

# Chapter 9

## Processing of Mass Spectrometry Data in Clinical Applications

Dario Di Silvestre, Pietro Brunetti, and Pier Luigi Mauri

**Abstract** Mass spectrometry-based proteomics has become the leading approach for analyzing complex biological samples at a large-scale level. Its importance for clinical applications is more and more increasing, thanks to the development of high-performing instruments which allow the discovery of disease-specific biomarkers and an automated and rapid protein profiling of the analyzed samples. In this scenario, the large-scale production of proteomic data has driven the development of specific bioinformatic tools to assist researchers during the discovery processes. Here, we discuss the main methods, algorithms, and procedures to identify and use biomarkers for clinical and research purposes. In particular, we have been focused on quantitative approaches, the identification of proteotypic peptides, and the classification of samples, using proteomic data. Finally, this chapter is concluded by reporting the integration of experimental data with network datasets, as valuable instrument for identifying alterations that underline the emergence of specific phenotypes. Based on our experience, we show some examples taking into consideration experimental data obtained by multidimensional protein identification technology (MudPIT) approach.

**Keywords** Mass spectrometry-based proteomics • Disease-specific biomarkers • Bioinformatic tools • Algorithms • Integration • Multidimensional protein identification technology

---

D. Di Silvestre • P. Brunetti • P.L. Mauri (✉)  
Proteomics and Metabolomics Laboratory, Institute  
for Biomedical Technologies – National Research Council,  
Via F.lli Cervi 93, 20090 Segrate, Milan, Italy  
e-mail: pierluigi.mauri@itb.cnr.it

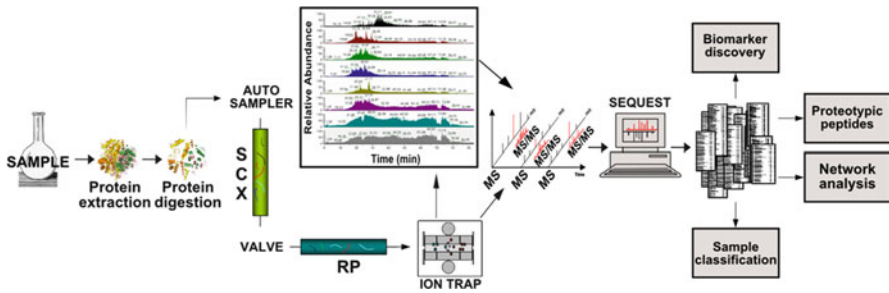
## 9.1 Introduction

The increasing availability of fully sequenced genomes is making the high-throughput proteomics research more and more possible. Developments in fractionation approaches coupled to advances in liquid chromatography (LC), mass spectrometry (MS), and bioinformatic tools have made proteomic approaches mature to analyze complex proteomes, such as *Homo sapiens* (Nilsson et al. 2010). In fact, although proteome complexity prevents the quantitative profiling of all proteins expressed in a cell or tissue at a given time, higher sensitivity, accuracy, and resolution of new MS instruments allow routine analysis, reaching limit of detection of attomole and dynamic range of  $1e^6$  (Yates et al. 2009).

High-throughput proteomics approaches allow to identify and quantify hundreds of proteins per sample, giving a snapshot of cells or tissues associated with different phenotypes. This wealth of data has driven strategies of investigation based on systems biology approaches, allowing insight into disease, taking into consideration functional relationship among proteins (Gstaiger and Aebersold 2009). In addition, highly specific biomarkers represent also key features for improving methods of diagnosis and prognosis or for monitoring disease progression under appropriate therapeutic approaches (Palmlblad et al. 2009; Simpson et al. 2009). In this context, MS has been introduced as a tool for enhancing the current clinical application practices and potentially for targeting the development of personalized medicine (Brambilla et al. 2012).

The ultimate success of MS-based proteomics analysis, both for research and for clinical applications, may be affected by several aspects. Like sample preparation, pre-fractionation methodologies, or instrument setup, data processing procedures represent an important step for obtaining good results and their correct interpretation. Evaluation of thousands of data by hand/eye is time consuming and subjected to biases and missed results. Therefore, to assist researchers during the different stages of analysis and to improve understanding of biological systems, an increasing number of tools and procedures are continually developing, giving rise to a specific bioinformatics area for proteomic applications (Di Silvestre et al. 2011).

In this chapter, we make an overview of the computational trends for processing proteomic data obtained by MS-based proteomics approaches. Based on our experience, we focused primarily on strategies related to multidimensional protein identification technology (MudPIT) approach (Mauri and Scigelova 2009), (Fig. 9.1). In particular, we have explored methods, algorithms, and procedures used for biomarker discovery, by means of label and label-free methods. In this context, we then introduce the main advances of the targeted proteomics (Lange et al. 2008) by investigating the bioinformatics aspects concerning the identification of proteotypic peptides (Craig et al. 2005; Kuster et al. 2005). In the second part of the chapter, we discuss recent advances regarding clinical proteomics application for discriminating sample, such as diseased and healthy. Finally, since most known mechanisms leading to disease involve multiple molecules, we conclude with a discussion of the integration of proteomic data with network datasets, as a promising framework for identifying subnetwork that underlines the emergence of specific phenotypes.



**Fig. 9.1** Multidimensional protein identification technology represents a fully automated technology that simultaneously allows separation of digested peptides, their sequencing, and identification of the corresponding proteins. Peptides are separated by means of strong ion exchange (SCX), using steps of increasing salt concentration, followed by C18 reverse phase (RP) chromatography, using an acetonitrile gradient. Finally, eluted peptides are directly analyzed by MS and raw spectra processed by specific algorithms and bioinformatics tools (see Supplemental Information Table 1). In this way, MudPIT permits simultaneous identification of hundreds, or even thousands, of proteins without limits related to pI, MW, or hydrophobicity. This huge amount of data represents a rich source of information, and their content may be exploited for discovery and classification approaches

## 9.2 Biomarker Discovery

Quantification of proteomic differences between samples at different biological condition, such as healthy and diseased, is a helpful strategy for providing important biological and physiological information concerning disease state (Simpson et al. 2009; Abu-Asab et al. 2011). For this purpose, MS-based approaches are applied for identifying proteins changing their abundance by comparing two or more samples. They consist of different strategies basically belonging to two categories which rely on stable isotope-labeling and label-free methodologies (Domon and Aebersold 2010).

As for labeling approaches, isotopes are introduced in the peptides to create a specific mass tag recognized by MS (Kline and Sussman 2010). Accordingly, quantification is achieved by measuring the ratio of the signal intensities between the unlabeled peptide and its identical counterpart enriched with isotopes (further details on stable isotope-labeling methods are reported in Supplemental Information). Absolute measurements of protein concentration may be achieved with spiked synthetic peptides, as in QconCAT (Mirzaei et al. 2008), AQUA (Gerber et al. 2003), SISCAPA (Anderson et al. 2004), VICAT (Lu et al. 2007), and PC-IDMS (Barnidge et al. 2004). Quantification is obtained by adding into the sample a known amount of an isotopically labeled peptide. In this way, the level of the endogenous form of peptide can be calculated. Of course, the identity of the peptide must be known prior to analysis by MS. Sometimes, if the  $m/z$  ratio of the spiked standard is the same of other peptides, it may lead to an inaccurate quantification. In this case, the ambiguity of the results may be minimized combining these approaches with the selected reaction monitoring (SRM) (Lange et al. 2008).

Although the approaches by labeling, with or without internal standard, allow a highly reproducible and accurate quantification of proteins, most of them have potential limitations, such as the complexity of sample preparation, the requirement of a large amount of time, the requirement of specific bioinformatics tools, and the high cost. As opposite, a simpler alternative concerns label-free approaches (Zhu et al. 2010). They are basically based on counting of peptides identified by means of tandem mass spectrometry (MS/MS) or by evaluating the signal intensity of peptides. Spectral sampling is directly proportional to the relative abundance of the protein in the mixture and therefore represents an attractive methodology, thanks to their intrinsic simplicity, throughput, and low cost.

For these reasons, researchers are increasingly turning to label-free shotgun proteomics approaches (Zhu et al. 2010). Even if they are less accurate, due to the systematic and nonsystematic variations between the experiments, they represent an attractive alternative for their high-throughput setting that also allows the comparison of an unlimited number of experiments with less time consumed. However, efforts should be made to improve experimentally reproducibility and so consequently the reliability of differentially expressed proteins.

A variety of label-free methodologies for semiquantitative evaluation of proteins have been described in literature by reporting a direct relationship between the protein abundance and the sampling parameters associated with identified proteins and peptides (Florens et al. 2002; Gao et al. 2003; Wang et al. 2003; Bridges et al. 2007). One of the most diffused approaches uses the spectral count (SpC) value (Liu et al. 2004) and is based on the empirical observation that more is the quantity of a protein in a sample and more tandem MS spectra may be collected for its peptides. In this context, the normalized spectral abundance factor (NSAF), or its natural log transformation, has been used for the quantitative evaluation with t-test analysis (Zybailov et al. 2006). Other authors have used the protein abundance index (PAI or emPAI) that is calculated by dividing, for each protein, the number of observed peptides by the number of all possible detectable tryptic peptides (Ishihama et al. 2005), while Zhang and colleagues processed SpC values by means of the statistical G-test as previously described (Zhang et al. 2006).

The need to automate the procedure for identifying biomarkers has driven many research groups to develop algorithms and in-house software for identification, visualization, and quantification of mass spectrometry data. Census (Park et al. 2008) and MSQuant (Mortensen et al. 2010) software allows protein quantification by processing MS and MS/MS spectra and they are compatible with label and label-free analysis as well as with high- and low-resolution MS data. In addition, Protein-Quant Suite (Mann et al. 2008) and ProtQuant (Bridges et al. 2007) software are attractive because they allow processing of data in different file formats, therefore collected by different types of mass spectrometers. This aspect focuses attention on the standardization of mass spectrometric data for their sharing and dissemination. In fact, over the years, MS instrument manufacturers have developed proprietary data formats, making it difficult. However, to address this limitation, several tools, such as Trans-Proteomic Pipeline (Deutsch et al. 2010),

allow the conversion of MS data in standard format, like mzData, mzXML, or mzML (Orchard et al. 2010).

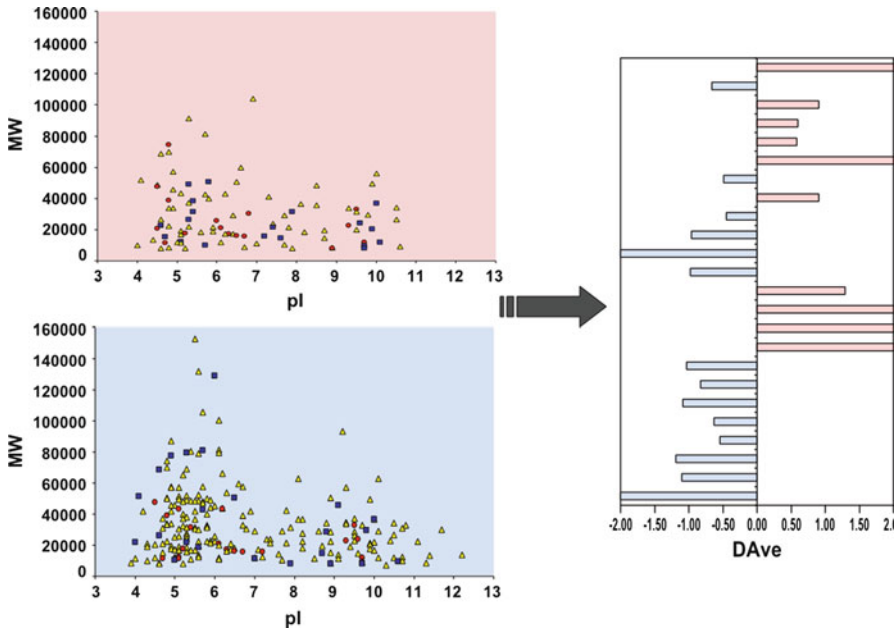
The list of computational tools developed for label-free quantitative analysis, by using LC-MS data, is very long. In addition to Corra (Brusniak et al. 2008) and APEX (Braisted et al. 2008) tools, PatternLab (Carvalho et al. 2008) allows different data normalization strategies, such as Total Signal, log preprocessing (by ln), Z normalization, Maximum Signal, and Row Sigma, for implementing ACFold and nSVM (natural support vector machine) methods to identify protein expression differences.

Based on our experience on proteomic analysis based on MudPIT approach, we developed a simple tool, called MAProMA (Multidimensional Algorithm Protein Map) (Mauri and Dehø 2008). It is based on a label-free quantitative approach based on processing of score/SpC values, by means of Dave and DCI algorithms (see [Supplemental Information](#)). Its effectiveness has been demonstrated in various studies (Mauri et al. 2005; Regonesi et al. 2006; Bergamini et al. 2012; Simioniuc et al. 2011). In addition, MAProMa allows the comparison of up to 125 protein lists and data visualization in a format more comprehensible to biologists (Fig. 9.2).

### 9.3 Proteotypic Peptides

A limitation of shotgun proteomics is due to potential inference problem that may affect protein quantification (Nesvizhskii and Aebersold 2005). In addition, limit of detection “may” exclude the identification of biologically relevant molecules. For identifying, validating, and transferring them to the routine clinical analysis, targeted proteomics or “selected reaction monitoring” (SRM) has been recently developed (Lange et al. 2008; Shipkova et al. 2008; Yang and Lazar 2009). The robustness and the simplicity of its data analysis are ideally suited for detecting and quantifying with high confidence up to 100 proteins per sample. For this purpose, mass spectrometers and bioinformatics tools are set to explore a defined number of proteins of interest, following, for each one, a set of representative peptides with a known  $m/z$  value. They are fragmented, and the monitoring of a specific daughter fragments allow a combination precursor-product, called “transition,” that is highly specific for each amino acid sequence.

These peptides, called proteotypic peptides, describe something typical of a protein. Initially, they were defined as the most observed peptides by the current MS-based proteomics approaches (Craig et al. 2005). Then, other authors added the uniqueness condition for a protein (Kuster et al. 2005), while more recently an empirical definition that defines proteotypic peptide as a peptide observed in more than 50% of all identifications of the corresponding parent protein was appended (Mallick et al. 2007). In other words, these peptides have to be previously identified, with a known MS/MS fragmentation pattern and specific for each targeted protein.



**Fig. 9.2** Virtual 2D MAP tool of MAProMa software allows a rapid evaluation of proteins identified by MudPIT by presenting them in the usual form for biologists (maps). It automatically plots in a virtual 2D map the Mw vs. pI for each protein identified, assigning it a color/shape according to a range of a sampling statistics (score or SpC) derived by SEQUEST data handling. This representation permits to have a rapid visual of the proteins that change comparing two or more conditions. In addition, using DAVE and DCI algorithms, MAProMa reports a histogram that shows the differentially expressed proteins, their identifier, and their DAVE value

The identification of proteotypic peptides useful for targeted proteomics is based on three different methods:

1. By experimental MS/MS data
2. By searching in specific databases
3. By using predictive strategies

The first one is based on the selection of peptides by using the adopted definitions. It allows also the investigation of organisms with proteotypic peptide data not stored inside specific data repositories. In this context, several databases have been developed. In particular:

- Global Proteome Machine Database (Craig et al. 2004) allows users to quickly compare their experimental results with the results previously observed by other scientists. For each dataset, it is possible to view observed spectra for the design of SRM experiments. Query of data may be performed by protein name or Ensembl identifier with the possibility to restrict search to a specific data source, such as eukaryotes, prokaryotes, virus, or precise organism. In addition, further filters may be set by keywords comprising organs, cell location, protein function, or PubMed id.

GPM project is linked to X! Software series (Craig and Beavis 2004) and, of course, with X! P3, the algorithm that makes possible the use of their spectra for profiling proteotypic peptide.

- PeptideAtlas (Desiere 2006) is a publicly accessible source of peptides experimentally identified by tandem mass spectrometry. Raw data, search results, and full builds may be also downloaded. User may browse data, selecting different sources, and few of these need the permission to access. Protein may be searched by different protein identifiers, such as Ensembl and IPI. In addition to general information like GO terms, orthologs, or description, a graphical description indicates the unique peptides found and their occurrence. For each one, it is possible to reach information, like spectra, modification, or genome mapping.

PeptideAtlas is linked to Trans-Proteomic Pipeline (Deutsch et al. 2010) that is used for processing data passed to PeptideAtlas and SBEAMS (Marzolf et al. 2006). In particular, data are processed for deriving the probability of a correct identification and therefore for insuring a high-quality database.

Other databases, designed for data warehousing, store MS/MS spectra collected from proteomics experiments. Even if they are not useful to find proteotypic peptides, they may be used in the comparison with own experimental data.

In particular:

- PRIDE (Martens et al. 2005) stores experiments, identified proteins and peptides, unique peptides, and spectra. In addition to protein (name or various identifiers) and PRIDE experiment identifier, it is possible to browse PRIDE by species, tissue, cell type, GO terms, and disease.
- Proteome Commons (Hill et al. 2010) is a public proteomics database linked to the Tranche (Falkner and Andrews 2007), a powerful open-source web application designed to store and exchange data. A public access to free, open-source proteomics tools, articles, data, and annotations is provided.
- Proteomexchange (Hermjakob and Apweiler 2006) is a work package for encouraging the data exchange and dissemination. Its consortium has been set up to provide a single point of submission of MS data concerning to the main existing proteomics repositories (at the moment PRIDE, PeptideAtlas, and Tranche).

Experimental data stored in the described repositories represent a wealthy source of information, useful for bioinformaticians which attempt to design algorithms for predicting peptides most observable using MS. For this purpose, the STEPP software contains an implementation of a trained support vector machine (SVM) (Cristianini and Shawe-Taylor 2000; Vapnik 1999) that uses a simple descriptor space, based on 35 properties of amino acid, to compute a score representing how proteotypic a peptide is by LC-MS (Webb-Robertson 2009). Similarly to STEPP, a predictor was developed, called Peptide Sieve (Mallick et al. 2007), by studying physicochemical properties of more than 600,000 peptides identified by four different proteomic platforms. This predictor has the ability to accurately

identify proteotypic peptides from any protein sequence and offer starting points for generating a physical model describing the factors that govern elements of proteomic workflows such as digestion, chromatography, ionization, and fragmentation. Other authors, like Tang et al., used neural networks (Riedmiller and Braun 1993) to develop the DetectabilityPredictor software that uses 175 amino acid properties (Tang et al. 2006). In the same way, artificial neural networks were used to predict peptides potentially observable for a given set of experimental, instrumental, and analytical conditions concerning multidimensional protein identification technology datasets (Sanders et al. 2007). Finally, random forest (Breiman 2001) was used to develop enhanced signature peptide (ESP) predictor. It was specifically designed for facilitating the development of targeted MS-based assays for biomarker verification or any application where protein levels need to be measured (Fusaro et al. 2009).

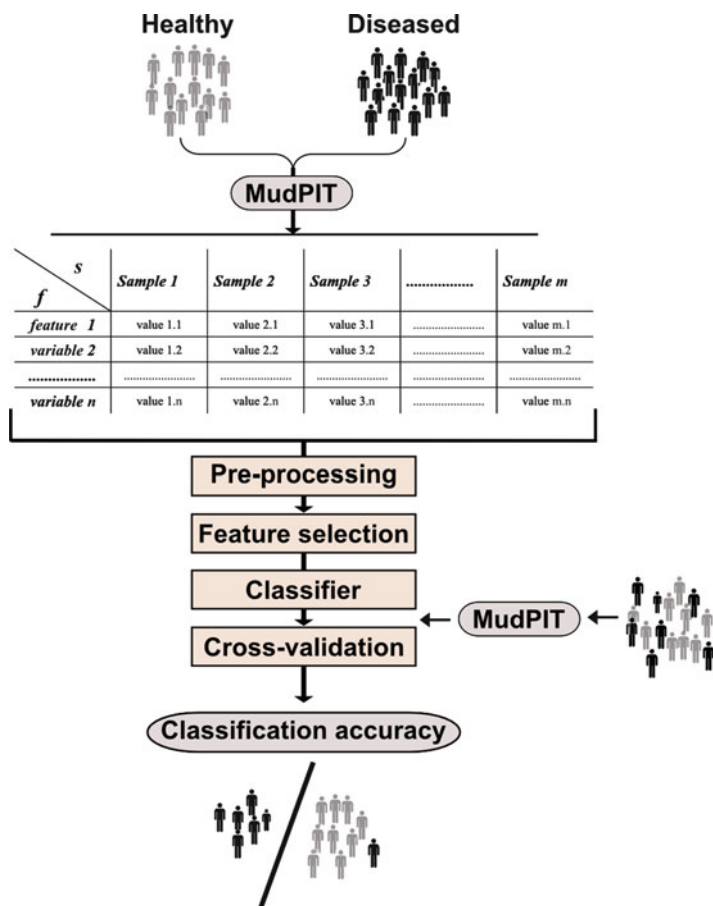
## 9.4 Classification and Clustering Algorithms

Clinical proteomics aims to use relevant data for improving disease diagnosis or for monitoring its progression (Palmlblad et al. 2009; Brambilla et al. 2012). In this context, biomarkers represent a key aspect to develop methods for classifying samples according to their phenotypes (e.g., healthy-diseased, early-late stage).

In addition, to address the biological questions, technologies for high-throughput proteomics allow long lists of spectra, sequenced peptides, and parent proteins that represent a wealthy source of data for identifying predictive biomarkers. For these purposes, most studies have used spectra, generated by MALDI and SELDI technology, in combination with a wide variety of prediction algorithms. On the contrary, fewer cases have taken into consideration data obtained by LC-MS analysis (see Supplemental Information Table 2). However, results of LC-MS analysis (or by MudPIT) are formatted in an  $m \times n$  matrix, with a structure very reminiscent of the output of microarray genomics experiments (Fig. 9.3). Hence, the software packages and the tools useful for analyzing genomics data may easily be used for proteomics (Ressom et al. 2008; Dakna et al. 2009).

Even if some properties of proteomic datasets are related to the analytical technology used to generate them, procedures for sample classification basically consist of four steps, such as data preprocessing, feature selection, classification, and cross-validation (Ressom et al. 2008; Dakna et al. 2009; Sampson et al. 2011; Barla et al. 2008). The first one aims to achieve reproducible results by minimizing errors due to the experimental-designed methodology. Mass spectral profiles may be influenced by several factors, such as baseline effects, shifts in mass-to-charge ratio, alignment problem, or differences in signal intensities that may be corrected by specific computational procedures (Yu et al. 2006; Arneberg et al. 2007; Pluskal et al. 2010). In the same way, variation of sampling parameters





**Fig. 9.3** Data matrix is obtained aligning features identified by analyzing sample by MudPIT approach. In this context, MAProMa software allows a rapid alignment of up to 125 protein lists. Rows in data matrix represent features (e.g., proteins, peptides, or  $m/z$  values) while columns indicate samples. Each cell of data matrix is represented by a value corresponding to parameter associated with features. In particular, spectral count, Xcorrelation (Xcorr), and signal intensity are used for protein, peptides, and  $m/z$  mass features, respectively

associated with sequenced protein, such as spectral count or score, is adjusted using related strategies of data normalization (e.g., Total Signal, log pre-processing (by  $\ln$ ), Z normalization, Maximum Signal, or Row Sigma) (Carvalho et al. 2008).

Typically, MudPIT analysis generates a number of variables usually bigger than the number of analyzed samples ( $f \gg s$ ). This complexity represents a key problem of computational proteomics, and most classification methods require the reducing of the dimensionality prior to classification. It is obtained by

discarding the irrelevant variables for obtaining a combination of features ( $f \ll s$ ), highly correlated and with a more informative lower dimensional space that maximizes the quality of the hypothesis learned from these features (Guyon et al. 2006).

Feature selection procedures may be classified in three different approaches based on different processes to rank features: filter, wrapper, and embedded (Levner 2005). A number of techniques have been used for the analysis of proteomic data, and these include methods such as support vector machines (SVM) and artificial neural networks (ANN) as well as approaches like partial least squares (PLS), principal component regression (PCR), and principal component analysis (PCA). A good overview of statistical and machine learning-based feature selection and pattern classification algorithms is reported by Resson and colleagues (Resson et al. 2008). Of course, different combinations of them show different sensitivity to noisy data and outliers as well as different susceptibility to the over-fitting problem (Sampson et al. 2011).

A limitation of many machine learning-based classification algorithms is that they are not based on a probabilistic model; therefore, there is no confidence associated with the predictions of new datasets. Inadequate performance could be attributed to different reasons (e.g., insufficient or redundant features, inappropriate model classifier, few or too many model parameters, under- or overtraining, and code error, as well as presence of highly nonlinear relationships, noise, and systematic bias). Thus, with the purpose of testing the adequacy/inadequacy of a classifier, after learning is completed, its performances are evaluated through validation set, previously unseen. For this purpose, various methods, such as  $k$ -fold cross-validation, bootstrapping, and holdout methods, have been used (Resson et al. 2008). The most common performance measures to evaluate the performances of classifiers are a confusion matrix and a receiver operating characteristic (ROC) curve. The first one shows information about actual and predicted classifications of a classifier and assesses its performances using standard indices, such as sensitivity, specificity, PPV, NPV, and accuracy values (see [Supplemental Information](#)). On the other hand, ROC is a plot of the sensitivity of a classifier against 1-specificity for multiple decision thresholds.

## 9.5 From Proteomics to Systems Biology

Proteomics is a holistic science that refers to the investigation of the entire systems. Before the advent of -omics technologies, reductionism has dominated the biological research for over a century by investigating individual cellular components. Despite its enormous success, it is more and more evident that most molecular functions occur from a concerted action of multiple molecules, and their investigation implies the examination of an ensemble of elements (Barabási and Oltvai 2004). In fact, biomolecular interactions play a role in the majority of cellular

processes that are regulated connecting numerous constituents, such as DNA, RNA, proteins, and small molecules.

Data abstraction in pathways or networks is the natural result of the desire to rationalize knowledge of complex systems. More recently, their use has changed from purely illustrative to an analytic purpose. In fact, even if it is purely virtual and not related to any intrinsic structure in the cell or organism, understanding how, where, and when single components interact is fundamental to facilitate the investigation of experimental data by taking into consideration the functional relationship among molecules.

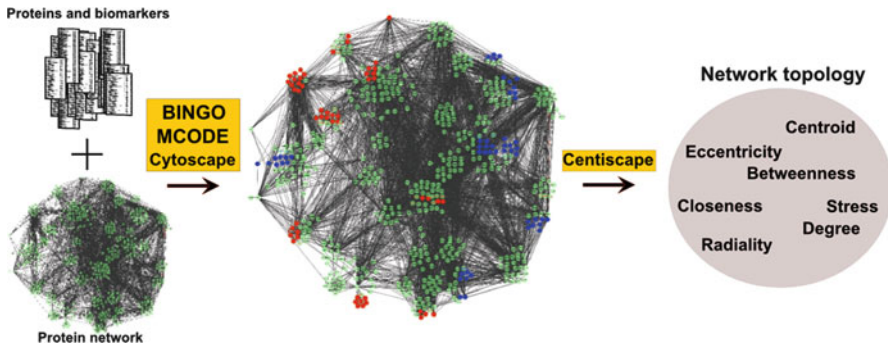
A major challenge for biologists and bioinformaticians is to gain tools, procedures, and skills for integrating data into accurate models that can be used to generate hypotheses for testing. This objective is partially the result of the confluence in systems biology of advances in computer science and -omics technologies. In this context, systems biology approaches have evolved in different strategies basically belonging to two categories, such as computational systems biology, which uses modeling and simulation tools (Barrett et al. 2006; Kim et al. 2012), and data-derived systems biology, which relies on “-omics” datasets (Rho et al. 2008; Li et al. 2009; Jianu et al. 2010; Pflieger et al. 2011).

For deciphering mechanisms of complex and multifactorial diseases, such as those concerning heart failure, recent studies have coupled proteomic and systems biology approaches (Wheelock et al. 2009; Isserlin et al. 2010; Arrell et al. 2011). From a standpoint of the data visualization, the possibility to map protein expression onto pathway or network reveals how they are modulated under different conditions, such as healthy and disease states (Gstaiger and Aebersold 2009; Sodek et al. 2008). About that, an unbiased procedure to identify subnetworks which change consistently between different states involves three key steps:

1. The execution of high-throughput proteomic experiments
2. The identification of candidate biomarkers by label or label-free methods
3. The integration of data into network model to identify clusters of proteins with under, over, and normal expression

In addition, subnetwork selected using experimental data may be analyzed by computing network centrality parameters (Scardoni et al. 2009) for identifying proteins with a relevant biological and topological significance (Fig. 9.4). However, some limitations concerning this kind of approaches could be represented by measurements that cover only a small fraction of the network or by organisms with a limited dataset of cataloged protein sequences and interactions.

To date, in order to visualize and analyze biological networks, a wide set of bioinformatic tools are available (Suderman and Hallett 2007) and include well-known examples, such as Cytoscape (Shannon et al. 2003), VisANT (Hu et al. 2005), Pathway Studio (Nikitin et al. 2003), PATIKA (Demir et al. 2002), Osprey (Breitkreutz et al. 2003), and ProViz (Iragne et al. 2005). Among these, Cytoscape is a Java application whose source code is released under the Lesser General Public License (LGPL). It is probably the most famous open-source software platform for



**Fig. 9.4** By means of Cytoscape software and its plugins, proteins and biomarkers identified by MudPIT are integrated in protein network for identifying pathways or subnetworks that underline the emergence of specific biological states. In addition, networks identified by experimental data may be analyzed by plugins, such as CentiScape, for calculating centrality parameters that indicate nodes with relevant biological and topological significance

visualizing network datasets and biological pathways and for integrating them with annotations or gene and protein expression profiles. Its core distribution provides a basic set of features. However, additional features are available as plugins, thanks to a big community of developers which uses the Cytoscape open API based on Java technology.

Most of the plugins are freely available and concern tasks like the importing and the visualizing of networks from various data formats, the generating of networks from literature searches, and the analysis or the filtering of them by selecting subsets of nodes and/or interactions in relation with topological parameters, GO annotation, or expression levels. In particular, for analyzing large set of proteomics data, we suggest some plugins, such as:

- CentiScape (Scardoni et al. 2009) that computes specific centrality parameters describing the network topology
- MCODE (Bader and Hogue 2003) that finds clusters or highly interconnected regions
- BiNGO (Maere et al. 2005) that determines the Gene Ontology (GO) categories statistically overrepresented in a set of genes or a subgraph of a biological network
- BioNetBuilder (Avila-Campillo et al. 2007) that offers a user-friendly interface to create biological networks integrated from several databases such as BIND (Alfarano et al. 2005), BioGRID (Stark et al. 2006), DIP (Xenarios et al. 2000), HPRD (Mishra et al. 2006), KEGG (Kanehisa et al. 2004), IntAct (Kerrien et al. 2007), MINT (Zanzoni et al. 2002), MPPI (Pagel et al. 2005), and Prolinks (Bowers et al. 2004) as well as interolog networks derived from these sources for all species represented in NCBI HomoloGene

Other important repositories for protein-protein interaction are STRING (von Mering et al. 2007), Reactome (Joshi-Tope et al. 2005), Pathway Commons (Cerami et al. 2011), and WikiPathways (Pico et al. 2008). However, an exhaustive overview of existing databases is available through the Pathguide website (<http://www.pathguide.org/>), a useful web resource where about 300 biological pathways and interaction database are described.

## 9.6 Conclusion

In the last few years, developments in MS instrumentation have increased both the number of identified proteins, reaching hundreds to thousands in a single experiment, and the confidence of such identifications. Thanks to this relevant amount of data, researchers are characterizing the discovery processes by integrating large set of experimental data into models used to generate hypothesis for testing. For this purpose, systems biology approaches provide a powerful strategy for linking biomarker expression with biological processes that can be segmented and linked to disease presentation. Mass spectrometry-based proteomics is emerging also as a powerful approach suitable to face clinical questions. Even if it is an area of still unrealized potential, clinical proteomics offers the promise of diagnosis, prognosis, and therapeutic follow-up of human diseases. However, given the current status of measurement reproducibility and lack of standardization, further comparative investigations are of great importance.

As widely emerged by this chapter, both for basic or clinical research, bioinformatics and statistical tools have a primary importance to support the discovery processes at various levels of sophistication, or for improving the performances of the technologies themselves. In particular, the relevant amount of data produced by the high-throughput proteomics technologies require powerful informatics supports for their organization and interpretation. In this context, several topics concerning data storage, their processing, their visualization, and their interpretation have been faced. However, the need of standards is considered fundamental, and some projects for sharing experimental data between research groups have been launched (e.g., MIAPE, CDISC, and HL7). These should increase meta-analysis, by using raw data from different centers, for helping the development that was grossly underestimated in the initial studies. In addition, as a proteomics community, we believe proteomics methodologies mature for tackling future challenges in clinical proteomics. However, the production of valuable data should rise in step with cooperation with medically focused groups.

**Acknowledgments** This study was supported by the Italian Ministry of Economy and Finance to the CNR for the Project “FaReBio di Qualita,” by Italian Ministry of University and Research for the Project FAR and by Fondazione Cariplo (2010-0653).

## Supplemental Information

### *Introduction*

**Table 1** Bioinformatics platforms for processing proteomics data

Name	Description	References
<b>Corra</b>	Frameworks for LC-MS analysis	<a href="#">Brusniak et al. (2008)</a>
<b>ATAQS</b>	Pipeline implemented for SRM	<a href="#">Brusniak et al. (2011)</a>
<b>Central Proteomics Facilities Pipeline</b>	Pipeline for the analysis of MS/MS proteomic data	<a href="#">Trudgian et al. (2010)</a>
<b>SASHIMI</b>	Suite of tools for MS/MS proteomics	<a href="#">Deutsch et al. (2010)</a>
<b>MS Data Miner</b>	A web-based software that accepts data from Mascot or other software	<a href="#">Dyrlund et al. (2012)</a>
<b>Katsura</b>	Overlays -omics empirical data onto metabolic pathways	<a href="#">Kanehisa et al. (2004)</a>
<b>ProteoWizard</b>	Set of libraries and tools to perform proteomics data analysis	<a href="#">Kessner et al. (2008)</a>
<b>ProteoConnections</b>	Web-based set of tools using for analyzing proteomic data	<a href="#">Courcelles et al. (2011)</a>
<b>Chipster</b>	A Java Web Start framework that organizes workflows for -omics data	<a href="#">Kallio et al. (2011)</a>
<b>Multiplierz</b>	A scriptable framework that access to manufacturer data files	<a href="#">Parikh et al. (2009)</a>
<b>Proteomatic</b>	A framework that permits to create concatenate scripts in pipeline	<a href="#">Specht et al. (2011)</a>
<b>DAnTE/Inferno</b>	Software to perform statistical analysis on proteomics data	<a href="#">Polpitiya et al. (2008)</a>

### *Biomarker Discovery (Stable Isotope Labeling)*

As for labeling approaches, the stable isotopes may be introduced in the peptide using different methods based on the metabolic, chemical, or enzymatic incorporation (Ong and Mann 2005). Metabolic labeling was described for marking protein of yeast by means of  $^{15}\text{N}$ -enriched cell culture medium (Oda et al. 1999). Since the number of labeled nitrogen atoms may vary from peptide to peptide, stable isotope labeling by amino acids in cell culture (SILAC) approach (Ong et al. 2002) was then introduced. In this case, culture medium contains  $^{13}\text{C}_6$ -Lys and  $^{13}\text{C}_6$  and  $^{15}\text{N}_4$ -Arg, ensuring at least one labeled amino acid of tryptic cleavage products. In addition, samples treated with different isotopes may be combined prior to sample preparation minimizing the potential errors introduced by their handling. However, a low cellular growth in adapted media may represent a potential drawback.

Methods which allow labeling by chemical or enzymatic incorporation overcome some limitations associated with metabolic labeling. They include isotope-coded affinity tags (ICAT), where free cysteine residues are tagged by a reagent containing eight or zero deuterium atoms (Gygi et al. 1999). Tags are linked to the biotin which may be exploited for enriching the labeled peptides using affinity purification prior to MS analysis. Although this strategy reduces the complexity of the peptide mixture, proteins that do not contain cysteine are excluded from the analysis. For this reason, other approaches use reactive residues that occur more frequently in proteins. At this class belong the isobaric tags for relative and absolute quantitation (iTRAQ) (Ross et al. 2004), the tandem mass tags (TMT) (Thompson et al. 1999), and the isotope-coded protein label (ICPL) (Schmidt et al. 2005). In particular, iTRAQ approach is widely used and it is based on the covalent labeling of the N-terminus and side chain amines of peptides. Multiplexing tagging allows the analysis of up to 8 samples per experiment (Choe et al. 2007). In fact, samples differently tagged are pooled and usually fractionated by LC and analyzed by MS/MS. However, problem of co-elution of peptides with similar mass could interfere with the quantification. Finally, as for enzymatic tagging of peptides, recently trypsin-catalyzed  $^{18}\text{O}$  labeling has grown in popularity due to its simplicity, its cost, and its ability to universally label peptides. Both C-terminal carboxyl group atoms of tryptic peptides can be enzymatically exchanged with  $^{18}\text{O}$  providing a labeled peptide with a 4-Da mass shift from the  $^{16}\text{O}$ -labeled sample (Qian et al. 2011).

### ***Biomarker Discovery (DAve and DCI Algorithms)***

A direct correlation between the SEQUEST-based score value and the relative abundance of the identified proteins has been previously demonstrated (Mauri et al. 2005; Regonesi et al. 2006). Based on this finding, protein profiles of healthy and diseased samples are semiquantitatively compared using a label-free proteomic approach based on DAve (differential average) and DCI (differential confidence index) algorithms of MAProMA software (Mauri and Dehò 2008).

In particular, DAve, which evaluates changes in protein expression, is defined as

$$(X - Y) / (X + Y) * 0.5,$$

while DCI, which describes the confidence of differential expression, is defined as

$$(X + Y) \times (X - Y) / 2$$

where  $X$  and  $Y$  represent the SEQUEST-based score or SpC values of a given protein in two compared samples.

Conventionally, signs (+/-) of DAve and DCI indicate if proteins are up-regulated in the first or in the second sample, respectively. A value of DAve  $>0.4$  (or  $\leq -0.4$ ) corresponds to SCORE ratio  $\geq 1.5$ . Coupled to a threshold value  $\geq 400$  (or  $\leq -400$ )

for DCI, it allows, with a good reliability, to identify differentially expressed (Mauri et al. 2005; Simioniuc et al. 2011; Bergamini et al. 2012) proteins. However, DAVE and DCI threshold values may be decreased when calculated, considering mean SCORE values derived from replicate analyses (DAVE  $\geq 0.2$  or  $< -0.2$  and DCI  $\geq 200$  and  $\leq -200$ ). On the contrary, when a single analysis per sample/condition is available, a better reliability of the differentially expressed proteins may be assured increasing the threshold values (DAVE  $\geq 0.8$  or  $< -0.8$  and DCI  $\geq 800$  and  $\leq -800$ ).

## Classification and Clustering Algorithms

**Table 2** Classification studies published in the last few years

Sample	References	Technology	Data	Algorithm	Biological condition
Urine	Dawson et al. (2012)	MALDI	Proteins/peptides	SVM	Ischemic stroke
Serum	Timms et al. (2011)	MALDI	–	–	Ovarian cancer
Virus	Wong et al. (2010)	MALDI	Spectra	Bayes classifiers	Influenza viruses
Tissues	Le Faouder et al. (2011)	MALDI-IMS	Protein peaks	SVM	Carcinoma
Tissues	M'Koma et al. (2011)	MALDI-IMS	Protein Peaks	k-nearest-neighbor	Colitis
Tissues	Djidja et al. (2010)	MALDI-IMS	Proteins/peptides	PCA-DA	Tumor
Acinetobacter spp.	Alvarez-Buylla et al. (2012)	MALDI	–	–	Acinetobacter spp.
Serum	Fan et al. (2012)	MALDI	Raw data	SVM	Breast cancer
Plasma	Fassbender et al. (2012)	MALDI	–	SVM	Endometriosis
Cerebrospinal fluid	Ishigami et al. (2012)	MALDI	Spectra	SVM-PCA	Parkinson
Cerebrospinal fluid	Komori et al. (2012)	MALDI	Spectra	SVM-PCA	Multiple sclerosis disorder
Pollen	Krause et al. (2012)	MALDI	Spectra	–	Pollen
Tissue	Meding et al. (2012)	MALDI	Spectra	SVM-RF	Tumor
Serum	Pecks et al. (2012)	MALDI	–	–	Preeclampsia
Bronchoalveolar fluid	Frenzel et al. (2011)	MALDI	Protein peaks	SVM	ALI/ARDS
Urine	Gao et al. (2011)	MALDI	Protein peaks	SVM	–

(continued)



**Table 2** (continued)

Sample	References	Technology	Data	Algorithm	Biological condition
<b>Serum</b>	Han (2010)	MALDI	<b>Spectra profile</b>	<b>SVM</b>	Hepatitis B
<b>Tissue</b>	Waloszczuk et al. (2011)	MALDI	<b>Protein peaks</b>	–	Lung cancer
<b>Urine</b>	Balog et al. (2010)	MALDI	<b>Peptide</b>	<b>SVM</b>	Schistosoma mansoni infection
<b>Tissue</b>	Kim et al. (2012)	MALDI	<b>Spectra</b>	<b>SVM-PCA</b>	Gastric cancer
<b>Bacterial suspensions</b>	Lasch et al. (2010)	MALDI	<b>Spectra</b>	<b>ANN</b>	Yersinia
<b>Tissue</b>	Liao et al. (2010)	MALDI	<b>Spectra</b>	<b>kNN</b>	Colorectal cancer
<b>Serum</b>	Camaggi et al. (2010)	MALDI	–	<b>RF</b>	Hepato carcinoma
<b>Plasma</b>	Lin et al. (2012)	SELDI	<b>Protein peaks</b>	<b>PCA</b>	Lung adenocarcinoma
<b>Serum</b>	Van Gorp et al. (2012)	SELDI	<b>Protein peaks</b>	<b>LS-SVM</b>	Lymph node status in cerebral cancer
<b>Serum</b>	Zhu et al. (2012)	SELDI	–	<b>SVM</b>	Pancreatic cancer
<b>Endometrial samples</b>	Kyama et al. (2011)	SELDI	<b>Protein peaks</b>	<b>SVM</b>	Endometriosis
<b>Serum</b>	Fan et al. (2010)	SELDI	<b>Protein peaks</b>	<b>SVM</b>	Breast cancer
<b>Serum</b>	Liu et al. (2010)	SELDI	<b>Spectra</b>	<b>Decision tree</b>	Tuberculosis
<b>Serum</b>	Tang et al. (2010)	SELDI	<b>Spectra</b>	<b>Kernel PLS models</b>	Ovarian cancer
<b>Tissues</b>	Wang et al. (2010)	SELDI	<b>Spectra</b>	<b>ANN</b>	Endometriosis
<b>Serum</b>	Song et al. (2012)	SELDI and ELISA	<b>Protein peaks</b>	<b>SVM</b>	Biliary atresia
<b>Serum</b>	Ahn et al. (2012)	Multiplex array	–	–	Gastric adenocarcinoma
<b>Plasma</b>	Izbicka et al. (2012)	Multiplex immunoassays	–	<b>SVM</b>	Lung cancer
<b>Tissue</b>	Lazova et al. (2012)	IMS	<b>Spectra</b>	–	Spitzoid malignant melanoma
<b>Serum</b>	Sui et al. (2010)	–	–	<b>Genetic algorithm</b>	Urinemia

*SVM* support virtual machine, *ANN* artificial neural network, *PCA* principal component analysis, *DA* discriminant analysis, *RF* random forest, *PLS* partial least square

## ***Classification and Clustering Algorithms (Sensitivity, Specificity, PPV, NPV, and Accuracy)***

A confusion matrix presents information about actual and predicted classifications made by a classifier. It assesses the classification performance of the classifier. TP, TN, FP, and FN indicate the number of true-positive, true-negative, false-positive, and false-negative samples, respectively. A false positive is when the outcome is incorrectly classified as positive. A false negative is when the outcome is incorrectly classified as negative. True positives and true negatives represent correct classifications. In particular:

$$\text{Sensitivity} = \text{TP}/(\text{TP} + \text{FN})$$

$$\text{Specificity} = \text{TN}/(\text{TN} + \text{FP})$$

$$\text{Positive predictive value} = \text{TP}/(\text{TP} + \text{FP})$$

$$\text{Negative predictive value} = \text{TN}/(\text{TN} + \text{FN})$$

$$\text{Overall classification accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN})$$

## **References**

- Abu-Asab MS, Chaouchi M, Alesci S, Galli S, Laassri M, Cheema AK, Atouf F, VanMeter J, Amri H. Biomarkers in the age of omics: time for a systems biology approach. *OMICS*. 2011;15:105–12.
- Alfarano C, Andrade CE, Anthony K, Bahroos N, Bajec M, Bantoft K, Betel D, Bobechko B, Boutillier K, Burgess E, Buzadzija K, Cavero R, D'Abreo C, Donaldson I, Dorairajoo D, Dumontier MJ, Dumontier MR, Earles V, Farrall R, Feldman H, Garderman E, Gong Y, Gonzaga R, Grytsan V, Gryz E, Gu V, Haldorsen E, Halupa A, Haw R, Hrvojic A, et al. The biomolecular interaction network database and related tools 2005 update. *Nucleic Acids Res*. 2005;33:D418–24.
- Anderson NL, Anderson NG, Haines LR, Hardie DB, Olafson RW, Pearson TW. Mass spectrometric quantitation of peptides and proteins using stable isotope standards and capture by anti-peptide antibodies (SISCAPA). *J Proteome Res*. 2004;3:235–44.
- Arneberg R, Rajalahti T, Flikka K, Berven FS, Kroksveen AC, Berle M, Myhr K-M, Vedeler CA, Ulvik RJ, Kvalheim OM. Pretreatment of mass spectral profiles: application to proteomic data. *Anal Chem*. 2007;79:7014–26.
- Arrell DK, Zlatkovic Lindor J, Yamada S, Terzic A. K(ATP) channel-dependent metaboproteome decoded: systems approach to heart failure prediction, diagnosis, and therapy. *Cardiovasc Res*. 2011;90:258–66.
- Avila-Campillo I, Drew K, Lin J, Reiss DJ, Bonneau R. BioNetBuilder: automatic integration of biological networks. *Bioinformatics*. 2007;23:392–3.
- Bader GD, Hogue CWV. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*. 2003;4:2.
- Barabási A-L, Oltvai ZN. Network biology: understanding the cell's functional organization. *Nat Rev Genet*. 2004;5:101–13.
- Barla A, Jurman G, Riccadonna S, Merler S, Chierici M, Furlanello C. Machine learning methods for predictive proteomics. *Brief Bioinform*. 2008;9:119–28.
- Barnidge DR, Hall GD, Stocker JL, Muddiman DC. Evaluation of a cleavable stable isotope labeled synthetic peptide for absolute protein quantification using LC-MS/MS. *J Proteome Res*. 2004;3:658–61.

- Barrett CL, Kim TY, Kim HU, Palsson BØ, Lee SY. Systems biology as a foundation for genome-scale synthetic biology. *Curr Opin Biotechnol.* 2006;17:488–92.
- Bergamini G, Di Silvestre D, Mauri P, Cigana C, Bragonzi A, De Palma A, Benazzi L, Döring G, Assael BM, Melotti P, Sorio C. MudPIT analysis of released proteins in *Pseudomonas aeruginosa* laboratory and clinical strains in relation to pro-inflammatory effects. *Integr Biol (Camb).* 2012;4:270–9.
- Bowers PM, Pellegrini M, Thompson MJ, Fierro J, Yeates TO, Eisenberg D. Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biol.* 2004;5:R35.
- Braisted JC, Kuntumalla S, Vogel C, Marcotte EM, Rodrigues AR, Wang R, Huang S-T, Ferlanti ES, Saeed AI, Fleischmann RD, Peterson SN, Pieper R. The APEX quantitative proteomics tool: generating protein quantitation estimates from LC-MS/MS proteomics results. *BMC Bioinformatics.* 2008;9:529.
- Brambilla F, Lavatelli F, Di Silvestre D, Valentini V, Rossi R, Palladini G, Obici L, Verga L, Mauri P, Merlini G. Reliable typing of systemic amyloidoses through proteomic analysis of subcutaneous adipose tissue. *Blood.* 2012;119:1844–7.
- Breiman L. Random forests. *Mach Learn.* 2001;45:5–32.
- Breitkreutz B-J, Stark C, Tyers M. Osprey: a network visualization system. *Genome Biol.* 2003;4:R22.
- Bridges SM, Magee GB, Wang N, Williams WP, Burgess SC, Nanduri B. ProtQuant: a tool for the label-free quantification of MudPIT proteomics data. *BMC Bioinformatics.* 2007;8 Suppl 7:S24.
- Brusniak M-Y, Bodenmiller B, Campbell D, Cooke K, Eddes J, Garbutt A, Lau H, Letarte S, Mueller LN, Sharma V, Vitek O, Zhang N, Aebersold R, Watts JD. Corra: computational framework and tools for LC-MS discovery and targeted mass spectrometry-based proteomics. *BMC Bioinformatics.* 2008;9:542.
- Carvalho PC, Fischer JSG, Chen EI, Yates 3rd JR, Barbosa VC. PatternLab for proteomics: a tool for differential shotgun proteomics. *BMC Bioinformatics.* 2008;9:316.
- Cerami EG, Gross BE, Demir E, Rodchenkov I, Babur O, Anwar N, Schultz N, Bader GD, Sander C. Pathway commons, a web resource for biological pathway data. *Nucleic Acids Res.* 2011;39:D685–90.
- Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics.* 2004;20:1466–7.
- Craig R, Cortens JP, Beavis RC. Open source system for analyzing, validating, and storing protein identification data. *J Proteome Res.* 2004;3:1234–42.
- Craig R, Cortens JP, Beavis RC. The use of proteotypic peptide libraries for protein identification. *Rapid Commun Mass Spectrom.* 2005;19:1844–50.
- Cristianini N, Shawe-Taylor J. An introduction to support vector machines and other kernel-based learning methods. Cambridge: Cambridge University Press; 2000.
- Dakna M, He Z, Yu WC, Mischak H, Kolch W. Technical, bioinformatical and statistical aspects of liquid chromatography-mass spectrometry (LC-MS) and capillary electrophoresis-mass spectrometry (CE-MS) based clinical proteomics: a critical assessment. *J Chromatogr B Analyt Technol Biomed Life Sci.* 2009;877:1250–8.
- Demir E, Babur O, Dogrusoz U, Gursoy A, Nisanci G, Cetin-Atalay R, Ozturk M. PATIKA: an integrated visual environment for collaborative construction and analysis of cellular pathways. *Bioinformatics.* 2002;18:996–1003.
- Desiere F. The PeptideAtlas project. *Nucleic Acids Res.* 2006;34:D655–8.
- Deutsch EW, Mendoza L, Shteynberg D, Farrah T, Lam H, Tasman N, Sun Z, Nilsson E, Pratt B, Prazen B, Eng JK, Martin DB, Nesvizhskii AI, Aebersold R. A guided tour of the trans-proteomic pipeline. *Proteomics.* 2010;10:1150–9.
- Di Silvestre D, Daminelli S, Brunetti P, Mauri P. Bioinformatics tools for mass spectrometry-based proteomics analysis. In: Li P, editor. *Reviews in pharmaceutical and biomedical analysis.* Bussum: Bentham Science Publishers; 2011. p. 30–51.
- Domon B, Aebersold R. Options and considerations when selecting a quantitative proteomics strategy. *Nat Biotechnol.* 2010;28:710–21.

- Falkner JA, Andrews PC. P6-T Tranche: secure decentralized data storage for the proteomics community. *J Biomol Technol.* 2007;18:3.
- Florens L, Washburn MP, Raine JD, Anthony RM, Grainger M, Haynes JD, Moch JK, Muster N, Sacci JB, Tabb DL, Witney AA, Wolters D, Wu Y, Gardner MJ, Holder AA, Sinden RE, Yates JR, Carucci DJ. A proteomic view of the *Plasmodium falciparum* life cycle. *Nature.* 2002;419:520–6.
- Fusaro VA, Mani DR, Mesirov JP, Carr SA. Prediction of high-responding peptides for targeted protein assays by mass spectrometry. *Nat Biotechnol.* 2009;27:190–8.
- Gao J, Opitck GJ, Friedrichs MS, Dongre AR, Hefta SA. Changes in the protein expression of yeast as a function of carbon source. *J Proteome Res.* 2003;2:643–9.
- Gerber SA, Rush J, Stemman O, Kirschner MW, Gygi SP. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proc Natl Acad Sci U S A.* 2003;100:6940–5.
- Gstaiger M, Aebersold R. Applying mass spectrometry-based proteomics to genetics, genomics and network biology. *Nat Rev Genet.* 2009;10:617–27.
- Guyon I, Gunn S, Nikravesh M, Zadeh LA. Feature extraction: foundations and applications. Berlin: Springer; 2006.
- Hermjakob H, Apweiler R. The Proteomics Identifications Database (PRIDE) and the ProteomeExchange Consortium: making proteomics data accessible. *Expert Rev Proteomics.* 2006;3:1–3.
- Hill JA, Smith BE, Papoulias PG, Andrews PC. ProteomeCommons.org collaborative annotation and project management resource integrated with the Tranche repository. *J Proteome Res.* 2010;9:2809–11.
- Hu Z, Mellor J, Wu J, Yamada T, Holloway D, Delisi C. VisANT: data-integrating visual framework for biological networks and modules. *Nucleic Acids Res.* 2005;33:W352–7.
- Iragne F, Nikolski M, Mathieu B, Auber D, Sherman D. ProViz: protein interaction visualization and exploration. *Bioinformatics.* 2005;21:272–4.
- Ishihama Y, Oda Y, Tabata T, Sato T, Nagasu T, Rappsilber J, Mann M. 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.* 2005;4:1265–72.
- Isserlin R, Merico D, Alikhani-Koupaei R, Gramolini A, Bader GD, Emili A. Pathway analysis of dilated cardiomyopathy using global proteomic profiling and enrichment maps. *Proteomics.* 2010;10:1316–27.
- Jianu R, Yu K, Cao L, Nguyen V, Salomon AR, Laidlaw DH. Visual integration of quantitative proteomic data, pathways, and protein interactions. *IEEE Trans Vis Comput Graph.* 2010;16:609–20.
- Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, Jassal B, Gopinath GR, Wu GR, Matthews L, Lewis S, Birney E, Stein L. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res.* 2005;33:D428–32.
- Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. *Nucleic Acids Res.* 2004;32:D277–80.
- Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, Dimmer E, Feuermann M, Friedrichsen A, Huntley R, Kohler C, Khadake J, Leroy C, Liban A, Lieftink C, Montecchi-Palazzi L, Orchard S, Risse J, Robbe K, Roechert B, Thorneycroft D, Zhang Y, Apweiler R, Hermjakob H. IntAct—open source resource for molecular interaction data. *Nucleic Acids Res.* 2007;35:D561–5.
- Kim HU, Sohn SB, Lee SY. Metabolic network modeling and simulation for drug targeting and discovery. *Biotechnol J.* 2012;7:330–42.
- Kline KG, Sussman MR. Protein quantitation using isotope-assisted mass spectrometry. *Annu Rev Biophys.* 2010;39:291–308.
- Kuster B, Schirle M, Mallick P, Aebersold R. Scoring proteomes with proteotypic peptide probes. *Nat Rev Mol Cell Biol.* 2005;6:577–83.
- Lange V, Picotti P, Doman B, Aebersold R. Selected reaction monitoring for quantitative proteomics: a tutorial. *Mol Syst Biol.* 2008;4:222.
- Levner I. Feature selection and nearest centroid classification for protein mass spectrometry. *BMC Bioinformatics.* 2005;6:68.

- Li J, Zimmerman LJ, Park B-H, Tabb DL, Liebler DC, Zhang B. Network-assisted protein identification and data interpretation in shotgun proteomics. *Mol Syst Biol.* 2009;5:303.
- Liu H, Sadygov RG, Yates 3rd JR. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal Chem.* 2004;76:4193–201.
- Lu Y, Bottari P, Aebersold R, Turecek F, Gelb MH. Absolute quantification of specific proteins in complex mixtures using visible isotope-coded affinity tags. *Methods Mol Biol.* 2007;359:159–76.
- Maere S, Heymans K, Kuiper M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics.* 2005;21:3448–9.
- Mallick P, Schirle M, Chen SS, Flory MR, Lee H, Martin D, Ranish J, Raught B, Schmitt R, Werner T, Kuster B, Aebersold R. Computational prediction of proteotypic peptides for quantitative proteomics. *Nat Biotechnol.* 2007;25:125–31.
- Mann B, Madera M, Sheng Q, Tang H, Mechref Y, Novotny MV. ProteinQuant suite: a bundle of automated software tools for label-free quantitative proteomics. *Rapid Commun Mass Spectrom.* 2008;22:3823–34.
- Martens L, Hermjakob H, Jones P, Adamski M, Taylor C, States D, Gevaert K, Vandekerckhove J, Apweiler R. PRIDE: the proteomics identifications database. *Proteomics.* 2005;5:3537–45.
- Marzolf B, Deutsch EW, Moss P, Campbell D, Johnson MH, Galitski T. SBEAMS-microarray: database software supporting genomic expression analyses for systems biology. *BMC Bioinformatics.* 2006;7:286.
- Mauri P, Dehò G. A proteomic approach to the analysis of RNA degradosome composition in *Escherichia coli*. *Methods Enzymol.* 2008;447:99–117.
- Mauri P, Scigelova M. Multidimensional protein identification technology for clinical proteomic analysis. *Clin Chem Lab Med.* 2009;47:636–46.
- Mauri P, Scarpa A, Nascimbeni AC, Benazzi L, Parmagnani E, Mafficini A, Peruta MD, Bassi C, Miyazaki K, Sorio C. Identification of proteins released by pancreatic cancer cells by multidimensional protein identification technology: a strategy for identification of novel cancer markers. *FASEB J.* 2005;19:1125–7.
- Mirzaei H, McBee JK, Watts J, Aebersold R. Comparative evaluation of current peptide production platforms used in absolute quantification in proteomics. *Mol Cell Proteomics.* 2008;7:813–23.
- Mishra GR, Suresh M, Kumaran K, Kannabiran N, Suresh S, Bala P, Shivakumar K, Anuradha N, Reddy R, Raghavan TM, Menon S, Hanumanthu G, Gupta M, Upendran S, Gupta S, Mahesh M, Jacob B, Mathew P, Chatterjee P, Arun KS, Sharma S, Chandrika KN, Deshpande N, Palvankar K, Raghavath R, Krishnakanth R, Karathia H, Rekha B, Nayak R, Vishnupriya G, et al. Human protein reference database–2006 update. *Nucleic Acids Res.* 2006;34:D411–14.
- Mortensen P, Gouw JW, Olsen JV, Ong S-E, Rigbolt KTG, Bunkenborg J, Cox J, Foster LJ, Heck AJR, Blagoev B, Andersen JS, Mann M. MSQuant, an open source platform for mass spectrometry-based quantitative proteomics. *J Proteome Res.* 2010;9:393–403.
- Nesvizhskii AI, Aebersold R. Interpretation of shotgun proteomic data the protein inference problem. *Mol Cell Proteomics.* 2005;4:1419–40.
- Nikitin A, Egorov S, Daraselia N, Mazo I. Pathway studio—the analysis and navigation of molecular networks. *Bioinformatics.* 2003;19:2155–7.
- Nilsson T, Mann M, Aebersold R, Yates 3rd JR, Bairoch A, Bergeron JJM. Mass spectrometry in high-throughput proteomics: ready for the big time. *Nat Methods.* 2010;7:681–5.
- Orchard S, Albar J-P, Deutsch EW, Eisenacher M, Binz P-A, Hermjakob H. Implementing data standards: a report on the HUPOPSI workshop September 2009, Toronto, Canada. *Proteomics.* 2010;10:1895–8.
- Pagel P, Kovac S, Oesterheld M, Brauner B, Dunger-Kaltenbach I, Frishman G, Montrone C, Mark P, Stümpflen V, Mewes H-W, Ruepp A, Frishman D. The MIPS mammalian protein-protein interaction database. *Bioinformatics.* 2005;21:832–4.
- Palmblad M, Tiss A, Cramer R. Mass spectrometry in clinical proteomics – from the present to the future. *Proteomics Clin Appl.* 2009;3:6–17.
- Park SK, Venable JD, Xu T, Yates 3rd JR. A quantitative analysis software tool for mass spectrometry-based proteomics. *Nat Methods.* 2008;5:319–22.

- Pflieger D, Gonnet F, de la Fuente van Bentem S, Hirt H, de la Fuente A. Linking the proteins – elucidation of proteome-scale networks using mass spectrometry. *Mass Spectrom Rev.* 2011;30:268–97.
- Pico AR, Kelder T, van Iersel MP, Hanspers K, Conklin BR, Evelo C. WikiPathways: pathway editing for the people. *PLoS Biol.* 2008;6:e184.
- Pluskal T, Castillo S, Villar-Briones A, Oresic M. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics.* 2010;11:395.
- Regonesi ME, Del Favero M, Basilico F, Briani F, Benazzi L, Tortora P, Mauri P, Dehò G. Analysis of the *Escherichia coli* RNA degradosome composition by a proteomic approach. *Biochimie.* 2006a;88:151–61.
- Ressom HW, Varghese RS, Zhang Z, Xuan J, Clarke R. Classification algorithms for phenotype prediction in genomics and proteomics. *Front Biosci.* 2008;13:691–708.
- Rho S, You S, Kim Y, Hwang D. From proteomics toward systems biology: integration of different types of proteomics data into network models. *BMB Rep.* 2008;41:184–93.
- Riedmiller M, Braun H. A direct adaptive method for faster backpropagation learning: the RPROP algorithm. In: *IEEE international conference on neural networks*, 1993, vol. 1. Piscataway: IEEE Service Center; 1993. p. 586–91.
- Sampson DL, Parker TJ, Upton Z, Hurst CP. A comparison of methods for classifying clinical samples based on proteomics data: a case study for statistical and machine learning approaches. *PLoS One.* 2011;6:e24973.
- Sanders WS, Bridges SM, McCarthy FM, Nanduri B, Burgess SC. Prediction of peptides observable by mass spectrometry applied at the experimental set level. *BMC Bioinformatics.* 2007;8 Suppl 7:S23.
- Scardoni G, Petterlini M, Laudanna C. Analyzing biological network parameters with CentiScaPe. *Bioinformatics.* 2009;25:2857–9.
- Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 2003;13:2498–504.
- Shipkova P, Drexler DM, Langish R, Smalley J, Salyan ME, Sanders M. Application of ion trap technology to liquid chromatography/mass spectrometry quantitation of large peptides. *Rapid Commun Mass Spectrom.* 2008;22:1359–66.
- Simioniuc A, Campan M, Lionetti V, Marinelli M, Aquaro GD, Cavallini C, Valente S, Di Silvestre D, Cantoni S, Bernini F, Simi C, Pardini S, Mauri P, Neglia D, Ventura C, Pasquinelli G, Recchia FA. Placental stem cells pre-treated with a hyaluronan mixed ester of butyric and retinoic acid to cure infarcted pig hearts: a multimodal study. *Cardiovasc Res.* 2011;90:546–56.
- Simpson KL, Whetton AD, Dive C. Quantitative mass spectrometry-based techniques for clinical use: biomarker identification and quantification. *J Chromatogr B Analyt Technol Biomed Life Sci.* 2009;877:1240–9.
- Sodek KL, Evangelou AI, Ignatchenko A, Agochiya M, Brown TJ, Ringuette MJ, Jurisica I, Kislinger T. Identification of pathways associated with invasive behavior by ovarian cancer cells using multi-dimensional protein identification technology (MudPIT). *Mol Biosyst.* 2008;4:762–73.
- Stark C, Breitkreutz B-J, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.* 2006;34:D535–9.
- Suderman M, Hallett M. Tools for visually exploring biological networks. *Bioinformatics.* 2007;23:2651–9.
- Tang H, Arnold RJ, Alves P, Xun Z, Clemmer DE, Novotny MV, Reilly JP, Radivojac P. A computational approach toward label-free protein quantification using predicted peptide detectability. *Bioinformatics.* 2006;22:e481–8.
- Vapnik V. *The nature of statistical learning theory.* New York: Springer; 1999.
- von Mering C, Jensen LJ, Kuhn M, Chaffron S, Doerks T, Krüger B, Snel B, Bork P. STRING 7—recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res.* 2007;35:D358–62.

- Wang W, Zhou H, Lin H, Roy S, Shaler TA, Hill LR, Norton S, Kumar P, Anderle M, Becker CH. Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. *Anal Chem*. 2003;75:4818–26.
- Webb-Robertson B-JM. Support vector machines for improved peptide identification from tandem mass spectrometry database search. *Methods Mol Biol*. 2009;492:453–60.
- Wheelock CE, Wheelock AM, Kawashima S, Diez D, Kanehisa M, van Erk M, Kleemann R, Haeggström JZ, Goto S. Systems biology approaches and pathway tools for investigating cardiovascular disease. *Mol Biosyst*. 2009;5:588–602.
- Xenarios I, Rice DW, Salwinski L, Baron MK, Marcotte EM, Eisenberg D. DIP: the database of interacting proteins. *Nucleic Acids Res*. 2000;28:289–91.
- Yang X, Lazar IM. MRM screening/biomarker discovery with linear ion trap MS: a library of human cancer-specific peptides. *BMC Cancer*. 2009;9:96.
- Yates JR, Ruse CI, Nakorchevsky A. Proteomics by mass spectrometry: approaches, advances, and applications. *Annu Rev Biomed Eng*. 2009;11:49–79.
- Yu W, Li X, Liu J, Wu B, Williams KR, Zhao H. Multiple peak alignment in sequential data analysis: a scale-space-based approach. *IEEE/ACM Trans Comput Biol Bioinform*. 2006;3:208–19.
- Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, Cesareni G. MINT: a Molecular INTERaction database. *FEBS Lett*. 2002;513:135–40.
- Zhang B, VerBerkmoes NC, Langston MA, Uberbacher E, Hettich RL, Samatova NF. Detecting differential and correlated protein expression in label-free shotgun proteomics. *J Proteome Res*. 2006;5:2909–18.
- Zhu W, Smith JW, Huang C-M. Mass spectrometry-based label-free quantitative proteomics. *J Biomed Biotechnol*. 2010;2010:840518.
- Zybailov B, Mosley AL, Sardi ME, Coleman MK, Florens L, Washburn MP. Statistical analysis of membrane proteome expression changes in *Saccharomyces cerevisiae*. *J Proteome Res*. 2006;5:2339–47.

## Supplementary Information References

- Ahn HS, Shin YS, Park PJ, Kang KN, Kim Y, Lee H-J, Yang H-K, Kim CW. Serum biomarker panels for the diagnosis of gastric adenocarcinoma. *Br J Cancer*. 2012;106:733–9.
- Alvarez-Buylla A, Culebras E, Picazo JJ. Identification of *Acinetobacter* species: is Bruker biotyper MALDI-TOF mass spectrometry a good alternative to molecular techniques? *Infect Genet Evol*. 2012;12:345–9.
- Balog CIA, Alexandrov T, Derks RJ, Hensbergen PJ, van Dam GJ, Tukahebw EM, Kabatereine NB, Thiele H, Vennervald BJ, Mayboroda OA, Deelder AM. The feasibility of MS and advanced data processing for monitoring *Schistosoma mansoni* infection. *Proteomics Clin Appl*. 2010;4:499–510.
- Bergamini G, Di Silvestre D, Mauri P, Cigana C, Bragonzi A, De Palma A, Benazzi L, Döring G, Assael BM, Melotti P, Sorio C. MudPIT analysis of released proteins in *Pseudomonas aeruginosa* laboratory and clinical strains in relation to pro-inflammatory effects. *Integr Biol (Camb)*. 2012;4:270–9.
- Brusniak M-Y, Bodenmiller B, Campbell D, Cooke K, Eddes J, Garbutt A, Lau H, Letarte S, Mueller LN, Sharma V, Vitek O, Zhang N, Aebersold R, Watts JD. Corra: computational framework and tools for LC-MS discovery and targeted mass spectrometry-based proteomics. *BMC Bioinformatics*. 2008;9:542.
- Brusniak M-YK, Kwok S-T, Christiansen M, Campbell D, Reiter L, Picotti P, Kusebauch U, Ramos H, Deutsch EW, Chen J, Moritz RL, Aebersold R. ATAQS: a computational software tool for high throughput transition optimization and validation for selected reaction monitoring mass spectrometry. *BMC Bioinformatics*. 2011;12:78.

- Camaggi CM, Zavatto E, Gramantieri L, Camaggi V, Strocchi E, Righini R, Merina L, Chieco P, Bolondi L. Serum albumin-bound proteomic signature for early detection and staging of hepatocarcinoma: sample variability and data classification. *Clin Chem Lab Med*. 2010; 48:1319–26.
- Choe L, D'Ascenzo M, Relkin NR, Pappin D, Ross P, Williamson B, Guertin S, Pribil P, Lee KH. 8-plex quantitation of changes in cerebrospinal fluid protein expression in subjects undergoing intravenous immunoglobulin treatment for Alzheimer's disease. *Proteomics*. 2007;7:3651–60.
- Courcelles M, Lemieux S, Voisin L, Meloche S, Thibault P. ProteoConnections: a bioinformatics platform to facilitate proteome and phosphoproteome analyses. *Proteomics*. 2011;11: 2654–71.
- Dawson J, Walters M, Delles C, Mischak H, Mullen W. Urinary proteomics to support diagnosis of stroke. *PLoS One*. 2012;7:e35879.
- Deutsch EW, Shteynberg D, Lam H, Sun Z, Eng JK, Carapito C, von Haller PD, Tasman N, Mendoza L, Farrah T, Aebersold R. Trans-proteomic pipeline supports and improves analysis of electron transfer dissociation data sets. *Proteomics*. 2010;10:1190–5.
- Djidja M-C, Claude E, Snel MF, Francese S, Scriven P, Carolan V, Clench MR. Novel molecular tumour classification using MALDI-mass spectrometry imaging of tissue micro-array. *Anal Bioanal Chem*. 2010;397:587–601.
- Dyrlund TF, Poulsen ET, Scavenius C, Sanggaard KW, Enghild JJ. MS Data Miner: a web-based software tool to analyze, compare and share mass spectrometry protein identifications. *Proteomics*. 2012;12(18):2792–6.
- Fan Y, Wang J, Yang Y, Liu Q, Fan Y, Yu J, Zheng S, Li M, Wang J. Detection and identification of potential biomarkers of breast cancer. *J Cancer Res Clin Oncol*. 2010;136:1243–54.
- Fan N-J, Gao C-F, Zhao G, Wang X-L, Liu Q-Y. Serum peptidome patterns of breast cancer based on magnetic bead separation and mass spectrometry analysis. *Diagn Pathol*. 2012;7:45.
- Fassbender A, Waelkens E, Verbeeck N, Kyama CM, Bokor A, Vodolazkaia A, Van de Plas R, Meuleman C, Peeraer K, Tomassetti C, Gevaert O, Ojeda F, De Moor B, D'Hooghe T. Proteomics analysis of plasma for early diagnosis of endometriosis. *Obstet Gynecol*. 2012;119:276–85.
- Frenzel J, Gessner C, Sandvoss T, Hammerschmidt S, Schellenberger W, Sack U, Eschrich K, Wirtz H. Outcome prediction in pneumonia induced ALI/ARDS by clinical features and peptide patterns of BALF determined by mass spectrometry. *PLoS One*. 2011;6:e25544.
- Gao B-X, Li M-X, Liu X-J, Cai J-F, Fan X-H, Yang X-L, Li X-M, Li X-W. Analyzing urinary proteome patterns of metabolic syndrome patients with early renal injury by magnet bead separation and matrix-assisted laser desorption ionization time-of-flight mass spectrometry. *Zhongguo Yi Xue Ke Xue Yuan Xue Bao*. 2011;33:511–16.
- Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol*. 1999;17:994–9.
- Han H. Nonnegative principal component analysis for mass spectral serum profiles and biomarker discovery. *BMC Bioinformatics*. 2010;11 Suppl 1:S1.
- Ishigami N, Tokuda T, Ikegawa M, Komori M, Kasai T, Kondo T, Matsuyama Y, Nirasawa T, Thiele H, Tashiro K, Nakagawa M. Cerebrospinal fluid proteomic patterns discriminate Parkinson's disease and multiple system atrophy. *Mov Disord*. 2012;27:851–7.
- Izbicka E, Streper RT, Michalek JE, Loudon CL, Diaz 3rd A, Campos DR. Plasma biomarkers distinguish non-small cell lung cancer from asthma and differ in men and women. *Cancer Genomics Proteomics*. 2012;9:27–35.
- Kallio MA, Tuimala JT, Hupponen T, Klemelä P, Gentile M, Scheinin I, Koski M, Käki J, Korpelainen EI. Chipster: user-friendly analysis software for microarray and other high-throughput data. *BMC Genomics*. 2011;12:507.
- Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. *Nucleic Acids Res*. 2004;32:D277–80.
- Kessner D, Chambers M, Burke R, Agus D, Mallick P. ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics*. 2008;24:2534–6.



- Kim HU, Sohn SB, Lee SY. Metabolic network modeling and simulation for drug targeting and discovery. *Biotechnol J*. 2012;7:330–42.
- Komori M, Matsuyama Y, Nirasawa T, Thiele H, Becker M, Alexandrov T, Saida T, Tanaka M, Matsuo H, Tomimoto H, Takahashi R, Tashiro K, Ikegawa M, Kondo T. Proteomic pattern analysis discriminates among multiple sclerosis-related disorders. *Ann Neurol*. 2012;71:614–23.
- Krause B, Seifert S, Panne U, Kneipp J, Weidner SM. Matrix-assisted laser desorption/ionization mass spectrometric investigation of pollen and their classification by multivariate statistics. *Rapid Commun Mass Spectrom*. 2012;26:1032–8.
- Kyama CM, Mihalyi A, Gevaert O, Waelkens E, Simsa P, Van de Plas R, Meuleman C, De Moor B, D’Hooghe TM. Evaluation of endometrial biomarkers for semi-invasive diagnosis of endometriosis. *Fertil Steril*. 2011;95:1338–43.e1–3.
- Lasch P, Drevinek M, Nattermann H, Grunow R, Stämmler M, Dieckmann R, Schwewecke T, Naumann D. Characterization of *Yersinia* using MALDI-TOF mass spectrometry and chemometrics. *Anal Chem*. 2010;82:8464–75.
- Lazova R, Seeley EH, Keenan M, Gueorguieva R, Caprioli RM. Imaging mass spectrometry – a new and promising method to differentiate Spitz nevi from Spitzoid malignant melanomas. *Am J Dermatopathol*. 2012;34:82–90.
- Le Faouder J, Laouirem S, Chapelle M, Albuquerque M, Belghiti J, Degos F, Paradis V, Camadro J-M, Bedossa P. Imaging mass spectrometry provides fingerprints for distinguishing hepatocellular carcinoma from cirrhosis. *J Proteome Res*. 2011;10:3755–65.
- Liao CCL, Ward N, Marsh S, Arulampalam T, Norton JD. Mass spectrometry protein expression profiles in colorectal cancer tissue associated with clinico-pathological features of disease. *BMC Cancer*. 2010;10:410.
- Lin Q, Peng Q, Yao F, Pan X-F, Xiong L-W, Wang Y, Geng J-F, Feng J-X, Han B-H, Bao G-L, Yang Y, Wang X, Jin L, Guo W, Wang J-C. A classification method based on principal components of SELDI spectra to diagnose of lung adenocarcinoma. *PLoS One*. 2012;7:e34457.
- Liu Q, Chen X, Hu C, Zhang R, Yue J, Wu G, Li X, Wu Y, Wen F. Serum protein profiling of smear-positive and smear-negative pulmonary tuberculosis using SELDI-TOF mass spectrometry. *Lung*. 2010;188:15–23.
- M’Koma AE, Seeley EH, Washington MK, Schwartz DA, Muldoon RL, Herline AJ, Wise PE, Caprioli RM. Proteomic profiling of mucosal and submucosal colonic tissues yields protein signatures that differentiate the inflammatory colitides. *Inflamm Bowel Dis*. 2011;17:875–83.
- Mauri P, Dehò G. A proteomic approach to the analysis of RNA degradosome composition in *Escherichia coli*. *Methods Enzymol*. 2008;447:99–117.
- Mauri P, Scarpa A, Nascimbeni AC, Benazzi L, Parmagnani E, Mafficini A, Peruta MD, Bassi C, Miyazaki K, Sorio C. Identification of proteins released by pancreatic cancer cells by multidimensional protein identification technology: a strategy for identification of novel cancer markers. *FASEB J*. 2005;19:1125–7.
- Meding S, Nitsche U, Balluff B, Elsner M, Rauser S, Schöne C, Nipp M, Maak M, Feith M, Ebert MP, Friess H, Langer R, Höfler H, Zitzelsberger H, Rosenberg R, Walch A. Tumor classification of six common cancer types based on proteomic profiling by MALDI imaging. *J Proteome Res*. 2012;11:1996–2003.
- Oda Y, Huang K, Cross FR, Cowburn D, Chait BT. Accurate quantitation of protein expression and site-specific phosphorylation. *Proc Natl Acad Sci U S A*. 1999;96:6591–6.
- Ong S-E, Mann M. Mass spectrometry-based proteomics turns quantitative. *Nat Chem Biol*. 2005;1:252–62.
- Ong S-E, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A, Mann M. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics*. 2002;1:376–86.
- Parikh JR, Askenazi M, Ficarro SB, Cashorali T, Webber JT, Blank NC, Zhang Y, Marto JA. Multiplierz: an extensible API based desktop environment for proteomics data analysis. *BMC Bioinformatics*. 2009;10:364.

- Pecks U, Schütt A, Röwer C, Reimer T, Schmidt M, Preschany S, Stepan H, Rath W, Glocker MO. A mass spectrometric multicenter study supports classification of preeclampsia as heterogeneous disorder. *Hypertens Pregnancy*. 2012;31:278–91.
- Polpitiya AD, Qian W-J, Jaitly N, Petyuk VA, Adkins JN, Camp 2nd DG, Anderson GA, Smith RD. DAnTE: a statistical tool for quantitative analysis of -omics data. *Bioinformatics*. 2008;24:1556–8.
- Qian W-J, Petritis BO, Nicora CD, Smith RD. Trypsin-catalyzed oxygen-18 labeling for quantitative proteomics. *Methods Mol Biol*. 2011;753:43–54.
- Regonesi ME, Del Favero M, Basilico F, Briani F, Benazzi L, Tortora P, Mauri P, Dehò G. Analysis of the *Escherichia coli* RNA degradosome composition by a proteomic approach. *Biochimie*. 2006;88:151–61.
- Ross PL, Huang YN, Marchese JN, Williamson B, Parker K, Hattan S, Khainovski N, Pillai S, Dey S, Daniels S, Purkayastha S, Juhasz P, Martin S, Bartlett-Jones M, He F, Jacobson A, Pappin DJ. Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. *Mol Cell Proteomics*. 2004;3:1154–69.
- Schmidt A, Kellermann J, Lottspeich F. A novel strategy for quantitative proteomics using isotope-coded protein labels. *Proteomics*. 2005;5:4–15.
- Simioniuc A, Campan M, Lionetti V, Marinelli M, Aquaro GD, Cavallini C, Valente S, Di Silvestre D, Cantoni S, Bernini F, Simi C, Pardini S, Mauri P, Neglia D, Ventura C, Pasquinelli G, Recchia FA. Placental stem cells pre-treated with a hyaluronan mixed ester of butyric and retinoic acid to cure infarcted pig hearts: a multimodal study. *Cardiovasc Res*. 2011;90:546–56.
- Song Z, Dong R, Fan Y, Zheng S. Identification of serum protein biomarkers in biliary atresia by mass spectrometry and ELISA. *J Pediatr Gastroenterol Nutr*. 2012;55(4):370–5.
- Specht M, Kuhlger S, Fufezan C, Hippler M. Proteomics to go: proteomatic enables the user-friendly creation of versatile MS/MS data evaluation workflows. *Bioinformatics*. 2011;27:1183–4.
- Sui W, Dai Y, Zhang Y, Chen J, Liu H, Huang H. Proteomic profiling of uremia in serum using magnetic bead-based sample fractionation and MALDI-TOF MS. *Ren Fail*. 2010;32:1153–9.
- Tang K-L, Li T-H, Xiong W-W, Chen K. Ovarian cancer classification based on dimensionality reduction for SELDI-TOF data. *BMC Bioinformatics*. 2010;11:109.
- Thompson D, Pepys MB, Wood SP. The physiological structure of human C-reactive protein and its complex with phosphocholine. *Structure*. 1999;7:169–77.
- Timms JF, Menon U, Devetyarov D, Tiss A, Camuzeaux S, McCurrie K, Nouretdinov I, Burford B, Smith C, Gentry-Maharaj A, Hallett R, Ford J, Luo Z, Vovk V, Gammerman A, Cramer R, Jacobs I. Early detection of ovarian cancer in samples pre-diagnosis using CA125 and MALDI-MS peaks. *Cancer Genomics Proteomics*. 2011;8:289–305.
- Trudgian DC, Thomas B, McGowan SJ, Kessler BM, Salek M, Acuto O. CFP: a central proteomics facilities pipeline. *Bioinformatics*. 2010;26:1131–2.
- Van Gorp T, Cadron I, Daemen A, De Moor B, Waelkens E, Vergote I. Proteomic biomarkers predicting lymph node involvement in serum of cervical cancer patients. Limitations of SELDI-TOF MS. *Proteome Sci*. 2012;10:41.
- Waloszczyk P, Janus T, Alchimowicz J, Grodzki T, Borowiak K. Proteomic patterns analysis with multivariate calculations as a promising tool for prompt differentiation of early stage lung tissue with cancer and unchanged tissue material. *Diagn Pathol*. 2011;6:22.
- Wang L, Zheng W, Ding X, Yu J, Jiang W, Zhang S. Identification biomarkers of eutopic endometrium in endometriosis using artificial neural networks and protein fingerprinting. *Fertil Steril*. 2010;93:2460–2.
- Wong JWH, Schwahn AB, Downard KM. FluTyper—an algorithm for automated typing and subtyping of the influenza virus from high resolution mass spectral data. *BMC Bioinformatics*. 2010;11:266.
- Zhu Y-W, Wang Y-D, Ye Z-Y, Hu X, Yu J-K. Application of serum protein fingerprint in diagnosis of pancreatic cancer. *Zhejiang Da Xue Xue Bao Yi Xue Ban*. 2012;41:289–97.



**Pier Luigi Mauri**, Principal Investigator, Italy

Dr. Mauri has a notable experience in the development and application of several technologies (HPLC, capillary electrophoresis, mass spectrometry) for the identification and quantification of several biomolecules in complex matrices. He has been involved in the study of the bioavailability of several drugs in humans and animals.

He has set up a proteomics laboratory based on both traditional proteomic technologies (2D gel and off-line identification LC-MS/MS) and innovative approaches. In particular, he is among the first Italian researchers to use the MudPIT methodology (2DC-MS/MS) and a parallel computing system for proteomics. The latter instrumentation allows for the proteomic profiling of biological samples, without limitations, according to protein molecular weight, pI, or hydrophobicity, and guarantees high productivity. Using this methodology, secreted proteins from tumor (*FASEB* 2005) and immune system cells have been characterized. Furthermore, direct analysis of biological samples (*JMS* 2007), enzymatic complexes (*Biochimie* 2006), development of novel methodological approaches for studying switch-redox proteins (*JBC* 2005 and 2006, *JMB* 2006), and structural and functional characterization of proteins (*Biol. Chem.* 2004, *FEBS* 2006) have been performed.

In addition, he has developed analytical methods for quantitative characterization of natural bioactive compounds (such as polyphenols and terpenes from Ginkgo biloba and other medicinal plants), and bioavailability studies in human and animals have been performed. He is coauthor of more than 100 scientific publications; he has participated in numerous international congresses and has taught in several metabolomic and proteomic courses.