Proteome Analysis

A New Challenge for Mass Spectrometry

by Fulvio Magni and Marzia Galli Kienle

As a rational consequence of the enormous outcome of human genome identification, interest in protein studies has increased greatly and today reference to the “proteome” is becoming a main topic in all fields of biomedical research. The method of choice for proteome identification implies electrophoretic separation of proteins followed by structural characterisation by mass spectrometry and database searching and combination of these techniques is now referred to as “proteomics”.

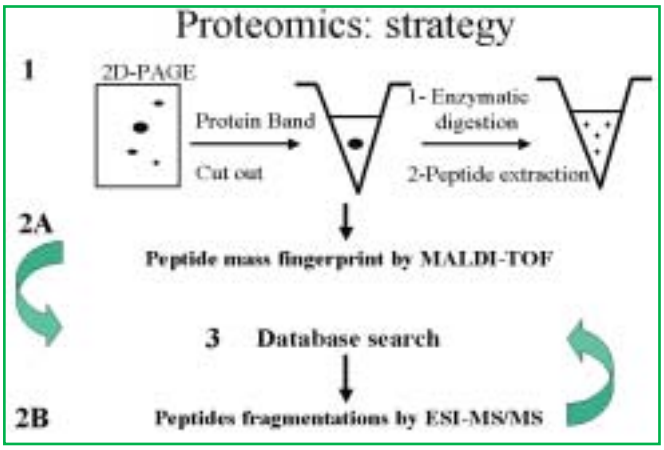

Figure 1 - Proteomics: strategy

During the last decade mass spectrometric instrumentation and related procedures for sample preparation have been improved to a remarkable degree in terms of sensitivity and accuracy of results, especially for protein analysis. The term “proteome” was first used in the 1994 to indicate the PROTEins expressed by a genOME. Proteomics was originally used to indicate the large scale characterization of the entire complement of proteins in a specific organism, tissue or cell type. Nowadays the term “proteome” is used in a more widely sense and it refers to two main groups of activities:

a) classical of proteomics, in which cell lysates are analysed by two-dimensional gels to visualize differential protein expression;
b) functional genomics or functional proteomics. In this last definition protein with common features are purified using specific approaches, such as affinity chromatography, isolation of either multi-protein complexes or single proteins from a specific part of the cell.

Additional importance of structural analysis is related to the identification of post-transcriptional control as well as post-translational modifications of proteins. Proteomics is expected to have a great impact on the Molecular Medicine and on the pharmaceutical industry (i.e. finding new molecular targets for both diagnosis and drug treatment). Proteome analysis, which relies, even if not exclusively, on the micro characterization of proteins separated by two-dimensional protein electrophoresis (2D-PAGE), can monitor synthesis rates, expression levels and post-translational modifications (PTMs) of proteins. A further interesting challenge for proteome is to establish effective connections between protein level and nucleic-acid level information about genes and gene networks. Proteome analysis includes three main steps:

1) extraction of all protein constituents from the target material (tissue, cell line or biological fluid) followed separation of the protein mixture by very powerful electrophoretic systems (two dimensional-gel or capillary electrophoresis; Figure 1, part 1);
2) generation of structural data by mass spectrometry (Maldi-Tof, LC-Esi or nanoEsi after enzymatic hydrolysis directly in gels; Figure 1, part 2A and 2B);
3) comparison of the masses of the measured proteolytic peptides with those expected to derive from the analysed protein with a specific enzyme (Figure 1, part 3).

Each protein sequence in the database is digested according to the specificity of the enzyme and the masses, the resulting peptides are calculated and a theoretical mass spectrum is constructed which is then compared with the measured. Possible protein structures in the database are sorted and the best fitting protein sequence is selected. The success of protein identification by peptide mapping is a result of certain characteristics of proteins, including the limited number of proteins for each organism, the large differences in amino acid sequences, and the large mass difference between different amino acids. At present the central role of mass spectrometry (MS) as the structural technique of choice is not questionable while the conventional choice of two-dimensional electrophoresis for protein mixture fractionation is open to discussion.

Sample presentation to the mass spectrometer

The success of experiments addressed to proteome analysis strictly depends upon sample preparation and introduction into the mass spectrometer. Although the term ‘proteomics’ was

F. Magni, M. Galli Kienle, Dipartimento di Medicina Sperimentale, Ambientale e Biotecnologie Mediche - Università di Milano Bicocca - Monza (MI). magni.fulvio@unimib.it

firstly used in 1996, the primary tools to display and to separate genome-wide protein expression has been available since 1975. Proteins are separated first by their charge using isoelectric focusing and then by size using SDS-PAGE (Figure 2). Several problems have been solved by the use of immobilized pH gradient (IPG) strips, which increase the reproducibility of isoelectric focusing by using pH gradients covalently bound to the polyacrylamide matrix. In addition to the almost complete elimination of pH drift, the use of IPG strips allows to adjust the pH separation to any range.

Other important technological improvements in 2DE include the development of sensitive protein stains as the silver stain which allow detection of proteins at or below nanogram quantities, and the use of in-gel sample application to IPG gradient gel strips. As compared to loading at either the anodic or cathodic ends of the gel, the in-gel sample loading permits the application of larger volumes and quantities of protein as well as reduction of focusing problems associated with protein precipitation. Despite these improvements, analytical 2DE has limited applicability because expression patterns of protein separated by charge and size does not allow identification of the protein in terms of the originating gene or genes. To this purpose, more recently, proteins are characterised by mass spectrometry through the identification of peptides deriving from enzymatic digestion of proteins obtained from 2DE gels [1]. For the separation of proteins preceding mass spectrometry in the mapping procedure, two-dimensional gel electrophoresis is most commonly employed even being aware that problems are faced with more hydrophobic proteins and that this denaturing technique necessarily destroys protein-protein interactions.

Recognising these limitations, several approaches have been attempted toward the use of multidimensional chromatographic sample preparation systems [2]. Yates *et al.* [3] have illustrated the value of combined solid-phase microextraction-multistep elution - capillary electrophoresis (CE) - electrospray tandem MS in the characterisation of the total tryptic digest of a ribosomal protein complex. This two-stage separation approach allowed high resolution separation of the peptides, together with the concentration of peptides into the low injection volumes (nL) required by CE. Other authors have also explored preconcentration approaches to the implementation of CE-MS in order to overcome the problem of limited sample volumes for CE [4].

An alternative method to the study of protein-protein interactions using the yeast nuclear pore complex Nup85p has been recently described [5]. Protein complexes were isolated using an affinity tag and carefully controlled crosslinking conditions allowed to couple only spatially adjacent proteins of the purified complex. The products of the crosslinking reaction were separated using SDS PAGE, digested within the gel and analysed by matrix-assisted laser desorption/ionisation (Maldi) time-of-flight (Tof) MS.

Proteome analyses can be highly focussed or concerned with profiling and comparisons of multiple constituents, aimed to correlate the proteins composition of sample to factors such as growth conditions, external influences and disease state. In profiling and comparative work including numerous samples the importance of achieving high-throughput via automation becomes an essential requirement. Several commercial systems are now available to perform robotic excision of gel spots followed by automated enzymatic digestion and loading

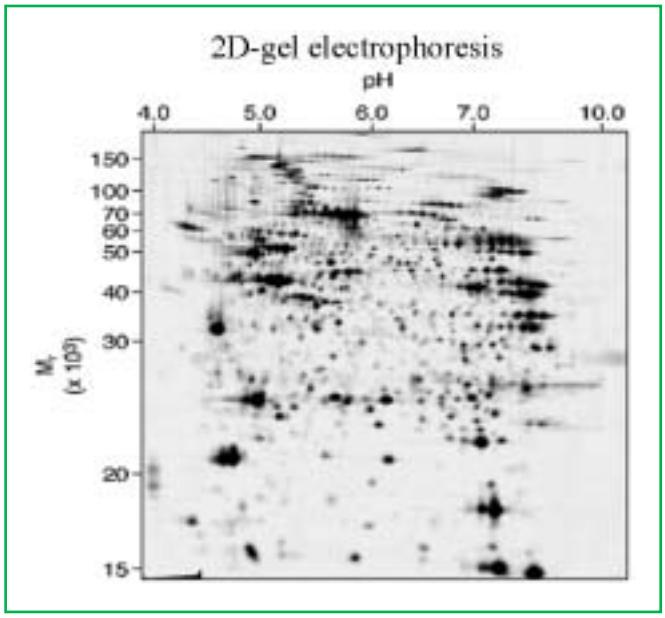

Figure 2 - 2D-gel electrophoresis

on Maldi MS sample plates. Analyte detection and enhanced sample throughput can be improved by the potential advantages of miniaturized systems. A high-throughput system able to digest and to perform peptide mapping of 100 protein samples within 3.5 hours has already been reported [6]. The main characteristic of this system is a microchip-immobilized enzyme reactor for on-line digestion of proteins within three minutes. Samples deposited into microvials containing crystallised matrix are then analysed automatically by a Maldi-Tof instrument.

A suggested miniaturization of devices for sample introduction into an electrospray MS source [7] may be a promise for integrated protein sample processing and mixture separation on a microchip that is coupled directly to an electrospray source. Gel-separated proteins are loaded onto the chip after enzymatic digestion with no sample clean-up step. On-chip CE is then used to achieve partial separation of the tryptic peptides. The device is coupled with automated tandem mass spectrometric analysis for protein identification [8].

Technology for mass spectrometry analysis of proteins and peptides

The most important improvement in proteomics has been the mass spectrometric identification of gel-separated proteins. Proteins are usually analysed by MS after enzymatic digestion for two main reasons: because problems are encountered in elution of separated proteins from the gel and because knowledge of the molecular weight of proteins is not usually sufficient for database identification. In contrast, peptides are easily eluted from gels and even a small set of peptide from a protein provides sufficient information for identification. The steps typically involved in the mass spectrometric analysis of a protein are illustrated in Figure 1. Ionisation techniques more frequently used for protein analysis are electrospray (or nanoEsi for smaller quantities) ionization (Esi) and matrix-assisted laser desorption ionization (Maldi). Mass analyzers are disparate and all have advantages and disadvantages. Maldi-Tof is more efficient for proteins with high

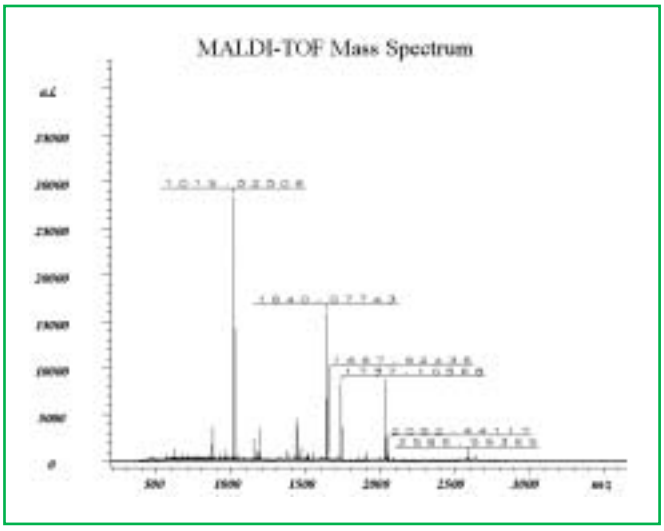

Figure 3 - Maldi-Tof mass spectrum

molecular mass while Esi-ion trap MS allows detection down to the femtomole level. For a more detailed review of Esi and Maldi techniques, the reader is referred to a review of the technologies.
MS analysis of gel-separated proteins and peptides gives different types of structural information. Direct information is obtained on the mass of a particular peptide and data from tandem mass spectra obtained either by postsource decay or collision-induced dissociation can generate amino acid sequences [9]. Further, mass shifts of peptide fragments as compared to those that would derived from the translated protein can provide data on glycosylation, phosphorylation and other post-translational modifications. Some developments in MS technology were directly addressed to proteomic analysis, particularly for the identification of proteins in macromolecular complexes [10] (Figure 2).
Two-dimensional chromatographic separation, peptide fragmentation, and comparison of amino acid sequences to genomic sequences represents a rapid method for protein identification that relies on the accuracy and predictive power of whole-genome sequences. A particular form of tandem mass spectrometry with sub-picomole sensitivity (0.3-3 pmol), high mass measurement accuracy (to +/-0.05%), and reduced sample-handling losses has been described in order to avoid the need for SDS-PAGE [11].

Peptides mass mapping or "mass fingerprint"
Maldi is the method of choice as a first step in protein analysis for several reasons. Many samples can be analysed at the same time on a single probe holder and measured in a single run. Combining Maldi with automated and highly sensitive preparation of the proteins for analysis, hundreds of proteins can be handled. The Maldi technique generates protonated molecular ions from analytes deposited on surfaces mixed with a matrix and this essential attribute has prompted several studies to evaluate the feasibility of protein or peptide Maldi analysis directly from electrophoresis gels or from membrane blots (Figure 3).
Strupat *et al.* [12] described the development of a method for obtaining the molecular masses of both proteins blotted onto membranes and proteolytic peptides arising from on-membrane digestion. Desorption from the membrane was achieved using an IR-Maldi that induces substantially less spontaneous fragmentation of glycosylated and phosphorylated peptides when compared to more widely diffused UV-Maldi and therefore facilitates detection of intact species. Maldi instruments equipped with reflectron Tof MS for the generation of structural data in peptide and protein analysis are in use.
Nevertheless, based on the following considerations, additional information obtained with this device is limited when compared to that deriving from Esi-MS. By the post-source decay method, decompositions between the ion source and the reflectron diminish the advantage deriving from controlled collisional activation of precursor ions. By stepping of the reflectron voltage, generation of product ion spectra results in poor resolution of the precursor and modest resolution of the products. Consequently time for spectrum acquisition is expanded up to 15 minutes.
Moreover, the Maldi process gives rise essentially to singly charged ions because the side chain lysine or arginine at the carboxy-terminal sequestrate a single proton away from the peptide backbone which is not useful for the promotion of fragmentation indicative of peptide sequence. A solution of this problem may be the derivatisation procedure proposed by Keough *et al.* [13], which implies the incorporation of a negatively charged group at the amino terminus.
As a consequence, gas-phase structures are produced by Maldi that incorporate, in addition to a proton at the carboxy-terminal residue an additional 'mobile' proton causing alternative cleavage processes at multiple points on the peptide backbone. However this approach has not received yet much attention, so that its applicability to diverse peptides is not available.
In order to overcome the above outlined limitations of conventional Maldi-Tof MS for structural analysis various attempts have been made. Both the incorporation of a Maldi into tandem quadrupole ion trap-Tof and tandem quadrupole-Tof instruments appear successful in promoting low energy collisionally activated decomposition (Cad) of peptide ions. More recently high energy collisional activation of peptide ions is obtained with a tandem Tof instrument with a new design of timed ion selector, that allows to select precursor ions with a window of four Thomsons (Th; m/z units).
By this approach leucine and isoleucine were differentiated through the formation of products originating form side chain cleavage. Orthogonal acceleration of ions by the combination of electrospray ionization into the Tof analyzer has increased the importance of these instruments in proteomics. In fact, as compared to a scanning linear quadrupole analyzer they can detect more ions among those produced in the ion source and consequently sensitivity is improved.

Peptide fragmentation
Unfortunately not all proteins can be identified by mass fingerprint alone. A large percentage of human proteins are still not represented full length in sequence database, small proteins sometimes do not result in a sufficient number of tryptic peptides for unambiguous identification and mixture of proteins can only be deconvoluted to their respective entries in the database with special interpretation. In many of those cases, a further analytical step by Esi-MS which is complementary to Maldi in proteome analysis.
A solution containing peptides deriving from the enzymatic digestion of gel-separated proteins (digestion carried out direct-

ly in-gel) is introduced with or without preliminary chromatographic separation through a capillary and dispersed at high voltage thus resulting a plume of droplets containing peptides. After desolvation, molecules remain charged by binding one or more protons.
After a mass spectrum is obtained, the instrument can automatically selectively choose one or more ions and promote their fragmentation by collision induced decomposition (Cid) with nitrogen or argon gas. Several series of product ions are then formed, from which complete or partial information on the amino acid sequence can be obtained. Particularly importance for proteomics has been the implementation of electrospray ionization on hybrid quadrupole/Tof instruments for tandem MS analyses.
The high detection efficiency of the Tof analyzer has permitted the recording of low energy Cid product ion spectra (of precursors selected by the quadrupole analyzer) with sensitivities up to two orders of magnitude superior to those achieved on tandem quadrupole instruments.
Instruments based on ion trapping appear to be potentially helpful for proteome applications, particularly due to the capability to perform multi-stage tandem MS, or MS n and to the rather limited costs. Moreover in these instruments ion/molecule reactions are easily achieved and this may represent a future improvement in structure evaluation.
Spectrometers based on Fourier-transform ion cyclotron resonance (FT-ICR) have also been tested for application in proteomics by several groups, and advantages have been demonstrated in terms of higher magnetic field strengths, separation of individual isotopic variants with high mass accuracy (±3 Da with a 112 kDa protein) and high sensitivity. Electron-capture dissociation (Ecd) [14] also show benefits for protein analysis because backbone cleavage occurs first, instead of cleavage of the residues introduced post-translationally such as glycosidic moieties.
Hydrolytic cleavage of the protein with enzymes, particularly trypsin, still now represents an almost obligatory step in proteome analysis. Nevertheless, in the future, improvement of instruments for tandem MS and of gas-phase ion chemistry may allow protein mapping without preliminary fragmentation of the molecule.

Informatics: softwares for protein identification

Considering the complex way to proteome identification, attention must be paid also to database searching that represents the third essential part of proteomics. Integration of experts in informatics represents the best for an efficient research in the proteome field.
Nevertheless, when direct involvement of informatics cannot be organised, a great help comes from a number of software and databases to which free access on Internet is given (Figure 4). Among these the Expert Protein Analysis System (ExPASy; http://www.expasy.ch) [15] is continuously updated and is at present the more complete.
There are various approaches to the identification of proteins from MS analytical data. Simple search from the molecular mass of the target protein is inappropriate and generally the search starts from the masses of proteolytic fragments. In this respect, accuracy in determining the mass of proteolytic fragments of the protein is relevant, particularly because incomplete purity of the analysed samples may result in the lack of

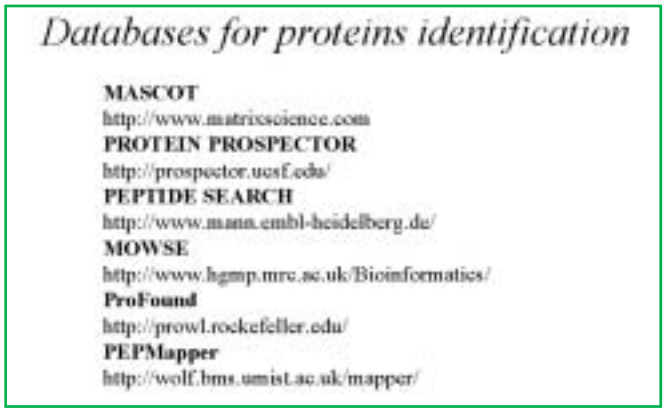

Figure 4 - Databases for proteins identification

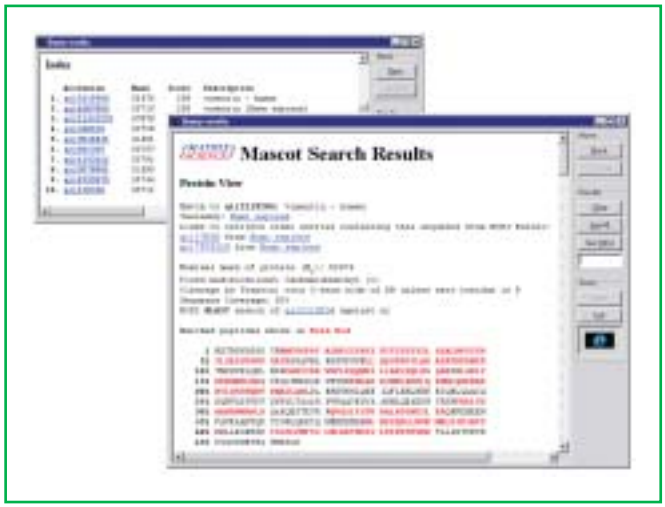

Figure 5 - Typical search result

specificity of the search. To solve this problem different approaches to database searching have been considered [16]. Simple mapping based on comparing peptide masses obtained by Maldi with those expected from the mass spectrum of each protein in the database usually results in higher scores to heavier proteins (typical search result is reported in Figure 5). Therefore search methods taking into account the protein size have been set up. There is also the need to improve tandem MS in order to limit the multiplicity of fragmentation pathways and to obtain sufficient sequence data at maximal sensitivities.
From this point of view in order to limit the excess of sequence data obtained in experiments carried out by tandem MS in the usual way derivatisation of the purified protein has been suggested in order to render more specific the fragmentation pattern has been suggested.
Peptide sequence data allow identification of complex protein mixtures because identification of a protein can be obtained from a single peptide, provided that data on the sequence of each peptide is obtained.

Applications

Referring to a proteome “standard cell condition”, a main goal in proteomics is to investigate modifications occurring under different states.
Increasing interest in this *differential proteome* analysis clear-

ly appears from the daily increase of literature data in this field. A great number of studies are focused on oncogenes, tumor suppressors, proteins regulating cell cycles and molecules involved in signal transduction in various cell types of different species [e.g. 17].
Important results are reported in cancer research, particularly related to leukaemia [18], breast cancer [e.g. 19] and bladder cancer [e.g. 20]. Other interesting results have been obtained on heart diseases [e.g. 21], neurological disorder [e.g. 22] and in the toxicology field [e.g. 23]. More recently efforts to define the *complete proteome* of biological fluids or tissues are reported [24-25].

Future expectations

A particular aspect of proteome research that has not been referred to in this review is the quantification of proteins. Isotope labelling appears to represent the best way to solve the problem. An used approach implies metabolic labelling of proteins with nitrogen-15.
This approach has limitations related to the amount of the target proteins in the analysed mixture. A new proposal for protein quantitation is based on the analysis of peptides obtained after derivatisation of cysteine residues of two different cell populations with a reagent either with the natural isotopic composition or labelled with stable isotopes [26]. Isotope distribution in peptides will represent a quantitative comparison of the same peptides in two different cell populations.

References

[1] P. Haynes *et al., Electrophoresis,* 1998, **19,** 1862.
[2] G.J. Opiteck *et al., Anal. BioChem.,* 1998, **258,** 349.
[3] W. Tong *et al., Anal. Chem.,* 1999, **71,** 2270.
[4] D. Figeys *et al., Anal. Chem.,* 1999, **71,** 2279.
[5] J. Rappsilber *et al., Anal. Chem.,* 2000, **72,** 267.
[6] S. Ekstrom *et al., Anal. Chem.,* 2000, **72,** 286.
[7] M. Wilm, M. Mann, *Anal. Chem.,* 1996, **68,** 1.
[8] J. Li *et al., Anal. Chem.,* 2000, **72,** 599.
[9] T. Keough *et al., Proc. Natl. Acad. Sci. USA,* 1999, **96,** 7131.
[10] A. Link *et al., Nat. Biotechnol.,* 1999, **17,** 676.
[11] J. Loo *et al., Electrophoresis,* 1999, **20,** 743.
[12] D. Schleuder *et al., Anal. Chem.,* 1999, **71,** 3238.
[13] T. Keough *et al., Proc. Natl. Acad. Sci.*, *USA,* 1999, **96,** 7131.
[14] R.A. Zubarev *et al., J. Am. Chem. Soc.,* 1998, **120,** 3265.
[15] ExPASy Molecular Biology Server on World Wide Web URL: http://www.expasy.ch/
[16] O.N. Jensen *et al., Anal. Chem.,* 1997, **69,** 4741.
[17] D. Hanahan, R.A. Weinberg, *Cell,* 2000, **100,** 57.
[18] R. Melhem *et al., Leukemia,* 1997, **11,** 1690.
[19] B. Franzen *et al., Br. J. Cancer,* 1996, **73,** 909.
[20] J.E. Celis *et al., Cancer Res.,* 1999, **59,** 3003.
[21] S. Pankuweit *et al., J. Mol. Cell. Cardiol.,* 1997, **29,** 77.
[22] P. Beaudry *et al., Dementia Geriatr. Cogn. Disord.,* 1999, **10,** 40.
[23] P. Cutler *et al., Electrophoresis,* 1999, **20,** 3647.
[24] C.S. Spahr *et al., Proteomics,* 2001, **1,** 108.
[25] P. Davidsson *et al., Proteomics,* 2001, **1,** 444.
[26] S. Gygi *et al., Nat. Biotechnol.,* 1999, **17,** 994.