3 Omics: Data Processing and Analysis

Saicharan Ghantasala, Shabarni Gupta, Vimala Ashok Mani, Vineeta Rai, Tumpa Raj Das, Panga Jaipal Reddy, and Veenita Grover Shah

Abstract

The innovations in genome sequencing technologies have emanated in better understanding of biosystems leading to the dawn of the "omics" era. Proteomics has been an integral interface in the post-genomic era, and has allowed researchers to explore other omics-based platforms like metabolomics, transcriptomics, phenomics, etc. In pursuit of obtaining a systemic understanding of biosystems, the scientific community is now largely incorporating a multi-omics-based workflow, with genomics and proteomics at the centre of this integrated approach. Techniques such as gel-based proteomics, mass spectrometry, protein microarrays and label-free platforms have emerged as powerful tools for high-throughput screening and discovery-based studies in many of these multi-omics disciplines. However, with increased throughput, large amount of data is generated, and analysis of huge data often poses a challenge to researchers. The automation in specialized software has been immensely helpful to researchers in data acquisition; however, the downstream workflow of these sophisticated technologies continues to disconcert scientists, embracing an integrated multi-omics approach. This chapter aims at providing an overview of various proteomics-based technologies and their data evaluation strategies in context to biological studies. Data storage in specialized databases also requires attention, but is beyond the scope of this chapter. Gel-based

S. Ghantasala • S. Gupta • V.A. Mani • V. Rai
T.R. Das • P.J. Reddy • V.G. Shah (✉)
Department of Biosciences and Bioengineering,
Indian Institute of Technology Bombay,
Powai, Mumbai 400076, Maharashtra, India
e-mail: veenita7shah@gmail.com

S. Srivastava (ed.), *Biomarker Discovery in the Developing World: Dissecting the Pipeline for Meeting the Challenges*, DOI 10.1007/978-81-322-2837-0_3

proteomics, mass spectrometry, protein microarrays and label-free technologies are some of the commonly employed techniques in metabolomics, interactomics, genomics and transcriptomics, thus encompassing a multi-omics perspective on data analysis.

Keywords

Omics • Data analysis • 2D-DIGE • MALDI-TOF • LC-MS-MS • Protein microarray • Surface Plasmon Resonance

3.1 Introduction

Omics refers to a field of biology aimed at elucidating the structure, function and dynamics of various biological entities that constitute a cell. Various omics technologies are used in different disciplines such as genomics, "the study of genes"; transcriptomics, "the study of expression and spatial distribution of gene at the mRNA level"; proteomics, "the study of gene function at the protein level"; and metabolomics, "the study of metabolites". The common objective of these omics approaches is the creation of large comprehensive data sets, which will help in a better understanding of the biology of organisms. Until a few decades ago, Sanger's sequencing was the sole method to gain an insight on primary structure of proteins, but recent advances in electrophoresis and chromatography, coupled with improvements in mass spectrometry, allow a better understanding of proteomes. Proteomics is the comprehensive analysis of protein structure, function and dynamics studied by investigating the protein abundance, modifications, interacting partners and pathways, in order to understand the cellular processes (Chandramouli and Qian 2009). Some of the sophisticated and high-throughput omics techniques, described in this chapter, such as 2-DE, DIGE, mass spectrometry, protein microarrays and surface plasmon resonance, make the quest of knowledge seem never ending. An overall schematic representing a typical proteomic analysis workflow using various omics tools is shown in Fig. 3.1.

With the advent of technology in proteomics, the data generation is fast, enormous and exploding in terms of size and complexity. The prime challenge today is to handle large data sets generated across the world and to analyse the same for relevant biological interpretations (Gomez-Cabrero et al. 2014). The accessibility of omics data to researchers is another matter of concern. Many proteomics data repositories have been established to safely store the vast data generated by researchers across the globe. Some of the known 2-DE data repositories include SWISS-2DPAGE, WORLD-2DPAGE Portal, EcoproDB, 2Dbase, GELBANK and proteome 2D-PAGE database. Recently, some mass spectrometry (MS)-based data repositories (PRIDE, GPMDB, PeptideAtlas, Tranche and NCBI peptidome) have also been established with generation of high-confidence data and coverage.

Besides proteomics, many other omics approaches have attained greater interest in research to offer a better understanding of biological systems. The emerging omics technologies such as transcriptomics, metabolomics and phenomics provide a critical insight into the origin, function and regulation of molecules and their pathways. With the influx of omics data, there is a growing need to know the subject better in order to combat the challenges in data management and analysis. Systems biology has undoubtedly evolved to manage, organize and process multi-omics data, in a methodical manner, by generating tools capable of analysing huge data sets (Gehlenborg et al. 1998). The current chapter

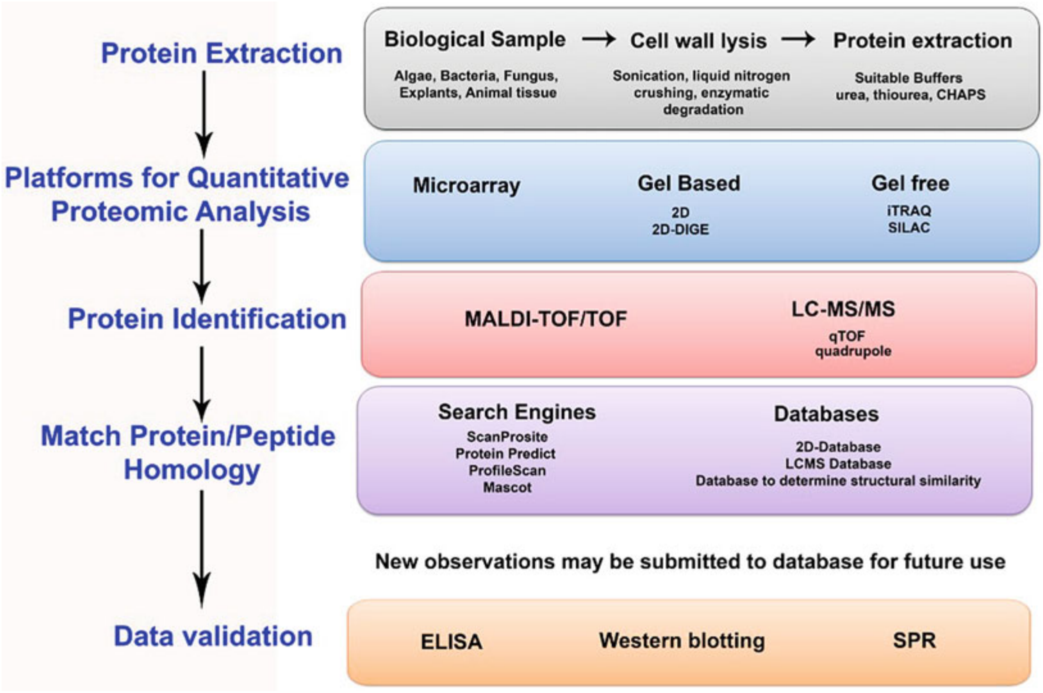

Protein Extraction
Platforms for Quantitative Proteomic Analysis
Protein Identification
Match Protein/Peptide Homology
Data validation
Biological Sample → Cell wall lysis → Protein extraction
Algae, Bacteria, Fungus, Explants, Animal tissue
Sonication, liquid nitrogen crushing, enzymatic degradation
Suitable Buffers urea, thiourea, CHAPS
Microarray
Gel Based
2D
2D-DIGE
Gel free
iTRAQ
SILAC
MALDI-TOF/TOF
LC-MS/MS
qTOF
quadrupole
Search Engines
ScanProsite
Protein Predict
ProfileScan
Mascot
Databases
2D-Database
LCMS Database
Database to determine structural similarity
New observations may be submitted to database for future use
ELISA
Western blotting
SPR

Fig. 3.1 Overall schematic representing a typical proteomic analysis workflow using various omics tools

focuses on providing an overview on some of the omics approaches, mentioned above, and their data evaluation strategies.

3.2 Two-Dimensional Gel Electrophoresis

Two-dimensional gel electrophoresis (2-DE) has been an effective gel-based method for global and quantitative proteomic analysis of various biospecimens. With its advent in the 1970s, this technique has paved the way for easy separation of complex protein mixtures, which was not possible through one-dimensional electrophoresis. The technique gained popularity in the 1990s with the introduction of immobilized pH gradient (IPG) strips which eased the protocols and increased its reproducibility and efficiency (Palzkill 2002; Beckett 2012). The basic principle involves the separation of a complex mixture of proteins based on two independent parameters: isoelectric point (pI) and molecular weight (Fig. 3.2a). The first-dimensional separation occurs through isoelectric focusing (IEF), a technique in which separation of proteins occurs based on their respective pI (the pH at which proteins form zwitterions), followed by a second-dimensional separation which is based on their molecular weight, similar to one-dimensional gel electrophoresis (1-DE). 2-DE approach can be used for global as well as differential protein profiling from biological samples (Fig. 3.2b). In the past, protein spot detection involved visual analysis of coomassie or silver-stained spots. Over the years, several software have been developed to enhance the sensitivity of protein spot detection (Table 3.1). The protein spots of interest (all spots in case of global proteomics and statistically significant spots in case of quantitative proteomics) can be excised, subjected to in-gel digestion followed by identification using mass spectrometry. The major advantage of 2-DE over 1-DE is the fact that it enables high-resolution separation of proteins with similar molecular weights but different pIs. Thus, 2-DE can be regarded as a milestone in

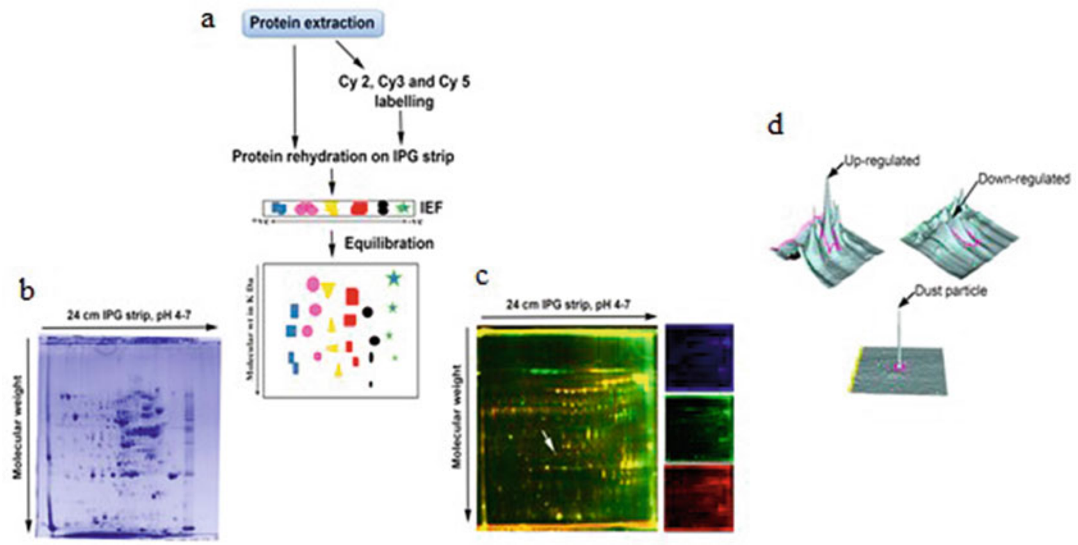


Fig. 3.2 Proteomic analysis using classical 2-DE and 2D-DIGE. A classical workflow demonstrating (**a**) 2-DE and 2D-DIGE experiment generating (**b**) coomassie-stained gel image of a complex protein mixture after classical 2-DE and (**c**) scanned image of a gel following 2D-DIGE with fluorescently labelled proteins as visible spots. 3D views for different (**d**) protein spots and dust particle as detected using DeCyder software

Table 3.1 List of commonly available software for 2-DE and 2D-DIGE analysis

S. No.	Software	Source	Applications
1	IMP7	www.gehealthcare.com	2-DE
2	DeCyder	www.gehealthcare.com	2D-DIGE
3	PDQuest	www.bio-rad.com	2-DE
4	Delta2D	www.decodon.com	2-DE and 2D-DIGE
5	GelScape	www.gelscape.ualberta.ca	2-DE
6	Flicker	http://open2dprot.sourceforge.net/Flicker	1-DE and 2-DE
7	REDFIN	www.ludesi.com/redfin	2-DE
8	Melanie	http://world-2dpage.expasy.org/melanie/	2-DE
9	ImageMaster	www.apbiotech.com	2-DE
10	ProteomeWeaver	www.defeniens-imaging.com	2-DE
11	Z3 2D-gel analysis system	www.compugen.co.il	2-DE
12	Progenesis and Phoretix 2D	www.nonlinear.com	1-DE and 2-DE

the proteomics era due to its ability to provide relatively higher protein coverage.

The limitations associated with 2-DE such as gel-to-gel variability and poor sensitivity led to the development of a relatively new technique with the ability to label proteins, which are then subjected to co-electrophoresis on a single gel. This technique, designed by Jon Menden's group, is referred to as two-dimensional difference gel electrophoresis (2D-DIGE) (Unlü et al. 1997). The methodology is similar to the conventional 2-DE except for the pre-electrophoretic labelling of samples with cyanine dyes (Fig. 3.2c). The cyanine dyes (Cy2, Cy3 and Cy5) bearing N-hydroxysuccinimidyl ester groups bind covalently with ε-amino groups of lysine residues in the protein of interest (Wilkins 2008). Each of these Cy dyes is matched for mass and charge, and is spectrally resolvable due to different excitation/emission wavelengths. Cy2 (excitation, 492 nm; emission, 510 nm) is preferably used to label the internal control, which is a pool of equal amounts of all samples loaded on the gel, whereas the samples under study (test and control) are labelled with Cy3 (excitation, 550 nm; emission, 570 nm) and Cy5 (excitation, 650 nm; emission, 670 nm) dyes. The internal control helps to eliminate the gel-to-gel variations. DIGE gels are usually run in biological replicates for gel consistency and better statistical analysis of protein spots.

3.2.1 Gel Analysis and Data Assimilation

The 2D-DIGE gels are scanned in a specialized variable mode laser scanner that enables visualization of a wide range of fluorescent wavelengths. The scanned images are further analysed using various software, some of which are enlisted in Table 3.1. These software align two or more gels in the same orientation, assign spot numbers to all protein spots in a gel and then overlay the gels. The differential expression (up- or down-regulation) in protein spots is enlisted based on their spot intensities. For instance, the DeCyder software provides options like differential in-gel analysis (DIA) for the analysis of an individual gel image and biological variation analysis (BVA) for the analysis of multiple gels (biological replicates) (GE Healthcare DeCyder 2D Software GE Healthcare 2-DE Principles and Methods 2004). DIA provides three-dimensional view of each protein along with its maximum slope and volume which aids in better understanding of the protein expression level in a given sample (Fig. 3.2d). Care must be taken to ensure meticulous gel cropping and overlaying to avoid elimination or mismatch of any protein spot. It is necessary to include certain filters during image analysis which aid in excluding some artifacts as the software cannot distinguish between protein spots and dust particles resulting in false 3D images (Fig. 3.2d). Additionally, manual curation is preferred to avoid false results and mismatching of protein spots. For the protein identification of 2D-DIGE analysis, a preparatory 2-DE gel run is preferred. However, a 2-DE gel may not be an exact replica of the corresponding DIGE gel making spot cutting a difficult task (Baggerman et al. 2005).

Owing to limitations such as limited protein coverage, higher sample requirement, low sensitivity in protein identification, low solubility of membrane-associated proteins, limited sample loading capacity of IPG strips and gel-to-gel variability, gel-based approaches (2-DE and 2D-DIGE) are increasingly being replaced by gel-free methods like iTRAQ (isobaric tag for relative and absolute quantitation) and SILAC (stable isotope labelling by amino acids in cell culture). These approaches directly label the peptides, which can be detected using mass spectrometry.

For many years, protein identification relied on a laborious technique called Edman degradation. However, there were several disadvantages linked to this method: (a) very slow and exhaustive process (only ten residues identified in 24 h), (b) required large amount of protein samples, (c) could not be performed if N-terminus of the protein was inaccessible (folded or modified) and (d) the reduction in efficiency after 50–60 residues.

The advent of high-throughput mass spectrometric technologies eased the task of protein identification and quantification with greater efficiency and accuracy. An overview of mass spectrometric analysis of proteins following the 2-DE procedure is described in the following sections.

3.3 Matrix-Assisted Laser Desorption/Ionization Time of Flight

Mass spectrometry is one of the key platforms in the field of proteomics. Out of the various mass spectrometric techniques available, matrix-assisted laser desorption/ionization time of flight (MALDI-TOF) gained popularity due to its ease of application for mass determination and protein identification. MALDI-TOF is a versatile approach for analysing proteins, peptides, oligonucleotides, glycans, polymers and organometallics, utilizing minimal reagents and easily accessible protocols.

MALDI-TOF like other mass spectrometric instruments consists of three functional components: an ion source to ionize and transfer analyte ions to gaseous phase, a mass analyser to separate ions based on their mass-to-charge ratio (m/z) and a detector to detect the separated analyte ions. Ionization techniques like MALDI and electrospray are soft-ionization approaches in mass spectrometry that allow large non-volatile biomolecules like proteins to ionize and vaporize readily (Croxatto et al. 2012).

MALDI is based on rapid photo-volatilization of analytes embedded in the matrix. A matrix is a compound capable of acquiring energy generated by laser and passing this energy to analyte molecules, thereby facilitating their desorption and ionization (Marvin et al. 2003). The excited matrix causes protonation of analytes resulting in formation of singly or doubly charged (sometimes multiply charged) ions in the gaseous phase. An electrostatic field further accelerates all analyte ions to attain equal kinetic energy. These ions further travel ahead in the flight tube to get separated by the analyser based only on their m/z ratio such that the ions with low m/z ratios travel faster compared to the ones with higher m/z. The separated ions are detected by the detector, which records the intensity of ions and generates a plot of m/z to relative ion intensity/abundance referred to as the mass spectrum. MALDI offers both positive and negative modes that can be selected during analysis depending on the charge attained by the analyte during ionization. Proteins and peptides are generally analysed using positive mode, whereas nucleic acids are analysed using negative mode.

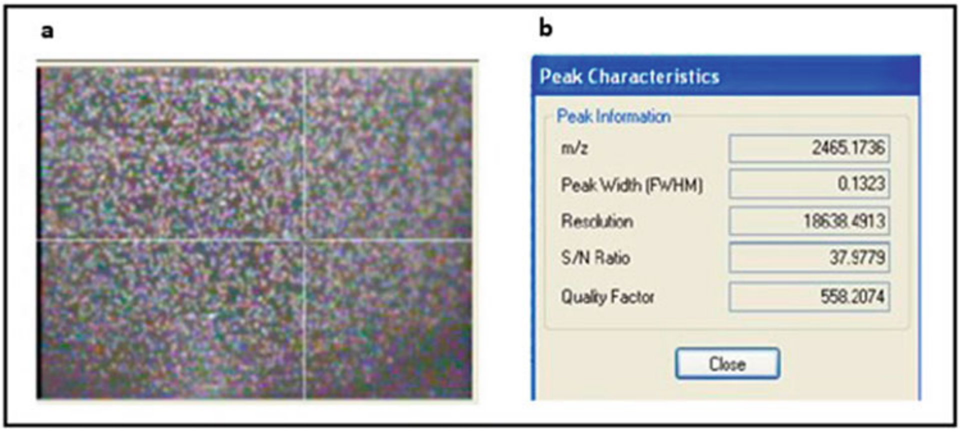
The selection of matrix is a crucial step in MALDI, and depends on the type of analyte and objective of analysis. Preparation of matrix solution involves dissolving the solid matrix in suitable organic solvents. Pure proteins and in-gel (trypsin or other suitable enzymes) digested proteins are analysed for mass determination and protein identification, respectively. Desalting is an important step performed to remove salts and other contaminants which may result in unwanted peaks and noise in the mass spectrum. Semi-purified proteins may generate large number of peaks causing uncertainty in the data obtained.

3.3.1 Data Acquisition and Analysis

Data acquisition refers to storing of electrical signals as mass spectrum after the ions are detected by the detector. The matrix and samples are loaded on the MALDI target resulting in sample crystallization followed by data acquisition by laser bombardment (Fig. 3.3a). In the generated mass spectrum, we may often observe noisy peaks having poor signal-to-noise ratio (S/N) and higher baseline for low m/z values. Data processing involves improving the S/N ratio of peaks and correcting the baseline by base smoothening and baseline subtraction, respectively (Fig. 3.3b). This is followed by peak picking (selection of suitable peaks), which is generally performed using algorithms such as SNAP (sophisticated numerical annotation procedure) and centroid. A modification of peak picking method called two-Gaussian algorithm is utilized for proteins showing low intensity in MS spectrum due to their poor expression (Kempka et al. 2004). Kernel matching pursuit (KMP) classifier is another novel algorithm that has been used to identify differentially expressed proteins in tissues of healthy controls and patients with lung cancer (Liu et al. 2003).

Bovine serum albumin (BSA) is extensively used as a control for MALDI studies. The molecular weight of BSA is determined in linear positive mode using sinapinic acid matrix. In linear mode, generally used for labile or high molecular weight molecules (proteins), the flight tube is straight with a detector at its end. In case of BSA, two major peaks are observed in the mass spectrum after initial processing of the raw data (smoothening and baseline correction). The peak with maximum intensity corresponds to the mass of intact singly charged BSA ion ~66.5 KDa (m/z value), whereas the smaller peak ~33 KDa denotes the doubly charged BSA ion (m/2Z) (Fig. 3.4a).

Mass determination for complex samples such as polysaccharides, polymers, glycans and



Peak Characteristics

Peak Information

m/z	2465.1736
Peak Width (FWHM)	0.1323
Resolution	18638.4913
S/N Ratio	37.9779
Quality Factor	558.2074

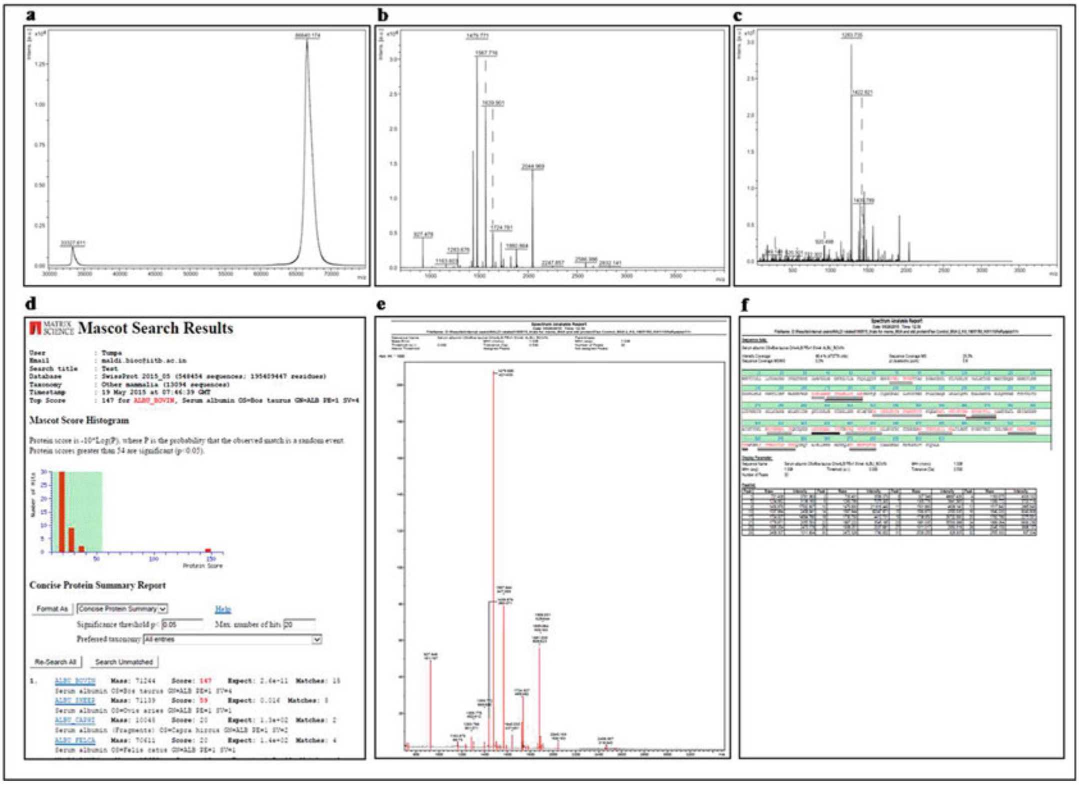
Fig. 3.3 Spotting and peak generation using MALDI-TOF. The representative figure illustrates (**a**) peptide-matrix crystals after spotting and (**b**) common peak characteristics obtained after peak generation

nucleic acids is more challenging compared to proteins. Polysaccharides (due to their high molecular weight) require additional salts and high concentration of matrix for better ionization and desorption. The MALDI analysis for oligosaccharides and polysaccharides may differ from each other due to their varying molecular weights and non-uniformity in their structure (Hao et al. 1998). For synthetic polymers, mass determination has always been questionable due to their molecular weight, polydisperse nature and the presence of additional end groups in their structure (Christian and Jackson 1997). However, MALDI-TOF plays a significant role in combating the challenges associated with mass determination of such complex samples.

For protein identification, the digested protein yields a mixture of peptides unique to the particular protein, which results in a peptide mass fingerprint (PMF) or peptide mass map. This PMF is further searched against different protein sequence databases identifying the correct match with the highest score (Webster and Oxley 2012). For BSA identification, after trypsin digestion, α-cyano-4-hydroxycinnamic acid (CHCA) is used as matrix to spot the BSA tryptic digest. The reflectron mode, designed for peptides and other small molecules, is used in such situations. In this mode, a reflectron (series of evenly spaced electrodes) is introduced in the flight tube at the end of the analyser to increase the flight length of the ions. Electrical field is applied over the reflectron so that the ions entering the reflectron undergo a continuous deceleration till they stop and leave the reflectron, where they get accelerated again to reach the detector. The reflectron thus facilitates in improving the resolution of small molecules by increasing their flight length.

The BSA peptide mass fingerprint shows many signature peptide peaks of different molecular masses: 927.4, 1,439.7, 1,479, 1,567.7, 1,724.7 and 2,044.9 (Fig. 3.4b). After processing the raw PMF, few peptide peaks with higher signal-to-noise ratios are selected for an extended step, the tandem mass spectrometry (MS-MS) (Fig. 3.4c). In tandem mass spectrometry, an additional analyser is incorporated, as observed in TOF/TOF with two TOF analysers or Q-TOF with one quadrupole and one TOF analyser. The two mass analysers in these instruments are separated by a collision cell and an ion deflection gate. The precursor ions, after being separated by the first analyser, get further fragmented in the collision cell, and the resultant daughter ions are analysed by the second analyser (Hoffman and Stroobant 2007). Tandem mass spectrometry thereby significantly increases the sensitivity of protein identification.

The MS-MS spectra for BSA with collective parent and daughter ions can then be searched



Mascot Search Results

```
User          : Tumpa
Email         : maldi.bio@iitb.ac.in
Search title  : Test
Database      : SwissProt 2015_05 (548454 sequences; 195409447 residues)
Taxonomy      : Other mammalia (13094 sequences)
Timestamp     : 19 May 2015 at 07:46:39 GMT
Top Score     : 147 for ALBU_BOVIN, Serum albumin OS=Bos taurus GN=ALB PE=1 SV=4
```

Mascot Score Histogram

Protein score is -10*Log(P), where P is the probability that the observed match is a random event. Protein scores greater than 54 are significant (p<0.05).

Concise Protein Summary Report

Significance threshold p< 0.05 Max. number of hits 20

Preferred taxonomy All entries

```
1.  ALBU_BOVIN   Mass: 71244   Score: 147   Expect: 2.4e-11  Matches: 15
    Serum albumin OS=Bos taurus GN=ALB PE=1 SV=4
    ALBU_SHEEP   Mass: 71139   Score: 59    Expect: 0.014    Matches: 8
    Serum albumin OS=Ovis aries GN=ALB PE=1 SV=1
    ALBU_CAPHI   Mass: 10048   Score: 20    Expect: 1.3e+02  Matches: 2
    Serum albumin (Fragments) OS=Capra hircus GN=ALB PE=1 SV=2
    ALBU_FELCA   Mass: 70611   Score: 20    Expect: 1.4e+02  Matches: 4
    Serum albumin OS=Felis catus GN=ALB PE=1 SV=1
```

Fig. 3.4 MALDI-TOF data analysis for mass determination and identification of bovine serum albumin. The panels demonstrate (**a**) mass spectrum of undigested BSA, (**b**) peptide mass fingerprint of trypsin-digested BSA and (**c**) MS/MS spectrum of BSA. Protein identification is performed using Swiss-Prot database to derive (**d**) Mascot search results for generating **(e)** Spectrum analysis report showing (**f**) BSA sequence data

against one of the various protein databases using mascot search engine (Fig. 3.4d). The search result also displays other serum albumins from closer taxonomical groups, due to protein sequence similarity in related species. The data retrieved from such databases provides a comprehensive information on molecular weight, isoelectric pH, protein sequence and the sequence coverage of protein under study (Fig. 3.4e, f). To achieve high sequence coverage, maximum number of peaks should be analysed while performing MS-MS.

Various databases, such as UniProt (Swiss-Prot), NCBInr, PlasmoDB and PlantEST (Table 3.2), help in protein identification via search engines like Matrix science (Mascot server). Though MALDI is a reliable and effective technique for rapid identification of proteins, the databases play a significant role in data generation and analysis. For this reason, studies associated with proteins from remote sample sources with no record of their information in existing databases face challenges in their true identification. Nevertheless, MALDI-TOF offers a promising platform for rapid molecular identification with extensive applications in clinical diagnostics, biomarker detection and tissue imaging studies.

3.4 Liquid Chromatography-Mass Spectrometry

The technique LC-MS is a result of a successful alliance between two techniques, the liquid chromatography and mass spectrometry. Liquid chromatography separates a highly complex mixture

Table 3.2 Few commonly used software/tools for protein identification from MS-MS data

S. No.	Software	Features	URL site	References
1	Mascot	Protein identification using mass spectrometry data	http://www.matrixscience.com/	Palagi et al. (2006)
2	MS-Fit	Mining the sequence of the protein from MS data	prospector.ucsf.edu	Palagi et al. (2006)
3	SEQUEST	Interpretation of tandem mass spectra data for protein identification and amino acid sequence	http://fields.scripps.edu/sequest/	Palagi et al. (2006)
4	X! Tandem	Protein identification using tandem MS data	http://www.thegpm.org/tandem/index.html	Palagi et al. (2006)
5	Sequit!	De novo sequencing of protein using tandem mass spectrum	http://www.sequit.org/	Palagi et al. (2006)
7	PEAKS	De novo sequencing of raw MS-MS data, label-free quantification	http://www.bioinfor.com/	Pevtsov et al. (2006)

into different subsets based on their physical and chemical properties. The separated mixtures are then ionized and injected into the mass spectrometer where ions get further separated under the influence of a strong electromagnetic field before getting detected. The data obtained is in the form of a spectrum, which can either be analysed manually (if the sample contains very few proteins) or with the use of specialized software developed for analysing high-throughput data (Table 3.2). Despite the superior capabilities of this technique, its use in protein studies picked up pace only in the 1980s, after the advent of ionization techniques such as electrospray ionization. This soft ionization method was demonstrated to generate charged ions from peptides in solution thereby paving the way for the emergence of LC-MS as a robust technique in proteomics. Over the last two decades, LC-MS has played a critical role in proteomics and contributed in sequencing of the yeast proteome (Peng et al. 2003), human tissues and cell lines (Phanstiel et al. 2011; Munoz et al. 2011; Geiger et al. 2012; Moghaddas Gholami et al. 2013), to name a few.

3.4.1 Data Acquisition and Analysis

Before the complex protein sample is injected into the mass spectrometer, a number of processing steps, aimed at reducing the complexity of samples, are involved. The process of enzymatic digestion ensures that the complex proteins are broken down into less complex peptides. MS instruments measure the m/z values of various fragments by reducing these sequences into patterns of numbers. Therefore, understanding the number pattern becomes very important in identifying the right peptide sequence leading to the right protein. A clear idea of the chemistry behind fragmentation is needed to help marvel the ability of these highly sensitive instruments. When the peptide ions collide with the neutral gas molecules in the collision cell, the most common cleavage involves that of the peptide bond forming the backbone. This cleavage results in the formation of two ions – most commonly referred to as the “y ion” and the “b ion”. The “y ion” is positively charged and represents the C-terminal end of the peptide, while the “b ion” is negatively charged ion and represents the N-terminal of the peptide. It is to be noted that in addition to the peptide bond, the other bonds of less significance also undergo fragmentation, albeit less frequently.

Manual analysis of the spectra and data interpretation is easier with a definite pattern of observed peptide peaks. In cases of complex peptide peak profiles, the process of understanding the data and being able to decipher the sequence can be time consuming. Moreover, the ambiguous fragmentation of certain peptides due to the variable amino acid fragmentation tendency and presence of proline, which does not facilitate easy fragmentation, can pose additional challenges, thereby making manual data analysis from MS-MS spectra an uphill task.

Owing to challenges associated with manual analysis of tandem mass spectrometry data, a number of software have been developed, each with their own merits and demerits. While most commercially available software come with a huge price tag, software like Mascot allow easy analysis and interpretation of data from tandem mass spectrometry experiments.

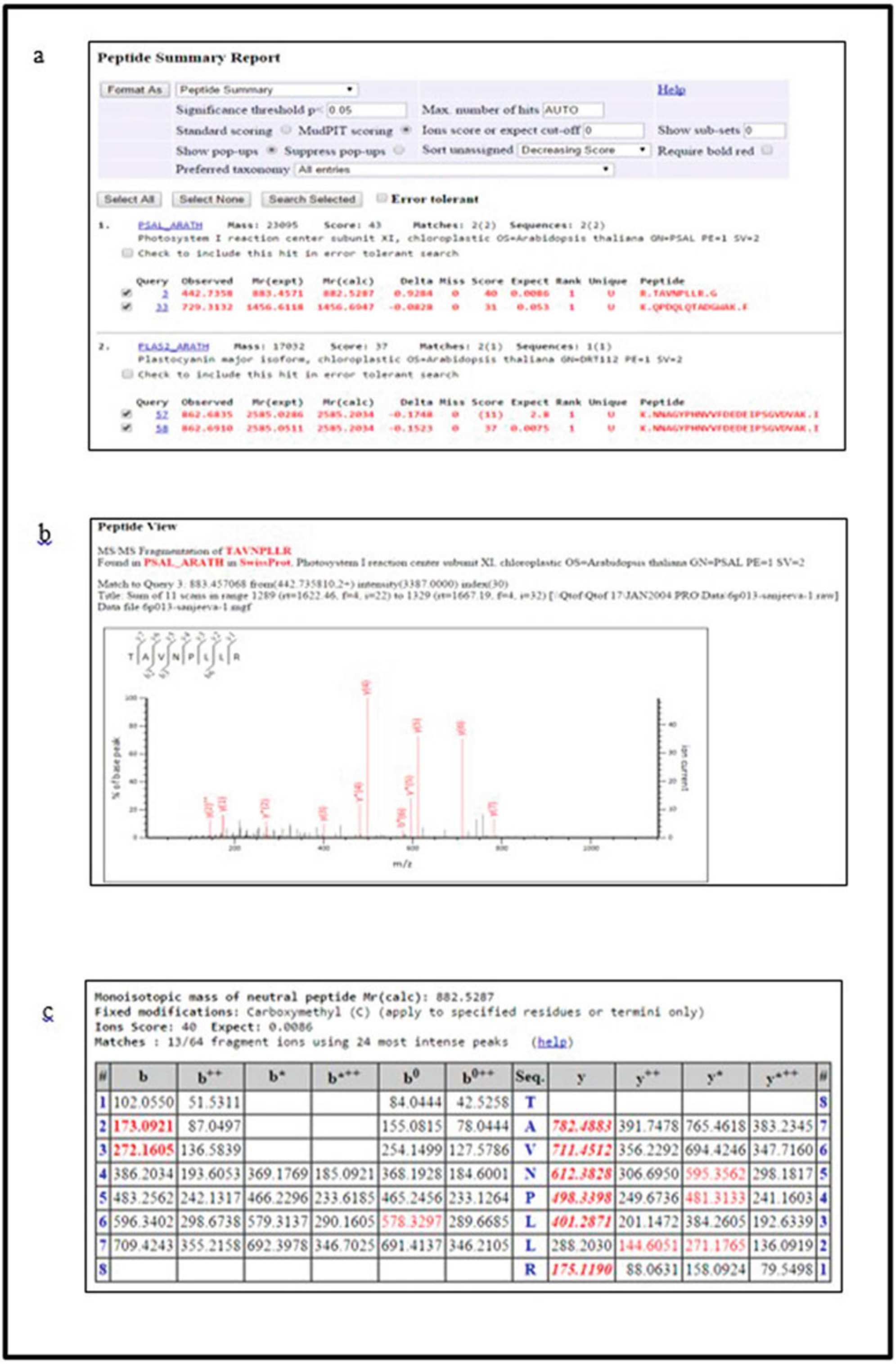
We now take up an example of an LC-MS run for a protein sample from *Arabidopsis thaliana* to better understand, analyse and interpret an LC-MS data set. The foremost step in analysing data using Mascot involves filling up of a MS-MS ion search form on the Mascot website, enquiring all important information such as the enzyme used during digestion, database used for search, number of missed cleavages, taxonomy, fixed and variable modifications, MS-MS tolerance, peptide charge, etc. After initiating the search, the software generates a result file containing a Mascot histogram and a peptide summary report. The protein hits falling outside the green region of the histogram are considered significant. The generated peptide summary report allows the user to select the significant peptides, in addition to other valuable information such as molecular weight, total ion score and the number of peptides matched (Fig. 3.5a). The peptide view contains all information regarding the peptide sequence derived from the corresponding “y” and “b” ions (Fig. 3.5b, c).

In the current analysis, a total of three significant hits were obtained. Of these, one hit corresponded to trypsin used for digestion of proteins. The other two peaks corresponded to the following proteins: photosystem I reaction centre subunit IV A (chloroplastic) and photosystem I reaction centre subunit IV B (chloroplastic), with molecular masses 14,958 Da and 15,188 Da, respectively.

3.4.2 Advancements in LC-MS-Based Quantification of Proteins

The technological advancements in mass spectrometry have led to increased use of these new-age instruments in quantitative proteomics. Quantification of peptides can either be achieved through differential labels such as iTRAQ and ICAT (isotope-coded affinity tag) or SILAC or label-free quantification approaches. The use of labels allows relative quantification of peptides among different biological samples in a single run. The iTRAQ-labelling approach enables calculation of peptide abundance based on intensities of fragment ions reported in the obtained MS-MS spectra (Chloe et al. 2007). The ICAT-based approach (Gygi et al. 1999) and SILAC (Ong et al. 2002), on the other hand, result in the generation of pairs of peptides with mass differences characteristic to the label used. The shift in masses and similarities in elution profiles form the basis for computing peptide ratios between the two forms of labels thereby allowing determination of the relative abundance of peptides. Despite their widespread usage, a major limitation associated with these approaches is the prejudice towards high-intensity peptide signals for the selection of the precursor ion, which results in under-sampling of low-abundant proteins in the sample mixture (Mueller et al. 2008).

The past few years have seen an upsurge in the label-free quantification strategies, which rely on correlating the abundance of a protein or peptide in a sample with the corresponding MS signal (Simpson et al. 2009). Determination of ion intensity using extracted ion chromatograms (XIC) is a popular method for the protein quantification. This involves summation of the number and intensity of the selected precursor ions at a specific m/z range and peak areas as a measure of the relative abundance (Old et al. 2005). Another approach that is increasingly being used for quantification is spectral counting of fragment ion spectra for a particular peptide. This is a semi-quantitative approach used for low to moderately mass-resolved LC-MS data. The approach relies



Peptide Summary Report

Format As: Peptide Summary ▾ — Help

Significance threshold p< 0.05 — Max. number of hits AUTO

Standard scoring ○ MudPIT scoring ● — Ions score or expect cut-off 0 — Show sub-sets 0

Show pop-ups ● Suppress pop-ups ○ — Sort unassigned Decreasing Score ▾ — Require bold red □

Preferred taxonomy All entries ▾

Select All — Select None — Search Selected — □ **Error tolerant**

```
1.     PSAL_ARATH     Mass: 23095     Score: 43     Matches: 2(2)  Sequences: 2(2)
       Photosystem I reaction center subunit XI, chloroplastic OS=Arabidopsis thaliana GN=PSAL PE=1 SV=2
     □ Check to include this hit in error tolerant search

       Query  Observed   Mr(expt)   Mr(calc)     Delta Miss Score Expect Rank Unique Peptide
     ☑    3  442.7358   883.4571   882.5287    0.9284    0    40  0.0086    1      U  R.TAVNPLLR.G
     ☑   33  729.3132  1456.6118  1456.6947   -0.0828    0    31   0.053    1      U  K.QPDQLQTADGWAK.F

2.     PLAS2_ARATH     Mass: 17032     Score: 37     Matches: 2(1)  Sequences: 1(1)
       Plastocyanin major isoform, chloroplastic OS=Arabidopsis thaliana GN=DRT112 PE=1 SV=2
     □ Check to include this hit in error tolerant search

       Query  Observed   Mr(expt)   Mr(calc)     Delta Miss Score Expect Rank Unique Peptide
     ☑   57  862.6835  2585.0286  2585.2034   -0.1748    0  (11)     2.8    1      U  K.NNAGYPHNVVFDEDEIPSGVDVAK.I
     ☑   58  862.6910  2585.0511  2585.2034   -0.1523    0    37  0.0075    1      U  K.NNAGYPHNVVFDEDEIPSGVDVAK.I
```

Peptide View

MS/MS Fragmentation of **TAVNPLLR**
Found in **PSAL_ARATH** in **SwissProt**, Photosystem I reaction center subunit XI, chloroplastic OS=Arabidopsis thaliana GN=PSAL PE=1 SV=2

Match to Query 3: 883.457068 from(442.735810,2+) intensity(3387.0000) index(30)
Title: Sum of 11 scans in range 1289 (rt=1622.46, f=4, i=22) to 1329 (rt=1667.19, f=4, i=32) [\\Qtof\Qtof 17\JAN2004.PRO\Data\6p013-sanjeeva-1.raw]
Data file 6p013-sanjeeva-1.mgf

```
Monoisotopic mass of neutral peptide Mr(calc): 882.5287
Fixed modifications: Carboxymethyl (C) (apply to specified residues or termini only)
Ions Score: 40  Expect: 0.0086
Matches : 13/64 fragment ions using 24 most intense peaks   (help)
```

#	b	b^{++}	b^{*}	b^{*++}	b^{0}	b^{0++}	Seq.	y	y^{++}	y^{*}	y^{*++}	#
1	102.0550	51.5311			84.0444	42.5258	T					8
2	**173.0921**	87.0497			155.0815	78.0444	A	***782.4883***	391.7478	765.4618	383.2345	7
3	**272.1605**	136.5839			254.1499	127.5786	V	***711.4512***	356.2292	694.4246	347.7160	6
4	386.2034	193.6053	369.1769	185.0921	368.1928	184.6001	N	***612.3828***	306.6950	595.3562	298.1817	5
5	483.2562	242.1317	466.2296	233.6185	465.2456	233.1264	P	***498.3398***	249.6736	481.3133	241.1603	4
6	596.3402	298.6738	579.3137	290.1605	578.3297	289.6685	L	***401.2871***	201.1472	384.2605	192.6339	3
7	709.4243	355.2158	692.3978	346.7025	691.4137	346.2105	L	288.2030	144.6051	271.1765	136.0919	2
8							R	***175.1190***	88.0631	158.0924	79.5498	1

Fig. 3.5 MS-MS search results of an in-gel digested protein sample mixture using Mascot. The panels demonstrate (**a**) peptide summary report containing details of significant hits, (**b**) peptide view of a significant hit indicating the peptide sequence information and (**c**) the information regarding residual masses of amino acids deducing the peptide sequence

on the assumption that the frequency of a particular precursor ion getting selected in a large data set is proportional to the abundance of the peptide in a sample. The spectral counts from peptides are averaged and an abundance index of a protein is generated (Liu et al. 2004; Gao et al. 2003; Colinge et al. 2005; Ishihama et al. 2005). The label-free approaches mentioned above facilitate quantification of peptides without using expensive labels and performing additional sample processing steps. They are, hence, becoming the choice of most researchers despite their inherent limitations (Simpson et al. 2009). However, these approaches are still evolving, and are believed to improve greatly in the years to come (Mueller et al. 2008).

3.5 Protein Microarrays

Protein microarrays have been widely accepted as a high-throughput technique to achieve systemic understanding of protein-protein interactions, functional analysis of proteins and autoantibody screening in various systems (Mitchell 2002; MacBeath 2002). This technology essentially relies on proteins immobilized on glass slides, traditionally coated with PVDF (polyvinylidene fluoride), nitrocellulose or polystyrene. The approach has now evolved into incorporating soft lithography techniques to enhance surface chemistry for the immobilization of proteins (Hu et al. 2011). These protein arrays are subjected to a set of probe molecules, and are classified on the basis of the biological question to be answered. For instance, functional protein arrays are arrays where immobilized proteins are subjected to probing by query DNA, RNA, peptides, small molecules, glycans or protein molecules to observe their interaction with ligands on the chip (Phizicky et al. 2003; Hu et al. 2011). Analytical protein arrays are arrays on which ligands like allergens, aptamers, antibodies or antigens are printed to perform protein profiling or clinical diagnostics (Phizicky et al. 2003). Reverse-phase protein microarrays is another popular kind of biomarker validation platform, where a large number of clinical samples such as, biofluids or cell lysates are printed on the chip and probed with antibodies for target biomarkers for large-scale screening in clinical cohorts (Zha et al. 2004; Tibes et al. 2006).

Although there are several types of protein microarrays with varied applications for each, autoantibody screening from biofluids like serum and CSF has been one of the most popular applications for elucidating novel biomarkers in infectious diseases or cancers (Song et al. 2010; Anderson et al. 2011; Hu et al. 2012). Autoantibody production is a response of the immune system against certain aberrant self-proteins, also termed as tumour-associated antigens (TAAs) in case of cancer (Anderson et al. 2011). Biofluids can be screened for the presence of autoantibodies by high-throughput protein microarrays harbouring human peptides or whole proteins (Fig. 3.6a, b). If antibodies are produced against an aberrant protein, they would bind to the antigen and detected using Cy dye-labelled secondary antibodies displaying fluorescent signals. The spot intensity of the protein would indicate the strength of the immunogenic response against a particular protein, which enables relative quantification of the protein. The signal intensities are measured by scanning the chip using a microarray scanner at appropriate channels depending on the absorbance wavelength of the Cy dye employed in the assay (Fig. 3.6a). The data exported is usually in .gpr, .cel or .txt format, depending on the scanner used to generate the output file, which contains information regarding image acquisition and each protein spot (feature). As against tissue biopsy, a highly invasive diagnostic approach, the autoantibody profiling using serum is a minimal invasive approach aiding in early diagnosis of cancer.

DNA microarrays provided the foundation for data analysis strategies for protein microarrays. (Hu et al. 2011; An et al. 2014). Here, we will focus on the generic data analysis approach applicable across all platforms of protein microarrays. The workflow of protein microarrays for screening autoantibodies is very similar to other immunological assays like western blotting and ELISA. However, the staggering difference in microarray throughput as against to these

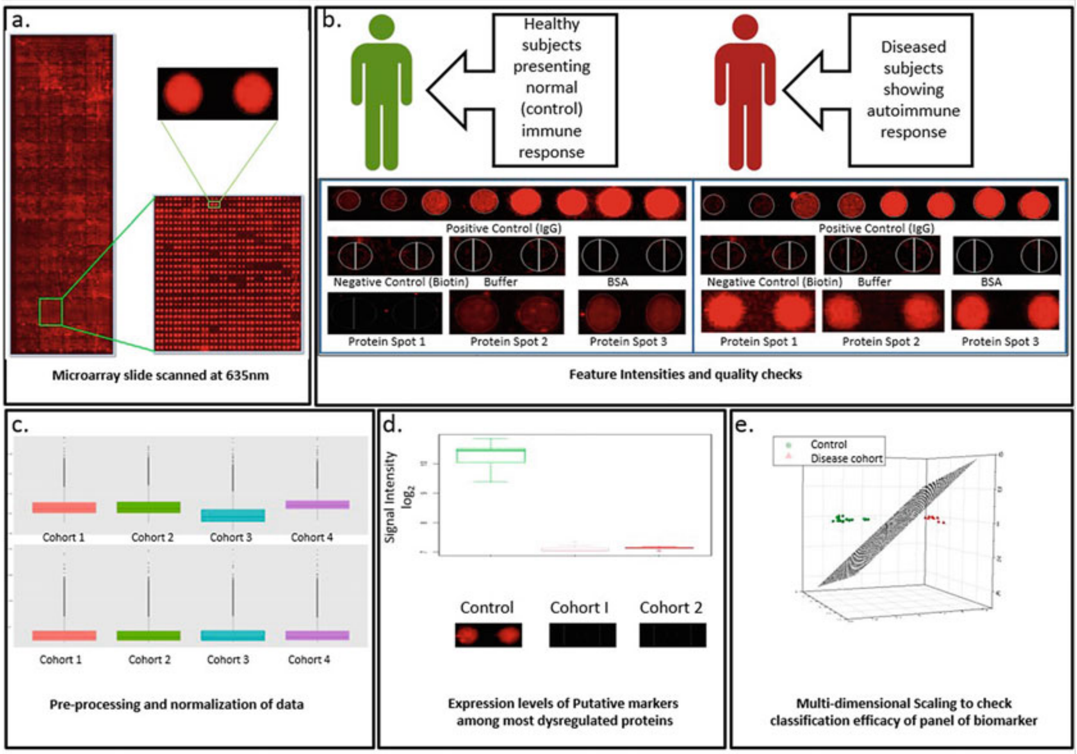

a.
Microarray slide scanned at 635nm
b.
Healthy subjects presenting normal (control) immune response
Diseased subjects showing autoimmune response
Positive Control (IgG)
Negative Control (Biotin)
Buffer
BSA
Protein Spot 1
Protein Spot 2
Protein Spot 3
Feature Intensities and quality checks
c.
Cohort 1
Cohort 2
Cohort 3
Cohort 4
Pre-processing and normalization of data
d.
Signal Intensity log₂
Control
Cohort I
Cohort 2
Expression levels of Putative markers among most dysregulated proteins
e.
Control
Disease cohort
Multi-dimensional Scaling to check classification efficacy of panel of biomarker

Fig. 3.6 An overview of protein microarray data analysis. The figures represent (**a**) a processed microarray slide scanned at 635 nm for Cy5. This results in illumination of features, signal intensities of which are extracted during data acquisition. The autoantibody profile of a healthy against diseased subject is shown in panel (**b**). Quality control features like positive controls and negative controls aid as landmarks in the normalization process. Differentially expressed proteins can visually be observed showing opposite trends in the two cohorts. The visual representation of unnormalized (*top panel*) and normalized data (*bottom panel*) is shown in panel (**c**). The differential expression of a putative protein marker emerging from the data, and its relative fold change across the cohorts is shown in (**d**) upper panel, whereas the bottom panel shows the visual spot intensities. Panel (**e**) represents the classification of subjects in two distinct cohorts based on the group of classifiers deduced using mathematical algorithms

traditional techniques, accompanied with other variables influencing the experimental outcome, makes protein microarray data analysis extremely challenging. Specialized software used for data acquisition from the microarray scanner, statistical models and robust computational support followed by systems biology approaches are the fundamentals of protein microarray data analysis (An et al. 2014). The primary statistical and computational elements of data analysis in a protein microarray experiment can be broadly divided into following stages.

3.5.1 Pre-processing of the Data: Background Correction and Normalization

The use of Cy dyes often results in false positives (noise), in addition to the true positives, due to non-specific binding. The true estimate of a spot intensity is obtained by subtracting the background intensity from the foreground, called background correction. In order to study the effects of different background correction methods, log-normalized foreground and background intensities are plotted for different samples without performing any correction at first. One of the methods (normexp + offset) from LIMMA (linear model for analysing differential expression) model is used to normally distribute the background intensities treating the foreground signal as an exponential distribution while stabilizing any resulting variance (Syed et al. 2015).

The underlying assumption of any microarray experiment is that the majority of proteins display the same expression levels across arrays. In order to study the biological differences, the technical variation that may arise due to dye bias, print-tip effects or day-to-day variations must first be optimized. It is therefore important that the data is normalized for an unbiased analysis (Fig. 3.6c). Some of the commonly used normalization strategies include quantile normalization, variance-stabilizing normalization, cyclic loess and robust-linear-model normalization (An et al. 2014). These are essentially mathematical algorithms aimed at distributing the variance arising in a set of arrays to normalize the signal intensities. Each of these strategies may be used under different set of conditions, depending on the nature of experiment in consultation with the biologists, clinicians and statisticians analysing the data. A comparative analysis of these normalization strategies has been described in one of the previous studies (Kingsmore 2006).

3.5.2 Differentially Expressed Proteins

Protein microarrays are generally used to comprehend the differential protein expression levels across any two cohorts (diseased vs. healthy) (Fig. 3.6d). Determination of differentially expressed proteins between two sets of samples involves statistical tests with a null hypothesis that no gene is differentially expressed. In order to screen for significant biomarkers that could differentiate these cohorts, a robust analysis of protein expression levels is required. The student's *t*-test (assuming normal distribution of data), rank product (non-parametric), Wilcoxon rank-sum test (assuming nonparametric, normal distribution approximation), significance analysis of microarrays (SAM), LIMMA and M-statistic are commonly employed for generating a list of differentially expressed proteins (An et al. 2014). However, since protein microarrays are highly dynamic, statisticians often choose one or a combination of these tools or alternatively develop complementary approaches to improve the stringency of data, especially if there are underlying assumptions of data distribution.

3.5.3 Shortlisting Differentially Regulated Proteins

Correction methods, like Benjamini-Hochberg, shortlist proteins based on their statistical significance providing a *p*-value cut-off. Another way to improve the stringency of shortlisting proteins is to employ a fold-change cut-off further to the corrected *p*-value cut-off. This is a manual way of examining if the data shortlisted qualifies the threshold cut-off, which may be of interest to a biologist. These shortlisted proteins could be studied further or subjected to algorithms like correspondence analysis (CA) (Syed et al. 2015), a model analogous to principal component analysis (PCA), to understand the degree of classification that can be achieved to segregate two cohorts. CA is a dimensionality reduction technique, which helps in narrowing down the long list of differentially expressed proteins. The list from correspondence analysis can be treated as set of markers whose values are associated with classes in a statistically significant manner rather than by mere chance. In order to select a panel of significant classifier proteins differentiating control samples from diseased samples, recursive feature elimination using models like support vector machine can be used (Syed et al. 2015). The efficacy of these models can be visually validated using multidimensional scaling plots (Fig. 3.6e) (Syed et al. 2015). Figure 3.7 describes the basic workflow of such an autoantibody screening experiment along with a general pipeline for protein microarray data analysis.

A biologist could use a systemic approach to understand the dysregulated pathways emerging

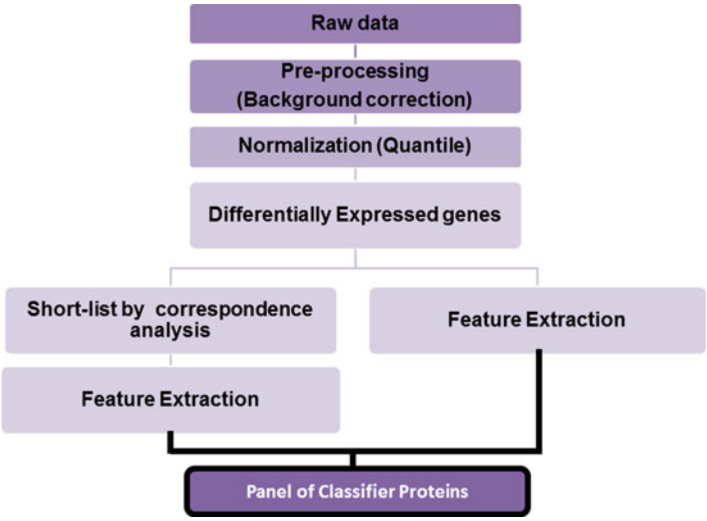

Raw data
Pre-processing (Background correction)
Normalization (Quantile)
Differentially Expressed genes
Short-list by correspondence analysis
Feature Extraction
Feature Extraction
Panel of Classifier Proteins

Fig. 3.7 Schematic representation of the protein microarray data analysis strategy

from the list of significantly dysregulated proteins. In addition to this, protein interaction and metabolic networks, gene ontology and gene set enrichment analysis can be performed to completely understand pathobiology of the disease under study (Syed et al. 2015). Classifier proteins with high fold-change values can be validated in clinical diagnostics. Thus protein microarray platform, with indispensable computational and statistical support for robust data analysis, is a powerful discovery tool for biomarker studies.

3.6 Surface Plasmon Resonance

Surface plasmon resonance (SPR) is an optical method to monitor changes in the refractive index of materials in the near vicinity of the metal surface. It is a phenomenon that occurs when polarized light, under the condition of total internal reflection, strikes an electrically conducting thin metal film at the interface between media of different refractive index: the glass of the sensor chip surface and the sample solution.

As the plane polarized light strikes through a high refractive index prism, the light becomes totally internally reflected and generates an evanescent wave that penetrate the thin metal film. At a certain angle of incidence, the incident light excites surface plasmons (electron charge density waves) on the metallic film. As a result, there is a characteristic absorption of energy via the evanescent wave field, and a drop in the intensity of reflected light at a specific angle known as the resonance angle. These surface plasmon waves are extremely sensitive to the refractive index of the solution within the effective penetration depth of the evanescent field. Interaction of biomolecules produces a change in the refractive index near the metal surface, leading to a shift in the resonance angle, which is monitored in real time by detecting changes in the intensity of the reflected light. The apparent rate constants for the association (K_a) and dissociation (K_d) can be analysed from the rate of change of the SPR signal.

SPR-based biosensors are now routinely used as an established platform for validating biomolecular interactions and performing concentration analysis (Pattnaik 2005; Helmerhorst et al. 2012; Berggård et al. 2007; Boozer et al. 2006; Shah et al. 2015). The technology allows analysis of these interactions in real time with high sensitivity and low sample requirement in a label-free environment. The method is not restricted to the usage of protein-protein interactions, but the generality extends to all kinds of molecules including protein-lipid, protein-RNA and protein-nanoparticle studies (Katsamba et al. 2002; Cedervall et al. 2007; Navratilova et al. 2006). Briefly, one of the interacting molecules (ligand) is bound on a sensor chip surface and the other interacting partner (analyte) is injected over the surface. The amount of analyte bound is continuously monitored as a function of time showing the progress of interaction. This plot of response against time is called sensorgram. The SPR response is proportional to the mass of analyte bound at the sensor surface. The analyte injection is followed by an increase in binding response which enables the determination of rate of complex formation (K_a). As the analyte injection is replaced by buffer flow, the rate of dissociation of the complex (K_d) can be monitored. The complex may not dissociate completely in many cases, wherein regeneration of the surface is performed with mild acidic or basic washing conditions.

3.6.1 SPR Data Processing and Analysis

In SPR, the data is collected continuously over time so that the kinetic parameters can be determined with accuracy and precision. SPR analysis of biomolecular interactions involves crucial experimental design and depends on several experimental factors such as optimum buffers, pH conditions, immobilization chemistry, ligand density, regeneration solutions, flow rate and temperature. In single-cycle kinetics, analytes with increasing concentrations are injected one after the other in a single cycle without the need to regenerate the surface between sample injections. However, in multi-cycle kinetics, different analyte concentrations are run as different cycles which may require regeneration of the surface after every individual cycle, depending on the dissociation pattern of the analyte molecule. Figure 3.8 demonstrates an example of protein interaction with a small drug molecule performed using both multi-cycle kinetics and single-cycle kinetics. The latter approach reduces the experimental time involved and seems aptly suitable for situations where optimum regeneration conditions cannot be achieved. It is suggested to immobilize low amount of ligand for kinetic analysis of macromolecular interactions to achieve surface saturation and avoid mass transport limitation effect and aggregation. Mass transport limitation occurs when the binding rate of analyte to ligand is faster than the diffusion rate of analyte to the ligand surface. Low surface immobilization and high flow rates can minimize this effect allowing better fit of models.

After the data is collected, several processing steps need to be performed before any quantitative information can be extracted. A number of software programs are available for processing SPR data, including BIAevaluation, Scrubber and CLAMP (Morton and Myszka 1998). The initial steps in data processing involve zeroing on the

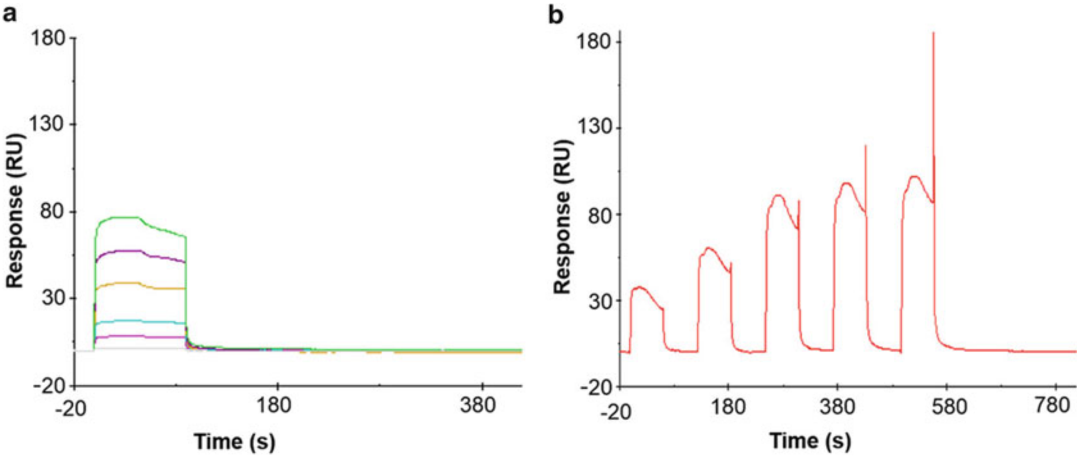


Fig. 3.8 Surface plasmon resonance analysis for a protein-drug interaction study. The figures illustrate protein interaction with a small drug molecule performed using (**a**) multi-cycle kinetics and (**b**) single-cycle kinetics

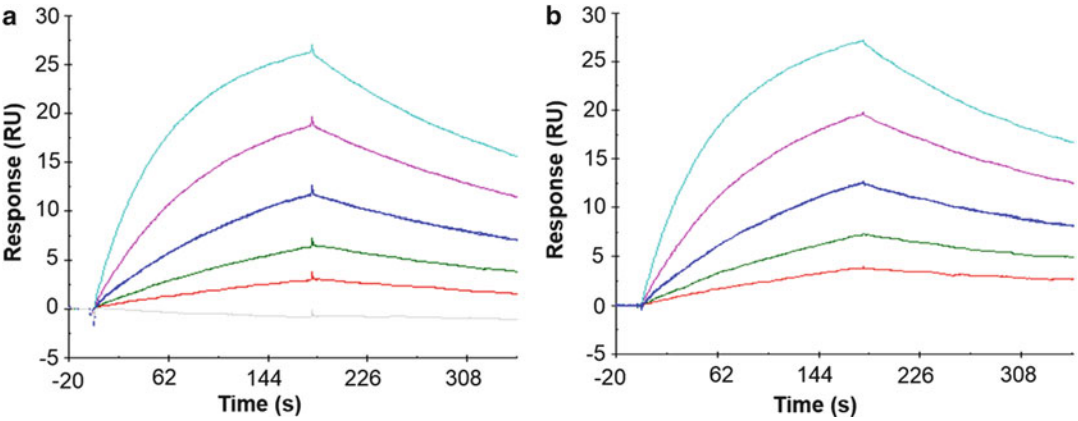


Fig. 3.9 An example illustrating blank subtraction for subtracting bulk effects and checking the specificity. A protein-protein kinetic sensorgram showing subtraction of an ideal baseline response (shown in *grey*) before sample injection from the obtained sample response of different concentrations (shown in different colours). The panel (a) shows the unsubtracted sensorgram, whereas panel (**b**) displays the blank-subtracted sensorgram

x-axis (time) and y-axis (response units). This aligns the beginning of injections with respect to each other and allows the responses observed from each flow channel to be compared with one other. The referencing steps help in minimizing artifacts, and also correcting for any bulk shift resulting from buffer mismatch in sample buffer and running buffer. In the first referencing step, the reference flow cell sensorgram is subtracted from the active flow cell to produce a sensorgram removing bulk shift contributions. In the second referencing step, as exemplified in Fig. 3.9 from a protein-protein interaction study, the effect of buffer injections is nullified by subtracting the baseline response before sample injection from the obtained sample response. These two referencing steps, known as double referencing, remove the systematic shifts and drifts in baseline, frequently observed in sensorgrams (Myszka 1999).

The data now becomes ready for fitting to appropriate models using a mathematical algorithm, and to further determine the kinetic parameters (K_a, K_d and K_D) and characterize the interaction. The chosen analyte concentration range should be wide enough to achieve surface saturation, the highest concentration being approximately five to ten times the $\mathrm{K_D}$ value. Purity of the ligand and analyte, immobilization heterogeneity, mass transfer effects, rebinding of analytes to ligand, buffer mismatch, inappropriate analyte range and complexity of biological systems can greatly influence the fitting of models. There are a number of kinetic models available that can fit the acquired data such as 1:1 Langmuir fit model, heterogeneity model, bivalent analyte model, conformational change model, etc. In general, the simplest model, the 1:1 Langmuir fit model (one ligand molecule interacts with one analyte molecule assuming that the interaction rate is not limited by mass transport) should be tried as the first attempt since most of the biological interactions occur in a 1:1 stoichiometry. There should be a valid justification for the use of other models, the results of which should be confirmed with other supporting experiments. Many factors need to be considered while deciding the correct fit model.

In a global fit, both association and dissociation data, for all analyte samples, are fit at the same time using a sum of squared residuals over every data point. The global approach is more efficient because there are fewer adjustable parameters, whereas a different R_{max} value is calculated for each curve in a local fit. After the fit is made, the curves need to be studied well to understand the accuracy of fit for the association and dissociation phase, and examine if the calculated R_{max} is within the expected range. R_{max} is the maximum feasible response that can be obtained for a specific interaction. The theoretical R_{max} for an interaction can be calculated based on the following formula: $\mathrm{R_{max}}$ = analyteMW/ligandMW

*R$_L$*Sm, where R_{max} is the maximum binding response (RU), R_L is the immobilization level, S_m is the stoichiometric ratio (number of binding sites per ligand) and MW is the molecular weight. In concept, the theoretical R_{max} should be the same for different analyte concentrations injected one after the other, if we do not consider the binding site loss due to harsh regeneration or incomplete regeneration. R_{eq} is the response obtained when binding between ligand and analyte reaches equilibrium. The equilibrium constant, K_D, can be calculated directly using steady-state or equilibrium analysis, where the rate of association equals the rate of dissociation. The response at equilibrium, R_{eq}, is measured over a given range of analyte concentrations, and the values are plotted against those analyte concentrations. The kinetic and equilibrium analyses performed on the same data set should ideally produce similar K_D values, which can reflect the confidence level of the data obtained.

Differences in sample buffer and running buffer result in bulk signal, which does not allow models to fit well to the data (Rich and Myszka 2010). Many times, when the analyte concentrations are not accurately known, the curve fittings using software can be misleading. One such example is demonstrated from a protein-small molecule interaction study (Fig. 3.10a). In such cases, a good fit of the analysed kinetic data can be confirmed by low chi^2 (less than 10 % of R_{max}), which is the average of squared differences between the measured data points and the corresponding fitted values. One chi^2 value, which gives a measure of the closeness of model fit, is determined for all curves fitted simultaneously. Residual plots, generated by some of the software, determine the accuracy of fitting even better than chi^2 values. The residual plot shows the difference in response between each data point for the experimental curves and the calculated curves. The shape and distribution of the residual plot give a better insight on the quality of fit to the chosen model. If there are systematic deviations between the experimental and fitted curves, the plot will indicate them by displacement from the zero line. Figure 3.10b demonstrates an ideal residual plot obtained from an antibody-protein interaction using BIAevaluation software. The guidelines are drawn in green to indicate the range of acceptability. Ideally, the noise level in the plot should be on the order of ±2 RU (Drescher et al. 2009).

3.7 Concluding Remarks

Omics platforms have emerged as powerful tools to help researchers look at biosystems with a global perspective. Innovations in technologies have broadened our existing knowledge, revealing the interplay of various biomolecules at the

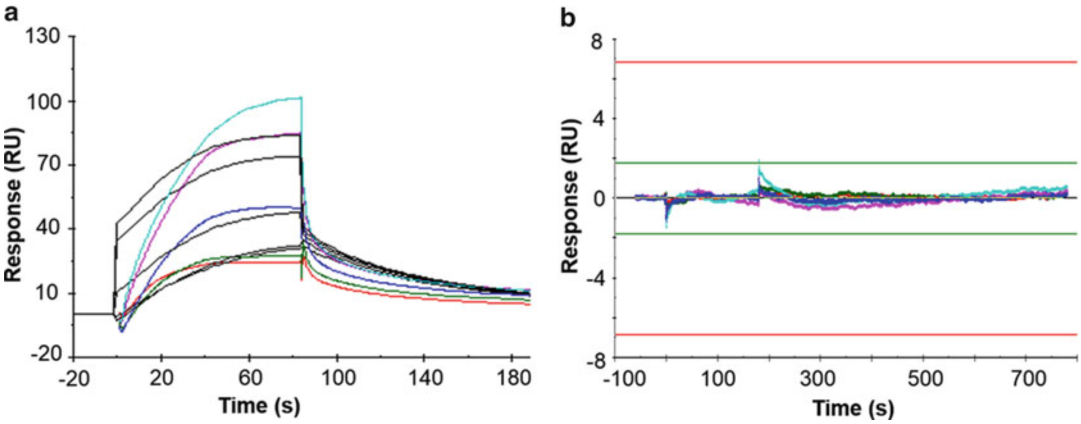


Fig. 3.10 An example illustrating a poor curve fitting and an ideal residual plot. A protein-small molecule interaction study resulting in (**a**) a poor curve fitting. An ideal (**b**) residual plot from an antibody-protein interaction showing differences in response between experimental and calculated curves, demonstrating the quality of fit to the chosen model

systemic level. However, the quantity of data assimilated through these technologies employs new challenges on data processing and analysis. High-throughput techniques aiming at decoding the complexity of biosystems have led to a surge of data, albeit with many downstream hurdles in the form of data storage and data analysis. In the last few decades, huge efforts have been devoted towards creation of database repositories for different data sets where researchers are encouraged to share their data associated with scientific publications. This has enabled researchers around the world to reproduce and validate the studies, as well as analyse the data in innovative ways using different methodologies. With multi-omics technologies routinely being used for various studies, it is important to appraise the challenges of data analysis associated with these sophisticated platforms. Data exploitation requires vital support from sophisticated software and explorative tools, employing statistical methods and visualization aids, to analyse heterogeneous data sets.

Lately, there have been significant advancements in proteomic techniques offering greater sensitivity and rapidity, complementing the traditional methods. Data processing and analysis in proteomics are certainly a complex multistep process. Common proteomic techniques like gel-based approaches, mass spectrometry, protein microarrays and label-free technologies find overlapping applications in multi-omics disciplines. As discussed above, these techniques are often employed for proteome profiling, identification of post-translational modifications, comparative expression analysis of proteins and studying molecular interactions. Accurate and reliable data processing and analysis are the fundamentals of these proteomic approaches to generate factual biological insights. Hence, data processing and analysis of heterogeneous data types is presently an active field of research where biologists and biostatisticians are persistently working together towards improving data utilization in research and discovery.

References

An LTT, Pursiheimo A, Moulder R et al (2014) Statistical analysis of protein microarray data: a case study in type 1 diabetes research. J Proteomics Bioinform S12:003. doi:10.4172/jpb.S12-003

Anderson KS, Sibani S, Wallstrom G et al (2011) Protein microarray signature of autoantibody biomarkers for the early detection of breast cancer. J Proteome Res 10:85–96. doi:10.1021/pr100686b

Beckett P (2012) The basics of 2D DIGE. In: Cramer R, Westermeier R (eds) Difference Gel Electrophoresis (DIGE): methods and protocols. Springer Protocols, New York, pp 9–19

Baggerman G, Vierstraete E, De Loof A, Schoofs L (2005) Gel-based versus gel-free proteomics: a review. Comb Chem High Throughput Screen 8:669–677

Berggård T, Linse S, James P (2007) Methods for the detection and analysis of protein–protein interactions. Proteomics 7:2833–2842. doi:10.1002/pmic.200700131

Boozer C, Kim G, Cong S et al (2006) Looking towards label-free biomolecular interaction analysis in a high-throughput format: a review of new surface plasmon resonance technologies. Curr Opin Biotechnol 17:400–405. doi:10.1016/j.copbio.2006.06.012

Cedervall T, Lynch I, Lindman S et al (2007) Understanding the nanoparticle-protein corona using methods to quantify exchange rates and affinities of proteins for nanoparticles. Proc Natl Acad Sci USA 104:2050–2055. doi:10.1073/pnas.0608582104

Chandramouli K, Qian P-Y (2009) Proteomics: challenges, techniques and possibilities to overcome biological sample complexity. Hum Genomics Proteomics HGP. doi:10.4061/2009/239204

Choe L, Ascenzo MD', Relkin NR et al (2007) 8-plex quantitation of changes in cerebrospinal fluid protein expression in subjects undergoing intravenous immunoglobulin treatment for Alzheimer's disease. Proteomics 7:3651–3660. doi:10.1002/pmic.200700316

Colinge J, Chiappe D, Lagache S et al (2005) Differential proteomics via probabilistic peptide identification scores. Anal Chem 77:596–606. doi:10.1021/ac0488513

Croxatto A, Prod'hom G, Greub G (2012) Applications of MALDI-TOF mass spectrometry in clinical diagnostic microbiology. FEMS Microbiol Rev 36:380–407. doi:10.1111/j.1574-6976.2011.00298.x

de Hoffmann E, Stroobant V (2007) Mass spectrometry: principles and applications. Wiley, New York

Drescher DG, Ramakrishnan NA, Drescher MJ (2009) Surface plasmon resonance (SPR) analysis of binding interactions of proteins in inner-ear sensory epithelia. Methods Mol Biol Clifton NJ 493:323–343. doi:10.1007/978-1-59745-523-7_20

Gao J, Opiteck GJ, Friedrichs MS et al (2003) Changes in the protein expression of yeast as a function of carbon source. J Proteome Res 2:643–649

GE Healthcare 2-D Electrophoresis Principles and Methods (2004) (online) http://www.med.unc.edu/pharm/sondeklab/Lab%20Resources/protein_purification_handbooks/2D%20electrophoresis.pdf

GE Healthcare DeCyder 2D Software, Version 6.5 User Manual (online) https://www.gelifesciences.com/gehcls_images/GELS/Related%20Content/Files/1314750913712/litdoc28401006_20131103235809.pdf

Gehlenborg N, O'Donoghue SI, Baliga NS et al (1998) Visualization of omics data for systems biology. Nat Methods 7:S56–S68. doi:10.1038/nmeth.1436

Geiger T, Wehner A, Schaab C et al (2012) Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. Mol Cell Proteomics 11:M111.014050. doi:10.1074/mcp.M111.014050

Gomez-Cabrero D, Abugessaisa I, Maier D et al (2014) Data integration in the era of omics: current and future challenges. BMC Syst Biol 8:I1. doi:10.1186/1752-0509-8-S2-I1

Gygi SP, Rist B, Gerber SA et al (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nat Biotechnol 17:994–999. doi:10.1038/13690

Hao C, Ma X, Fang S et al (1998) Positive- and negative-ion matrix-assisted laser desorption/ionization mass spectrometry of saccharides. Rapid Commun Mass Spectrom 12:345–348. doi:10.1002/(SICI)1097-0231(19980415)12:7<345::AID-RCM165>3.0.CO;2-B

Helmerhorst E, Chandler DJ, Nussio M, Mamotte CD (2012) Real-time and label-free bio-sensing of molecular interactions by surface plasmon resonance: a laboratory medicine perspective. Clin Biochem Rev Aust Assoc Clin Biochem 33:161–173

Hu S, Xie Z, Qian J et al (2011) Functional protein microarray technology. Wiley Interdiscip Rev Syst Biol Med 3:255–268. doi:10.1002/wsbm.118

Hu C-J, Song G, Huang W et al (2012) Identification of new autoantigens for primary biliary cirrhosis using human proteome microarrays. Mol Cell Proteomics 11:669–680. doi:10.1074/mcp.M111.015529

Ishihama Y, Oda Y, Tabata T et al (2005) Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. Mol Cell Proteomics 4:1265–1272. doi:10.1074/mcp.M500061-MCP200

Christian A Jackson WJS (1997) Application of mass spectrometry to the characterization of polymers. Curr Opin Solid State Mater Sci 661–667. doi: 10.1016/S1359-0286(97)80006-X

Katsamba PS, Park S, Laird-Offringa IA (2002) Kinetic studies of RNA-protein interactions using surface plasmon resonance. Methods San Diego Calif 26:95–104. doi:10.1016/S1046-2023(02)00012-9

Kempka M, Sjödahl J, Björk A, Roeraade J (2004) Improved method for peak picking in matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. Rapid Commun Mass Spectrom 18:1208–1212. doi:10.1002/rcm.1467

Kingsmore SF (2006) Multiplexed protein measurement: technologies and applications of protein and antibody arrays. Nat Rev Drug Discov 5:310–320. doi:10.1038/nrd2006

Liu Q, Krishnapuram B, Pratapa P, et al (2003) Identification of differentially expressed proteins using MALDI-TOF mass spectra. In: Conference record of the thirty-seventh asilomar conference on signals, systems and computers, 2004. pp 1323–1327 vol.2

Liu H, Sadygov RG, Yates JR (2004) A model for random sampling and estimation of relative protein abundance in shotgun proteomics. Anal Chem 76:4193–4201. doi:10.1021/ac0498563

MacBeath G (2002) Protein microarrays and proteomics. Nat Genet 32(Suppl):526–532. doi:10.1038/ng1037

Marvin LF, Roberts MA, Fay LB (2003) Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry in clinical chemistry. Clin Chim Acta Int J Clin Chem 337:11–21

Mitchell P (2002) A perspective on protein microarrays. Nat Biotechnol 20:225–229. doi:10.1038/nbt0302-225

Moghaddas Gholami A, Hahne H, Wu Z et al (2013) Global proteome analysis of the NCI-60 cell line panel. Cell Rep 4:609–620. doi:10.1016/j.celrep.2013.07.018

Morton TA, Myszka DG (1998) Kinetic analysis of macromolecular interactions using surface plasmon resonance biosensors. Methods Enzymol 295:268–294

Mueller LN, Brusniak M-Y, Mani DR, Aebersold R (2008) An assessment of software solutions for the analysis of mass spectrometry based quantitative proteomics data. J Proteome Res 7:51–61. doi:10.1021/pr700758r

Munoz J, Low TY, Kok YJ et al (2011) The quantitative proteomes of human-induced pluripotent stem cells and embryonic stem cells. Mol Syst Biol 7:550. doi:10.1038/msb.2011.84

Myszka DG (1999) Improving biosensor analysis. J Mol Recognit 12:279–284. doi:10.1002/(SICI)1099-1352(199909/10)12:5<279::AID-JMR473>3.0.CO;2-3

Navratilova I, Dioszegi M, Myszka DG (2006) Analyzing ligand and small molecule binding activity of solubilized GPCRs using biosensor technology. Anal Biochem 355:132–139. doi:10.1016/j.ab.2006.04.021

Old WM, Meyer-Arendt K, Aveline-Wolf L et al (2005) Comparison of label-free methods for quantifying human proteins by shotgun proteomics. Mol Cell Proteomics 4:1487–1502. doi:10.1074/mcp.M500084-MCP200

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

Palagi PM, Hernandez P, Walther D, Appel RD (2006) Proteome informatics I: bioinformatics tools for processing experimental data. Proteomics 6:5435–5444. doi:10.1002/pmic.200600273

Palzkill T (2002) Proteomics. Kluwer Academic Publishers, New York

Pattnaik P (2005) Surface plasmon resonance: applications in understanding receptor-ligand interaction. Appl Biochem Biotechnol 126:79–92

Peng J, Elias JE, Thoreen CC et al (2003) Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. J Proteome Res 2:43–50

Pevtsov S, Fedulova I, Mirzaei H et al (2006) Performance evaluation of existing de novo sequencing algorithms. J Proteome Res 5:3018–3028. doi:10.1021/pr060222h

Phanstiel DH, Brumbaugh J, Wenger CD et al (2011) Proteomic and phosphoproteomic comparison of human ES and iPS cells. Nat Methods 8:821–827. doi:10.1038/nmeth.1699

Phizicky E, Bastiaens PIH, Zhu H et al (2003) Protein analysis on a proteomic scale. Nature 422:208–215. doi:10.1038/nature01512

Rich RL, Myszka DG (2010) Grading the commercial optical biosensor literature-class of 2008: “The Mighty Binders”. J Mol Recognit JMR 23:1–64. doi:10.1002/jmr.1004

Shah VG, Ray S, Karlsson R, Srivastava S (2015) Calibration-free concentration analysis of protein biomarkers in human serum using surface plasmon resonance. Talanta 144:801–808. doi:10.1016/j.talanta.2015.06.074

Simpson KL, Whetton AD, Dive C (2009) Quantitative mass spectrometry-based techniques for clinical use: biomarker identification and quantification. J Chromatogr B Anal Technol Biomed Life Sci 877:1240–1249. doi:10.1016/j.jchromb.2008.11.023

Song Q, Liu G, Hu S et al (2010) Novel autoimmune hepatitis-specific autoantigens identified using protein microarray technology. J Proteome Res 9:30–39. doi:10.1021/pr900131e

Syed P, Gupta S, Choudhary S, et al (2015) Autoantibody profiling of gliomas to identify biomarkers using human proteome arrays. Sci Rep (in press)

Tibes R, Qiu Y, Lu Y et al (2006) Reverse phase protein array: validation of a novel proteomic technology and utility for analysis of primary leukemia specimens and hematopoietic stem cells. Mol Cancer Ther 5:2512–2521. doi:10.1158/1535-7163.MCT-06-0334

Unlü M, Morgan ME, Minden JS (1997) Difference gel electrophoresis: a single gel method for detecting changes in protein extracts. Electrophoresis 18(11):2071–2077

Webster J, Oxley D (2012) Protein identification by MALDI-TOF mass spectrometry. In: Zanders ED (ed) Chemical genomics and proteomics. Humana Press, Totowa, pp 227–240

Wilkins MR (2008) Proteome research: concepts, technology and application. Springer, New York

Zha H, Raffeld M, Charboneau L et al (2004) Similarities of prosurvival signals in Bcl-2-positive and Bcl-2-negative follicular lymphomas identified by reverse phase protein microarray. Lab Invest 84:235–244. doi:10.1038/labinvest.3700051