Protein Cell 2012, 3(9): 641–647
DOI 10.1007/s13238-012-2079-5

Protein & Cell

MINI-REVIEW

Mass spectrometry based proteomics, background, status and future needs

Peter Roepstorff ✉

Department of Biochemistry and Molecular Biology, University of Southern Denmark Campusvej 55, DK 5230 Odense M, Denmark
✉ Correspondence: roe@bmb.sdu.dk

Received August 9, 2012 Accepted August 15, 2012

ABSTRACT

An overview of the background for proteomics and a description of the present state of art are given with a description of the main strategies in proteomics. The advantages and limitations of the two major strategies, 2D-gel based and LC-MS based, are discussed and a combination for the two, CeLC-MS is described. A number of challenging problems which have been solved using different proteomics strategies including the advantage of organell enrichment or modifications specific peptide isolation to get deeper into the proteome are described. Finally the present status and future needs discussed.

KEYWORDS proteomics, 2D-PAGE, LC-MS, isoforms, phosphorylation

INTRODUCTION

The concept of proteomics dates back long time before the term proteomics was introduced. The first descriptions of an approach which today would be called proteomics dates back to the early 1980's where Julio Celis developed protein separation by highly efficient 2D-PAGE followed by identification of the proteins in the spots by Edman degradation and comparison with available protein sequence databases (Bauw et al., 1989). During the 1980's a wealth of DNA sequence information was also generated mainly from mRNA's and it became obvious that mass spectrometric molecular weight information could be combined whit DNA sequence information to identify proteins and peptides derived from the mRNA's. One of the first attempts was termed peptide charting and the mass spectrometric molecular weight information obtained from different neuropeptides was used to identify these and their modifications by comparison with the precursor m RNA sequences (Feistner et al., 1989), At that time, the mass spectrometric techniques that allowed direct analysis of peptides and proteins were Fast Atom Bombardment and Plasma Desorption Mass spectrometry. Although allowing demonstration of the principles, these method suffered by several limitations which prevented their widespread use in biological studies. With the advent of MALDI (Karas and Hillenkamp, 1988) and ESI MS (Fenn et al., 1989) these limitations were overcome and shortly after a number of papers demonstrating the general use of MS for protein identification were published within a period of three months. They were all based on measuring the mass of the peptides obtained by tryptic digestion of the proteins (Mann et al., 1993; James et al., 1993; Pappin et al., 1993) and included the identification of proteins separated by 2D-PAGE after in gel digestion (Henzel et al., 1993), so called peptide mass fingerprinting. Soon it became clear that with the increasing size of the protein sequence databases the molecular weight information of the tryptic peptides was insufficient for confident protein identification. Therefore, Mann and Wilm (1994) introduced the concept of sequence tags that combined molecular mass information with partial sequence information generated by MS/MS for protein identification. The term proteomics in analogy with the term genomics was introduced rather late as the analysis of the complete protein complement from a cell, a tissue or an organism (Wilkins et al., 1996). Presently the omics concept is widely used for a large number of different types of analyses, e.g., glycomics, metabolomics and lipidomics. It also became clear that the complexity of the peptide mixtures derived by tryptic digestion of complex protein mixtures became far too complex for analysis without separation prior to the mass spectrometric analyses. Separation can take place on two levels, either the protein or the peptide level. These needs resulted in a boom in instrumental development for separation on both levels and in development of methods for selective enrichment of specific proteins or specific types of protein modifications. Also the mass spectrometers and the associated software underwent a dramatic development in terms of performance as well as user friendliness. The present state of art of proteomics is a result of a number of technological and conceptual developments and has in a surprisingly short time due to the demands and many challenges from the biological problems grown to a very high level.

PROTEOMICS STRATEGIES

Proteomics is performed on three different levels (Fig. 1). In the present account only the two first levels will be discussed.

Two dominant strategies are used in Expression proteomics (Fig. 2). The gel based strategy (Fig. 2 left side) is the oldest one but still has its justification for many studies. The advantage of 2D-PAGE is that it is by far the most efficient method for protein separation in terms of resolution and sensitivity. It also in many cases allows separation of different isoforms of the proteins, forms which can be either due to sequence variants or due to post translational modifications. Using a range of narrow range IPG strips in the first isoelectric focusing dimension up to 15,000 protein spots can be resolved. This, however, might only represent around five thousand gene products because most proteins are present in several forms which will be separated in distinct spots. The limitation is that the dynamic range is limited and that quantification of the proteins is normally based on the intensity of the spots as measured by image analysis. The latter also means that relative quantification between two different situations most frequently is performed based on two sets of gels unless quantification is performed using DIGE, which

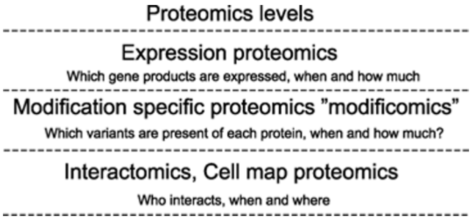
Proteomics levels
Expression proteomics
Which gene products are expressed, when and how much
Modification specific proteomics "modificomics"
Which variants are present of each protein, when and how much?
Interactomics, Cell map proteomics
Who interacts, when and where

Figure 1. The different levels of proteomics. Only the two first will be dealt with in this account.

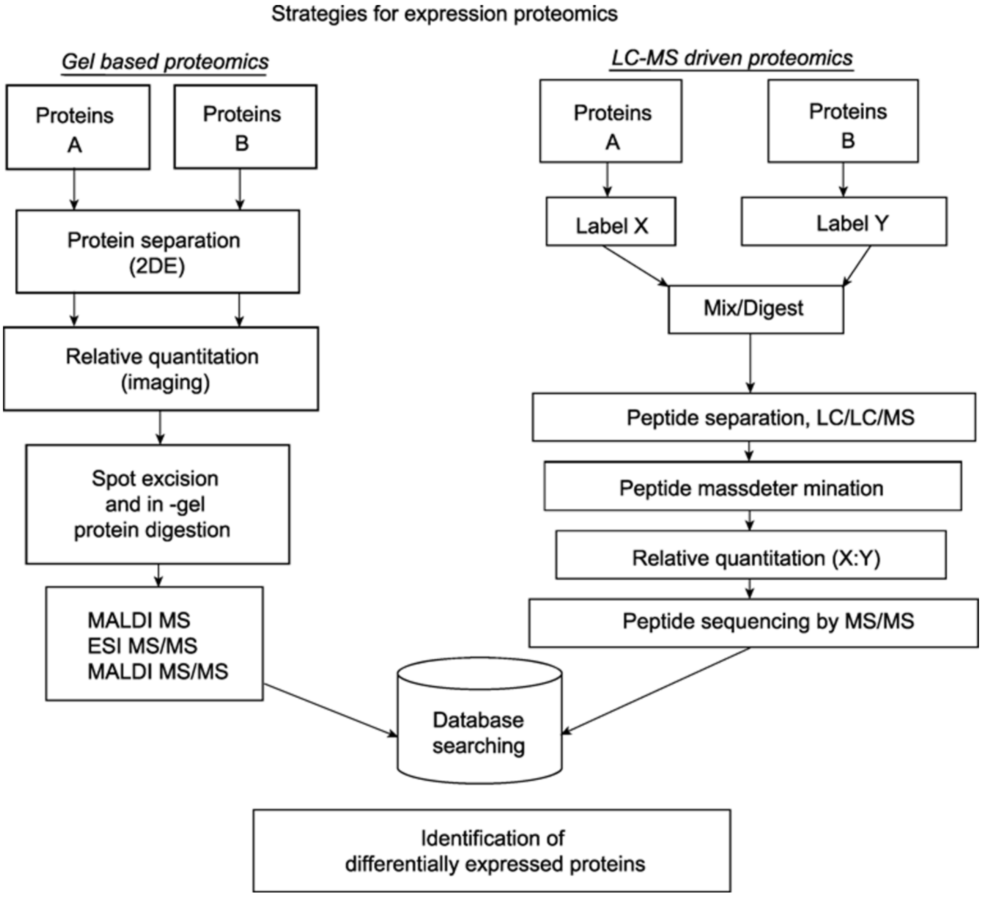
Strategies for expression proteomics
Gel based proteomics
Proteins A
Proteins B
Protein separation (2DE)
Relative quantitation (imaging)
Spot excision and in -gel protein digestion
MALDI MS
ESI MS/MS
MALDI MS/MS
LC-MS driven proteomics
Proteins A
Proteins B
Label X
Label Y
Mix/Digest
Peptide separation, LC/LC/MS
Peptide massdeter mination
Relative quantitation (X:Y)
Peptide sequencing by MS/MS
Database searching
Identification of differentially expressed proteins

Figure 2. The two major strategies in expression proteomics.

involves labeling of the different samples with different fluorophores which can then be quantified separately in one gel. Quantification might also be compromised if a gel spot contains more that one protein which in spite of the high resolution might often be the case. Protein identification in this strategy is performed after in-gel digestion and peptide mass fingerprinting by MS supplemented by sequencing of selected peptides by MS/MS. MALDI ToF/ToF instruments are ideal for protein identification in this strategy and also allow identification of several proteins in one spot if they are not separated in the electrophoretic procedure. In spite of the limitations of the gel based strategy, it has proven to be very efficient in solving biological questions and it is often a very good initial approach in a study.

In the LC-MS based strategy (Fig. 2 right side) the proteins are not separated prior to identification. The entire batch of proteins is enzymatically digested and the resulting peptides submitted to separation by liquid chromatography prior to mass spectrometric analysis. Since the resulting peptide mixture contains thousands of peptides a single chromatographic step is not sufficient and frequently multiple chromatographic steps are needed. The most common combination is ion exchange chromatography followed by RP-HPLC either directly on line with the mass spectrometer or in a two step procedure with fraction collection between the chromatographic steps. Relative quantification between two different situations is performed by stable isotope labeling either on the protein level or on the peptide level and the samples combined prior to separation. A list of commonly used labeling reagents is given in Table 1. In LC-MS based proteomics, protein identification is performed based on fragmentation of the peptides by MS/MS and subsequent comparison of the experimentally obtained spectra with the theoretical MS/MS spectra generated from the databases. The draw back in this strategy, also termed shot gun proteomics, is that most proteins are identified based on only a few peptides with the result that isoforms of the proteins most frequently cannot be distinguished and post translational modifications are often missed. The dynamic range is somewhat better than for the gel-based strategy, but due to instrumental limitations low abundant proteins often will be lost. The problem arises because the mass spectrometer can only sequence a limited number of peptides in a given time window and these will normally be the most intense peaks, whereas low abundant peaks will not be sequenced. It is possible to get deeper in the proteome by introducing exclusion lists, so that peaks which are sequenced in a first run are excluded in a second run. Technical replicates often only shows limited overlap because the selection of peptides to sequence might not be the same in consecutive runs. It is possible to get deeper in the proteome by applying separation procedures prior to the LC-MS based analysis. In our laboratory a preferred strategy (GeLC-MS/MS) is to separate the proteins by 1D-PAGE, slice the gel and perform in gel digestion of the proteins followed by LC-MS/MS, (Fig. 3). An alternative is to perform organelle isolation prior to the proteome analysis,. An example of that is given later in this paper.

MODIFICATION SPECIFIC PROTEOMICS

Analysis of the posttranslational modifications represents a special challenge because a given site in a protein might be only partially modified. In addition, many post translational modifications result in lower ionization yields of the modified peptides and consequently poorer sensitivity for their detection compared to the corresponding unmodified peptide, and many modifications are labile under MS/MS conditions with the result that information about their position might be lost in the sequencing step (Mann and Jensen, 2003). This means that specific enrichment of the modified peptides (or proteins) is needed for their detection as well as specific mass spectrometric conditions for maintenance of the positional information (Zao and Jensen, 2009). We have in our laboratory developed methods for the analysis of a number of modifications, the method for enrichment and the stability of the modification under MS/MS conditions are given in Table 2. The stability of the modifications upon MS/MS refers to fragmentation by collision induced dissociation (CID). New fragmentation methods, electron capture dissociation (ECD) and electron transfer dissociation (ETD), that allows peptide bond fragmentation without loss of the modifications, are now available. Of these ETD is especially relevant for proteomics since it can be performed on a number of mass spectrometers currently used in proteomics. The enrichment of post translationally modified proteins and peptides has the advantage that it allows more in dept analysis of the proteome because low abundant modified proteins that are not detected in expression proteomics might be identified after enrichment of the modified form.

CHALLENGES IN PROTEOMICS

In spite of the dramatic development in instrumentation and software for proteomics there are still a considerable number of challenges. In the following some examples from recent studies in our research group will be given.

Detection of low abundant proteins after organelle enrichment

Castor beans are a source of lipids, mainly of 12-hydroxy-oleate, of high value for biodiesel production. The residue after extraction of the lipids is used as animal food, but unfortunately contains the toxic protein ricin as well as a strong allergen, 2S-Albumin. To understand the processes in development of the castor beans, proteomics analysis of different stages were performed. As with most seeds the proteome is dominated by a rather small number of very abundant storage proteins (Fig. 4A) and to get deeper in the proteome,

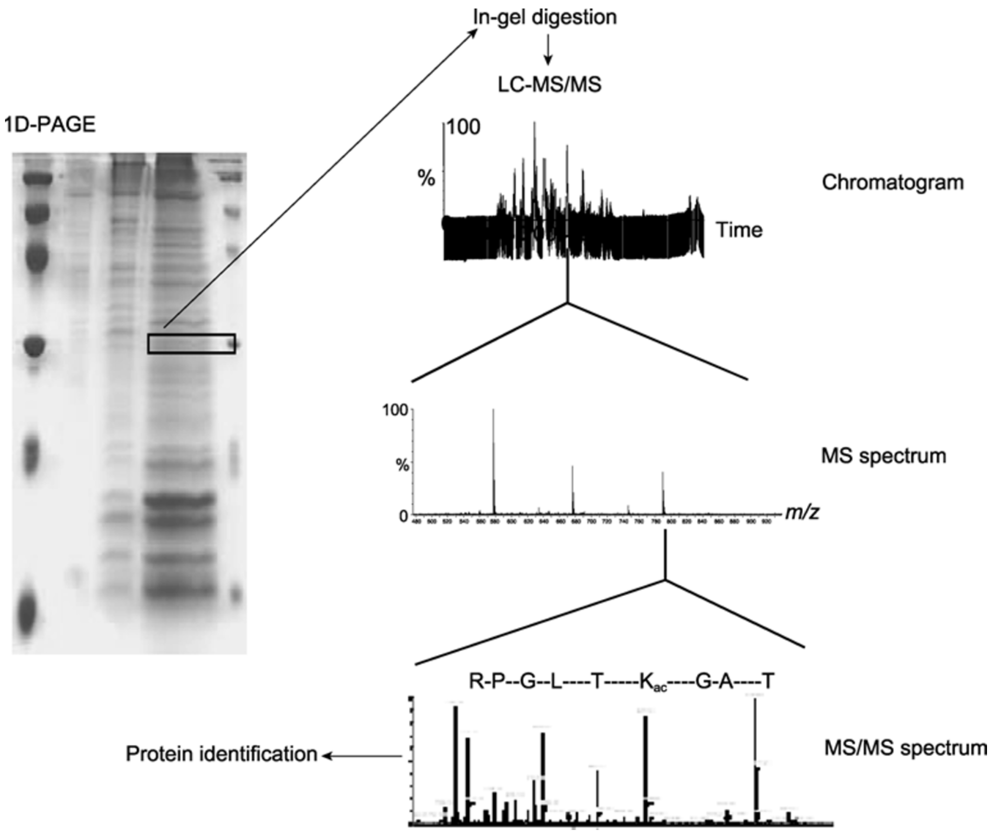

Figure 3. Combined gel-LC-MS based strategy (GeLC-MS).

Table 1. Some commonly used reagents used for isotope based quantification in LC-MS-based proteomics

Labeling reagent	Labeling level	Residue affected	Quantification by
SILAC (Stable isotope labeling in cell culture, ^{15}N, ^{13}C or ^{2}H)	Cell level	Many possible	MS
ICAT (isotope coded affinity tag)	Protein level	Cysteine	MS
Mass Tag, SPITC, CAF dimethyl, ICPL	Peptide level	N-terminal and Lysine	MS
^{18}O	Peptide level	C-terminal, Asp upon deglycosylation of Asn	MS
iTRAC, TMT, DiArt	Peptide level	N-terminal and Lysine	MS/MS

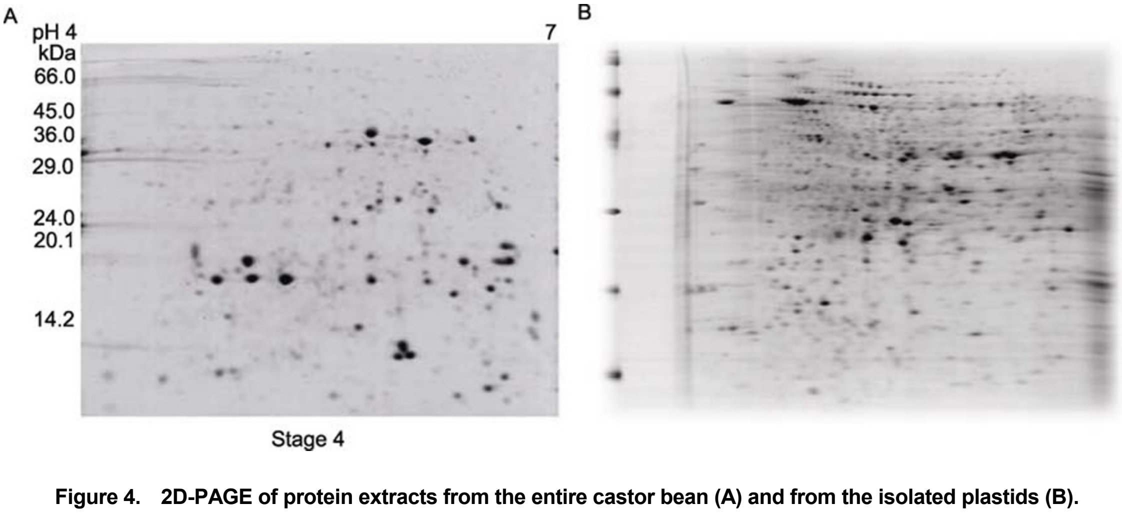

Figure 4. 2D-PAGE of protein extracts from the entire castor bean (A) and from the isolated plastids (B).

Table 2. List of modifications which we frequently analyze for in our research group and the methods used

Modification	Amino acid residues	Selective enrichment	Stable under MS or MS/MS conditions[1]
Phosphorylation	Ser, The, Tyr, His	IMAC or TiO_2	No, except Tyr
Phosphorylation	Tyr	Antibodies	Yes
Oxidation	Met, Cys, Trp	None	Yes
Methylation	Lys, Arg, His	Antibodies	Yes
Acetylation	N-terminal, Lys	Antibodies	Yes
Hydroxylation	Pro, Asp	None	Yes
Carboxylation	Glu	None	No
Glycosylation	Asn, Thr, Ser	HILIC, TiO_2, Hydrazide chemistry[2]	No
PyroGlutamate	N-terminal Glu	None	Yes
Nitration	Tyr	Reduction and biotinylation or COFRADIC[3]	No
Halogenation	Aromatic residues	None	Yes
Carbonylation	Many residues	Hydrazide chemistry	Yes

[1] Refers to fragmentation by CID. Most are stable under ETD fragmentation.
[2] An appropriate tag is attached by hydrazide chemistry, followed by affinity purification.
[3] COFRADIC refers a two-step chromatographic purification with an intermediate reaction to change the retention pattern of the modified peptides (Larsen et al., 2011).

plastids which are crucial for the production of the lipids were isolated allowing detection of a large number of proteins which were not visible in the gels obtained with the entire seeds (Fig. 4B). This study was followed by a study of the nucellus isolated from the same stages of castor bean development using GeLC-MS/MS approach (Nogueira et al., 2012). The studies of the plastids and the nucelli yielded highly complementary information about the castor bean development and revealed that ricin in spite of previous assumptions was already present in the nucelli.

Determination of isoforms

2D-PAGE based proteomics is the ideal strategy for investigation of protein isoforms and processing. In beer production the barley cultivars used for malt production might influence the quality of the beer. In order to investigate if the different barely cultivars could be distinguished based on their content of isoforms, 2D-Page was performed on the proteins isolated during seed development, from mature seeds and during germination. The study was concentrated on peroxidase isoforms because the barley genome contains genes for at least 6 peroxidases. In total 13 spots (Fig. 5) were found to contain products from the peroxidase genes representing two known peroxidase genes for one of which the protein had not previously been found. In addition a third protein matching a putative peroxidase gene was identified having 86% identity to a wheat peroxidase indicating that it most likely was also an active peroxidase. Additional variants of these proteins were found to be glycosylated or truncated. To identify specific protein isoforms and their truncated forms it is crucial to obtain full-sequence coverage. 99%–100% sequence coverage was obtained for all but one of the spots by in gel digestion with three different enzymes, trypsin and endoproteinases AspN and LysC. The study demonstrated that the peroxidase isoform pattern was dependent on the tissue, the developmental stage of the grain, the time after germination and the cultivars (Laugesen et al., 2007).

Quantitative analysis of protein modification events

Quantitative time resolved analysis of modification events also represents a major challenge. In a recent study the time resolved quantitative phosphorylation in signaling cascades as function of stimulation with angiotensin 1–7 was investigated. Human aortic endothelial cells were stimulated with angiotensin 1–7 and samples taken at 0, 3, 5 and 20 min, digested in solution and labeled with four plex iTRAQ reagents, followed by specific enrichment of the phosphopep-

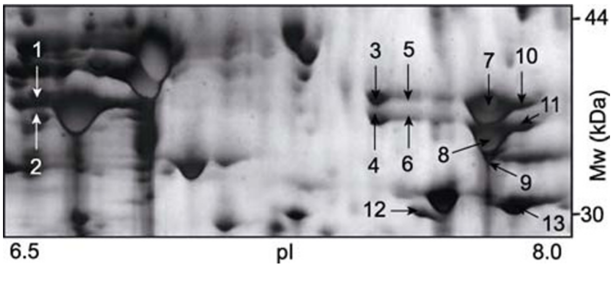

Figure 5. Section of a 2D-gel containing the spots corresponding to barley peroxidase isoforms. Spots 1 and 2 are identified as the major barley peroxidase isoform BP1, spots 3-1 to the hither too never observed isoform BSSP1 and spots 12 and 13 to the novel isoform homologous to the known wheat peroxidase. The variants with the highest molecular weight of each isoform are glycosylated.

tides and analyzed by LC-MS/MS on an orbitrap XL mass spectrometer using peptide fragmentation by both low and high energy collision. The phosphopeptides were isolated using a recently in our research group developed strategy for highly efficient phosphopeptide enrichment (Engholm-Keller et al., 2011) resulting in 99% of all isolated peptides being phosphopeptides. 1288 unique phosphosites on 699 different proteins were identified, of which 121 sites on 79 proteins were differentially regulated indicating that two different signaling pathways were affected by angiotensin 1–7 stimulation. Some sites were up regulated and others were down regulated as function of time whereas others showed transient regulation. The site specific assignment was important since different phosphorylation sites in a protein might be differently regulated. The manuscript also describes some of the many pitfalls when interpreting quantitative proteomics data (Verano-Braga et al., 2012).

Proteomics of organisms without known genome

Protein identification by proteomics is based on sequence information for the studied organism is available in databases. Unfortunately, this is not the case for a major part of all living organisms. Therefore working with such organisms represents a real challenge. Recently, we were challenged by a group of biologists who studied the leaf-cutting ant that grows fungi on the leaves in their underground nest gardens. Prior to inoculation with the fungi the ants place fecal droplets on the leaves. The question was why? Fecal droplets from 50 ants were collected and submitted to 1D-PAGE. Surprisingly, several proteins seemed to intact in the fecal droplets. The corresponding bands were in gel digested and the digests analyzed by MALDI-MS/MS. Since no genomic information was available for the ant and the fungus, manual *de novo* sequencing was performed and a number of reliable sequences obtained. Upon BLAST search they all seemed to be related to known fungal pectinolytic peptides, so most likely the proteins were derived from the fungi and passed undigested through the intestines of the ants. Based on the obtained sequences probes were constructed and the corresponding genes sequenced and cloned. They were as expected related to known fungal pectinolytic enzymes showing a symbotic development between the ants and the fungi they eat. The ants eat the fungi and up concentrates the plant cell wall degrading enzymes from the fungus in the fecal droplets which then are used to facilitate the fungal infection of the leaves in the garden. To confirm this, some ants were fed with sugar and others with fungi. The activity of a number of relevant pectinolytic enzymes in the ant feces was then tested with standard enzymatic assays. No enzymatic activity was observed for ants fed with sugar and high activity for ants fed with fungi (Schiøtt et al., 2010).

Status and future needs

Proteomics has come off its age and is now a standard tool in many biological studies. The fast development is partly due to the rapidly increased amount of genomic data which are essential for proteomics and partly to the rapid instrument development which has been strongly influenced by the acceptance of proteomics in the biological community. Improved electrophoretic equipment and high performance capillary chromatographic systems are now available and not least a large number of new mass spectrometers designed for use in the biological laboratory and not requiring experts in mass spectrometry. Typical examples are MALDI-ToF/ToF, Q-ToF and orbitrap instruments which have pushed the limits dramatically. The associated software has also undergone a dramatic improvement over the years. This means that the methods are available for investigating numerous basic and applied questions in biology and medicine. However, many scientists have not fully appreciated the importance of addressing the right question, of selecting the appropriate strategies for sample preparation and proteomics strategy. Proteins are individuals and cannot be handled with standard kits like the ones used in genomics. Therefore any project needs a well defined question, a careful planning of the experimental conditions and the strategy to be used and not least a validation of the results to understand the biological context of the observations. It in not enough just to get a list of identified proteins or phosphorylation sites without relating it to the biological question.

In spite of the rapid development of proteomics there are still needs for the future. A major problem is that the present dominating bottom up strategy rarely comes close to a complete sequence coverage, which means that information about splice variants, isoforms and post translational modifications often is lost. There is increasing evidence that these are very important for the function of the living organisms. The only way to ensure complete coverage of the sequences and observation of all the variants of a protein is by analysis of the intact molecules, also termed top down proteomics. Unfortunately, this has only to a limited extend been possible until recently and only with the very expensive and expertise demanding FT-ICR instruments (Boyne et al., 2006). Recently it seems that the orbitrap instrument can achieve top down analysis of at least medium size proteins and generate sufficient sequence information to identify the protein and regions containing modifications (Ahlf et al., 2012). However, another obstacle is that the methods for high resolution and sensitivity separation of the proteins for top down proteomics either are not compatible with the subsequent mass spectrometric analysis, e.g. 2D-PAGE, or introduce too big losses of protein or poor resolution, most chromatographic techniques. There is a big need for development of improved methods for protein separation compatible with mass spectrometric analysis. The available methods for enrichment of modified proteins and peptides only cover a small fraction of all the more that 300 known protein modifications and additional methods need to be developed. The software for data analysis also need improvement to be able to handle the huge amount of data generated in proteomics experiments. This is especially true for software to assignment of protein modifications and development of reliable software for de novo sequencing, which in spite of several software packages on the market still can only be reliably performed by very time consuming manual interpretation of the spectra. In conclusion, proteomics has come far during the past two decades and has gained broad acceptance in the scientific community. There is still need for improvements and also a need for education of the users of proteomics techniques.

REFERENCES

Ahlf, D.R, Compton, P.D., Tran, J.C., Early, B.P, Thomas, P.M., and Kelleher, N.L. (2012). Evaluation of the compact high-field orbitrap for top-down proteomics of human cells. J Proteome Res. (In press).

Bauw, G., Vandamme, J., Puype, M., Vandekerchove, J., Gesser, B., Ratz, G.P., Lauritsen, J.B., and Celis, J.E. (1989). Protein-electroblotting and protein-microsequencing strategies in generating protein data-bases from two-dimensional gels. (computerized protein data-bases human genome sequencing). Proc Natl Acad Sci U S A 86, 7701–7705.

Boyne II, M.T., Pesavento, J.J., Mizzen, C.A., and Kelleher, N.L. (2006). Precise characterization of human histones in the H2A gene family by top down mass spectrometry. J Proteome Res 5, 248–253.

Engholm-Keller, K., Hansen, T.A., Palmisano, G., and Larsen, M.R. (2011). Multidimensional strategy for sensitive phosphoproteomics incorporating protein prefractionation combined with SIMAC, HILIC, and TiO(2) chromatography applied to proximal EGF signaling. J Proteome Res 10, 5383–5397.

Feistner, G.J., Højrup, P., Evans, C.J., Barofsky, D.F., Faull, K.F., and Roepstorff, P. (1989). Mass spectrometric charting of bovine posterior/interior pituitary peptides. Proc Natl Acad Sci U S A 86, 6013–6017.

Fenn, J.B., Mann, M., Meng, C.K., Wong, S.F., and Whitehouse, C.M. (1989). Electrospray ionization for the mass spectrometry of large biomolecules. Science 246, 64–71.

Henzel, W.J., Billeci, T.M., Stults, J.T., and Wong, S.C. (1993). Identifying proteins from 2-dimensional gels by molecular mass searching of peptide-fragments in protein sequence databases. Proc Natl Acad Sci U S A 90, 5011–5015.

James, P., Quadroni, M., Carafoli, E., and Gonnet, G. (1993). Protein identification by mass profile fingerprinting. Biochem Biophys Res Commun 195, 58–64.

Karas, M., and Hillenkamp, F. (1988). Laser desorption ionization of proteins with molecular masses exceeding 10000 daltons. Anal Chem 60, 1299–2301.

Mann, M., Højrup, P., and Roepstorff, P. (1993). Use of mass spectrometric molecular weight information to identify proteins in sequence databases. Biol Mass Spectrom 22, 338-345.

Mann, M., and Wilm, M. (1994). Error tolerant identification of peptides in sequence databases by peptide sequence tags. Anal Chem 66, 4390–4399.

Mann, M., and Jensen, O.N. (2003). Proteomic analysis of post-translational modifications. Nat Biotechnol 21, 255–261.

Larsen, T.R., Bache, N., Gramsbergen, J.B., and Roepstorff, P. (2011). Identification of nitrotyrosine containing peptides using combined fractional diagonal chromatography (COFRADIC) and off-line nano-LC-MALDI. J Am Soc Mass Spectrom 22, 989–996.

Laugesen, S., Bak-Jensen, K.S., Hägglund, P., Henriksen, A., Finnie, C., Svensson, B., and Roepstorff, P. (2007). Barley peroxidase isozymes.Expression and post-translational modification in mature seeds as identified by two-dimensional gel electrophoresis and mass spectrometry. Intl J Mass Spectrometry 268, 244–253.

Nogueira, F.C., Palmisano, G., Soares, E.L., Shah, M., Soares, A.A., Roepstorff, P., Campos, F.A., and Domont, G.B. (2012). Proteomic profile of the nucellus of castor bean (Ricinus communis L.) seeds during development. J Proteomics 75, 1933–1939.

Pappin, D.J.C., Højrup, P., and Bleasby, A.J. (1993) Rapid identification of proteins by peptide-mass finger printing. Curr Biol 3, 327–332

Schiøtt, M., Rogowska-Wrzesinska, A., Roepstorff, P., and Boomsma, J.J. (2010). Leaf-cutting ant fungi produce cell wall degrading pectinase complexes reminiscent of phytopathogenic fungi. BMC Biol 8, 156–168.

Wilkins, M.R., Pasquali, C., Appel, R.D., Ou, K., Golaz, O., Sanchez, J.C., Yan, J.X., Gooley, A.A., Hughes, G., Humphery-Smith, I., et al. (1996). From proteins to proteomes: Large scale protein identification by two-dimensional electrophoresis and amino acid analysis. Biotechnology 14, 61–65.

Verano-Braga, T., Schwämmle, V., Sylvester, M., Passos-Silva, D.G., Peluso, A.A., Etelvino, G.M., Santos, R.A., and Roepstorff, P. (2012). Time-resolved quantitative phosphoproteomics: new insights into angiotensin-(1-7) signaling networks in human endothelial cells. J Proteome Res 11, 3370–3381.

Zhao, Y., and Jensen, O.N. (2009). Modification-specific proteomics: Strategies for characterization of post-translational modifications using enrichment techniques. Proteomics 9, 4632–4641.