mikesh et al., 2006). ETD uses a radio frequency quadropole ion trapping instrument that is not only inexpensive and low maintenance but easily accessible. However, ECD requires FT-ICR, a high mass accuracy instrument that is expensive and least accessible (Mikesh et al., 2006). While modifications on cellular proteins are widespread, any alterations on protein are an indication of its role in biological phenomenon. PTM's provide insights into the function and role of a protein; therefore analyzing them is critical for disease investigations (Mikesh et al., 2006).

Protein identification depends upon the PMF characteristic of a protein and the pattern of masses generated by a MS. The fingerprint is then searched against a proteome database; best matches of experimentally obtained peptide map to theoretical PMF of individual proteins within a database, leads to the identity of unknown protein. Several factors that can affect peptide mapping results can be grouped together as either fingerprint–constructing or fingerprint-searching factors (Veenstra and Yates, 2006). Factors that influence fingerprint construction include i) Noise level of a peptide set ii) The number of peptides in a given fingerprint and iii) Mass accuracy based on instrument calibration. When searching the database, characteristics of the organism being studied and specific PTM's are two factors that need attention. The query PMF is compared to every sequence that exists in the specified database. A match is evaluated based on various algorithms which returns a probability based score. If the fragment is a result of tryptic digest, every fragment between K and R in a protein's theoretical sequence is quantified by the weight of amino acids in that fragment (in Daltons) and if the mass of peak submitted as a query matches to this calculated mass, a random chance or true protein identification (Veenstra and Yates, 2006). The accuracy for the mass of an unknown sample can vary anywhere from between 1 to 1000 ppm, with 100 to 400 ppm being the most typical for many laboratories. Depending upon how well is the instrument calibrated; a more stringent search will lead a searching algorithm to fail to match some observed peaks to database fragment masses resulting in no identification. The percentage length coverage of a peptide is an index of how well is a protein represented in the query PMF. Fragments outside the allowable range of 600Da and above 3000Da and below an intensity threshold are not included in the query PMF in database search. PMF can sometimes have limitations such as some peptides tend to ionize over the expense of others or signal for modified peptides can be observed that are not predicted by *in silico* digestion which cannot be matched unless accounted for the modification (Palcy and Chevet, 2006) Features such as database size, distribution frequency of a peptide mass for a given protein, and distribution of mass accuracy are some parameters that can influence the specificity of a search Therefore the choice of a parameter helps user to get high specificity without missing a true protein positive (Veenstra and Yates, 2006).The intensities of spectrum do not correlate with the amount of peptide in the samples due to suppression effect and ionization bias, therefore it is relevant to evaluate if the intense peaks have been used toward protein identification or not. For large proteins, large amount of peptides should be matched whereas for a small protein low number of peptide matches can result in reasonable coverage (Veenstra and Yates, 2006).

Tandem MS (MS/MS), on the other hand reveals additional information on peptide sequences. Peptide samples can be separated by one or multi-dimensional LC and subjected to tandem mass spectrometry for peptide sequencing. Database search parameters include types of ion selected, method of mass calculation, peptide charge state, and parent ion tolerance. Types of ions selected for generating theoretical data can depend upon the kind of instrument used for fragmentation. Mass spectrometers such as ion trap, quadrupole and Q-TOF result in *b* and *y* ions whereas instrument with high energy collision induce dissociation (CID) can generate a, c, x and z ions as well. For the calculation of peptide mass, monoisotopic or average method can be used as mass spectrometers do no measure mass of peptides but instead mass to charge values (Veenstra and Yates, 2006). For a given protein, the monoisotopic mass is the mass of the isotopic peak whose elemental composition is composed of the most abundant isotopes of those elements. Average mass is the weighted average of all the isotopic masses abundant of that element. High resolution mass spectrometers can use monoisotopic determination for mass whereas with ion traps it is better to use average mass (low resolution). Peptide ion charge state can be determined in high resolution instruments by the isotopic distribution patterns observed in MS spectrum. With low resolution instruments it is not possible to tell the exact charge, though single and multiple charged ions can be easily distinguished. Parent ion tolerance allows certain measured peptides selected from sequence database to be scored against the experimental spectra along with the choice of enzyme used for peptide digestion. The number of candidate peptides needed for analysis is reduced with the specification of an enzyme which reduces search time significantly. Modifications such as reduction and alkylation for gel based proteins are incorporated prior to analysis; Other modifications categorized as static or variable are incorporated as search parameters in database search.; static modifications are where all occurrences of a residue are modified whereas variable modifications are when some residues may or may not be modified (Veenstra and Yates, 2006).

For protein mixtures as complex as > 10,000 proteins, fragment-ion matching technology is used instead of PMF. Peptides from protein digest are dissociated into fragments using mass spectrometers, the mass spectral of which is measured and searched against a database to determine the resulting precursor peptide mass. This approach is particularly useful when a peptide sequence is unique to possibly identify the protein origin based only on MS/ MS fragmentation. Researchers have found that tandem mass spectrometry has a higher success rate in protein identification than MS-based identification (Gulcicek et al., 2005). Another method for database searching involves the "sequence tag" approach which uses short amino acid tags generated after tandem MS interpretation against peptides in protein

databases for the same enzymatic cleavage. For a protein with no previous sequence information, *de novo* interpretation is considered useful at the tandem MS level (Veenstra and Yates, 2006). For a given peptide sequence all fragment ions and masses can be specified, which is exactly how de novo sequencing tries to assemble amino acids sequences for a peptide based on spectral pattern.

In order to search databases, several MS search engines have been developed for peptide identification by searching experimental mass spectra against MS data of *in silico* digested protein databases. SEQUEST (BioWorks), MASCOT and ProteinProspector are some of the algorithms used for peptide identification. Peptide, matching a protein entry are clustered together and reported as a protein hit (Palcy and Chevet, 2006). The database score is computed according to some scoring function that measures the degree of similarity between experimental spectra and the peptide pattern observed for theoretical fragmentation. SEQUEST one of the most commonly used programs calculates cross correlation score for all peptides queried. In addition to X-corr a derivative score which computes the relative difference between the best and second best X-corr is computed which is useful for discriminating between correct and incorrect identifications. MASCOT, a probability based score estimates the probability of matches occurring by chance for the number of peaks in an experimental spectra and the distribution of a predicted ions. With the SEQUSET algorithm, manual review of data is needed to avoid choosing false positives. Since MASCOT uses probability based scoring that assigns score to all identifications, it depends entirely upon the researcher to consider which protein identification as significant (Gulcicek et al., 2005).

Alternatively, spectral library (instead of theoretical spectra) can be used for database searches. Here, peptide mass spectral libraries (MS/MS spectra) become standardized resource for robust peptide identification as they are based on actual physical measurements of peptides already identified in previous experiments (Kienhuis and Geerdink, 2002). Many advantages exist in using this approach in comparison to traditional approach: firstly it is fast as it uses experimental spectra only rather than searching against all possible peptide sequences generated from genomic sequence. Secondly, peptides and proteins are identified with higher sensitivity because an experimental spectrum is more likely to match to a library spectrum better (i.e. with higher confidence) than to a theoretically predicted spectrum. Ultimately, spectrum libraries can provide a common reference point allowing researchers to objectively analyze and compare datasets generated in different experiments (Kienhuis and Geerdink, 2002).

LC–MS Based Quantitative Assessment

Mass spectrometry identifies unknown biomolecules based on their accurate mass and fragmentation pattern. However, for proteomic studies this is possible only if sample is a simple mixture or has been previously divided into simpler parts by high resolution separation methods. Liquid chromatography interfaced with tandem mass spectrometry is shown to be well suited for such quantitation purposes without the use of gels. However, LC-MS generated data is contingent upon factors such as instrument sensitivity, detection coverage, dynamic range, mass accuracy and resolution (Listgarten and Emili, 2005). One such high resolution method that accommodates most of the above mentioned separation methodologies, is nano-liquid chromatography (nano-LC). While detecting proteins present at reasonable levels was easy to achieve, measuring small quantity proteins in complex mixtures has always been very challenging. In order to achieve maximum sensitivity, concentration of low quantity proteins needs to be within detection limits of the instrument pushing the need for small volumes. This approach has led to the development of nano-LC where LC pumping devices capable of delivering samples at nl/ min flowrate separate components on columns of diameter size < 100µm (Qian et al., 2006). Additionally samples can be enriched for low abundant protein by means of depletion, pre-fractionation and concentration techniques prior to nano-LC-MS detection. This is greatly beneficial for samples that have proteins in concentration range of $10^5 – 10^{10}$ or even more (tissue, plasma, serum etc). Nano-LC offers several advantages such as: low sample volumes, improved electrospray efficiency due to small droplet formation, high efficiency packed columns that can be operated with MS friendly solvent systems, reproducible delivery of solvents to nano columns etc.(Qian et al., 2006) However certain disadvantages limit its potential as a high throughput technique. These include the not so robust LC-MS interface as nano-flow components and analytical column interfacing to nano-spray emitter (5µm) are frequently prone to failure, small volume leaks that go undetected, dispersion due to dead volume between components and of course longer runs needed for better separation (Qian et al., 2006). Overall, in absence of a single platform for global profiling of a proteome, nano-LC definitely has the upper hand with maximizing number of identifications being reported; provided samples are pre-fractionated prior to analysis. Importantly other new techniques like fast LC, gas phase separations and better nano-ESI interfaces also present a promising future for discovery applications (Qian et al., 2006).

Also, since mass spectrometry is not intrinsically quantitative, strategies have been developed that allow differential mass-labeling of analytes prior to mass spectrometry. Although these methods are “gold standards” for protein quantitation, they have not been widely used for large-scale multiplexed analyses. This is mainly due to their relatively high cost, limited availability of different mass-coded labels and frequent under-sampling associated with MS/MS. Besides, since peptide identification seems to precede quantification, numerous peptides are identified that are unchanged in abundance between samples. As a result, a lot of instrument time and data analysis is consumed by proteins that may have less biological significance. Discussed below are commonly used methods absolute quantification and relative quantification (using stable isotope labeling and label-free methods) for quantitative assessment of protein expression (Fig. 2.).

Absolute quantitation

Absolute quantitation of proteins, commonly known as AQUA, is achieved by adding internal standards of known quantity to a protein digest that is subsequently compared to the mass spectrometric signal of peptide present in the sample. It uses synthetic peptides that have some kind of differential isotopic label used for spiking purposes prior to LC-MS. These synthetic peptides can match to the experimentally observed sequence but are synthesized with heavy analogues of amino acids. Quan-

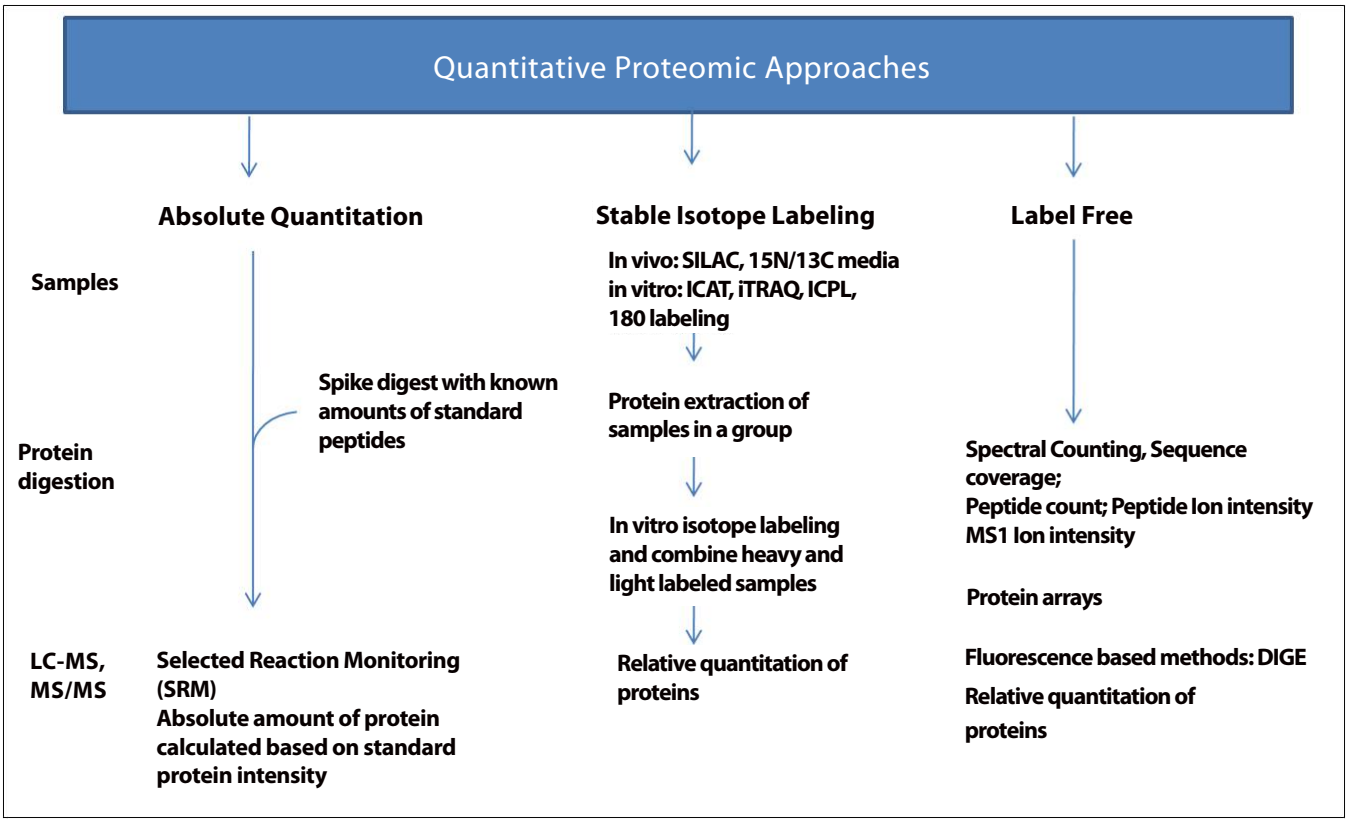


Figure 2: Different approaches for quantitative assessment of proteins in biological samples.

tification is achieved by calculating intensity ratio of the endogenous peptide (light) to the intensity of the reference peptide (heavy), that share same physicochemical properties including chromatographic elution, ionization efficiency, fragmentation pattern but are distinguished by mass difference (Pan et al., 2009). While this approach is attractive for validation purposes, a number of limitations exist such as: very few proteins can be quantified, amount of labeled standard needs to be determined before spiking, ambiguity due to presence of multiple isobaric peptides in mixture etc. Some of these issues can be resolved by selected reaction monitoring (SRM) method that compares the intensity of precursor and fragment ions of "heavy" standard to "light" peptides of a test protein (Raghothama, 2007). The combination of peptide retention time, mass and fragment mass removes any ambiguities in peptide assignment and generally broadens the quantification range (Bantscheff et al., 2007). SRM has been applied to studying low abundant yeast proteins involved in gene silencing (Bantscheff et al., 2007). Variation to SRM is multiple reaction monitoring (MRM) which uses multiple peptides for multiplexed quantitation. Samples include internal standard peptides with light and heavy label such that it is applicable to pair peptides in an identical fashion (Okulate et al., 2007). In addition to using individual tryptic peptides, one can use a recombinant protein made up of many tryptic peptides, as a standard protein. While having prior information on protein helps decide what masses to look for, it is important to realize that amount of protein determined in an experiment may not reflect its true expression levels in a cell (Bantscheff et al., 2007). Protein standard absolute quantification (PSAQ) is another method that uses an intact protein with stable isotope label as an internal standard for quantification. PSAQ has been successfully applied to quantify staphylococcal super antigen toxins in urine and drinking water and determine absolute levels of alcohol dehydrogenase in human liver samples (Bantscheff et al., 2007). Metal coded tags (MeCAT) is another method that utilizes metal bound by MeCAT reagent to a protein or biomolecule, in combination with element mass spectrometry inductively coupled plasma mass spectrometry (ICP-MS), for first time absolute quantification (Pan et al., 2009).

Stable isotope labeling

For past several decades, stable isotope labeling has been recognized as an accurate method for MS based protein quantification. While several methods exist, each method is distinguishable by the way heavy labels are introduced into a peptide/protein. Based on the introduction of stable isotope labels into analytes, isotope labeling is categorized as metabolic (*in vivo* or culture) and chemical (*in vitro*).

Metabolic labeling relies on growing cells in culture media with isotopically labeled amino acids and nutrients that allow *in vivo* protein labeling during cell growth process. Relative quantitation is determined by comparison of "heavy" with "light" labeled cells. Peptides identical in sequence (between samples) but labeled with different mass show up as a distinct mass shift on MS. There are different kinds of metabolic labeling.

Stable isotope labeling by means of metabolites (^{15}N or ^{13}C) is achieved by enriching media with ^{15}N ammonium salts to replace all ^{14}N nitrogen atoms or ^{13}C glucose to replace ^{12}C atoms respectively. Both "light" and "heavy" samples are combined in 1:1 ratio to exclude for any experimental variations during cell growth. The $^{14}N/^{15}N$ or $^{12}C/^{13}C$ labeled peptides are identified in mass spectra as doublet ion cluster, separated by the mass shift introduced by heavy nitrogen isotope. Comparison based

on peak intensities or peak areas is used to relatively quantify protein samples. Depending upon the length of a peptide and varying number of N or C atoms, the heavy isotope leads to varying mass shift which results in highly complex mass spectral data. ^{15}N labeling has been used for simpler organisms such as bacteria, yeast and has been applied to *A. thaliana* and mammalian cell culture as well (Gulcicek et al., 2005).

Stable isotope labeling by amino acids in cell culture (SILAC) is another labeling approach that uses amino acid as a labeling precursor, added to the culture media during cell growth (Monteoliva and Albar, 2004). Originally developed to generate mass-tagged peptides for accurate and specific protein identification via peptide fingerprinting, this method is now established for quantitative proteomics. This approach makes tandem MS interpretation much easier as labeled and unlabelled peptides mass differences are easily predictable (Monteoliva and Albar, 2004). Unlike $^{14}N/^{15}N$ peptide pair comparison, peptides in SILAC exhibit mass differences defined by the combination of isotopically labeled amino acids used for labeling purpose. First described for yeast model, SILAC has been applied towards studying protein-protein interactions, identifying post translational modifications and assessing protein expression levels (Monteoliva and Albar, 2004). To achieve complete labeling of proteome, only amino acids that are essential to the organism or contribute to genetically auxotrophic state are generally used. Amino acids with relatively high abundance such as arginine, leucine and lysine are employed to result in high number of labeled peptides, thus providing information on multiple peptide pairs. Isotopes generally used are: ^{13}C and ^{2}H. SILAC being a simple process, has been used for identifying and analyzing post translational modifications and signal transduction events in yeast pheromone pathway as well (Monteoliva and Albar, 2004).

Several studies have effectively applied metabolic labeling with stable isotopes towards comparative proteomic investigations. Some of these studies include investigation of human Hela cells labeled with $^{13}C_6$Arg and $^{2}H_3$Leu labels followed by LC-MS/MS analysis, *S. cerevisae* cells labeled with $^{2}H_{10}$Leu for protein identification with MALDI, mouse fed with $^{3}C_2$Gly labeled diet to map peptides to a protein based on mass and leucine content information, etc. (Beynon and Pratt, 2005). SILAC has been successful with mammalian cell culture labeling where isotope labeled essential amino acids have been fully incorporated into the proteome, plant cell culturing with 70%-80% incorporation and with auxotrophic yeast mutants (Engelsberger et al., 2006). Elements of quantitative proteomics that have greatly benefited from SILAC include the following: formation of signal-dependent protein complexes, modification-dependent protein-protein interaction screens, analyses of the dynamics of signal-dependent phosphorylation events etc (Engelsberger et al., 2006). Major advantages over other stable isotope labeling strategies include: labels that are biosynthesized and are present in live cells, compatibility with cell culture conditions, no affinity purification step, samples or cells from two states can be mixed into one providing an internal control for finding real proteomic differences independent of variability from processing steps, absence of side reactions due to highly specific enzymes, etc. Limitations include: a small subset of tryptic peptides being tagged, substantial incorporation of isotope needed for effective labeling, experimental variability introduced during labeling processes, etc. (Beynon and Pratt, 2005).

For samples that are less amenable to metabolic labeling, chemical reactions have been exploited to introduce isotope encoded tags into proteins/peptides. The chemical tag chosen is targeted to a specific functional group of an amino acid residue to which it is covalently bonded. Choice of a labeling method depends upon factors such as sample complexity, protein quantity and downstream instrumentation employed. Chemical tagging can be categorized as:

Isotope coded affinity tags (ICAT), a widely accepted quantitative technique allows protein quantification by using light or heavy isotopes that bind to sulphydryl groups of amino acid residues that can be identifiable by micro-capillary LC/ESI/MS/MS (Monteoliva and Albar, 2004). Chemical incorporation of isotope tags is typically pursued after protein extraction, with control and experimental samples being derivatized with light and heavy ICAT reagent followed by trypsin digestion (Monteoliva and Albar, 2004). The ICAT reagent has three components: a biotin tag, an oxyethylene linker region and a thiol specific iodoacetyl group that derivatizes Cys residues in proteins. Labeled peptides are fractionated using strong-cation exchange liquid chromatography followed by RP-HPLC and tandem mass spectrometry analysis to identify and quantify ICAT peptide pairs (Gulcicek et al., 2005). ICAT is analogous to microarray use of two different dyes or DIGE protein expression analysis, as changes in expression are determined by differences in intensity observed. ICAT technology has been reportedly applied to several proteomic studies including total proteome characterization of yeast, P*seudomonas aeruginosa*, etc. (Monteoliva and Albar, 2004). Labeling is dependent upon the presence of cysteine residues as its sulphydryl groups are chemically labeled in proteins; which makes this technique less suitable to proteins that lack Cys residues. The limitation arises at the mass spectra and database search levels, as proteins that lack cysteine cannot be included in the analysis (Monteoliva and Albar, 2004). The advantage is however, enrichment for Cys containing peptides which reduce sample complexity before MS analysis (Monteoliva and Albar, 2004).

Another strategy in quantitative MS-based proteomics is the derivatization of primary amines including amino termini of proteins/peptides. Two techniques currently established under this category are iTRAQ and ICPL:

Isobaric tag for relative and absolute quantification (iTRAQ) specifically aims at multiplexing sample without generating clusters of peptide pairs. This technique utilizes labeled amine modifying chemistry with MS/MS based quantification mode. iTRAQ reagents consist of three principal components: a reporter group based on N methyl piperazine, carbonyl balance group and a peptide reactive group (McMaster, 2005). Owing to selective inclusion of ^{13}C, ^{15}N and ^{18}O atoms, differentially labeled peptides appear as single peaks in MS spectra which can be quantified based on their iTRAQ reporter ion information received in the second step of mass analysis. Since iTRAQ labeled peptides need to be analyzed with tandem MS, the strategy relies heavily on only these reporter groups that can be observed in MS/MS scans. Due to appearance of reporter ions in low m/z range, iTRAQ cannot be used with conventional ion trap instruments.

Isotope coded protein labeling (ICPL) uses an isotope coded N- nicotinoyl oxysuccinamide tag to allow incorporation of amine reactive tags inside intact proteins. The samples are reduced and alkylated before derivatizing with ICPL however; one need to consider that tryptic cleavage of ICPL labeled proteins would occur only to C terminal Arg residues and not at the modified Lys residues. Use of ICPL for MS-based quantitative proteomics has been demonstrated in various differential analysis studies including rat hepatoma cell exposure to carcinogenic toxin, halobacterium membrane proteome etc. (McMaster, 2005).

^{18}O labeling: Peptide C termini can be selectively derivatized by incorporating heavy oxygen atoms (^{18}O) using serine proteases in combination with protein digestion or after completion of amide bond hydrolysis with heavy labeled water. For relative quantification, two samples digested parallel in H_2O^{18} and H_2O^{16} are mixed in 1: 1 ratio prior to chromatographic separation and MS analysis. Relative abundance is determined by comparing signal intensities or peak areas of ^{16}O/ ^{18}O encoded peptide pairs (McMaster, 2005).

Chemical labeling of protein samples is mainly achieved by post-synthetic modification of proteins and tryptic peptides, by chemical and enzymatic derivatization. While chemical labeling has been advantageous for highly complex samples, it is prone to certain limitations. Certain concerns that limit its application are: incomplete labeling of peptides that incorporate label at different rates making data analysis a formidable task, use of cysteine and lysine residues in ICAT making this technique less suitable for proteins that have no or few lysine or cysteine residues or for identifying post translational modifications and splice isoforms, labeling kinetics dependent upon protein turnover, modified lysine is not digested by trypsin resulting in longer peptides that obscure MS analysis, high labeling efficiency needed prior to separation as incomplete labeling impairs resolving power and of course side reaction during labeling that can lead to unforeseen products that confound data interpretation (Bantscheff et al., 2007).

Label-free

While labeling protocols (e.g., ICAT, iTRAQ, ^{18}O- or ^{15}N-labeling, etc.) remain the core technologies used in MS-based proteomic quantification, increasing efforts have been directed to the label-free approaches. Label-free method is attractive to investigators due to cost effectiveness, simpler experimental protocols, fewer measurement artifacts, and limited availability of isotope labeled references (Goodlett and Yi, 2003; Lill, 2003). The most common label-free methods include the following:

Spectral count method, is where the total number of MS/MS spectra taken on peptides from a given protein in a given LC-MS/MS analysis is used to compare differential abundance between cases and controls (Old et al., 2005). This approach is based on the fact that more a protein present in a sample; more of MS/MS spectra is collected for its peptides. This method simply counts the number of spectra identified for a given peptide in different samples and integrates results of all measured peptides for the protein quantified. It can be used for quantitative protein profiling as extensive MS/MS data is collected across chromatographic time scale. *Sequence coverage method* uses information on total coverage of a protein sequence by its identified peptides (Florens et al., 2002). The *peptide count method* uses the total number of peptides identified from a protein (Gao et al., 2003). *Peptide ion intensity method* measures peptide ion intensity by integrating area under the curve and comparing them for their relative abundance. It requires MS data to collect under "data dependent" mode (MS scan, Zoom Scan and MS/MS scan). *Comparison of ion intensities*, is a method where LC-MS runs are compared to identify differentially abundant ions at specific *m/z* and retention time (RT) points. This approach is based on precursor signal intensity (MS), applicable to data derived from high mass precision spectrometers. The high resolution facilitates extraction of peptide signal at the MS1 level and thus uncouples quantification from the identification process. It is based on the observation that intensity in ESI-MS is linearly proportional to the concentration of the ions being detected. The key factor to label free method is in the reproducibility of its LC-MS runs and proper alignment of LC runs, for reliable detection of differences. Since label free methods can go beyond pair-wise comparison they rely heavily on computational analysis.

The first three methods relate the relative protein abundance to the observed sampling statistics from tandem MS/MS. However, these methods are not fast enough to probe every ion detected in the first stage of mass spectrometry and much of the information available in that stage is discarded, especially for low-abundance ions. Direct comparison of LC-MS peaks without using the corresponding MS/MS data provides the opportunity to examine all biomolecules present in the entire LC-MS profiles. To estimate relative abundance of biomolecules from multiple LC-MS runs, some investigators apply direct comparison methods using MS1 ions and the entire retention profiles, (Prakash et al., 2006; Radulovic et al., 2004) while others use monoisotopic masses and the peak apex of elution profiles.(Kearney and Thibault, 2003; Pierce et al., 2005; Wang et al., 2003) It is based on the principle that the relative abundance of the same biomolecule in different samples can be estimated by the precursor ion signal intensity across consecutive LC-MS runs, given that the measurements are performed under identical conditions (Kuhner and Gavin, 2007).

It also appears that labeling efficiency is not consistent as it varies between samples. Alternatively, label free methods are used to calculate relative abundance of a biomolecule by estimating precursor ion signal intensity across consecutive LC-MS runs. The assumption being all measurements are performed under identical conditions (Kuhner and Gavin, 2007). A critical challenge in using this method for biomarker discovery lies in normalizing and aligning the LC-MS data from various runs to ensure bias-free comparison of the same biological entities across multiple spectra.

With respect to advances made at instrumentation level, sample preparation, analytical platforms, computing power and bioinformatics algorithms, label free quantitation has been a successful technique for comparing ion intensities. However, certain constraints such as stability and reproducibility of analytical platform limit its full potential. As label-free LC-MS method relies heavily on pattern matching of LC-MS runs, obtaining a high degree of experimental reproducibility is essential. This is challenging, especially when large number of samples are involved. While true reasons for irreproducible LC-MS runs are incompletely understood, factors contributing towards this

are: i) Low data quality ii) Variability due to sample preparation process (depletion, trypsin digestion, etc.) iii) Variability due to analytical equipment performance (separation, ionization, ion suppression-detection) iv) Sytematic bias of different spectrometry methods or stochastic sampling etc. (Baggerly et al., 2004). Data reproducibility can be evaluated by means of quality control process where intensity and retention time reproducibility or pattern similarity of overlapping MS features is compared. Differences across runs can be minimized by measuring and controlling all sources of variation. Different ways to accomplish this are: careful experimental design, use of standard protocols, controlling experimental conditions when conducting studies, repeated measurements, validation after every shift in protocol and better methods of external calibration (Baggerly et al., 2004).

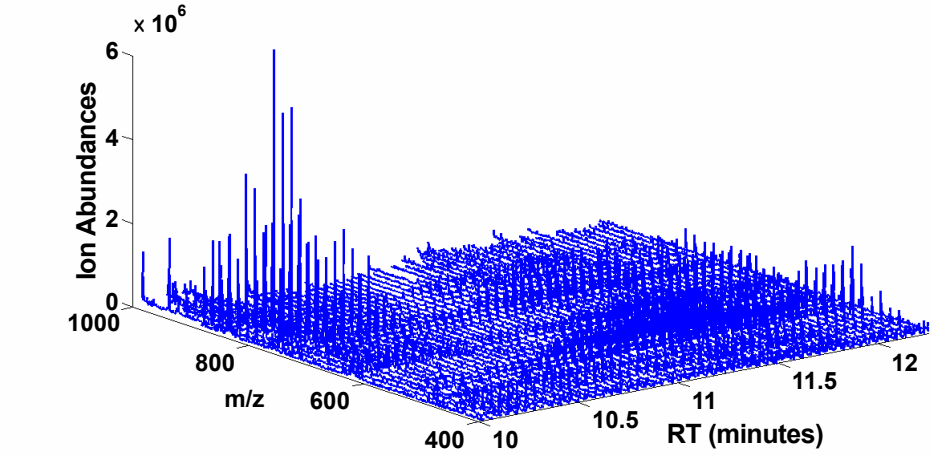
Computational Methods for Mass Spectrometry Data Analysis

Mass spectra contain true signal and electronic/chemical noise due to contaminants and matrix; which causes varying baseline (Malyarenko et al., 2005). In addition, mass spectra reflect variability in sample preparation and sample degradation. Previous quality-control experiments identified properties of mass spectrometric measurements that must be accounted for at analysis (Fung and Enderwick, 2002; Yasui et al., 2003). Thus, detection of differential protein expression through analysis of mass spectral data requires careful experimental design. It is important to take into account population sampling, matching of controls, protocols for unbiased sample collection, uniform sample preparation methods, and appropriate mass spectrometric analysis. Sorace and Zhan reported the possibility of experimental bias in their assessment of surface enhanced laser desorption/ionization time-of-flight (SELDI-TOF) analysis of ovarian cancer (Sorace and Zhan, 2003). Ransohoff indicated that bias will increasingly be recognized as the most important 'threat to validity' that must be addressed in the design, conduct and interpretation of such research (Ransohoff, 2005). Bias can occur if the case and control groups are handled in systematically different ways, introducing an apparent 'signal' into one group but not the other. Such differences might be introduced at several stages, including specimen collection, handling and storage, or during data generation. Diamandis questioned why the features and classification performance vary so drastically across studies (Diamandis, 2004). This concern is based on the observations that different SELDI-TOF approaches combined with different machine learning techniques for pattern recognition produce highly variable results in terms of relevant features and classification accuracy. Such variation may be attributed to a large number of features relevant to the task of discriminating healthy individuals from those afflicted with cancer. Baggerly et al., (2004) indicated the cause for inconsistent result could be the chemical/electronic noise and/or bias introduced during the acquisition of the MS spectra.

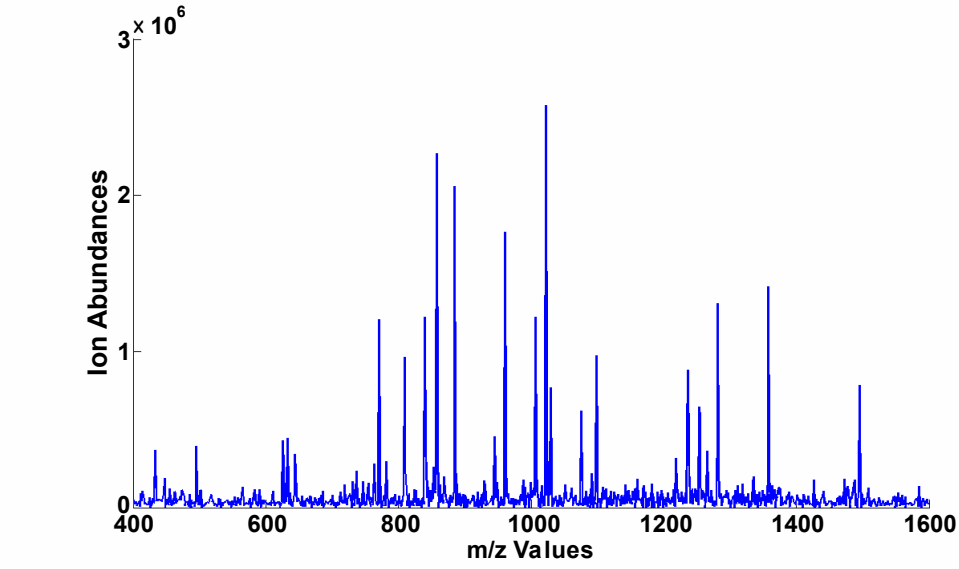
A mass spectrum is represented by a large sequence of paired values, where each pair contains the following: (1) a measured intensity, which depends on the abundance of the detected biomolecules and (2) a mass-to-charge ratio (m/z), which depends on the molecular mass of detected biomolecules. When obtaining a spectrum, we expect imperfect measurements caused by noise, peak broadening, instrument distortion and saturation, isotopes, miscalibration, and contaminants of various kinds. The impact of these artifacts can be minimized by preprocessing the raw spectra prior to selecting differentially abundant peaks. The purpose of preprocessing is to correct intensity and m/z values in order to: (1) reduce noise, (2) reduce amount of data, and (3) make the spectra comparable to each other. For example, outlier screening removes spectra whose data distributions substantially differ from others. Binning reduces the dimension of a spectrum by grouping intensity measurements at adjacent m/z values into bins. Smoothing is a process by which data points are averaged with their neighbors as in a time-series data to increase signal-to-noise ratio. Baseline correction flattens the base profile of a spectrum to minimize the impact of varying baseline caused by the chemical noise in the matrix or by ion overloading; drifting baseline introduces serious distortion of ion intensities without adequate correction. Normalization reduces systematic variation that may be caused by varying amounts of protein, sample degradation over time, or variation in the sensitivity of the MS ion detector. Peak detection deals with the identification peaks that display a reasonable intensity compared to those that may be just noise. The simple peak finding algorithm provides the locations of potential peaks and their associated left-hand and right-hand bases. Peak calibration allows correction of drifts that do not reflect any real sample variation. Without peak calibration, the same peak (e.g. the same protein) can have different *m/z* values across samples. To allow an easy and effective comparison of different spectra, peak alignment methods find a common set of peak locations (i.e. *m/z* values) in a set of spectra, in such a way that all spectra have common m/z values for the same biological entities.

A critical challenge in using LC-MS for detecting differential protein expression lies in normalizing and aligning the LC-MS data from various runs to ensure bias-free comparison of the same biological entities across multiple runs. This is particularly important in label-free quantification and comparison of analytes by LC-MS. The output of an LCMS experiment consists of three dimensions: (1) the elution time, also called retention (RT) point, (2) the *m/z* value, and (3) the intensity (ion abundance). Figure 3a presents three-dimensional data derived from a typical LC-MS experiment for a single run (Listgarten et al., 2005). As shown in the figure 3, each LC-MS run generates spectra comprised of hundreds of peak intensities for peptides with specific RT and *m/z* values. Figure 3b shows a mass spectrum (ion abundance vs. *m/z*) at a particular RT point (RT in the figure is 10 minutes). Figure 3c depicts the total ion current (TIC) obtained by calculating the sum of the ion abundances across the *m/z* dimension for each RT point. Although RT is a continuous variable, the LCMS system produces mass spectra at a discrete set of RT points, usually a few seconds apart. It is typical to represent RT points by scan indices, since there is a one-to-one correspondence between RT points and total MS scan numbers.

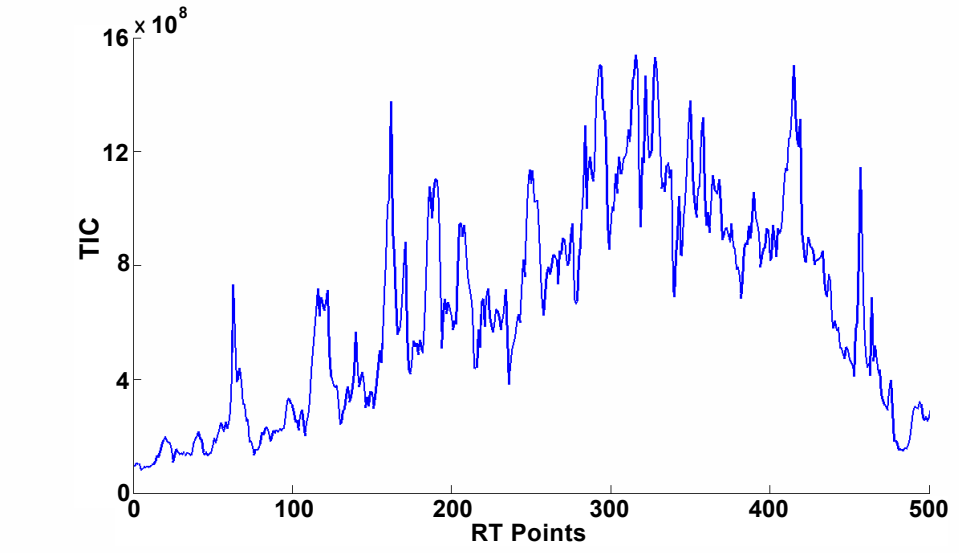
In differential protein expression studies, multiple LC-MS runs are compared to identify differentially abundant peptides between distinct biological groups. This is a challenging task because of the following reasons: (1) substantial variation in RT across multiple runs due to the LC instrument conditions and the variable complexity of peptide mixtures, (2) variation in *m/z* values due to occasional drift in the calibration of the mass



(a) Three-dimensional LC-MS data of a sample for RT points between 10 and 12 minutes and m/z values between 400 and 1000.



(b) Mass spectrum in the range between 400 and 1600 m/z at RT=10 minutes.



(c) TIC plot of the LC-MS data between 10 to 55 minutes of RT.

Figure 3: Data derived from a typical LC-MS experiment (Listgarten et al., 2005).

spectrometry instrument, and (3) variation in peak intensities due to spray conditions (in most cases this is proportional to concentration of peptides in the sample). Thus, efficient and robust normalization and alignment algorithms are needed for quantitative comparison of multiple LC-MS runs. Figure 4 presents a typical LC-MS run of a sample on a Qstar Elite instrument (Q-TOF). For visualization purpose, TIC and extracted ion current (XIC) are plotted. The former is a plot of the sum ion count across the entire *m/z* range vs. retention time. The latter is a plot of the sum of the ion signal for a particular *m/z* value vs. the retention time. Figure 5 depicts three TIC profiles obtained from the same subject. Overlaying profiles from replicate LC-MS data allow us to assess the reproducibility of the sample preparation and LC-MS data generation process.

The increasing demand and challenges for label-free quantification of analytes through LC-MS have led to the development of a number of software packages including OpenMS (Kohlbacher et al., 2007), CPM (Listgarten et al., 2007), LCMSWARP (Jaitly et al., 2006; Umar et al., 2007), MapQuant, (Leptos et al., 2006), Msight (Palagi et al., 2005), msInspect (Bellew et al., 2006), SpecArray,(Li et al., 2005), SuperHirn, (Mueller et al., 2007), mzMine (Katajamaa and Oresic, 2005), and Xalign (Zhang et al., 2005). Commercial software tools include MSView (Wang et al., 2003), Spectromania (Tammen et al., 2003), MosaiquesVisu (Wittke et al., 2004), MATLAB Bioinformatics Toolbox 3.0 (The Mathworks), MassHunter (Agilent), metAlign (PlanResearch International B.V.), MS Resolver (Pattern Recognition Systems), Rosetta Elucidator (Rosetta Biosoftware), DeCyder (GE Healthcare AB), Sieve (Thermo Fisher Scientific), MarkerView (Applied Biosystems), and MassLynx (Waters). Most of these software tools have their own computational requirements and implicit challenges. Some are instrument specific, others are proprietary. Thus, they lack the flexibility to analyze data generated from different instruments and the options to further optimize the algorithms.

LC-MS data preprocessing

Various data preprocessing steps are conducted before LC-MS runs can be compared for differential protein expression. These include deconvolution of multiple charged peaks and isotope clusters using the maximum entropy approach (Zhongqi et al., 1997). Other preprocessing steps include outlier screening, binning, baseline correction, smoothing, alignment, and normalization. In the following, we briefly discuss alignment and normalization methods.

Alignment is necessary to correct for chromatographic and mass spectrometric drifts that do not reflect real sample variation. Alignment methods find a common set of features across

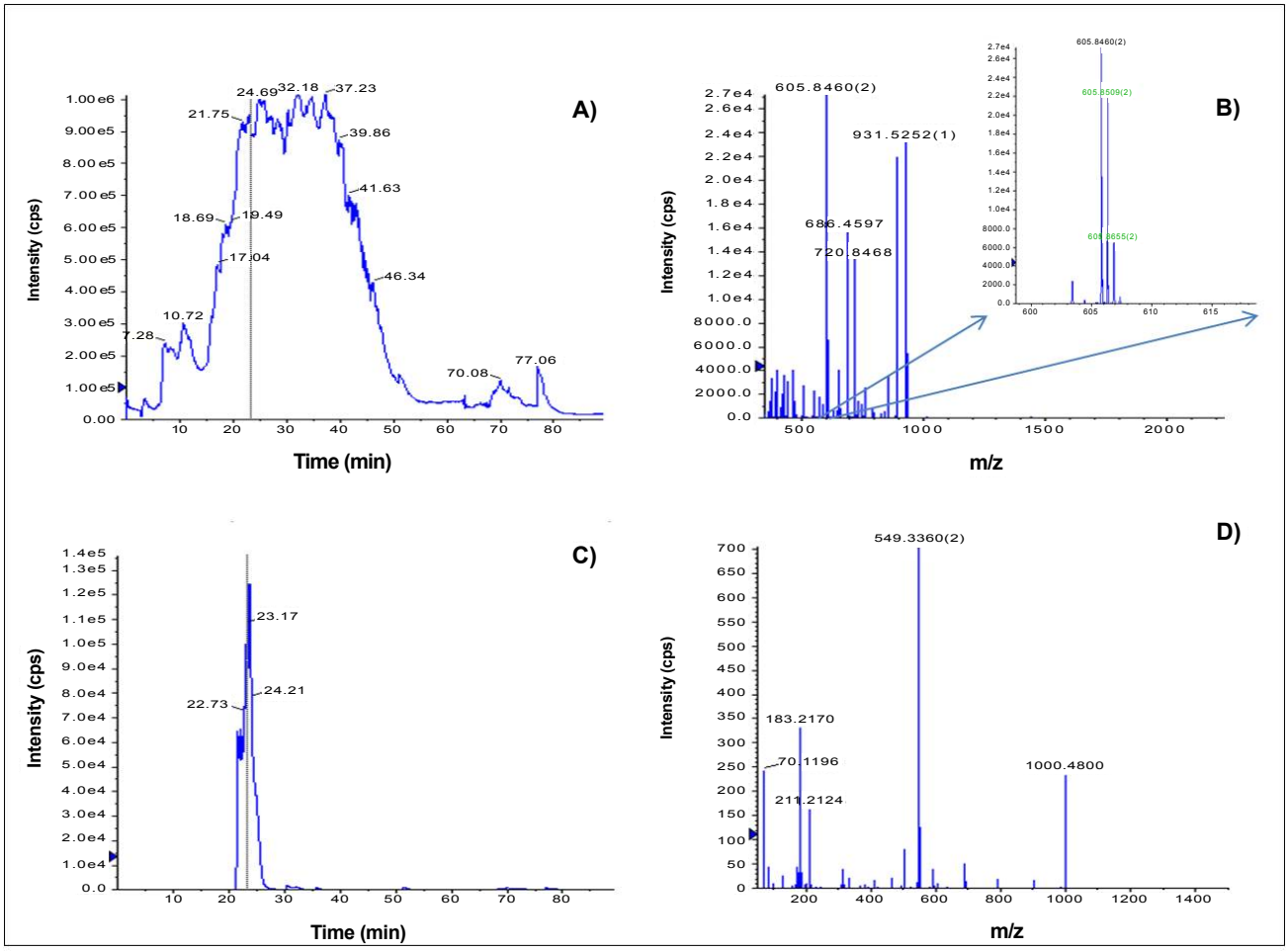


Figure 4: Typical LC-MS data from Qstar Elite instrument: A) Total ion current scan for a 90 min run from a quadrupole-time-of-flight hybrid mass spectrometer; B) MS (+TOF) survey scan at 23.44 min (Inset: Magnified view of mass region 600-620 m/z); C)Extracted ion current of 605.84 m/z peak; D) TOF LC- product of 605.84 m/z peak.

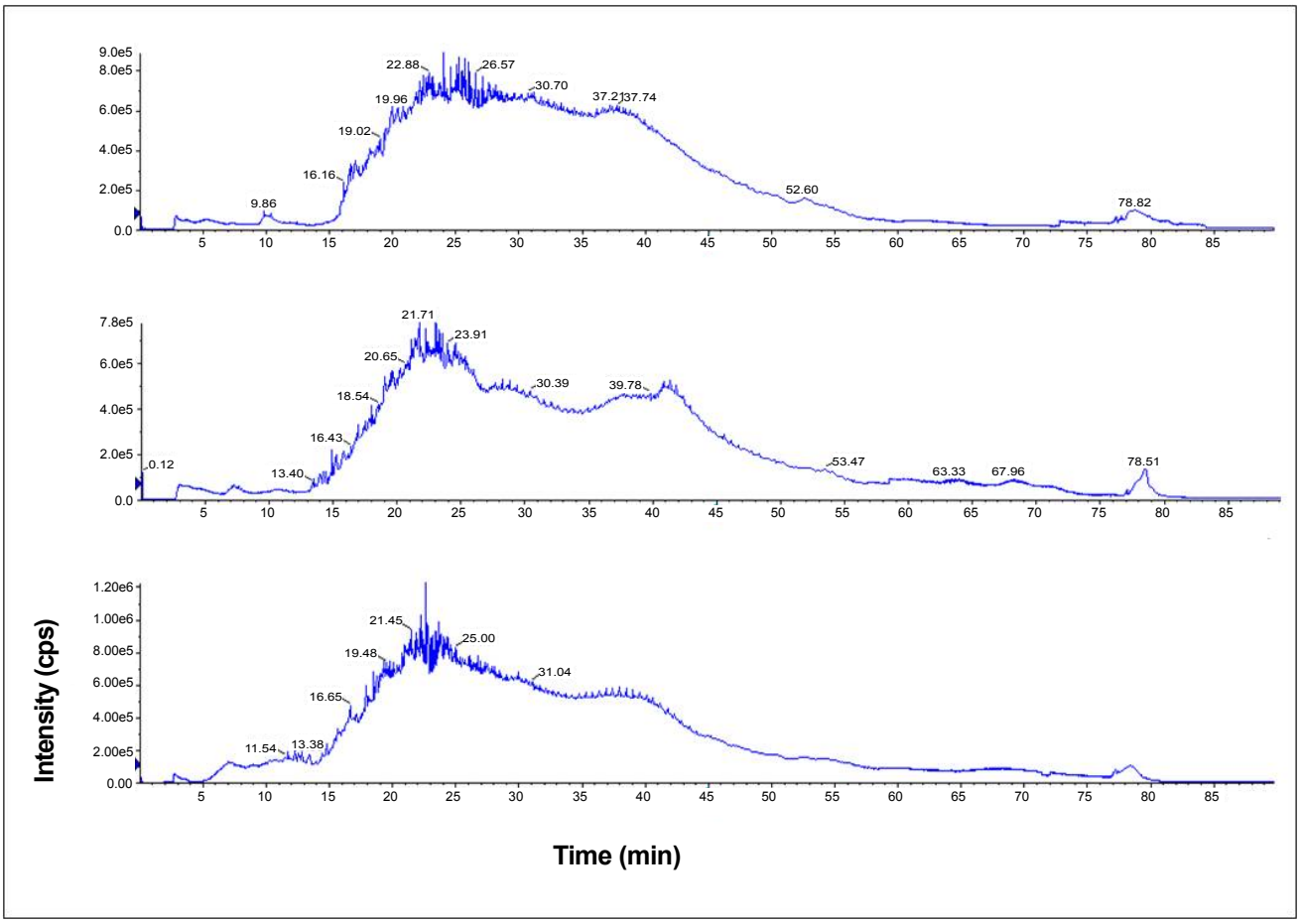


Figure 5: TIC chromatograms from replicated runs of human IgG and Albumin depleted serum. SCX fraction (0.1M) (10µl) was loaded for 90 min with a flow rate of 300nl/min in 1% mobile phase B (98% ACN, 2% water, 0.1% formic acid).

LC-MS runs to allow quantitative comparison of the same biological entities. Without alignment, the same biomolecule can have different *m/z* or retention time point across multiple runs. Thus, alignment with respect to both *m/z* and retention time is a prerequisite for quantitative comparison of proteins/peptides by LC-MS.

Alignment algorithms have traditionally been used on data points and/or feature vectors of fixed dimension (Ramsay and Silverman, 2002). Applications of these algorithms for LC-MS data alignment have been reported in the literature (America et al., 2006; Horvatovich et al., 2007; Jaitly et al., 2006; Listgarten et al., 2007; Mueller et al., 2007; Pierce et al., 2005; Prakash et al., 2006; Radulovic et al., 2004; Sadygov et al., 2006; Wang et al., 2007; Wiener et al., 2004). The most common approaches for aligning LC-MS data are based on the identification of landmarks or structural points (referring to the unique charge species in data) and the use of internal standards, respectively. The landmarks are usually associated with maxima, minima, or other critical or inflection points. Multiple LC-MS runs are then aligned so that the landmarks are synchronized. In this framework, the most widely used algorithm is dynamic time warping (DTW) that performs the alignment in time axis by stretching or shrinking the time series data. Another common method is correlation optimized warping (COW), which computes a piecewise linear transformation by dividing the time series into segments and then performing a linear warp within each segment to optimize overlap while constraining segment boundaries. The parameters for the best linear transformation are determined by maximizing the sum of correlation coefficients or covariance between data segments in pairs of samples. Most of the existing algorithms including DTW and COW are either limited to a consensus combination of pair-wise alignment or use a reference (template) for alignment. This limitation leads to suboptimal results compared to global alignment techniques.

Normalization is one of the important preprocessing tasks needed to separate interesting biological variation from obscuring sources of variability introduced in LC-MS–based studies. In particular, when data are available from multiple LC-MS experiments in which expressions of different types of biomolecules are measured for the same participants, the integration of the data is nontrivial. For example, many clustering algorithms measure profile dissimilarity by Euclidean distance. If one experimental platform produces large numbers and the others produce small numbers, then profiles from the former experiment will dominate (and potentially distort) the identification of clusters. To mitigate such difficulties, it is desirable to begin by converting the different measurements to a common scale. Due to lack of reliable methods, internal standards spiked in biological samples are typically used for normalization. For example, the mzMine toolbox utilizes multiple internal standard compounds injected to samples to calculate a set of normalization factors, one for each standard compound based on either searching for a standard compound peak closest to the peak or using weighted contribution of each standard compound (Katajamaa and Oresic, 2005). However, as the authors themselves noted that this method suffers from the ad hoc assign-

ments of internal standards for each component based on a subset of relevant chemical properties (Sysi-Aho et al., 2007). Also, in the context of the need for universally applicable analytical tools and since internal standards can vary depending on the instrument used and samples under study, it is desired to develop a normalization method that does not rely on internal standards.

Alignment and normalization methods that rely on optimization of global fitting function provide an alternative solution to address the above challenges without requiring landmarks or internal standards. For example, a recently introduced method called continuous profile model (CPM) has been applied for alignment and normalization of continuous time-series data and for detection of differences in multiple LC-MS data (Listgarten et al., 2005; Listgarten et al., 2007).

Difference detection

Difference detection deals with the identification of peaks that represent differentially abundant biomolecules. Various unsupervised and supervised methods have been proposed for peak selection from LC-MS data. For example, principal component analysis (PCA) transforms the spectral data to a new coordinate system such that the variables in the new data space (known as scores or principal components) are orthogonal and are sorted in the decreasing order of their variances. The peaks that contribute to the top factors are identified by using the eigen-value plot (Purohit and Rocke, 2003). A similar approach has been used in a supervised way (e.g., partial least squares, PLS), where the training examples with known disease status are used to calculate the factors. The weight plot obtained from this PLS analysis provides a tool to select useful peaks (Chen et al., 2007; Purohit and Rocke, 2003).

Another commonly used supervised approach applies statistical analyses such as t-test, shrinkage t-statistic, and weighting factor,(Golub et al., 1999) which recognize differentially abundant peaks between two groups with multiple subjects. For example, in a pair-wise comparison between patient and control subgroups, we calculate the shrinkage t-statistic for each feature (with a specific retention time point and *m/z* value) in the preprocessed LC-MS data. The shrinkage t-statistic is a regularized t-statistic that is based on a model-free shrinkage estimator of the variance vector across peptides/glycans (Opgen-Rhein and Strimmer, 2007). To calculate non-parametric p-values, a permutation method can be used by randomly reassigning the class labels and computing the corresponding t-statistics. The resulting p-values are utilized to control the false discovery rate. Alternatively, multivariate permutation tests are used for controlling the number and proportion of false discoveries.(Korn et al., 2004) The permutation tests are based on permutations of the labels of which samples are in which classes. For each permutation, the shrinkage t-statistics are recomputed to determine a measure of the extent it appears differentially expressed between the random classes determined by the random permutation. The peaks are then ranked by their shrinkage t-statistic for the permutation. This process is repeated for a large number of permutations. Consequently, for any threshold, we compute the distribution of the number of peptides/glycans that would have t-statistic better than that threshold for permutations. That is the distribution of the number of false discoveries, since peptides/glycans that are significant for random permutations are false discoveries.

The selected peaks are typically used as inputs to a pattern classification algorithm such as random forest and support vector machine (SVM). Random Forest is an ensemble of unpruned classification or regression trees, induced from bootstrap samples of the training data, using random feature selection in the tree induction process. It is a classification method based on "growing" an ensemble of decision tree classifiers. In order to classify a new object, the input is analyzed using each of the classification trees in the forest. Each tree gives a classification, "voting" for that class. The forest chooses the classification having the most votes (over all the trees in the forest). A measure of the importance of classification variables is also calculated by considering the difference between the results from original and randomly permuted versions of the data set. Prediction is made by aggregating (majority vote for classification or averaging for regression) the predictions of the ensemble. Random forest generally exhibits a substantial performance improvement over the single tree classifier such as classification and regression tree (CART). Izmirlian (Izmirlian, 2004) discussed how the random forest approach can be successfully applied for in proteomics profiling study to construct a classifier and discover peak intensities most likely responsible for the separation between classes.

The SVM recursive feature elimination (SVM-RFE) algorithm recursively classifies samples with SVM and selects peaks according to their SVM weights (Guyon et al., 2002). Benefiting from the good performance of SVMs in high-dimensional gene expression data, SVM-RFE is often considered as one of the best feature selection algorithms in the literature. Also, stochastic global optimization methods such as genetic algorithms, simulated annealing, and swarm intelligence methods have been used to systematically select features from a high-dimensional search space without the need for an exhaustive search. We previously developed a hybrid of SVM and ant colony optimization (ACO) to select a panel of optimal peaks (Ressom et al., 2007). To evaluate the generalization capability of the peaks and the SVM classifier determined by the training data set, we test the SVM classifier using a blind validation set, i.e., a test set that is set aside during the process of data preprocessing, peak selection, and building the SVM classifier. An important weakness of many machine learning-based classification algorithms is that they are not based on a probabilistic model. There is no probability level or confidence interval associated with predictions derived from using them to classify a new set of data. The confidence that an analyst can have in the accuracy of the results produced by a given classifier is based purely on its historical accuracy—how well it has predicted the desired response in other, similar circumstances. Thus, after learning is completed, a machine-learned paradigm is evaluated for its performance through previously unseen testing data set (also known as a blind validation set). The purpose of this evaluation is to prove the adequacy or to detect the inadequacy of selected peaks or a classifier. Inadequate performance could be attributed to insufficient or redundant peaks, inappropriate selection of model structure for the classifier, too few or too many model parameters, insufficient training, overtraining, error in the program code, or complexity of the underlying system such as presence of highly nonlinear

relationships, noise, and systematic bias. The aim of evaluating is a classifier is to insure that it serves as a general model. A general model is one whose input-output relationships (derived from the training data set) apply equally well to new sets of data (previously unseen test data) from the same problem not included in the training set. Thus, the main goal of machine learning-based modeling is thus the generalization to new data of the relationships learned on the training set (Wang et al., 2006b).

Various methods have been used to test the generalization capability of a classifier. These include the *k*-fold cross-validation, bootstrapping, and hold-out methods. In *k*-fold cross-validation, we divide the data into *k* subsets of (approximately) equal size. We train the model *k* times, each time leaving out one of the subsets from training, but using only the omitted subset to compute the classification error. If *k* equals the sample size, this is called "leave-one-out" cross-validation. In the leave-one-out method, one sample is selected as a validation sample and feature selection and classifier building are performed using the remaining data set. The resulting model is tested on the validation sample. The process is repeated until all samples appear in the validation set. In the hold-out method, only a single subset (also known as validation set) is used to estimate the generalization error. Thus, the hold-out method does not involve crossing. In bootstrapping, a sub-sample is randomly selected from the full training data set with replacement. Common bootstrapping methods include bagging and boosting. Bagging can be used with many classification methods and regression methods to reduce the variance associated with prediction, and thereby improve the prediction process. In bagging, many bootstrap samples are drawn from the available data, some prediction method is applied to each bootstrap sample, and then the results are combined by voting. Boosting can also be used to improve the accuracy of classification. Unlike bagging, the samples used at each step are not all drawn in the same way from the same population, but rather the incorrectly predicted cases from a given step are given increased weight during the next step. Hence, boosting uses a weighted average of results obtained from applying a prediction method to various samples.

Whole population based approaches (e.g., SVM-RFE and ACO-SVM) enable us to select a panel of peaks that lead to good classification accuracy. Although a subset of the peaks identified by these methods may be attributed to a subgroup of subjects, neither the subset of peaks nor the subgroup of subjects could be isolated due to the nonlinear interaction of the peaks. Methods that search for subgroup-specific peaks and discover unknown subgroups are needed. Such methods are expected to give insight into the relationship between the selected peaks and the corresponding subgroup of subjects. For example, in gene expression data analysis, methods for capturing genes differentially expressed in only a subset of patients have been explored, instead of trying to identify differentially expressed genes at the whole-population level (e.g., comparison of sample means) (Lyons-Weiler et al., 2004; Pavlidis and Poirazi, 2006). These types of methods offer a more patient-specific approach for marker identification, and can select markers that exhibit complex patterns that are missed by metrics that work under the comparison of two prelabeled phenotypic groups (Friedman and Meulman, 2004; Kim et al., 2006; Tadesse et al., 2005).

Challenges and Future Outlook

It is clearly evident that a single technology or method alone cannot address issues associated with dynamic constituents of a proteome; however, improvements made so far have definitely broadened utility of "proteomics" as a tool for biological understanding. One of the remaining challenges in proteomics is to quantify all protein entities in a single measurement. What is desirable is a fully integrated multifunctional system that would allow comprehensive quantification of a wide spectrum of proteins (Jayaraman, 2002).

Challenging as it is, experimental design would need to include suitable sample preparation, labeling and detection methods conducive to downstream quantification (Zhiyuan et al., 2007). Factors such as sample preparation steps that impact accuracy and precision of data acquired would need to be controlled. For example, depletion helps increase range of detection but consequently leads to sampling of a small subset of proteins. Ultimately, only a fraction of all proteins detected gets quantified (Zhiyuan et al., 2007). Although each sample preparation step is known to enhance data quality, sometimes excessive steps can result in selective loss of analytes. While sample complexity is considerably reduced, the exact state of a proteome is misrepresented. Similarly, proteomics discovery is limited by low data sampling rates which results in low analytical throughput. Even though many analytes can be detected by a single mass spectrometric measurement, time needed for efficient chromatographic separation of numerous peptides restricts high throughput analysis.

While multiplexing of protein profiling is desirable, current tools in discovery experiments impede simultaneous monitoring of all proteins. With advances in microfabrication technique, current bottlenecks perceived in proteome profiling schemes can be avoided. Profiling schemes such as microfluidics-based isoelectric focusing system, or microdialysis of proteins have successfully emerged to deliver potential improvements needed in this area (Jayaraman, 2002). For instance, microfluidics has been successfully applied for protein/peptide separation using chromatographic and/or electrokinetic-based principles (Veenstra and Yates, 2006). Successful separation of yeast cell protein lysate has been demonstrated by using multi-dimensional system (Veenstra and Yates, 2006). While microfluidics is capable of dealing with minute sample amounts, it would be unrealistic to expect that certain problems encountered for un-miniaturized setups will be resolved with microfluidics (Veenstra and Yates, 2006). Issues such as co-migration or co-elution of proteins will still need to be resolved. Since most of the issues faced depend upon the intrinsic nature of the proteins and not the analytical tool itself, such factors need to be resolved at the separation level.

As large volume and high dimensional data are being generated by the rapidly expanding use of mass spectrometric technologies, the number of reported applications of proteomic pattern recognition algorithms is expected to increase. However, with increasing demand comes the need for further improvements that can make implementation of these algorithms for high dimensional LC-MS data analysis more efficient. Key improvements include: (i) careful study design to minimize the effect of factors that may introduce bias to the data; (ii) enhanced com-

putational power to handle the high dimensionality and large volume data; (iii) improved high-throughput technologies with less background noise and technical variability; (iv) enhanced quality control and protocol development/implementation; (v) improved data preprocessing methods to minimize the impact of background noise, sample degradation, and variability in sample preparation and instrument settings (v) improved visualization tools to assess data quality and interpret results; (vi) adequate data storage and retrieval systems; (vii) advances in statistical and machine learning methods to enhance their speed and make them more accessible to the user.

Careful study design is needed to make sure that a protocol is in place that enables appropriate randomization and replication to avoid bias in sample collection and sample preparation (Zhang and Chan, 2005). Zhang (Zhang, 2005) noted that systematic biases from pre-analytical variability, which are attributed to samples could be collected under different protocols for different purposes, and analytical variability caused by sample preparation methods are often specific to institutions (sites). Hence, the use of specimens from multiple institutions combined with sound study is suggested as a means to address such biases. It is also indicated that the typical way of pooling multiple data sets together, followed by randomly dividing them into training and testing sets may still turn out to be overly optimistic with results unsustainable in actual "field use." With the large number of simultaneously measured variables, it is possible for a complex multivariate model to pick up from a pooled dataset the different types of systematic biases that existed in the original individual data sets. Hence, unless the number of sites is large and diverse enough to form a true representative sample of the target population, the "mix-and-split" use of multi-site samples is not recommended. An alternative and more conservative approach is to conduct independent discovery sessions using the data sets separately, followed by inter-institution validation.

Acknowledgements

This work was supported in part by the National Science Foundation Grant IIS-0812246, the National Cancer Institute (NCI) R21CA130837 Grant, NCI Early Detection Research Network Associate Membership Grant, and the Prevent Cancer Foundation Grant awarded to HWR.

References

1. America AH, Cordewener JH, van Geffen MH, Lommen A, Vissers JP, et al. (2006) Alignment and statistical difference analysis of complex peptide data sets generated by multidimensional LC-MS. Proteomics 6: 641-53. » CrossRef » Pubmed » Google Scholar
2. Baggerly KA, Morris JS, Coombes KR (2004) Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments. Bioinformatics 20: 777-85. » CrossRef » Pubmed » Google Scholar
3. Bantscheff M, Schirle M, Sweetman G, Rick J, Kuster B (2007) Quantitative mass spectrometry in proteomics: A critical review. Analytical and Bioanalytical Chemistry 389: 1017-1031. » CrossRef » Pubmed » Google Scholar
4. Bellew M, Coram M, Fitzgibbon M, Igra M, Randolph T, et al. (2006) A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC-MS. Bioinformatics 22: 1902-9. » CrossRef » Pubmed » Google Scholar
5. Beynon RJ, Pratt JM (2005) Metabolic labeling of proteins for proteomics. Mol Cell Proteomics 4: 857-72. » CrossRef » Pubmed » Google Scholar
6. Builder SE, Ed."Hydrophobic Interaction Chromatography: Principles and Methods" Amersham Pharmacia Biotech; ISBN: 9197049042. » CrossRef » Pubmed » Google Scholar
7. Chambers G, Lawrie, Laura, Cash P, Murray, et al. (2000) Proteomics:a new approach to teh study of disease. Journal of Pathology 192: 280-288. » CrossRef » Pubmed » Google Scholar
8. Chen C, Gonzalez FJ, Idle JR (2007) LC-MS-based metabolomics in drug metabolism. Drug Metab Rev 39: 581-97. » CrossRef » Pubmed » Google Scholar
9. Delmotte N, Lasaosa M, Tholey A, Heinzle E, Huber CG (2007) Two-dimensional reversed-phase x ion-pair reversed-phase HPLC: an alternative approach to high-resolution peptide separation for shotgun proteome analysis. J Proteome Res 6: 4363-73. » CrossRef » Pubmed » Google Scholar
10. Diamandis EP (2004) Mass spectrometry as a diagnostic and a cancer biomarker discovery tool: opportunities and potential limitations. Mol Cell Proteomics 3: 367-378. » CrossRef » Pubmed » Google Scholar
11. Engelsberger WR, Erban A, Kopka J, Schulze WX (2006) Metabolic labeling of plant cell cultures with K(15)NO3 as a tool for quantitative analysis of proteins and metabolites. Plant Methods 2: 14. » CrossRef » Pubmed » Google Scholar
12. Florens L, Washburn MP, Raine JD, Anthony RM, Grainger M, et al. (2002) A proteomic view of the Plasmodium falciparum life cycle. Nature 419: 520-6. » CrossRef » Pubmed » Google Scholar
13. Friedman JH, Meulman JJ (2004) Clustering objects on subsets of attributes. Journal of the Royal Statistical Society: Series B 66: 815-849. » CrossRef » Google Scholar
14. Fung ET, Enderwick C (2002) ProteinChip clinical proteomics: computational challenges and solutions. Biotechniques 32: 34-41. » CrossRef » Pubmed » Google Scholar
15. Gao J, Opiteck GJ, Friedrichs MS, Dongre AR, Hefta SA (2003) Changes in the protein expression of yeast as a function of carbon source. J Proteome Res 2: 643-9. » Pubmed » Google Scholar
16. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, et al. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286: 531-7. » CrossRef » Pubmed » Google Scholar
17. Goodlett DR, Yi EC (2003) Stable isotopic labeling and mass spectrometry as a means to determine differences in protein expression. TrAC Trends in Analytical Chemistry 22: 282-290.
18. Graves PR, Haystead TA (2002) Molecular biologist's guide to proteomics. Microbiol Mol Biol Rev 66: 39-63. » CrossRef » Pubmed » Google Scholar
19. Gulcicek EE, Colangelo CM, McMurray W, Stone K, Williams K, et al. (2005) Proteomics and the analysis of

proteomic data: an overview of current protein-profiling technologies. Curr Protoc Bioinformatics 13: Unit 13.1. » CrossRef » Pubmed

20. Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene Selection for cancer classification using support vector machines. Machine learning 46: 389-422. » CrossRef » Google Scholar

21. Hall DA, Ptacek J, Snyder M (2007) Protein microarray technology. Mech Ageing Dev 128: 161-7. » CrossRef » Pubmed » Google Scholar

22. Horvatovich P, Govorukhina NI, Reijmers TH, van der Zee AGJ, Suits F, et al. (2007) Chip-LC-MS for label-free profiling of human serum. Electrophoresis 28: 4493-4505. » CrossRef » Pubmed » Google Scholar

23. Izmirlian G (2004) Application of the random forest classification algorithm to a SELDI-TOF proteomics study in the setting of a cancer prevention trial. Ann N Y Acad Sci 1020: 154-74. » CrossRef » Pubmed » Google Scholar

24. Jaitly N, Monroe ME, Petyuk VA, Clauss TR, Adkins JN, et al. (2006) Robust algorithm for alignment of liquid chromatography-mass spectrometry analyses in an accurate mass and time tag data analysis pipeline. Anal Chem 78: 7397-409. » CrossRef » Pubmed » Google Scholar

25. Jayaraman ML (2002) Advances in proteomic technologies. Annu Rev Biomed eng 4: 347-372.

26. Katajamaa M, Oresic M (2005) Processing methods for differential analysis of LC/MS profile data. *BMC* Bioinformatics 6: 179. » CrossRef » Pubmed » Google Scholar

27. Kearney P, Thibault P (2003) Bioinformatics meets proteomics—bridging the gap between mass spectrometry data analysis and cell biology. J Bioinform Comput Biol 1: 183-200. » CrossRef » Pubmed » Google Scholar

28. Kienhuis PGM, Geerdink RB (2002) A mass spectral library based on chemical ionization and collision-induced dissociation. Journal of Chromatography A 974: 161-168. » CrossRef » Pubmed » Google Scholar

29. Kim S, Tadesse MG, Vannucci M (2006) Variable selection in clustering via Dirichlet process mixture models. Biometrika 93: 877-893. » CrossRef » Google Scholar

30. Kohlbacher O, Reinert K, Gropl C, Lange E, Pfeifer N, et al. (2007) TOPP—the OpenMS proteomics pipeline. Bioinformatics 23: e191-7. » CrossRef » Pubmed » Google Scholar

31. Korn EL, Troendle JF, Simon R (2004) Controlling the number of false discoveries: application to high-dimensional genomic data. Journal of Statistical Planning and Inference 124: 379-398. » CrossRef » Google Scholar

32. Kuhner S, Gavin AC (2007) Towards quantitative analysis of proteome dynamics. Nat Biotechnol 25: 298-300. » CrossRef » Pubmed » Google Scholar

33. Leptos KC, Sarracino DA, Jaffe JD, Krastins B, Church GM (2006) MapQuant: open-source software for large-scale protein quantification. Proteomics 6: 1770-82. » CrossRef » Pubmed » Google Scholar

34. Li XJ, Yi EC, Kemp CJ, Zhang H, Aebersold R (2005) A software suite for the generation and comparison of peptide arrays from sets of data collected by liquid chromatography-mass spectrometry. Mol Cell Proteomics 4: 1328-40. » CrossRef » Pubmed » Google Scholar

35. Lill J (2003) Proteomic tools for quantitation by mass spectrometry. Mass Spectrom Rev 22: 182-94. » CrossRef » Pubmed » Google Scholar

36. Listgarten J, Emili A (2005) Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry. Mol Cell Proteomics 4: 419-34. » CrossRef » Pubmed » Google Scholar

37. Listgarten J, Neal RM, Roweis ST, Emily A (2005) Multiple Alignment of Continuous Time Series. Neural Information Processing Systems 17: 817-824. » Google Scholar

38. Listgarten J, Neal RM, Roweis ST, Wong P, Emili A (2007) Difference detection in LC-MS data for protein biomarker discovery. Bioinformatics 23: e198-204. » CrossRef » Pubmed » Google Scholar

39. Lyons-Weiler J, Patel S, Becich MJ, Godfrey TE (2004) Tests for finding complex patterns of differential expression in cancers: towards individualized medicine. BMC Bioinformatics 5: 110. » CrossRef » Pubmed » Google Scholar

40. Malyarenko DI, Cooke WE, Adam BL, Malik G, Chen H, et al. (2005). Enhancement of sensitivity and resolution of surface-enhanced laser desorption/ionization time-of-flight mass spectrometric records for serum peptides using time-series analysis techniques. Clin Chem 51: 65-74. » CrossRef » Pubmed » Google Scholar

41. McMaster M (2005) LC/MS: A Practical User's Guide. » CrossRef » Pubmed » Google Scholar

42. Mikesh LM, Ueberheide B, Chi A, Coon JJ, Syka JE, et al. (2006) The utility of ETD mass spectrometry in proteomic analysis. Biochim Biophys Acta 1764: 1811-22. » CrossRef » Pubmed » Google Scholar

43. Monteoliva L, Albar JP (2004) Differential proteomics: An overview of gel and non-gel based approaches. Brief Funct Genomic Proteomic 3: 220-239. » CrossRef » Pubmed » Google Scholar

44. Mueller LN, Rinner O, Schmidt A, Letarte S, Bodenmiller B, et al. (2007) SuperHirn - a novel tool for high resolution LC-MS-based peptide/protein profiling. Proteomics 7: 3470-80. » CrossRef » Pubmed » Google Scholar

45. Okulate MA, Kalume DE, Reddy R, Kristiansen T, Bhattacharyya M, et al. (2007) Identification and molecular characterization of a novel protein Saglin as a target of monoclonal antibodies affecting salivary gland infectivity of Plasmodium sporozoites. Insect Mol Biol 16: 711-22. » CrossRef » Pubmed » Google Scholar

46. Old WM, Meyer-Arendt K, Aveline-Wolf L, Pierce KG, Mendoza A, et al. (2005) Comparison of label-free methods for quantifying human proteins by shotgun proteomics. Mol Cell Proteomics 4: 1487-502. » CrossRef » Pubmed » Google Scholar

47. Opgen-Rhein R, Strimmer K (2007) Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach. Stat Appl Genet Mol Biol 6: Article9. » CrossRef » Pubmed » Google Scholar

48. Orvisky E, Drake SK, Martin BM, Abdel-Hamid M, Ressom HW, et al. (2006) Enrichment of low molecular weight fraction of serum for MS analysis of peptides associated with hepatocellular carcinoma. Proteomics 6: 2895-902. » CrossRef » Pubmed » Google Scholar

49. Palagi PM, Walther D, Quadroni M, Catherinet S, Burgess J, Zimmermann-Ivol CG, et al.. (2005) MSight: an image analysis software for liquid chromatography-mass spectrometry. Proteomics 5: 2381-4. » CrossRef » Pubmed » Google Scholar

50. Palcy S, Chevet E (2006) Integrating forward and reverse proteomics to unravel protein function. Proteomics 6: 5467-80. » CrossRef » Pubmed » Google Scholar

51. Pan S, Aebersold R, Chen R, Rush J, Goodlett DR, et al. (2009) Mass spectrometry based targeted protein quantification: methods and applications. J Proteome Res 8: 787-97. » CrossRef » Pubmed » Google Scholar

52. Panchaud A, Affolter M, Moreillon P, Kussmann M (2008) Experimental and computational approaches to quantitative proteomics: status quo and outlook. J Proteomics 71: 19-33. » CrossRef » Pubmed » Google Scholar

53. Pavlidis P, Poirazi P (2006) Individualized markers optimize class prediction of microarray data. BMC Bioinformatics 7: 345. » CrossRef » Pubmed » Google Scholar

54. Pierce KM, Wood LF, Wright BW, Synovec RE (2005) A comprehensive two-dimensional retention time alignment algorithm to enhance chemometric analysis of comprehensive two-dimensional separation data. Anal Chem 77: 7735-43. » CrossRef » Pubmed » Google Scholar

55. Prakash A, Mallick P, Whiteaker J, Zhang H, Paulovich A, et al. (2006) Signal maps for mass spectrometry-based comparative proteomics. Mol Cell Proteomics 5: 423-32. » CrossRef » Pubmed » Google Scholar

56. Purohit PV, Rocke DM (2003) Discriminant models for high-throughput proteomics mass spectrometer data. Proteomics 3: 1699-703. » CrossRef » Pubmed » Google Scholar

57. Qian WJ, Jacobs JM, Liu T, Camp DG 2nd, Smith RD (2006) Advances and challenges in liquid chromatography-mass spectrometry-based proteomics profiling for clinical applications. Mol Cell Proteomics 5: 1727-44. » CrossRef » Pubmed » Google Scholar

58. Radulovic D, Jelveh S, Ryu S, Hamilton TG, Foss E, et al. (2004) Informatics platform for global proteomic profiling and biomarker discovery using liquid chromatography-tandem mass spectrometry. Mol Cell Proteomics 3: 984-97. » CrossRef » Pubmed » Google Scholar

59. Raghothama CAP (2007) Quantitative proteomics for identification of cancer biomarkers. Proteomics-Clinical Applications 1: 1080-1089.

60. Ramsay JO, Silverman BW (2002) Applied functional data analysis : methods and case studies, Springer, New York. » CrossRef » Google Scholar

61. Ransohoff DF (2005) Bias as a threat to the validity of cancer molecular-marker research. Nat Rev Cancer 5: 142-9. » CrossRef » Pubmed » Google Scholar

62. Ressom HW, Varghese RS, Drake SK, Hortin GL, Abdel-Hamid M, et al. (2007) Peak selection from MALDI-TOF mass spectra using ant colony optimization. Bioinformatics 23: 619-26. » CrossRef » Pubmed » Google Scholar

63. Rieux L (2006) A nanoLC-MS-based platform for peptide analysis.

64. Ruelle V, Falisse-Poirrier N, Elmoualij B, Zorzi D, Pierard O, et al. (2007) An immuno-PF2D-MS/MS proteomic approach for bacterial antigenic characterization: to Bacillus and beyond. J Proteome Res 6: 2168-75. » CrossRef » Pubmed » Google Scholar

65. Sadygov RG, Maroto FM, Huhmer AF (2006) ChromAlign: A two-step algorithmic procedure for time alignment of three-dimensional LC-MS chromatographic surfaces. Anal Chem 78: 8207-17. » CrossRef » Pubmed » Google Scholar

66. Scigelova M, Makarov A (2006) Orbitrap mass analyzer—overview and applications in proteomics. Proteomics 2: 16-21. » CrossRef » Pubmed » Google Scholar

67. Sorace JM, Zhan M (2003) A data review and re-assessment of ovarian cancer serum proteomic profiling. BMC Bioinformatics 4: 24. » CrossRef » Pubmed » Google Scholar

68. Souchelnytskyi S (2005) Bridging proteomics and systems biology: What are the roads to be traveled. Proteome 5: 4123-4137. » CrossRef » Pubmed » Google Scholar

69. Stanton P (2003) HPLC of Peptides and Proteins: Methods and Protocols.

70. Sysi-Aho M, Katajamaa M, Yetukuri L, Oresic M (2007) Normalization method for metabolomics data using optimal selection of multiple internal standards. BMC Bioinformatics 8: 93. » CrossRef » Pubmed » Google Scholar

71. Tadesse MG, Sha N, Vannucci M (2005) Bayesian Variable Selection in Clustering High-Dimensional Data. Journal of the American Statistical Association 100: 602-617. » CrossRef » Google Scholar

72. Tammen H, Kreipe H, Hess R, Kellmann M, Lehmann U, et al. (2003) Expression profiling of breast cancer cells by differential peptide display. Breast Cancer Res Treat 79: 83-93. » CrossRef » Pubmed » Google Scholar

73. Umar A, Luider TM, Foekens JA, Pasa-Tolic L (2007) NanoLC-FT-ICR MS improves proteome coverage attainable for approximately 3000 laser-microdissected breast carcinoma cells. Proteomics 7: 323-9. » CrossRef » Pubmed

74. Veenstra TD, Yates JR (2006) Proteomics for Biological Discovery: Automation in Proteomics. Wiley, John & Sons. » CrossRef » Google Scholar

75. Wang G, Wu WW, Zeng W, Chou CL, Shen RF (2006a) Label-free protein quantification using LC-coupled ion trap or FT mass spectrometry: Reproducibility, linearity, and application with complex proteomes. Journal of Proteome Research 5: 1214-1223. » CrossRef » Pubmed » Google Scholar

76. Wang P, Tang H, Fitzgibbon MP, McIntosh M, Coram M, et al. (2007) A statistical method for chromatographic alignment of LC-MS data. Biostatistics 8: 357-67. »CrossRef » Pubmed » Google Scholar

77. Wang W, Zhou H, Lin H, Roy S, Shaler TA, et al. (2003) Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. Anal Chem 75: 4818-26.» CrossRef » Pubmed » Google Scholar

78. Wang Z, Wang Y, Xuan J, Dong Y, Bakay M, et al. (2006b) Optimized multilayer perceptrons for molecular classification and diagnosis using genomic data. Bioinformatics 22: 755-61. » CrossRef » Pubmed » Google Scholar

79. Wells DA, Weil, David A (2003) Directions in Automated Sample Preparation of Proteins. Pharma Genomics.

80. Wiener MC, Sachs JR, Deyanova EG, Yates NA (2004) Differential mass spectrometry: a label-free LC-MS method for finding significant differences in complex peptide and protein mixtures. Anal Chem 76: 6085-96. »CrossRef » Pubmed » Google Scholar

81. Wittke S, Kaiser T, Mischak H (2004) Differential polypeptide display: the search for the elusive target. J Chromatogr B Analyt Technol Biomed Life Sci 803: 17-26. » CrossRef » Pubmed » Google Scholar

82. Yasui Y, Pepe M, Thompson ML, Adam BL, Wright GL Jr, et al. (2003) A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection. Biostatistics 4: 449-63.» CrossRef » Pubmed » Google Scholar

83. Yates JR, Ruse CI, Nakorchevsky A (2009) Proteomics by mass spectrometry: approaches, advances, and applications. Annu Rev Biomed Eng 11: 49-79.» CrossRef » Pubmed » Google Scholar

84. Zhang X, Asara JM, Adamec J, Ouzzani M, Elmagarmid AK (2005) Data pre-processing in liquid chromatography-mass spectrometry-based proteomics. Bioinformatics 21: 4054-9. » CrossRef » Pubmed » Google Scholar

85. Zhang Z (2005) Bioinformatics tools for differential analysis of proteomic expression profiling data from clinical samples. Taylor & Francis CRC Press.

86. Zhang Z, Chan DW (2005) Cancer proteomics: in pursuit of "true" biomarker discovery. Cancer Epidemiol Biomarkers Prev 14: 2283-6.» CrossRef » Pubmed » Google Scholar

87. Zhiyuan H, Leroy H, Qiang T (2007) Quantitative proteomic approaches for biomarker discovery. Proteomics - Clinical Applications 1: 1036-1041.

88. Zhongqi Z, Shenheng G, Marshall AG (1997) Enhancement of the Effective Resolution of Mass Spectra of High-Mass Biomolecules by Maximum Entropy-Based Deconvolution to Eliminate the Isotopic Natural Abundance Distribution. Journal of the American Society for Mass Spectrometry 8: 659-670. » CrossRef » Google Scholar