2

Bioinformatics Applied to Proteomics

Simone Cristoni[1] and Silvia Mazzuca[2]
[1]*Ion Source Biotechnologies srl, Milano,*
[2]*Plant Cell Physiology laboratory, Università della Calabria, Rende, Italy*

1. Introduction

Proteomics is a fundamental science in which many sciences in the world are directing their efforts. The proteins play a key role in the biological function and their studies make possible to understand the mechanisms that occur in many biological events (human or animal diseases, factor that influence plant and bacterial grown). Due to the complexity of the investigation approach that involve various technologies, a high amount of data are produced. In fact, proteomics has known a strong evolution and now we are in a phase of unparalleled growth that is reflected by the amount of data generated from each experiment. That approach has provided, for the first time, unprecedented opportunities to address biology of humans, animals, plants as well as micro-organisms at system level. Bioinformatics applied to proteomics offered the management, data elaboration and integration of these huge amount of data. It is with this philosophy that this chapter was born.

Thus, the role of bioinformatics is fundamental in order to reduce the analysis time and to provide statistically significant results. To process data efficiently, new software packages and algorithms are continuously being developed to improve protein identification, characterization and quantification in terms of high-throughput and statistical accuracy. However, many limitations exist concerning bioinformatic spectral data elaboration. In particular, for the analysis of plant proteins extensive data elaboration is necessary due to the lack of structural information in the proteomic and genomic public databases. The main focus of this chapter is to describe in detail the status of bioinformatics applied to proteomic studies. Moreover, the elaboration strategies and algorithms that have been adopted to overcome the well known limitations of the protein analysis without database structural information are described and disclosed.

This chapter will get rid of light on recent developments in bioinformatic and data-mining approaches, and their limitations when applied to proteomic data sets, in order to reinforce the interdependence between proteomic technologies and bioinformatics tools. Proteomic studies involve the identification as well as qualitative and quantitative comparison of proteins expressed under different conditions, together with description of their properties and functions, usually in a large-scale, high-throughput format. The high dimensionality of data generated from these studies will require the development of improved bioinformatics tools and data-mining approaches for efficient and accurate data analysis of various

biological systems (for reviews see, Li et al, 2009; Matthiesen & Jensen, 2008; Wright et al, 2009). After a rapid moving on the wide theme of the genomic and proteomic sciences, in which bioinformatics find their wider applications for the studies of biological systems, the chapter will focus on mass spectrometry that has become the prominent analytical method for the study of proteins and proteomes in post-genome era. The high volumes of complex spectra and data generated from such experiments represent new challenges for the field of bioinformatics. The past decade has seen an explosion of informatics tools targeted towards the processing, analysis, storage, and integration of mass spectrometry based proteomic data. In this chapter, some of the more recent developments in proteome informatics will be discussed. This includes new tools for predicting the properties of proteins and peptides which can be exploited in experimental proteomic design, and tools for the identification of peptides and proteins from their mass spectra. Similarly, informatics approaches are required for the move towards quantitative proteomics which are also briefly discussed. Finally, the growing number of proteomic data repositories and emerging data standards developed for the field are highlighted. These tools and technologies point the way towards the next phase of experimental proteomic and informatics challenges that the proteomics community will face.
The majority of the chapter is devoted to the description of bioinformatics technologies (hardware and data management and applications) with particular emphasis on the bioinformatics improvements that have made possible to obtain significant results in the study of proteomics. Particular attention is focused on the emerging statistic semantic, network learning technologies and data sharing that is the essential core of system biology data elaboration.
Finally, many examples of bioinformatics applied to biological systems are distributed along the different section of the chapter so to lead the reader to completely fill and understand the benefits of bioinformatics applied to system biology.

2. Genomics versus proteomics

There have been two major diversification paths appeared in the development of bioinformatics in terms of project concepts and organization, the -omics and the bio-. These two historically reflect the general trend of modern biology. One is to go into molecular level resolution. As one of the -omics and bio- proponents, the -omics trend is one of the most important conceptual revolutions in science. Genetic, microbiology, mycology and agriculture became effectively molecular biology since 1970s. At the same time, these fields are now absorbing omics approach to understand their problems more as complex systems. Omics is a general term for a broad discipline of science and engineering for analyzing the interactions of biological information objects in various omes. These include genome, proteome, metabolome, expressome, and interactome. The main focus is on mapping information objects such as genes, proteins, and ligands finding interaction relationships among the objects, engineering the networks and objects to understand and manipulate the regulatory mechanisms and integrating various omes and omics subfields.
This was often done by researchers who have taken up the large scale data analysis and holistic way of solving bio-problems. However, the flood of such -omics trends did not occur until late 1990s. Until that time, it was by a relatively small number of informatics advanced people in Europe and the USA. They included Medical Research Council [MRC] Cambridge, Sanger centre, European Bioinformatics Institute [EBI], European Molecular

Biology Laboratory [EMBL], Harvard, Stanford and others. We could clearly see some people took up the underlying idea of -ome(s) and -omics quickly, as biology was heading for a more holistic approach in understanding the mechanism of life. Whether the suffix is linguistically correct or not, the -omics suffix changed in the way many biologists view their research activity. The most profound one is that biologists became freshly aware of the fact that biology is an information science more than they have thought before.

In general terms, genomics is the -omics science that deals with the discovery and noting of all the sequences in the entire genome of a particular organism. The genome can be defined as the complete set of genes inside a cell. Genomics, is, therefore, the study of the genetic make-up of organisms. Determining the genomic sequence, however, is only the beginning of genomics. Once this is done, the genomic sequence is used to study the function of the numerous genes (functional genomics), to compare the genes in one organism with those of another (comparative genomics), or to generate the 3-D structure of one or more proteins from each protein family, thus offering clues to their function (structural genomics). At today a list of sequenced eukaryotic genomes contains all the eukaryotes known to have publicly available complete nuclear and organelle genome sequences that have been assembled, annotated and published. Starting from the first eukaryote organism *Saccharomyces cerevisiae* to have its genome completely sequenced at 1998, further genomes from 131 eukaryotic organisms were released at today. Among them 33 are Protists, 16 are Higher plants, 26 are Fungi, 17 are Mammals Humans included, 9 are non-mammal animals ,10 are Insects, 4 Nematodes, remaining 11 genomes are from other animals and as we write this chapter, others are still to be sequenced and will be published during the editing of this book. A special note should be paid to the efforts of several research teams around the world for the sequencing of more than 284 different Eubacteria, whose numbers increased by 2-3% if we consider the sequencing of different strains for a single species; also a list of sequenced archaeal genomes contains 28 Archeobacteria known to have available complete genome sequences that have been assembled, annotated and deposited in public databases.

A striking example of the power of this kind of -omics and knowledge that it reveals is that the full sequencing of the human genome has dramatically accelerated biomedical research and diagnosis forecast; very recently Eric S. Lander (2011) explored its impact, in the decade since its publication, on our understanding of the biological functions encoded in the human genome, on the biological basis of inherited diseases and cancer, and on the evolution and history of the human species; also he foresaw the road ahead in fulfilling the promise of genomics for medicine.

In the other side of living kingdoms, genomics and biotechnology are also the modern tools for understanding plant behavior at the various biological and environmental levels. In The Arabidopsis Information Resource [TAIR] a continuously updated database of genetic and molecular biology data for the model higher plant *Arabidopsis thaliana* is maintained (TAIR Database, 2009)

This data available from TAIR include the complete genome sequence along with gene structure, gene product information, metabolism, gene expression, DNA and seed stocks, genome maps, genetic and physical markers, publications, and information about the *Arabidopsis* research community. Gene product function data is updated every two weeks from the latest published research literature and community data submissions. Gene structures are updated 1-2 times per year using computational and manual methods as well as community submissions of new and updated genes.

Genomics provides also boosting to classical plant breeding techniques, well summarized in the Plants for the Future technology platform (http://www.epsoweb.eu/catalog/tp/tpcom_home.htm). A selection of novel technologies come out that are now permitting researchers to identify the genetic background of crop improvement, explicitly the genes that contribute to the improved productivity and quality of modern crop varieties. The genetic modification (GM) of plants is not the only technology in the toolbox of modern plant biotechnologies. Application of these technologies will substantially improve plant breeding, farming and food processing. In particular, the new technologies will enhance the ability to improve crops further and, not only will make them more traceable, but also will enable different varieties to exist side by side, enhancing the consumer's freedom to choose between conventional, organic and GM food. In these contexts agronomical important genes may be identified and targeted to produce more nourishing and safe food; proteomics can provide information on the expression of transgenic proteins and their interactions within the cellular metabolism that affects the quality, healthy and safety of food. Taking advantage of the genetic diversity of plants will not only give consumers a wider choice of food, but it will also expand the range of plant derived products, including novel forms of pharmaceuticals, biodegradable plastics, bio-energy, paper, and more. In this view, plant genomics and biotechnology could potentially transform agriculture into a more knowledge-based business to address a number of socio-economic challenges.

In systems biology (evolutionary and/or functionally) a central challenge of genomics is to identify genes underlying important traits and describe the fitness consequences of variation at these loci (Stinchcombe et al., 2008). We do not intend to give a comprehensive overview of all available methods and technical advances potentially useful for identifying functional DNA polymorphisms, but rather we explore briefly some of promising recent developments of genomic tools from which proteomics taken its rise during the last twenty years, applicable also to non model organisms.

The genome scan, became one of the most promising molecular genetics (Oetjen et al., 2010). Genome scans use a large number of molecular markers coupled with statistical tests in order to identify genetic loci influenced by selection (Stinchombe & Hoekstra, 2008). This approach is based on the concept of 'genetic hitch-hiking' (Maynard Smith & Haigh, 1974) that predicts that when neutral molecular markers are physically linked to functionally important and polymorphic genes, divergent selection acting on such genes also affects the flanking neutral variation. By genotyping large numbers of markers in sets of individuals taken from one or more populations or species, it is possible to identify genomic regions or 'outlier loci' that exhibit patterns of variation that deviate from the rest of the genome due to the effects of selection or treats (Vasemägi & Primmer 2005). An efficient way of increasing the reliability of genome scans, which does not depend on the information of the genomic location of the markers, is to exploit polymorphisms tightly linked to the coding sequences, such as expressed sequence tag (EST) linked microsatellites (Vigouroux et al., 2002; Vasemägi et al., 2005). Because simple repeat sequences can serve as promoter binding sites, some microsatellite polymorphisms directly upstream of genes may have a direct functional significance (Li et al., 2004).

EST libraries represent sequence collections of all mRNA (converted into complementary or cDNA) that is transcribed at a given point in time in a specific tissue (Bouck & Vision, 2007). EST libraries have been constructed and are currently being analyzed for many species whose genomes are not completed. EST library also provide the sequence data for

expression analysis using *Quantitative real-time PCR (QPCR),* as well as for transcription profiling using *microarrays* and, finally, the EST database can be a valuable tool for identifying new candidate polymorphism in proteins of specific interest. QPCR is a method that can measure the abundance of mRNA (converted in cDNA) of specific genes (Heid et al., 1996). The expression of a target gene can be related to the total RNA input, or it can be quantified in relation to the expression of a reference gene, the housekeeping gene (HKG, i.e. gene always expressed at the same level). Unfortunately, a universal reference gene that is expressed uniformly, in all biological conditions in all tissues, does not exist. For each experimental setup using QPCR, the choice of HKG must reflect the tissue used and the experimental treatment.

While QPCR can only handle a few candidate genes, *microrrays technology* quantifies the expression level of hundreds to thousands of genes simultaneously, providing a powerful approach for the analysis of global transcriptional response (Yauk & Berndt, 2007). For example, the analysis of mRNA via genomic arrays is one approach to finding the genes differentially expressed across two kind of tissue or sample obtained under two experimental conditions or to finding the genes that matter to organisms undergoing environmental stress. Additionally, microarray data can be used to distinguish between neutral and adaptive evolutionary processes affecting gene expression (e.g. Gibson, 2002; Feder & Walser, 2005; Whitehead & Crawford, 2006). Nevertheless, a sequencing revolution is currently driven by new technologies, collectively referred to as either 'next-generation' sequencing, 'highthroughput' sequencing, 'ultra-deep' sequencing or 'massively parallel' sequencing. These technologies allow us the large scale generation of ESTs efficiently and cost-effectively available at the National Centre Biotechnology Information database [NCBI-dbEST] (http://www.ncbi.nlm.nih.gov/dbEST); Shendure et al., 2005). There are increasing studies in which 454 technologies, combined or not with Solexa/Illumina, are used to characterize transcriptomes in several plant and animal species (Emrich et al., 2007; Metzker, 2010; Eveland et al., 2008; Bellin et al., 2009). To give an idea of the potential implications of these sequencing technologies it is enough to know that the pyrosequencing delivers the microbial genome sequence in 1 hour, thus upsetting perspectives in basic research, phylogenetic analysis, diagnostics as in industrial applications (Clarke, 2005; Hamady et al., 2010; Yang et al., 2010; Claesson et al., 2009). Even in full sequenced genomes, such as in *Arabidopsis* or humans, this deep sequencing is allowing to identify new transcripts not present in previous ESTs collections (Weber et al., 2007; Sultan et al., 2010). Also specific transcriptomes are being generated in species for which previous genomic resources lacked because of the large size of their genomes (Alagna et al., 2009; Wang et al., 2009; Craft et al., 2010) The new transcripts are also being used for microarrays design (Bellin et al.,2009), and also for high throughput SSRs or SNPs identification. SNP detection is performed by aligning raw reads from different genotypes to a reference genome or transcriptome previously available in plants (Barbazuk et al., 2006), as in plants, (Trick et al., 2009; Guo et al., 2010), animals (Satkoski et al., 2008) and humans (Nilsson et al., 2004).

De novo assembly of raw sequences coming from a set of genotypes, followed by pairwise comparison of the overlapping assembled reads has also successfully used in species lacking any significant genomic or transcriptomic resources (Novaes et al., 2008). The principle behind these applications (as termed sequence census methods) is simple: complex DNA or RNA samples are directly sequenced to determine their content, without the requirement

for DNA cloning. Thus, these novel technologies allow the direct and cost-effective sequencing of complex samples at unprecedented scale and speed, making feasible to sequence not only genomes, but also entire transcriptomes expressed under different conditions. Moreover, a unique feature of sequence census technologies is their ability to identify, without prior knowledge, spliced transcripts by detecting the presence of sequence reads spanning exon-exon junctions. Hence, next-generation sequencing delivers much more information at affordable costs, which will increasingly supersede microarray based approaches (Marguerat et al., 2008).

It is noteworthy, however, that transcription profiling has been questioned as an effective tool for the discovery of genes that are functionally important and display variable expression (e.g. Feder & Walser, 2005). In fact, the vast majority of genes implicated by transcriptomics can be expected to have no phenotype. Furthermore, even if the synthesis of mature protein is closely linked to the abundance of its corresponding mRNA, the concentration of mature protein is the net of its synthesis and degradation. Degradation mechanisms and rates can vary substantially and lead to corresponding variation in protein abundance (Feder & Walser, 2005). The physiological measurements of protein abundance for selected gene candidate could be a valuable addition to pure transcriptomic studies (Jovanovic et al., 2010).

It is reasonable that a method should measure the most relevant output of gene expression, namely dependent changes in protein amounts from potential target genes. Moreover, to be worthwhile, the method should be easy to use, fast, sensitive, reproducible, quantitative and scalable, as several hundred proteins have to be tested. A technique that promises to fulfill most of those criteria is proteomics which is experiencing considerable progress after the massive sequencing of many genomes from yeast to humans for both basic biology and clinical research (Tyers & Mann, 2003). For identifying and understanding the proteins and their functions from a cell to a whole organism, proteomics is a necessity in the assortment of –omics technologies.

Historically, the term *proteome* was coined by Mark Wilkins first in 1994 as a blend of proteins and genome and Wilkins used it to describe the entire complement of proteins expressed by a genome, cell, tissue or organism. Subsequently this term has been specified to contain all the expressed proteins at a given time point under defined conditions and it has been applied to several different types of biological systems (Doyle, 2011; Ioannidis, 2010; Heazlewood, 2011; Prokopi & Mayr, 2011; Wienkoop et al, 2010).

In a basic view, a cellular proteome is the collection of proteins found in a particular cell type under a particular set of conditions such as differentiation stage, exposure to hormone stimulation inside tissues or changing of physical parameters in an environment. It can also be useful to consider an organism's complete proteome, which can be conceptualized as the complete set of proteins from all of the various cellular proteomes. This is very roughly the protein equivalent of the genome. The term "proteome" has also been used to refer to the collection of proteins in certain sub-cellular biological systems. For example, all of the proteins in a virus can be called a viral proteome. The proteome is larger than the genome, especially in eukaryotes, in the sense that there are more proteins than genes. This is due to alternative splicing of genes and post-translational modifications like glycosylation or phosphorylation. Moreover the proteome has at least two levels of complexity lacking in the genome. When the genome is defined by the sequence of nucleotides, the proteome cannot

be limited to the sum of the sequences of the proteins present. Knowledge of the proteome requires knowledge of the structure of the proteins in the proteome and the functional interaction between the proteins.

The escalating sequencing of genomes and the development of large EST databases have provided genomic bases to explore the diversity, cellular evolution and adaption ability of organisms. However, by themselves, these data are of limited use when they try to fully understand processes such as development, physiology and environmental adaptation. Taking advances from genomic information the proteomics can assign function to proteins and elucidate the related metabolism in which the proteins act (Costenoble et al., 2011; Chen et al., 2010; Joyard et al., 2010, Tweedie-Cullen & Mansuy, 2010).

In a wide-ranging functional view, proteomics is matching to genomics: through the use of pure genome sequences, open reading frames (ORFs) can be predicted, but they cannot be used to determine if or when transcription takes place. Proteomics, indicating at what level a protein is expressed, can also provide information about the conditions under which a protein might be expressed, its cellular location (Agrawal et al., 2010; Jamet et al., 2006; Rossignol et al., 2006; Tyers & Mann, 2003), the relative quantities (Yao et al., 2001; Molloy et al., 2005), and what protein–protein interactions take place (Giot et al., 2003; Schweitzer et al., 2003). Genomics, in essence, demonstrates which genes are involved, whereas proteomics can show clearer relationships by illustrating functional similarities and phenotypic variances.

Because the environments in which organisms live is dynamic, the success of a species depends on its ability to rapidly adapt to varying limiting factors such as light (for plants above all), temperature, diet or nutrient sources. Since the proteome of each living cell is dynamic, proteomics allows investigators to clarify if and to what extent various pathways are utilized under varying conditions, triggered by the action of the environment on the system, and relative protein-level response times. In other words how organisms are able to biochemically survive to conditions imposed by environment.

Huge amount of data have been accumulated and organized in world-wide web sites served for proteomics as main proteomics-related web sites have been lunched (Tab 1).

For example The Human Protein Reference Database represents a centralized platform to visually illustrate and integrate information pertaining to domain architecture, post-translational modifications, interaction networks and disease association for each protein in the human proteome; on the ExPASy Proteomics site, tools are available locally to the server or are developed and hosted on other servers.

As concerning plant proteomics, the research community is well served by a number of online proteomics resources that hold an abundance of functional information. Recently, members of the *Arabidopsis* proteomics community involved in developing many of these resources decided to develop a summary aggregation portal that is capable of retrieving proteomics data from a series of online resources (Joshi et al., 2010, http://gator.masc-proteomics.org/). This means that information is always up to date and displays the latest datasets. The site also provides hyperlinks back to the source information hosted at each of the curated databases to facilitate analysis of the primary data. Deep analyses have also performed on organelle proteomics as in protists, animals and plants. A well-known database, launched in 2004, is devoted to proteomics of mitochondria in yeast (Ohlmeier et al., 2004; http://www.biochem.oulu.fi/proteomics/ymp.html), while the Nuclear Protein Database [NPD] is a curated database that contains information on more than 1300

World-wide Web Sites served for Proteomics		
Name	Web site	Characteristics
WORLD-2DPAGE Index to federated 2-D PAGE database	http://www.expasy.ch/ch2d/2d-index.htm	integrated proteome database for use in cancer research by two-dimensional difference gel electrophoresis (2D-DIGE)
2D GEL DATABASES WORLD-WIDE (GeMDBJ Database Link Station)	https://gemdbj.nibio.go.jp/dqdb/diqe/servlet/DiqeLinkTo2dDatabaseCountryServlet	
ExPASy Proteomics tools	http://www.expasy.ch/tools/	Protein identification and characterization with peptide mass fingerprinting data
Mascot Search	http://www.matrixscience.com/search_form_select.html	search engine which uses mass spectrometry data to identify proteins from primary sequence databases
ProteinProspector	http://prospector.ucsf.edu/prospector/mshome.htm	Proteomics tools for mining sequence databases in conjunction with Mass Spectrometry experiments.
MASCP Gator	http://gator.masc-proteomics.org/ MASCP Gator	the portal provides hyperlinks back to the source information hosted at each of the curated databases
PPDB	http://ppdb.tc.cornell.edu/ The Plant Proteome Database	database dedicated to the whole plant proteome
AT_Chloro	http://www.grenoble.prabi.fr/at_chloro/ AT_CHLORO Database	database dedicated to the chloroplast proteome from Arabidopsis thaliana
Main Proteomics-related Web Sites		
The Japanese Electrophoresis Society Home Page	http://www.jes1950.jp/english/	it promotes the development of electrophoretic technologies and their applications
DNA Data Bank of Japan (DDBJ)	http://www.ddbj.nig.ac.jp/	is the sole nucleotide sequence data bank in Asia, which is officially certified to collect nucleotide sequences from researchers
HUPO (Human Proteome Organisation)	http://www.hupo.org/	international scientific organization representing and promoting proteomics through international cooperation and collaborations
Electrophoresis (Wiley-VCH)	http://www.wiley-vch.de/publish/en/journals/alphabeticIndex/2027/	is one of the world's leading journals for new analytical and preparative methods and for innovative applications on all aspects of electrophoresis
UniProtKB/Swiss-Prot Protein Knowledgebase	http://us.expasy.org/sprot/relnotes/spwrnew.html	provide comprehensive and non-redundant complete proteome sets for all species that are currently covered

Table 1. List of the main proteome databases.

vertebrate proteins that are thought, or are known, to localize to the cell nucleus. The database can be accessed at http://npd.hgu.mrc.ac.uk and is updated monthly. Very recently, plant organelle proteomics has experienced a rapid growth in the field of functional proteomics (see the review, Agrawal et al., 2010); from this efforts gave rise seven main websites of which two are devoted to the plastid (Plant Proteomic DataBase [PPDB] http://ppdb.tc.cornell.edu/), two are specific for mitochondria (Arabidopsis Mitochondrial Protein DataBase [AMPDB], http://plantenergy.uwa.edu.au/application/ampdb/; Arabidopsis Mitochondrial Protein Project [AMPP], http://gartenbau.unihannover.de/genetic/AMPP), one is an accurate database of comprehensive chloroplast proteome (AT_Chloro, http://www.grenoble.prabi.fr/protehome/grenoble-plant-proteomics/).

An area of study within proteomics is 'expression proteomics', which is defined as the use of quantitative protein-level measurements of gene expression to characterize biological processes and deduce the mechanisms of gene expression control. Expression proteomics allows researchers to obtain a quantitative description of protein expression and its changes under the influence of biological perturbations, the occurrence of post-translational modifications and the distribution of specific proteins within cells (Baginsky et al., 2010; Roth et al., 2010).

As an example of high technological potential of expression proteomics, in the last ten years plant proteomics research has been conducted in several land species achieving a high degree of knowledge of the dynamics of the proteome in many model plants (Agrawal & Rakwal, 2005; Baerenfaller et al., 2008; Grimplet et al., 2009; Komatsu, 2008; Plomion et al., 2006) and thereafter translating this knowledge in other species whose genome sequence is still under construction. The most successful studies are those which use separation of subcellular compartments (Haynes & Roberts, 2007; Dunkley et al., 2006; Agrawal et al., 2010) such as mitochondria (Heazlewood et al., 2005), chloroplast (Ferro et al., 2010), endoplasmic reticulum (Maltman et al., 2007), peroxisomes (Fukao et al., 2002), plastoglobules (Grennan, 2008), vacuoles (Jaquinod et al., 2007), nucleus (Repetto et al., 2008) since they contain a limited number of proteins thus helping the protein identification.

Since 30 years, the greater part of research into the plant proteome has utilized two-dimensional sodium dodecyl sulphate–polyacrylamide gel electrophoresis (2D SDS–PAGE) for the protein separation step, which is usually followed by protein identification by mass spectrometry (MS). Proteomics, the study of the proteome, has largely been practiced through the separation of proteins by two dimensional gel electrophoresis. In the first dimension, the proteins are separated by isoelectric focusing, which resolves proteins on the basis of charge. In the second dimension, proteins are separated by molecular weight using SDS-PAGE. The gel is dyed to visualize the proteins and the spots on the gel are proteins that have migrated to specific locations.

The number of spots resolved in plant proteomics 2D projects depends on the chosen tissue and plant species as well as the protein nature (i.e. basic or acid, soluble or membrane-associated; Tsugita & Kamo, 1994; Porubleva et al., 2001). The gel plugs, containing the proteins of interest are collected to further analyses by mass MS approaches and database searches (Chevalier, 2010; Yates *et al*., 2009; Zhao &d Lin, 2010). This method where proteins are analyzed after enzymatic digestion is widely used for high complexity samples in large scale analyses and it is known as "bottom up approach" that was discussed in detail in the next paragraph. Attention must given to the importance of sound statistical treatment of the resultant quantifications in the search for differential expression. Despite wide availability of

proteomics software, a number of challenges have yet to be overcome regarding algorithm accuracy, objectivity and automation, generally due to deterministic spot-centric approaches that discard information early in the pipeline, propagating errors. We review recent advances in signal and image analysis algorithms in 2-DE, MS, LC/MS and Imaging MS.

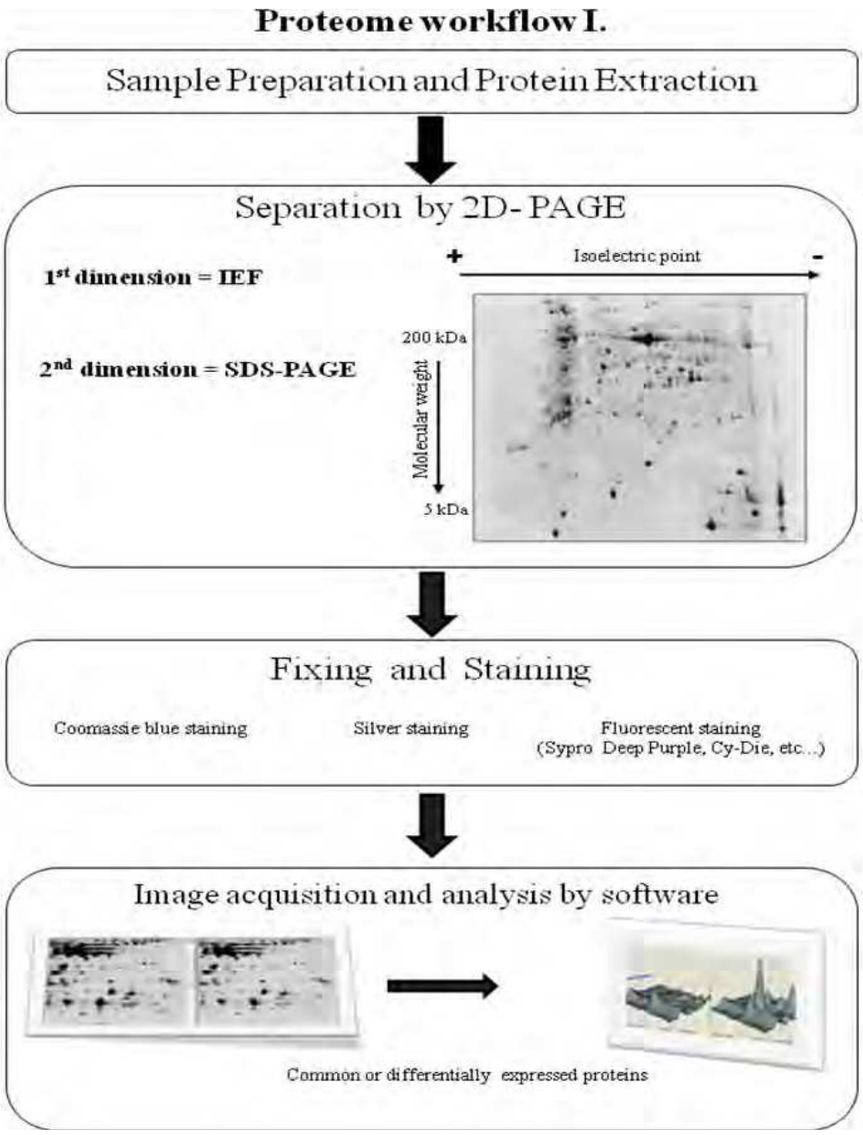


Fig. 1. Proteome workflow I: after sample preparation and protein extraction, proteins are initially separated by isoelectric focusing (IEF) in which they migrate along an IEF strip which has a pH gradient between a cathode and an anode; the migration of each protein ends when it reaches its isoelectric point in the gradient. This strip is then applied to a SDS polyacrylamide gel in which the second dimension of the separation occurs according to molecular weights. After fixation, the gel is stained by different techniques and its digital image is acquired to be further analyzed by specific softwares, in order to found the significant differentially expressed proteins.
With permission of Nova Science Publishers, Inc.

3. Bioinformatics in proteomics

Mass spectrometry became a very important tool in proteomics: it has made rapid progresses as an analytical technique, particularly over the last decade, with many new types of hardware being introduced (Molloy et al, 2005; Matthiesen and Jensen, 2008, Yates et al, 2009). Moreover, constant improvements have increased the levels of MS sensitivity, selectivity as well as mass measurement accuracy. The principles of mass spectrometry can be envisaged by the following four functions of the mass spectrometer: i) peptide ionization; ii) peptide ions analyses according to their mass/charge ratio (*m/z*) values ; iii) acquisition of ion mass data ; iv) measurement of relative ion abundance. Ionization is fundamental as the physics of MS relies upon the molecule of interest being charged, resulting in the formation of positive ions, and, depending on the ionization method, fragment ions. These ion species are visualized according to their corresponding *m/z* ratio(s), and their masses assigned. Finally, the measurement of relative ion abundance, based on either peak height or peak area of sample(s) and internal standard(s), leads to a semi-quantitative request.

3.1 Typical procedure for proteome analysis

Proteome data elaboration procedure is different depending of the study target. In general the studies can be qualitative in order to characterize the organisms expressed proteome and quantitative to detect potential biomarker related to disease or other organism proprieties. The principal proteomics studies are:

i. Full proteomics (qualitative);
ii. Functional proteomics (relative quantitation studies);
iii. Post translational modification functional proteomics (qualitative and relative quantitation studies)

3.2 Data elaboration for full proteome analysis

In full proteomics analysis (Armengaud et al. 2010) the proteins are usually extracted and qualitatively identified. These studies are usually performed in order to understand what proteins are expressed by the genome of the organism of interest. The general analytical scheme is reported in Figure 2.

Basically, after protein separation, mainly through gel electrophoresis or other separation approaches (liquid chromatography etc.), proteins are identified by means of mass spectrometric technique. Two kind of data processing algorithms can be employed depending by the analytical technology used to analyze the proteins. The two approaches are:

i. Bottom up approach. It is used to identify the protein of interest after enzymatic or chemical digestion;
ii. Top down approach. In this case proteins are not digested but directly analyzed by mass spectrometric approaches;

In the former case (bottom up) the protein are digested by means of enzymatic or chemical reaction and the specific peptides produced are then analyzed to identify the protein of interest. This results can be obtained using mass spectrometric mass analyzer that can operate in two conditions: a) full scan peptide mass fingerprint (MS) and b) tandem mass spectrometry (MS/MS). In the case a) the mass/charge (m/z) ratio of the peptide is obtained using high resolution and mass accurate analyzer (time of flight, FTICR; see

Cristoni et al., 2004, 2003). The combination of the high accurate m/z ratio of the detected peptides is checked against the theoretical one generated by virtual digestion of the proteins present in the known database. A list of protein candidates is so obtained with relative statistical identification score, correlated to the number of peptides detected, per proteins and peptide mass accuracy. The principal software package used for this kind of data elaboration are reported in table 2.

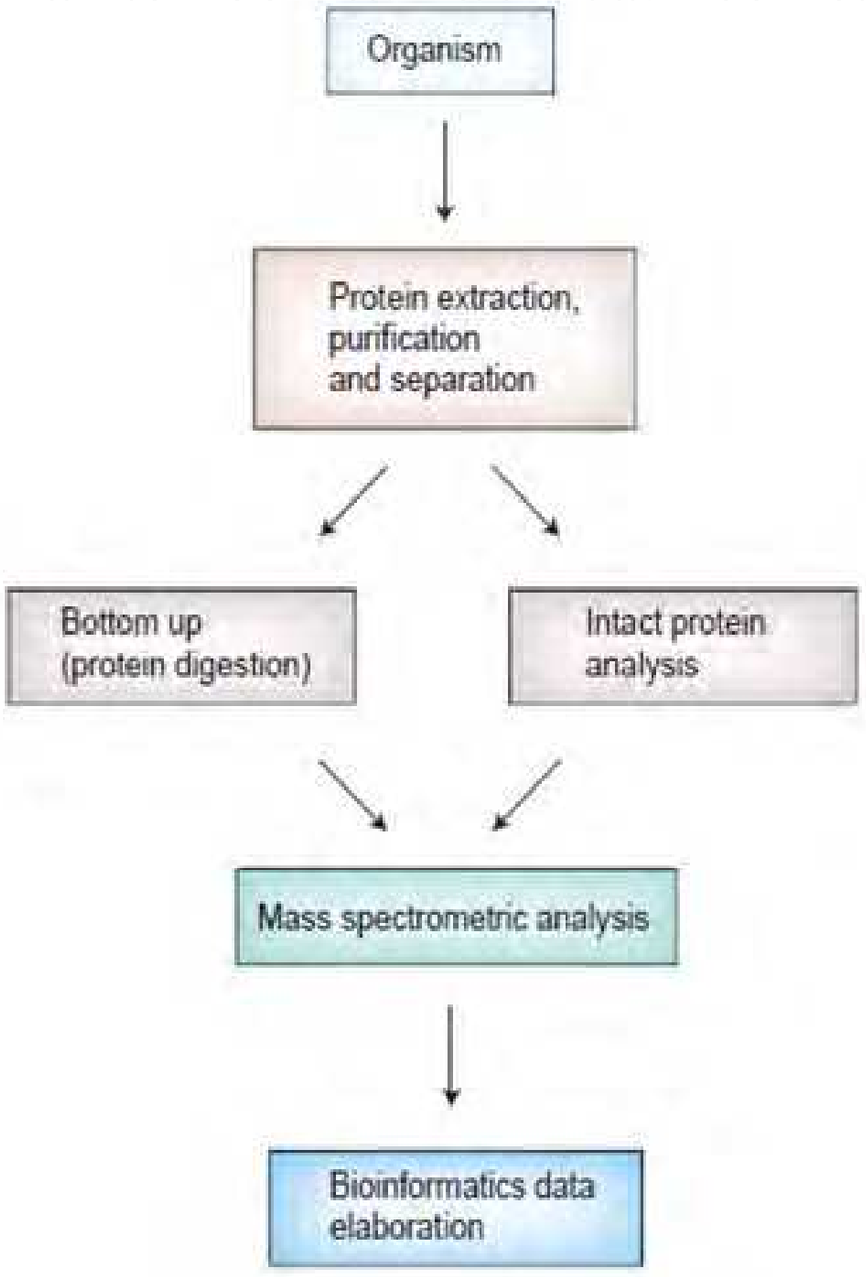


Fig. 2. General analytical scheme of Full proteomic analysis.

Software name	Availability	Web	Categor
Aldent	fre	http://www.expasy.org/tools/aldente/	database searching and PMF
FindPep	Open source,free	http://www.expasy.ch/tools/findpept.html	database searching and PMF
FindMo	Open source,free	http://www.expasy.ch/tools/findmod	database searching , PM and
MASCO	Commercial	http://www.matrixscience.com	database searching and PMF
ProFoun	Open source,free	http://prowl.rockefeller.edu/prowl-cgi/profound.exe	database searching and PMF
PepNovo	Ope source,free	http://peptide.ucsd.edu/pepnovo.py	DeNov
PEAK	Commercial	http://www.bioinfor.com/peaksonline	DeNov
Lutefis	Open source,free	http://www.hairyfatguy.com/Lutefisk/	DeNov
SEQUES	Commercial	http://fields.scripps.edu/sequest	databas searchin
XTandem	Open source,free	http://www.thegpm.org/tandem	database searching
OMSS	Open source,free	http://pubchem.ncbi.nlm.nih.gov/omssa	database searching
PHENY	Commercial	http://www.phenyx-ms.co	database searching
ProbI	Open source,free	http://www.systemsbiology.org/research/probid/	database searching
Popita	Open source,free	http://www.expasy.org/people/pig/heuristic.html	database searching
Interac	Open source,free	http://www.proteomecenter.org/software.php	database searching
DTAselec	Open source,free	http://fields.scripps.edu/DTASelect/index.html	database searching
Chompe	Open source,free	http://www.ludwig.edu.au/jpsl/jpslhome.html	database searching
ProteinProphet	Open source,free	http://www.proteomecenter.org/software.php	database searching
FindPep	Open source,free	http://www.expasy.ch/tools/findpept.html	database searching
GutenTa	Open source,free	http://fields.scripps.edu/	Denovo and PTM
M -Convolution	fre	http://prospector.ucsf.edu/prospector/mshome.htm	DeNov and
M -Alignment	fre	http://prospector.ucsf.edu/prospector/mshome.htm	DeNov and

Table 2. Summary of the most recognized softwares employed for protein analysis.

For instance one of the most employed algorithm for PMF is Aldente (http://www.expasy.ch). This software allows the protein identification in multi-step way. In the first step the most statistically significant proteins are identified on the basis of accurate peptide m/z combination. In the second one the peptide m/z ion leading to the first identification are not considered and other spectra m/z signal combination are considered in order to identify other proteins. The step is reaped since the identification statistic is good enough in order to identify the protein candidates. In the case b) (MS/MS) the peptides are fragmented using different kind of chemical physical reactions [collision induced dissociation (CID), electron transfer dissociation (ETD), ecc]. The m/z ratio of the peptide fragments is then analyzed in order to obtain peptide structural information. Two approaches are usually employed in order to elaborate the fragmentation spectra: database search and *de novo* sequence. In the case of database search, the peptide MS/MS fragmentation spectra are matched against the theoretical one extracted from public or private repositories. The peptide sequence identification is obtained on the basis of a similarity score among the experimental MS/MS and the theoretical MS/MS spectra. The main limitation of this approach is that only known proteins, reported in the database can be identified. For instance, Thegpm (The Global Proteome Machine; http://www.thegpm.org) is an open source project aims to provide a wide range of tools for proteome identification. In particular X!Tandem software (Muth et al., 2010) is widely employed for database search protein identification. When the protein sequence is not perfectly known, denovo sequence method can be used. In this case, the sequence is obtained directly from the MS/MS spectra avoiding the step of database spectrum search. The obtained

sequences are then compared with those contained in the database so to detect homologies. Even in this case a statistical protein identification score is calculated on the basis of the number of homologues fragments obtained for each protein candidate. The software usually employed in the case of database search approach are classified in table 2 together with those employed for *de novo*. An example of software used for de novo porpoises is PepNovo (Frank et al., 2005). It has been presented a novel scoring method for de novo interpretation of peptides from tandem mass spectrometry data. Our scoring method uses a probabilistic network whose structure reflects the chemical and physical rules that govern the peptide fragmentation. The algorithm was tested on ion trap data and achieved results were comparable and in some cases superior to classical database search algorithms. Moreover, different elaborative approaches have been developed in order to increase the sample throughput and statistical accuracy of the identification process (Jacob et al., 2010). Various statistical validation algorithms have been translated into binary programs and are freely distributed on the internet (table 1). Others are not freely available while some have been theoretically described but have not been translated into a freely available or commercial binary program (table 1). It must be stressed that open-source and freely available programs are capable of highly accurate statistical analysis. For example, an interesting free program referred to as ProbID (Zhang et al., 2002) is freely available for evaluation. This program is based on a new probabilistic model and score function that ranks the quality of the match between the peptide. ProbID software has been shown to reach performance levels comparable with industry standard software. A variety of other software, based on heuristic or similar to ProbID Bayesian approach have been developed (Jacob et al., 2010). Some of these software are reported in table 2. It must be stressed that many of these software packages require a web server to operate (e.g., ProbID). This fact introduces some problems related to the difficulty to administrate a server, especially from a security point of view in the case of cracker informatic attacks to a chemstation connected to the internet.

The analysis of intact proteins (top down approach) can be an alternative to bottom up one (Cristoni et. al., 2004). In the first step of data elaboration, the molecular weight of the protein is obtained using dedicated deconvolution algorithm. For instance, Zheng and coworkers have proposed a new algorithm for the deconvolution of ESI mass spectra based on direct assignment of charge to the measured signal at each m/z value in order consequently indirectly to obtain the protein molecular weight (Zheng H, et al. 2003). Another interesting deconvolution approaches is based on the free software named MoWeD (Lu et al., 2011). It can be used to rapidly process LC/ESI/MS data to assign a molecular weight to peptides and proteins. It must be stressed that, the list of found components can also be compared with a user defined list of target molecular weight values making it easy to identify the different proteins present in the analyzed samples. However, when the protein sample mixture is highly complicated, these software tools could fail. This occurs especially if the analysis is performed using low mass accuracy instruments (e.g., IT) and if the chromatographic separation performed before MS analysis is not optimal. Thus, the molecular weight data must be integrated with protein sequence information. In this case, intact proteins ions are analyzed and fragmented by means of high resolution and mass accuracy mass analyzer (e.g.: FTICR, orbitrap, QTOF etc;). The mass spectra obtained are matched directly with the theoretical one present in the database and a statistical score, based on the spectra similarity, is associated with the protein identification. The main advantage of this technology is the ability to investigate intact proteins sequence directly avoiding time consuming digestion steps. On the other hand the majority of algorithm are

usually developed for bottom up approach. In fact for different chemical physical reasons, that are not related to this chapter theme, the sensitivity in detecting high molecular weight proteins is definitely lower with respect to that obtained by detecting low molecular weight peptide after protein digestion. An example of algorithm for protein identification, by intact protein ion fragmentation, has been proposed by McLafferty and co-workers (Sze et al., 2002). A free web interface to be used to analyze proteins MS data using the top-down algorithm is available free of charge for academic use. In the top-down MS approach, the multicharged ions of proteins are dissociated and the obtained fragment ions are matched against those predicted from the database protein sequence. This is also a very powerful tool to characterize proteins when complex mixtures are available.

3.3 Data elaboration for functional proteome

Functional proteome (May et al., 2011) is related to both identify differentially expressed proteins among different sample lines and obtain their relative quantitation. For instance, it is possible to compare differentially expressed proteins among control and unhealthy subjects affected by different diseases (Nair et al., 2004). The classical approach (Figure 3) is based on the protein separation by means of 2D-GEL electrophoresis.

The protein are then colored by using specific reagent (e.g. blue coumassie, silver stain etc) and the gel images are obtained by means of a normal or laser fluorescence scanner. Specific software are then employed in order to overlap the images and detect the differentially expressed proteins on the basis of the color intensities. This approach, has strong limitations mainly in terms of elaboration time needed to obtain the match. Nowadays some apparatus have been developed in order to mark, with different label fluorescence reagents, the proteins extracted from different spots. Thus it is possible to run more samples at the same time and detect the proteins of more spots, separately, by means of different fluorescence laser. Distinctive images relative to different gradient of fluorescence are so simultaneously obtained, this results in differentially expressed proteins.

High innovative shut-gun technology based on liquid chromatography coupled to high resolution mass spectrometry, have been recently developed and employed for functional proteomics purposes. In particular, to compare a complex protein mixture of different experimental lines, the obtained peptides after digestion have been analyzed by means of Surface Activated Chemical Ionization (SACI; Cristoni et al. 2007) technologies coupled to high relation and mass accuracy mass analyzer (e.g. Orbitrap, QTOF etc). Very recently *SACI* technology has been applied in seagrass proteomics (Finiguerra et al., 2010). In fact, the increasing sensitivity of this ionization device improves peptides detection thus recovering the limited sea grass genome resources. SACI leads to benefits, in complex plant protein mixture analysis, in terms of quantitative accuracy, precision, and matrix effect reduction, that have been widely demonstrated (Cristoni et al., 2009). As regard peptide analysis, it was observed that, by changing in-source ionization conditions, one could selectively produce both in-source singly and doubly charged species (Cristoni et al., 2007), which are both of interest. This technologic approach yields maximum benefits when data are acquired using a high mass-accuracy and high-resolution mass analyzer that can operate in both full-scan and tandem mass spectrometry (MS/MS) acquisition conditions. The SACI technique strongly increased the number of detectable proteins and of assigned peptides for each protein. For example, with SACI technology application, it was possible to identify a previously identified protein (a heat shock cognate protein), 1000 fold over expressed in

deeper plants (-27 m) in comparison with the more shallow plants (-5m), detecting four peptides respect to only two detected by micro-ESI (Finiguerra et al., 2010).

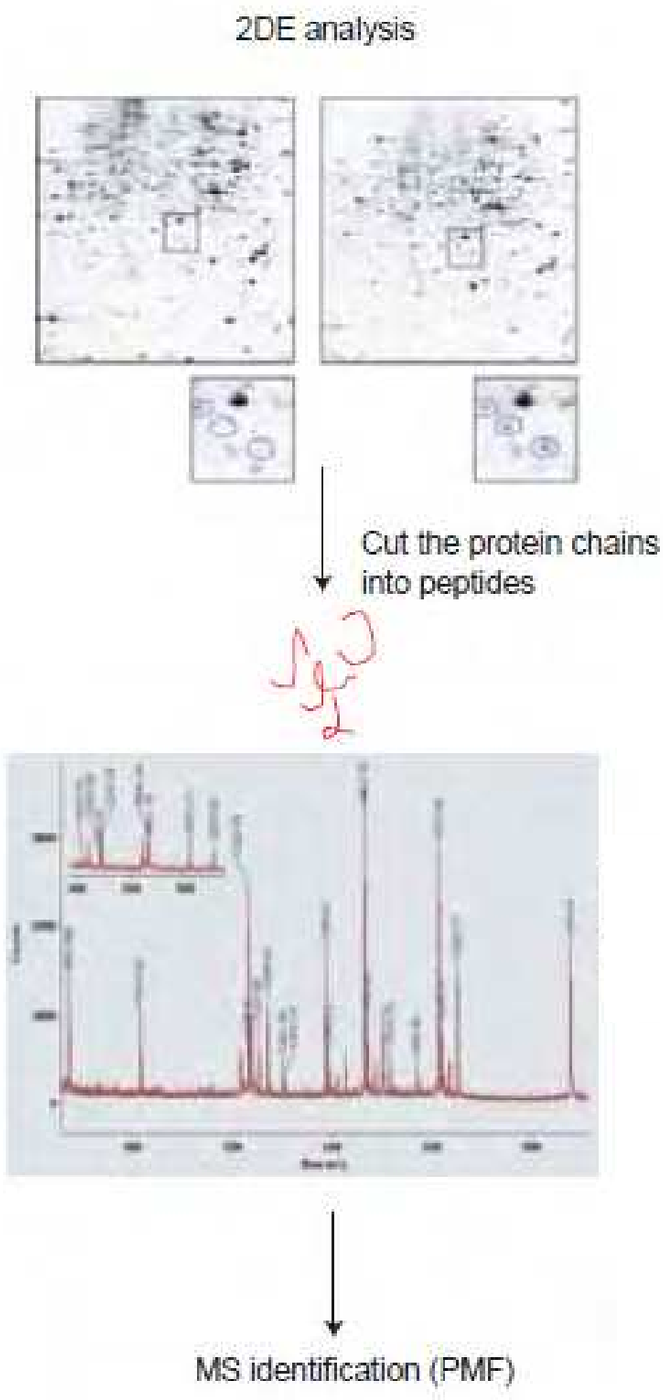


Fig. 3. Classical approach for functional proteome analysis.

The differentially expressed peptides are compared using specific chromatographic alignment software (Sandin et al., 2011). One of the most effective software is named XCMS (Margaria et al., 2008; Kind et al., 2007). The peptide mass fingerprint of the differentially expressed peptides followed by database search and de novo sequencing approach lead to the rapid identifications of differentially expressed proteins.

3.4 Data elaboration for the study of post translational modification

Post translational modification (PTM) detection and quantization is one of the most difficult task in proteomics research. Usually, they are detected through bottom up approaches. For example, considering that phosphorylated peptides do not show a high quality signal intensity, consequently leading to lower sensitivity, dedicated informatics tools have been developed in order to characterize the phosforilation sites on the basis of peptides fragmentation spectra. Different tools developed for this purpose are reported in table 1. All tools detect the modified sites on the basis of fragmentation spectra. Basically, the MS/MS spectra of the theoretical modified peptides are calculated and matched with the experimental one. Even in this case the similarity score is used in order to identified the peptides and the corresponding phosphorylation site.

In the case of PTM the classical PMF and database search approaches cannot be used due to the fact that the modifications and the mutations cause shifts in the MS/MS peaks. Several approaches use an exhaustive search method to identify and characterize the mutated peptides. A virtual database of all modified peptides for a small set of modifications is generated and the peptide MS/MS spectrum is matched against the theoretical spectra of the virtually modified peptides. However, the MS/MS peak shifts result in an increase in the search space and a long elaboration time could be required. To solve this problem some algorithms have been developed (Cristoni S, Bernardi LR. et al. 2004). A good method to detect the mutated peptides is based on the tags approach. For example, the GutenTag software developed by Yates and coworkers use this strategy to identify and characterize the mutated peptides (Tabb DL, et al. 2003). This software infers partial sequences ('tags') directly from the fragment ions present in each spectrum, examines a sequence database to assemble a list of candidate peptides and evaluates the returned peptide sequences against the spectrum to determine which is the best match. The software, written in the Java programming language, runs equally well under Microsoft Windows, Linux and other operating systems. GutenTag is specific to doubly charged peptides. Pevzner and coworkers have also developed some interesting algorithms for the analysis of peptide mutations (Pevzner PA, et al. 2001). In this case, two software packages (MS-CONVOLUTION and MS-ALIGNMENT) that implement the spectra (table 2) convolution and spectral alignment approaches, to identify peptides obtained through enzymatic digestion, have been used to identify and characterize peptides differing by up to two mutations/modifications from the related peptide in the database. This is a two-stage approach to MS/MS database searching. At the first stage, the spectral alignment is used as a filter to identify *t,* top-scoring peptides in the database, where *t* is chosen in such a way that it is almost guaranteed that a correct hit is present among the top *t* list. These top t hits form a small database of candidate peptides subject to further analysis at the second stage. At the verification stage, each of these t peptides can be mutated (as suggested by spectral alignment) and compared against the experimental spectrum. However, the peptide mutation or modification can produce low informative fragmentation behavior (Cristoni et al., 2004), in which case the protein modification identification may fail. It is also possible to use the PMF approach to

characterize mutations and modifications (Cristoni et al., 2004). In this case, it is necessary to use a high mass accuracy mass spectrometer since the characterization of a mutation or modification is based on the identification of the accurate m/z ratios of digested peptide. Freeware software to identify protein modifications and mutations using database search and PMF are reported in the Information Resources section. For example, the software FindPept is capable of identifying peptides that result from nonspecific cleavage of proteins from their experimental masses, taking into account artifactual chemical modifications, PTM and protease autolytic cleavage. If autolysis is to be taken into account, an enzyme entry must be specified from the drop-down list of enzymes for which the sequence is known. Furthermore, this is a web application installed on the expasy website and therefore it is not necessary to install and administrate it on a local server. Another field in which different algorithms have been employed is the characterization of disulfide cross-link locations (Cristoni et al., 2004). For instance, some tools available on a public website http://www.expasy.ch were recently developed for this purpose. This software is referred to as Protein Disulfide Linkage Modeler and it permits the rapid analysis of mass spectromic disulfide cross-link mapping experiments. The tool can be used to determine disulfide linkages in proteins that have either been completely or partially digested with enzymes. The masses of all possible disulfide-linked multichain peptide combinations are calculated from the known protein sequence and compared with the experimentally determined masses of disulfide-linked multichain peptides. Thus, this software is based on the fragmentation behavior of the cross-linked peptides obtained by enzymatic digestion. However, several issues may occur despite the fact that this algorithm executes its work very well. The major issue is that proteins containing linker cysteines have domains that are very resistant to proteolysis. Furthermore, the fragmentation of the cross-linked peptide ions may lead to a spectra that is difficult to elaborate even if specific algorithms are used. This is due to the high chemical noise level that is present in the fragmentation spectra of their multicharged ions (Craig et al., 2003).

4. Data management - The advent of semantic technologies and machine learning methods for proteomics

For Systems Biology the integration of multi-level Omics profiles (also across species) is considered as central element. Due to the complexity of each specific Omics technique, specialization of experimental and bioinformatics research groups have become necessary, in turn demanding collaborative efforts for effectively implementing cross-Omics (Wiesinger M, et al. 2011).

In recent years large amounts of information have been accumulated in proteomic, genetic and metabolic databases. Much effort has been dedicated to developing methods that successfully exploit, organize and structure this information. In fact semantic is the study of meaning. In the case of proteomics it can be used in order to find specific relations among proteins and metabolomics, genomics and ionomics networks. For instance the group of Masaneta-Villa and co-workers (Massanet-Vila et al., 2010) has developed a high-throughput software package to retrieve information from publicly available databases, such as the Gene Ontology Annotation (GOA) database and the Human Proteome Resource Database (HPRD) and structure their information. This information is presented to the user as groups of semantically described dense interaction subnetworks that interact with a target protein. Another interesting technology in the semantic field has been proposed by

the group of Mewes HW. and co-workers (Mewes et al., 2011). This group has many years of experience in providing annotated collections of biological data. Selected data sets of high relevance, such as model genomes, are subjected to careful manual curation, while the bulk of high-throughput data is annotated by automatic means. This is, in fact an important point, manual curation is essential for semantic technology purposes. The data mean must be carefully checked before of the insertion in the semantic database otherwise serious meaning error can occurs during the research phase. High-quality reference resources developed in the past and still actively maintained include Saccharomyces cerevisiae, Neurospora crassa and Arabidopsis thaliana genome databases as well as several protein interaction data sets (MPACT, MPPI and CORUM). More recent projects are PhenomiR, the database on microRNA-related phenotypes, and MIPS PlantsDB for integrative and comparative plant genome research. The interlinked resources SIMAP and PEDANT provide homology relationships as well as up-to-date and consistent annotation for 38,000,000 protein sequences. PPLIPS and CCancer are versatile tools for proteomics and functional genomics interfacing to a database of compilations from gene lists extracted from literature. A novel literature-mining tool, EXCERBT, gives access to structured information on classified relations between genes, proteins, phenotypes and diseases extracted from Medline abstracts by semantic analysis.

Another interesting semantic application has been shown by Handcock J. and co-workers (Handcock, et al., 2010). This group has semantically correlate proteomics information to specific clinical diseases. They have produced a database mspecLINE. Given a disease, the tool will display proteins and peptides that may be associated with the disease. It will also display relevant literature from MEDLINE. Furthermore, mspecLINE allows researchers to select proteotypic peptides for specific protein targets in a mass spectrometry assay.

Another interesting semantic technology is based on machine learning and is employed for biomarker discovery purposes (Barla et al., 2008).The search for predictive biomarkers of disease from high-throughput mass spectrometry (MS) data requires a complex analysis path. Preprocessing and machine-learning modules are pipelined, starting from raw spectra, to set up a predictive classifier based on a shortlist of candidate features. As a machine-learning problem, proteomic profiling on MS data needs caution like the microarray case. The risk of over fitting and of selection bias effects is in fact, pervasive.

Summarizing semantic technologies can be useful both to correlate the different omics sciences information and to correlate the single omics (e.g. proteomics) to specific information like clinical disease correlated to differentially expressed proteins between control and unhealthy groups (biomarker discovery).

5. Conclusions

Bioinformatics for proteomics has grown significantly in the recent years. The ability of process an high amount of data together with the high specificity and precision of the new algorithm in the protein identification, characterization and quantization make now possible to obtain an high amount of elaborated data.

The main problem remain the data management of a so high amount of data. Find the correlation among different proteomic data and the other omics sciences (metabolomics, genomics, ionomics) still remain a difficult task. However, database technology together with new semantic statistical algorithm are in evolution powerful tools useful to overcome this problem.

6. Acknowledgment

This is in memory of Anna Malomo-Mazzuca and Remo Cristoni, mother and father of the authors.

7. References

Alagna, F.; D'Agostino, N.; Torchia, L.; Servili, M.; Rao, R.; Pietrella, M.; Giuliano. G.; Chiusano, M.L.; Baldoni, L. & Perrotta, G. (2009). Comparative 454 pyrosequencing of transcripts from two olive genotypes during fruit development. *BMC Genomics,* vol. 10, p.399

Agrawal, G.K. & Rakwa, R. (2005). Rice proteomics: a cornerstone for cereal food crop proteomes. *Mass Spectrom. Review*, vol. 25, pp.1–53

Agrawal, G.K.; Bourguignon, J.; Rolland, N.; Ephritikhine, G.;, Ferro, M.; Jaquinod, M.; Alexiou, K.G.; Chardot, T.; Chakraborty, N.; Jolivet, P.; Doonan, J.H. & Rakwal, R. (2010). Plant organelle proteomics: collaborating for optimal cell function. *Mass Spectrometry Reviews*, 30: n/a. doi: 10.1002/mas.20301

Andacht, T.M. & Winn, R.N.(2006). Incorporating proteomics into aquatic toxicology. *Marine Environmental Research.*, vol. 62, pp.156–186

Apraiz, I.; Mi, J. & Cristobal, S.; Identification of proteomic signatures of exposure to marine pollutants in mussels (*Mytilus edulis*). *Molecular and Cellular Proteomics*, (2006), No (5),pp. 1274–1285

Armengaud, J. (2010). Proteogenomics and systems biology: quest for the ultimate missing parts. *Expert Rev Proteomics*, vol. 7, No.1, pp. 65-77.

Baerenfaller, K.; Grossmann, J.; Grobei, MA.; Hull, R.; Hirsch-Hoffmann, M.; Yalovsky, S.; Zimmermann, P.; Grossniklaus, U.; Gruissem, W. & Baginsky, S. (2008). Genome-scale proteomics reveals *Arabidopsis thaliana* gene models and proteome dynamics. *Science*, vol. 16, pp. 938-941

Baginsky, S. (2009). Plant proteomics: concepts, applications, and novel strategies for data interpretation. *Mass Spectrometry Review,* vol.28, pp.93-120

Baginsky, S.; Hennig, L.; Zimmermann, P. & Gruissem W. (2010). Gene expression analysis, proteomics, and network discovery. *Plant Physiol*. Vol.152, No.(2), pp. 402-10

Barbazuk, W.B.; Emrich, S.J.; Chen, H.D.; Li, L. & Schnable, P.S.(2006). SNP discovery via 454 transcriptome sequencing. *Plant J.* vol. 51, pp. 910-918

Barla, A. Jurman, G. Riccadonna, S. Merler, S. Chierici, M. Furlanello, C. 2008. *Brief Bioinform,* 9, 2, pp. 119-128.

Bellin, D.; Ferrarini, A.; Chimento, A.; Kaiser, O.; Levenkova, N.; Bouffard, P. & Delledonne, M. (2009). Combining next-generation pyrosequencing with microarray for large scale expression analysis in non-model species. *BMC Genomics* vol. 24, pp.55.

Bouck, A.M.Y.; Vision, T. (2007). The molecular ecologist's guide to expressed sequence tags *Molecular Ecology,* vol.16, pp.907–924

Claesson, M.J.; O'Sullivan, O.; Wang, Q.; Nikkilä, J.; Marchesi, J.R.; Smidt, H.; de Vos, W.M.; Ross, R.P. & O'Toole. (2009). Comparative analysis of pyrosequencing and a phylogenetic microarray for exploring microbial community structures in the human distal intestine. *PW.PLoS One, vol.* 20, No. 4, pp.e6669

Chen, X.; Karnovsky, A.; Sans, M.D.; Andrews, P.C. & Williams, J.A. (2010). Molecular characterization of the endoplasmic reticulum: insights from proteomic studies. *Proteomics*, vol. 10, pp. 4040-52. doi: 10.1002/pmic.201000234

Chevalier, F. (2010). Highlights on the capabilities of «Gel-based» proteomics. *Proteome Science*, vol. 8, No.(23), doi:10.1186/1477-5956-8-23

Clarke, S.C. (2005). Pyrosequencing: nucleotide sequencing technology with bacterial genotyping applications. *Expert Rev Mol Diagn*., vol. 5, No.(6), pp.947-53

Costenoble, R.; Picotti, P.; Reiter, L.; Stallmach, R.; Heinemann, M.; Sauer, U. & Aebersold, R. (2011). Comprehensive quantitative analysis of central carbon and amino-acid metabolism in *Saccharomyces cerevisiae* under multiple conditions by targeted proteomics. *Molecular Systems Biology, vol.* 7, pp. 464

Craft, J.A.; Gilbert, J.A.; Temperton, B.; Dempsey, K.E.; Ashelford, K.; et al. (2010). Pyrosequencing of *Mytilus galloprovincialis* cDNAs: Tissue-Specific Expression Patterns. *PLoS ONE*, vol. 5, No. (1): pp.e8875. doi:10.1371/journal.pone.0008875

Craig, R. Krokhin, O. Wilkins, J. Beavis, RC. 2003. Implementation of an algorithm for modeling disulfide bond patterns using mass spectrometry. *J. Proteome Res,* 2, 6, pp. 657–661.

Cristoni, S.; Zingaro, L.; Canton, C.; Cremonesi, P.; Castiglioni, B.; Morandi, S.; Brasca, M.; Luzzana, M. & Battaglia, C. (2009). Surface-activated chemical ionization and cation exchange chromatography for the analysis of enterotoxin A. *Journal of Mass Spectrometry*, vol. 44,pp. 1482-1488

Cristoni, S.; Rubini, S. & Bernardi, L.R. (2007). Development and applications of surface-activated chemical ionization. *Mass Spectrometry Reviews*, vol. 26, pp. 645-656

Cristoni, S. & Bernardi, L.R. (2004). Bioinformatics in mass spectrometry data analysis for proteomics studies. *Expert Rev Proteomics,* vol. 1, No.4, pp. 469-83.

Cristoni, S. & Bernardi, L,R. (2003). Development of new methodologies for the mass spectrometry study of bioorganic macromolecules. *Mass Spectrom Rev,* vol.22, No.6, pp.369-406.

Dunkley, T.P.J.; Hester, S.; Shadforth, I.P.; Runions, J.; Weimar, T.; Hanton, S.L.; Griffin, J.L.; Bessant, C.; Brandizzi, F.; Hawes, C.; Watson, R.B.; Dupree, P.& Lilley, KS. *(*2006). Mapping the *Arabidopsis* organelle proteome. *PNAS,* vol. 103, pp. 6518-6523

Doyle, S. (2011). Fungal proteomics: from identification to function. *FEMS Microbiol Lett*. doi: 10.1111/j.1574-6968.2011.02292

Emrich, S.J.; Barbazuk, W.B.; Li, L. & Schnable, P.S. (2007). Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome Res*, vol.17, pp.69-73

Eveland, A.L.; McCarty, D.R. & Koch, K.E. (2008). Transcript Profiling by 3'-Untranslated Region Sequencing Resolves Expression of Gene Families. *Plant Physiol*, vol. 146, pp. 32-44

Feder, M.E. & Walser, J.C. (2005).The biological limitations of transcriptomics in elucidating stress and stress responses. *Journal of Evolutionary Biology*, vol.18, pp.901–910

Ferro, M.; Brugière, S.; Salvi, D.; Seigneurin-Berny, D.; Court, M.; Moyet, L.; Ramus, C.; Miras, S.; Mellal, M.; Le Gall, S.; Kieffer-Jaquinod, S.; Bruley, C.; Garin, J.; Joyard, J.; Masselon, C. & Rolland, N. (2010). AT_CHLORO: A comprehensive chloroplast proteome database with subplastidial localization and curated information on envelope proteins. *Molecular & Cellular Proteomics*, vol. 9, pp. 1063-1084

Finiguerra, A.; Spadafora, A.; Filadoro, D. & Mazzuca, S. (2010). Surface-activated chemical ionization time-of-flight mass spectrometry and labeling-free approach: two powerful tools for the analysis of complex plant functional proteome profiles. *Rapid Communication in Mass Spectrometry*, vol. 24,pp.1155-1160

Frank A, & Pevzner P. (2005). PepNovo: de novo peptide sequencing via probabilistic network modeling. *Anal Chem,* vol.15, No.77, pp. 964-973.

Fukao, Y.; Hayashi, M. & Nishimura, M. (2002). Proteomic analysis of leaf peroxisomal proteins in greening cotyledons of *Arabidopsis thaliana*. *Plant and Cell Physiology*, vol. 43, pp. 689-696.

Gibson, G. (2003).Microarrays in ecology and evolution: a preview. *Molecular Ecology,* vol. 11, pp. 17-24

Giot, L.; Bader, J.S.; Brouwer, C.; Chaudhuri, A.; & 44 others. (2003). A Protein Interaction Map of *Drosophila melanogaster*. *Science*, vol. 302 , pp.1727–1736

Grennan, A.K. (2008). Plastoglobule proteome. *Plant Physiology*, vol. 147, pp. 443–445

Grimplet, J.; Cramer, G.R.; Dickerson, J.A.; Mathiason, K.; Van Hemert, J.; et al. (2009).VitisNet: ''Omics'' Integration through Grapevine Molecular Networks. *PLoSONE,* vol. 4,no.(12), pp.e8365. doi:10.1371/journal.pone.0008365

Guo, S.; Zheng, Y.; Joung, J.G.; Liu, S.; Zhang, Z.; Crasta, O.R.; Sobral, B.W.; Xu, Y.; Huang, S. & Fei, Z. (2010) Transcriptome sequencing and comparative analysis of cucumber flowers with different sex types. *BMC Genomics*, vol.11, pp.384

Hamady, M.; Lozupone, C. (2007). Knight R Fast UniFrac: Facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. *ISME J., vol.* 4, No.(1), pp.17-27

Handcock, J. Deutsch, EW. Boyle, J. 2010. mspecLINE: bridging knowledge of human disease with the proteome. *BMC Med. Genomics,* 10, 3, pp. 7.

Haynes, P. A; Roberts, TH. Subcellular shotgun proteomics in plants: Looking beyond the usual suspects. *Proteomics,* , 7, 2963 – 2975

Heazlewood, J.L. & Millar, H. (2005). AMPDB: the *Arabidopsis* Mitochondrial protein Database. *Nucleid Acids Research*. Vol. 33,pp.605–610.

Heazlewood J.L. (2011). The Green proteome: challenges in plant proteomics. Front. Plant Sci., vol. 2, pp6. doi:10.3389/fpls.2011.00006;

Heid, C.A.; Stevens, J.; Livak, K.J. & Williams, PM.(1996). Real time quantitative PCR. *Genome Research,* , 6, 986–994

Ioannidis JP.2010. A roadmap for successful applications of clinical proteomics. *Proteomics Clin Appl.* doi: 10.1002/prca.201000096;

Jacob, RJ. (2010). Bioinformatics for LC-MS/MS-based proteomics. *Methods Mol Biol,* vol. 658, pp. 61-91

Jaquinod, M.; Villiers, F.; Kieffer-Jaquinod, S.; Hugouvieux, V.; Bruley, C.; Garin, J. & Bourguignon, J. (2007). A Proteomics Dissection of *Arabidopsis thaliana* Vacuoles Isolated from Cell Culture. *Molecular and Cellular Proteomics, 20, vol.* 6, 394-412

Jamet, E.; Canut, H.; Boudart, G. & Pont-Lezica, R.F. (2006).Cell wall proteins: a new insight through proteomics *Trends in Plant Science*, vol.11, pp. 33-39

Johnson, S.C. & Browman, H.I. (2007). Introducing genomics, proteomics and metabolomics in marine ecology. *Marine Ecology Progress Series*, vol. 33,pp. 247–248

Joyard, J.; Ferro, M.; Masselon, C.; Seigneurin-Berny, D.; Salvi, D.; Garin, J. & Rolland, N. (2010). Chloroplast proteomics highlights the subcellular compartmentation of lipid metabolism. *Prog Lipid Res*. Vol.49, No. (2), pp.128-58

Joshi H.J.; Hirsch-Hoffmann, M.; Baerenfaller, K.; Gruissem, W.; Baginsky, S.; Schmidt, Robert.; Schulze, W. X.; Sun, Q.; van Wijk K.J.; Egelhofer V.; Wienkoop, S.; Weckwerth, W.; Bruley, C.; Rolland,N.; Toyoda, T.; Nakagami, H.; Jones, A.M.; Briggs, S.P.; Castleden, I.; Tanz, S.K.; A. Millar, H; & Heazlewood. J.L. (2011). MASCP Gator: An Aggregation Portal for theVisualization of Arabidopsis Proteomics Data. *Plant Physiology*, Vol. 155, pp. 259–270

Jovanovic, M.; Reiter, L.; Picotti, P.; Lange, V.; Bogan, E.: Hurschler, B.A.; Blenkiron, C.; Lehrbach, N.J.; Ding, X.C.; Weiss, M.; Schrimpf, S.P.; Miska, E.A.; Großhans, H.; Aebersold, R. & Hengartner, M.O. (2010). A quantitative targeted proteomics approach to validate predicted microRNA targets in *C. elegans* . *Nature methods*, vol.7, N.(10),pp. 837-845

Kind, T. Tolstikov, V. Fiehn, O. Weiss, RH. 2007. A comprehensive urinary metabolomic approach for identifying kidney cancerr. *Anal Biochem,* vol.363, No.2, pp. 185-195.

Lander, E.S. (2011). Initial impact of the sequencing of the human genome. *Nature*, vol. 470, pp.187–197, doi:10.1038/nature09792

Li, Y.C.; Korol, A.B.; Fahima, T. & Nevo, E. (2004). Microsatellites within genes: structure, function, and evolution. *Molecular Biology and Evolution*, vol. 21, pp.991-1007

Li, X.; Pizarro, A.; Grosser, T. (2009). Elective affinities--bioinformatic analysis of proteomic mass spectrometry data. *Arch Physiol Biochem*., vol. 115, No(5), pp.311-9

Lopez, J.L. (2007). Applications of proteomics in marine ecology. *Marine Ecology Progress Series*, vol.332, pp. 275–279

Lu, W; Callahan, J.H.; Fry, FS.; Andrzejewski, D.; Musser, SM. & Harrington, P.B. (2011). A discriminant based charge deconvolution analysis pipeline for protein profiling of whole cell extracts using liquid chromatography-electrospray ionization-quadrupole time-of-flight mass spectrometry. *Talanta,* vol. 30, No.84, pp. 1180-1187

Maltman, D.J.; Gadd, S.M.; Simon, W.J. & Slabas, A.R.(2007). Differential proteomic analysis of the endoplasmic reticulum from developing and germinating seeds of castor (*Ricinus communis*) identifies seed protein precursor as significant component of the endoplasmic reticulum. *Proteomics.* Vel. 7, pp.1513–1528

Margaria, T. Kubczak, C. Steffen, B. 2008. Bio-jETI: a service integration, design, and provisioning platform for orchestrated bioinformatics processes. *BMC Bioinformatics,* vol. 9, No. 4, pp. S12.

Marguerat, S.; Brian, T.; Wilhelm, B.T. & Bahler, J. (2008). Next-generation sequencing: applications beyond genomes *Biochemical Society Transactions*, vol. 36, pp.1091-1096.

Maynard Smith, J., & Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genetical Research,* vol. 23, pp.23–35

May, C.; Brosseron, F; Chartowski, P.; Schumbrutzki, C.; Schoenebeck, B. & Marcus, K. (2011). Instruments and methods in proteomics. *Methods Mol Bio,* vol. 696, pp. 3-26.

Matthiesen, R. & Jensen, O.N. (2008). Methods Analysis of mass spectrometry data in proteomics. *Mol Biol*. Vol. 453, pp.105-22

Massanet-Vila, R. Gallardo-Chacon, JJ. Caminal, P. Perera, A. 2010. An information theory-based tool for characterizing the interaction environment of a protein. *Conf Proc IEEE Eng Med Biol Soc,* 2010, pp. 5529-5532

Metzker, M.L. (2010). Sequencing technologies-the next generation. *Nature Reviews Genetics*, vol. 11, pp.31-46| doi:10.1038/nrg2626

Mewes, HW. Ruepp, A. Theis, F. Rattei, T. Walter, M. Frishman, D. Suhre, K. Spannagl, M. Mayer, KF. Stümpflen, V. Antonov, A. 2011. MIPS: curated databases and comprehensive secondary data resources in 2010. *Nucleic Acids Res,* 39, pp. 220-224.

Molloy, M.P.; Donohoe, S.; Brzezinski, E.E.; Kilby, G.W.; Stevenson, T.I.; Baker, J.D.; Goodlett, D.R. & Gage, D.A. (2005). Large-scale evaluation of quantitative reproducibility and proteome coverage using acid cleavable isotope coded affinity tag mass spectrometry for proteomic profiling. *Proteomics,* vol. 5, pp. 1204-1208, NCBI-dbEST database [http://www.ncbi.nlm.nih.gov/dbEST]

Muth, T.; Vaudel, M.; Barsnes, H.; Martens, L. & Sickmann, A. (2010). XTandem Parser: an open-source library to parse and analyse X!Tandem MS/MS search results. *Proteomics,* vol.10, No.7, pp. 1522-1524.

Nilsson, T,K. & Johansson, C,A. (2004). A novel method for diagnosis of adult hypolactasia by genotyping of the -13910 C/T polymorphism with Pyrosequencing technology. Scand. *J. Gastroenterol*. Vol. 39, pp.287–290

Nair, KS. Jaleel, A. Asmann, YW. Short, KR. Raghavakaimal, S. (2004). Proteomic research: potential opportunities for clinical and physiological investigators. *Am J Physiol Endocrinol Metab,* vol. 286, No.6, pp. 863-874

Novaes, E.; Drost, D.; Farmerie, W.; Pappas, G.; Grattapaglia, D.; Sederoff, R.R. & Kirst, M. (2008).High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. *BMC Genomics* vol. 9, pp.312

Nunn, B.L; & Timperman, T.A. (20079. Marine proteomics. *Marine Ecology Progress Series*, vol. 332, pp. 281-289

Oetjen, K.; Ferber, S.; Dankert, I. & Reusch, T.B.H. (2010). New evidence for habitat-specic selection in Wadden Sea *Zostera marina* populations revealed by genome scanning using SNP and microsatellite markers *Marine Biology* vol. 157, pp. 81–89

Ohlmeier S.; Kastaniotis A. J.; Hiltunen J. K. & Bergmann U. (2004) The yeast mitochondrial proteome - A study of fermentative and respiratory growth. *J Biol Chem.,* vol. 279, pp. 3956-3979

Pevzner, PA. Mulyukov, Z. Dancik, V. Tang CL. 2001. Efficiency of database search for identification of mutated and modified proteins via mass spectrometry. *Genome Res,* 11, 2, pp. 290-299.

Powell, M.J; Sutton, J.N.; Del Castillo, C.E & Timperman, AI. (2005). Marine proteomics: generation of sequence tags for dissolved proteins in seawater using tandem mass spectrometry. *Marine Chemistry*, vol. 95, pp.183-198

Porubleva, L.; VanderVelden, K.; Kothari, S.; Oliver, DJ. & Chitnis, PR. (2001).The proteome of maize leaves: use of gene sequences and expressed sequence tag data foridentification of proteins with peptide mass fingerprinting. *Electrophoresis*, vol. 22, pp. 1724-1738

Plomion, C.; Lalanne, C.; Clavero, S.; Meddour, H.; Kohler, A.; and others. (2006). Mapping the proteome of poplar and application to the discovery of drought-stress responsive proteins. *Proteomics* ,vol.6, pp. 6509–6527

Prokopi, , &, Mayr, M. (2011). Proteomics: a reality-check for putative stem cells. *Circ. Res.*, Vol.108, No. (4),pp.499-511

Repetto, O.; Rogniaux , H.; Firnhaber, C.; Zuber, H.; Küster, H.; Larré, C.; Thompson, R. & Gallardo, K. (2008). Exploring the nuclear proteome of *Medicago truncatula* at the switch towards seed filling. *Plant Journal,* vol. 56, pp. 398

Rossignoll, M.; Peltier, J.B.; Mock, H.P; Matros, A.; Maldonado, A.M. & Jorrín, J.V. (2006). Plant proteome analysis: A 2004–2006 update. *Proteomics*, vol. 6, pp.5529–5548

Roth, U.; Razawi, H.; Hommer, J.; Engelmann, K.; Schwientek, T.; Müller, S.; Baldus, S.E.; Patsos, G.; Corfield, A.P.; Paraskeva, C. & Hisch, F.G. (2010). Differential expression proteomics of human colorectal cancer based on a syngeneic cellular model for the progression of adenoma to carcinoma. Proteomics Clin Appl., vol. 4, no.(8-9),pp.748. doi: 10.1002/prca.201090028

Sá-Correia I, Teixeira MC 2010. 2D electrophoresis-based expression proteomics: a microbiologist's perspective. Expert Rev Proteomics. Dec;7(6):943-53;

Sandin, M. Krogh, M. Hansson, K. Levander, F. (2011) Generic workflow for quality assessment of quantitative label-free LC-MS analysis. *Proteomics,* vol. 11, No. 6, pp. 1114-1124

Satkoski, J.A.; Malhi, R.S.; Kanthaswamy, S.; Tito, R.Y.; Malladi V.S. & Smith, D.G. (2008). Pyrosequencing as a method for SNP identification in the rhesus macaque (*Macaca mulatta*). *BMC Genomics* vol.9, 256doi:10.1186/1471-2164-9-256

Schweitzer, B.; Predki, P. & Snyder, M. (2003). Microarrays to characterize protein interactions on a whole-proteome scale. *Proteomics*, vol.3, pp.2190-2199

Shendure, J.; Porreca, G.J.; Reppas, N.B.; Lin, X.; McCutcheon, J.P.; Rosenbaum, A.M.; Wang, M.D.; Zhang, K.; Mitra, R.D. & Church GM: Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, vol.309, pp.1728-1732

Stinchcombe, J.R.; & Hoekstra, H.E. (2008). Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. *Heredity*, vol.100, pp. 158-170.

Sultan M.; Schulz, M.H.; Richard, H.; Magen, A.; Klingenhoff, A.; Scherf, M.; Seifertet M. al. (2010). A Global View of Gene Activity and Alternative Splicing by Deep Sequencing of the Human Transcriptome *Science*, vol. 321, No. (5891): 956-960. *DOI:* 10.1126/science.1160342

Sze, SK. Ge, Y. Oh, H. McLafferty, FW. 2002. Top-down mass spectrometry of a 29-kDa protein for characterization of any posttranslational modification to within one residue. *Proc Natl Acad Sci U S A*, 99, 4, pp. 1774-1779.

TAIR Database: The Arabidopsis Information Resource [http://www.arabidopsis.org/] webcite Tair_9_pep_ release 2009 06 19)

Tabb, DL. Saraf, A. Yates, JR. 2003. GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. *Anal Chem,* vol.75, No. 23, pp. 6415-6421.

Tyers, M. & Mann, M. (2003). From genomics to proteomics. *Nature, vol. 422*, pp. 193–197

Trick, M.; Long, Y.; Meng, J. & Bancroft, I. (2009). Single nucleotide polymorphism (SNP) discovery in the polyploid Brassica napus using Solexa transcriptome sequencing. *Plant Biotechnol J*, vol. 7, pp. 334 -346.

Tsugita, A.; & Kamo, M. (1999). 2-D electrophoresis of plant proteins. *Methods in Molecular Bioogy,* vol.112,pp. 95-97.

Tweedie-Cullen, R.Y. & Mansuy, I.M. (2010). Towards a better understanding of nuclear processes based on proteomics. *Amino Acids*, vol. 39, No. (5), pp.1117-30.

Vasemägi, A. & Primmer, C.R. (2005). Challenges for identifying functionally important genetic variation: the promise of combining complementary research strategies. *Molecular Ecology,* vol. 14,pp.3623–3642

Vasemägi, A.; Nilsson, J. & Primmer, C.R. Expressed Sequence Tag-Linked Microsatellites as a Source of Gene-Associated Polymorphisms for Detecting Signatures of Divergent selection in Atlantic Salmon (*Salmo salar* L.). *Molecular Biology and Evolution*, pp. 1067-1073

Vigouroux, Y.; McMullen, M.; Hittinger, C.T.; Houchins, K.; Schulz, L.; Kresovich, S.; Matsuoka, Y. & Doebley, J. (2002). Identifying genes of agronomic importance in maize by screening microsatellites for evidence of selection during domestication. *Proceedings of the National Academy of Sciences,* USA, vol. 99, pp. 9650–9655

Wang, W.; Wang, Y.; Zhang, Q.; Qi, Y. & Guo, D. (2009). Global characterization of *Artemisia annua* glandular trichome transcriptome using 454 pyrosequencing. *BMC Genomics*, vol. 10, pp.465

Weber, A.P.; Weber, K.L.; Carr, K.; Wilkerson, C. & Ohlrogge, J.B. (2007). Sampling the *Arabidopsis* transcriptome with massively parallel pyrosequencing. *Plant Physiol*., vol 144, pp. 32-42

Wienkoop, S.; Baginsky, S. & Weckwerth, W.J. (2010). *Arabidopsis thaliana* as a model organism for plant proteome research. *Proteomics* , vol. 73, No.(11), pp.2239-48

Wiesinger, M. Haiduk, M. Behr, M. de Abreu Madeira, HL. Glöckler, G. Perco, P. Lukas, A. 2011. Data and knowledge management in cross-Omics research projects. *Methods Mol Biol,* 719, pp. 97-111

Whitehead, A.; & Crawford, D.L. (2006). Neutral and adaptive variation in gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, vol.103, pp.5425–5430

Wright, J.C.& Hubbard, S.J. (2009). Recent developments in proteome informatics for mass spectrometry analysis. *Comb. Chem. High Throughput Screen*., vol. 12, No.(2), pp.194-202

Yang, S.; Land, M.L.; Klingeman, D.M.; Pelletier, D.A.; Lu, T.Y.; Martin, S.L.; Guo, H.B. & Smith, J.C.; Brown, S.D. (2010).Paradigm for industrial strain improvement identifies sodium acetate tolerance loci in *Zymomonas mobilis* and *Saccharomyces cerevisiae*. *Proc Natl Acad Sci U S A*, vol.107, No. (23), pp10395-400

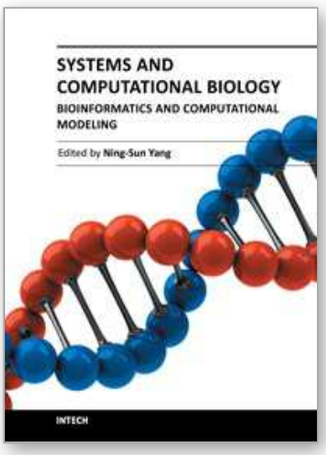
Yao, X.; Freas, A.; Ramirez, J.; Demirev, P.A. &Fenselau, C. (2001). Proteolytic ^{18}O labeling for comparative proteomics: model studies with two serotypes of adenovirus. *Analytical Chemistry*, vol. 73, pp.2836–2842

Yauk, C.L. & Lynn Berndt, M.L. (2007). Review of the Literature Examining the Correlation Among DNA Microarray Technologies. *Environmental and Molecular Mutagenesis*, vol. 48,pp. 380-394

Yates, J.R.; Ruse, C.I & Nakorchevsky, A. (2009). Proteomics by mass spectrometry : approaches, advances, and applications. *The Annual Review of Biomedical Engineering*, vol.11, pp. 49-79

Zhang, N.; Aebersold, R. & Schwikowski, B. (2002). ProbID: a probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. *Proteomics,* vol. 2, No. 10, pp.1406-1412

Zheng, H.; Ojha, P.C.; McClean, S.; Black, ND.; Hughes, JG. & Shaw, C. (2003). Heuristic charge assignment for deconvolution of electrospray ionization mass spectra. *Rapid Commun Mass Spectrom*, vol.17, No.5, pp. 429-436.



Systems and Computational Biology - Bioinformatics and Computational Modeling

Edited by Prof. Ning-Sun Yang

ISBN 978-953-307-875-5
Hard cover, 334 pages
Publisher InTech
Published online 12, September, 2011
Published in print edition September, 2011

Whereas some â€œmicroarrayâ€ or â€œbioinformaticsâ€ scientists among us may have been criticized as doing â€œcataloging researchâ€, the majority of us believe that we are sincerely exploring new scientific and technological systems to benefit human health, human food and animal feed production, and environmental protections. Indeed, we are humbled by the complexity, extent and beauty of cross-talks in various biological systems; on the other hand, we are becoming more educated and are able to start addressing honestly and skillfully the various important issues concerning translational medicine, global agriculture, and the environment. The two volumes of this book present a series of high-quality research or review articles in a timely fashion to this emerging research field of our scientific community.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Simone Cristoni and Silvia Mazzuca (2011). Bioinformatics Applied to Proteomics, Systems and Computational Biology - Bioinformatics and Computational Modeling, Prof. Ning-Sun Yang (Ed.), ISBN: 978-953-307-875-5, InTech, Available from: http://www.intechopen.com/books/systems-and-computational-biology-bioinformatics-and-computational-modeling/bioinformatics-applied-to-proteomics

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen