Mass spectrometry and proteomics
Steven P Gygi* and Ruedi Aebersold†

Proteomics is the systematic analysis of the proteins expressed by a cell or tissue, and mass spectrometry is its essential analytical tool. In the past two years, incremental advances in standard proteome technology have increased the speed of protein identification with higher levels of automation and sensitivity. Furthermore, new approaches have provided landmark advances in determining functionally relevant properties of proteins, including their quantity and involvement within protein complexes.

Introduction
A core component of proteomics is the ability to systematically identify every protein expressed in a cell or tissue as well as to determine the salient properties of each protein (e.g. abundance, state of modification, involvement in multi-protein complexes, etc.). The technology for such analyses integrates separation science for the separation of proteins and peptides, analytical science for the identification and quantification of the analytes, and bioinformatics for data management and analysis. Its initial implementation consisted of the combination of high-resolution two-dimensional gel electrophoresis (2DE), using IEF (isoelectric focusing)/SDS-PAGE gel, for the separation, detection and quantification of individual proteins present in a complex sample with mass spectrometry and sequence database searching for the identification of the separated proteins. A commonly used method is schematically illustrated in Figure 1. This technique and variations thereof (for review see [1]) have been used to identify and catalog large numbers of proteins present in a complex sample and to represent them in a proteome database, a process we refer to here as ‘descriptive proteomics’. For example, Shevchenko et al. [2] systematically identified 150 yeast proteins from 2D gels. Numerous such annotated databases are now accessible. The same techniques have also been used as a global discovery tool to detect dynamic changes in the proteome of a cell or tissue in response to external or internal perturbations. Because the detection of dynamic changes requires accurate quantification of each detected component, we use the term ‘quantitative proteomics’.

In this report we summarize the most significant developments related to proteomics and mass spectrometry as they have been reported from January 1999 to April 2000. Advances in core mass spectrometry technology have led to further refinements of the 2DE-based proteomics methods. They have also catalyzed alternative approaches to the traditional gel-based methods, such as the introduction of accurate protein quantification based on isotope dilution theory and the systematic analysis of protein complexes.

Advances in MS technology for proteome analysis
In this section we summarize advances in MS instruments, their control and operation, and progress in the searching tools used for the identification of proteins by correlating mass spectrometric data with sequence databases.

The performance of existing types of mass spectrometers for proteomics research has incrementally improved as new types of mass spectrometers were introduced. The instruments most commonly used throughout the review period can be grouped into two categories: single stage mass spectrometers and tandem MS-based systems. Single stage mass spectrometers, most notably the matrix-assisted laser desorption ionization (MALDI) time-of-flight (TOF) instruments, were used in numerous projects for large-scale protein identification by the peptide mass mapping technique. This method is particularly successful for the identification of proteins from species with smaller and completely sequenced genomes [3,4]. Tandem MS instruments such as the triple quadrupole, ion-trap, and the recently introduced hybrid quadrupole-time-of-flight (Q-TOF) were routinely applied in LC–MS/MS or nanospray experiments with electrospray ionization (ESI) to generate peptide fragment ion spectra suitable for protein identification by sequence database searching. The increased use of instrument control programs to automatically select specific peptide ions for collision-induced dissociation (CID) (data-dependent CID) was a notable trend with these MS/MS instruments.

Several new configurations of mass spectrometers with a high potential to impact proteomics research were introduced. Two groups have recently coupled a MALDI ionization source to a hybrid Q-TOF instrument [5,6]. The outstanding mass accuracy and sensitivity afforded by a Q-TOF instrument enhance the database search results and also make it the instrument of choice for de novo sequencing.
of MS/MS spectra. The MALDI–Q-TOF configuration provides exciting opportunities for automation and high-throughput applications and the possibility to archive samples on the sample plate for re-analysis at a later time. Medzihradszky et al. [7] described a different hybrid instrument referred to as MALDI TOF−TOF. This instrument shares many of the advantages of the MALDI–Q-TOF and is additionally capable of high-energy CID and very fast scan rates. The Fourier-transform ion cyclotron resonance (FT-ICR) MS is relatively new to proteomics. These instruments have extremely high sensitivity and resolution with mass accuracy that can exceed 1 ppm. These properties were used to measure and quantitate intact molecular masses of hundreds of proteins in a single analysis [8•]. Goodlett et al. [9•] showed that the accurate mass of a single peptide, measured by FT-MS, along with easily obtainable constraints can be used to identify proteins by sequence database searching.

Proteomics would be impossible without software tools to correlate mass spectrometric data with sequence databases. Existing database searching programs have become both more sophisticated and (Web) accessible. In addition, new algorithms were introduced. The major relevant programs are Sequest [10], MASCOT [11], PeptideSearch [12], PROWL [13], and Protein Prospector [14]. Of these, Sequest sets the benchmark for protein identification using CID spectra because it works best with marginal MS/MS data, is highly reliable, automatically analyses the data from entire LC-MS/MS experiments, and requires no user interpretation. Of the programs mentioned, however, Sequest alone does not allow searching on the internet. MASCOT is a new, speedy, Web accessible and versatile program with capabilities for peptide mass fingerprinting, database searching with partially interpreted (i.e. a sequence tag is manually interpreted), and uninterpreted CID spectra.
Advances in descriptive proteomics

At the beginning of the review period, essentially all proteome projects were based on a combination of 2DE for protein separation, visualization and quantification and mass spectrometry for protein identification. This approach has been advanced by the developments in MS described above, by incremental improvements to 2DE, and by innovative combinations of gel electrophoresis and MS. Improvements to 2DE include the introduction of new fluorescent staining methods providing higher sensitivity and larger dynamic range compared with silver staining [15•] and increased resolution by expanding the pI range of the first dimension (zoom gels) and prefractionation of complex protein samples prior to 2DE (for review see [16•]). Binz et al. [17] described a new method for the systematic analysis of proteins separated by 2DE. All the proteins in a 2D gel are concurrently digested by transblotting through a membrane that is covalently derivatized with trypsin. The generated peptides are then trapped on a membrane and identified by MALDI−TOF mass fingerprinting. The overall 2DE−MS approach has been used to generate annotated 2D gel databases for a myriad of cell types from different species. A partial list of such resources is available via the internet (www.lecb.ncifrf.gov/EP/table2Ddatabases.html).

Although these improvements advanced 2DE-based proteome technology, they did not address the fundamental limitations of the method for detecting specific classes of proteins, including those of low abundance, poor solubility, very small or large size and extreme pI. Several groups therefore explored the possibility of replacing one or both gel electrophoresis dimensions in proteome projects with alternative separation methods. Loo et al. [18] replaced the SDS-PAGE dimension of 2DE by scanning IEF gels directly with a MALDI−TOF mass spectrometer, thus generating a ‘virtual 2D gel’ image in which the protein mass is measured by mass spectrometry. Oda et al. [19••] replaced the IEF dimension of 2DE by preparative reversed phase (RP)-HPLC, and Wall et al. [20] used preparative IEF performed in solution followed by RP-HPLC on nonporous resins to separate proteins prior to analysis by MALDI−TOF-MS. Link et al. [21••] completely eliminated any protein separation. They analyzed complex protein mixtures by digesting the unseparated protein sample and analyzing the resulting peptide mixture by two-dimensional (strong cation exchange/RP) chromatography (LC/LC) coupled on-line to an ESI-MS/MS instrument. Using a similar LC/LC−MS/MS, we successfully detected low abundance yeast proteins and thus demonstrated that this approach can overcome the limited dynamic range of 2D gels [22••].

The objective of these advances has been a proteome analysis technology with higher throughput, greater automation and increased comprehensiveness. It can be expected that these developments will continue and possibly be accelerated by the application of microfabrication technology. Early examples of this include the construction of sample-handling devices [23,24] and surface-enhanced laser desorption/ionization (SELDI) protein chip arrays for isolating and analyzing proteins and peptides with specific properties [25]. Although these methods are likely to eventually detect and identify every protein in a sample, with the exception of 2DE they are inherently not quantitative.

Quantitative proteomics

To add a quantitative dimension to non-2DE-based proteome analyses, the venerable technique of stable-isotope labeling [26] has been adapted for protein analysis. The method involves the addition to a sample of chemically identical but stable isotopically labeled internal standards (e.g. using 2H, 13C, 15N, etc.). Because ionization efficiency is highly variable for different peptides, the only suitable internal standard for a candidate peptide is that very peptide labeled with stable isotopes. Quantitative protein profiling is therefore accomplished when a protein mixture (reference sample) is compared with a second sample containing the same proteins at different abundances and labeled with heavy stable isotopes. In theory, all the peptides in the sample then exist in analyte pairs of identical sequence but different mass. Because the peptide pairs have the same physico-chemical properties, they are expected to behave identically during isolation, separation and ionization. Thus, the ratio of intensities of the lower and upper mass components provides an accurate measure of the relative abundance of the peptides (and hence the protein) in the original protein mixtures. Three groups have independently reported measuring quantitative protein profiles based on stable isotopes [8•,19•,26••], and two others are preparing manuscripts (H Langen et al., personal communication; P James et al., personal communication). The techniques differ in the method of incorporation of heavy isotopes and in the analytical procedures used.

Oda et al. [19••] grew one yeast culture on medium containing the natural abundance of the isotopes of nitrogen (14N, 99.6%; 15N, 0.4%), while another culture was grown on the same medium enriched in 15N (96%). After an appropriate growing period, the cell pools were combined, and proteins of interest were extracted and separated by RP-HPLC and then by SDS-PAGE. In-gel digestion of excised spots of interest resulted in peptide fragments, which were identified by peptide mass mapping. Each 15N that was incorporated shifted the mass of any given peptide upwards, leading to a paired peak for each peptide. The authors measured protein expression of 42 high-abundance proteins derived from two pools of Schizosaccharomyces cerevisiae that differed only in their ability to express the G1 cyclin CLN2. The percentage error of the experimental technique was found to be excellent (±10%). The authors went on to measure differential phosphorylation states in the yeast protein, Ste20, by the same technique. Pasa-Tolic et al. [8•] used stable-isotope-depleted media to impart a specific isotope signature into proteins. They compared the cadmium stress response in Escherichia coli grown in normal and rare-isotope-depleted (13C-, 15N- and 2H-depleted) media. Intact
protein mass measurements were carried out by FT-ICR MS. While no protein was positively identified, the expression ratios for 200 different proteins were compared.

Clearly, stable isotope metabolic protein labeling using $^{15}$N-enriched or depleted media permits quantitative protein profiling either in conjunction with 2DE, or other separation techniques. However, this method has several disadvantages. First, the method does not allow for the analysis of proteins directly from tissue. Second, the stable-isotope-enriched media are costly and may themselves affect cellular growth and protein production. Third, the increase in nominal mass because of stable-isotope incorporation is not known until the sequence is determined. Therefore protein identification must necessarily precede quantification.

We have recently published a novel method for quantitative protein profiling based on isotope-coded affinity tags (ICAT) [27••]. In this method (Figure 2), the stable isotopes are incorporated post isolation by selective alkylation of cysteines with either a heavy (d8) or light (d0) reagent. The two protein mixtures are then mixed. At this point, any optional fractionation technique can be performed to enrich for low abundance proteins or to reduce the complexity of the mixture, while the relative quantities are strictly maintained. Prior to analysis, the protein mixture is digested with trypsin. ICAT-labeled peptides are isolated by biotin-affinity chromatography and then analyzed by online HPLC coupled to a tandem mass spectrometer. The ratio of the ion intensities for an ICAT-labeled pair quantifies the relative abundance of its parent protein in the original cell state. In addition, the tandem mass spectrum reveals the sequence of the peptide and unambiguously identifies the protein. This strategy results in the quantification and identification of all protein components in a mixture. It is, in theory, applicable to protein mixtures as complex as the entire proteome.

There are several advantages to the ICAT strategy. First, the method is compatible with any amount of protein harvested from bodily fluids, cells or tissues under any
growth conditions. Second, the alklylation reaction is highly specific and occurs in the presence of salts, detergents, and stabilizers (e.g. SDS, urea, guanidine-HCl). Third, the complexity of the peptide mixture is reduced by isolating only cysteine-containing peptides. Fourth, the ICAT strategy permits almost any type of biochemical, immunological, or physical fractionation, which makes it compatible with the analysis of low-abundance proteins. There are two disadvantages to the method. First, the size of the ICAT label (~500 Da) is a relatively large modification that remains on each peptide throughout the MS analysis. This can complicate the database searching algorithms, especially for small peptides (<7 amino acids). Second, the method fails for proteins that contain no cysteines. Only a small percentage of proteins are cysteine-free (8% in yeast), however, and ICAT reagents with specificities to groups other than thiols could be synthesized.

Analysis of protein complexes

Most cellular functions are not performed by individual proteins but rather by protein assemblies, also termed multi-protein complexes. It is rightly assumed that proteins which specifically interact also partake in the same function. The identification of specifically interacting proteins is, therefore, a critical component of the proteomics because it directly relates to protein function within biological processes. In general, the methods described above for the analysis of protein mixtures are also well suited for the analysis of protein complexes. Indeed, some of the most scientifically rewarding applications of protein mass spectrometry have been from this arena. Link et al. [21**] identified more than 70 proteins present in the yeast ribosome in a single analysis using LC/LC−MS/MS. Rout et al. [28**] exhaustively analyzed the composition, architecture and transport mechanism of the yeast nuclear pore complex; Rappaport et al. [29*] have utilized chemical cross-linking and MS to examine the spatial organization of multi-protein complexes; and Heller et al. [30] have examined the components of the T-cell receptor complex. Such projects critically depend on the ability to cleanly isolate the target complex in good yields. To this end, Bouvier et al. [31**] have developed a tandem affinity purification (TAP) method and demonstrated its impressive efficiency by examining the yeast spliceosome.

Conclusions

A main strength of proteomics is the ability to analyze the dynamics of biological processes by the systematic analysis of expressed proteins. The technical advances described in this review, in particular the ability to measure accurately the quantitative changes induced by perturbations on large numbers of proteins and the ability to analyze functional protein complexes, add significantly to our ability to study biological processes and systems from a global standpoint. The coming year will likely be both impressive and exciting in the field of proteomics. Continued technical advances will enable researchers to achieve true proteome analysis (i.e. the analysis of all proteins expressed in a cell) and enable further types of global measurements that are essential for the definition of the functional state of proteins.

Acknowledgements

This work was supported by grants from the National Institutes of Health (HG00041, RR11823, T32HG00035, CA84698, A114109), National Science Foundation (BIR 9214821) and Merck Genome Research Institute.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

* of special interest
** of outstanding interest


This paper describes sensitive, reversible protein staining using Sypro Tangerine.


22. Gygi SG, Corthals GL, Zhang Y, Roochon Y, Aebersold R: Evaluation of two-dimensional gel electrophoresis based proteome analysis technology. *Proc Natl Acad Sci USA* 2000, in press. This paper thoroughly assesses 2D-gel-based proteome analysis. The classes of proteins identified from 2D gels were examined. It concludes that low abundance proteins cannot be detected from proteome analysis of unfractionated cell lysates.