Mass spectrometry from genomics to proteomics

Large-scale DNA sequencing has stimulated the development of proteomics by providing a sequence infrastructure for protein analysis. Rapid and automated protein identification can be achieved by searching protein and nucleotide sequence databases directly with data generated by mass spectrometry. A high-throughput and large-scale approach to identifying proteins has been the result. These technological changes have advanced protein expression studies and the identification of proteins in complexes, two types of studies that are essential in deciphering the networks of proteins that are involved in biological processes.

The elucidation of an organism’s genome is the first and important step towards understanding its biology, and the data created by whole-genome sequencing have significant benefits in fields outside those of genomics and bioinformatics. One area to benefit is that of proteomics. The term proteomics, or more appropriately functional proteomics, describes the ability to apply global (proteome-wide or system-wide) experimental approaches to assess protein function. Proteomics has emerged as a new experimental approach in part because mass spectrometry has simplified protein analysis and characterization, and several important and recent innovations have extended the capability of mass spectrometry.

Mass spectrometry of biological molecules

Mass spectrometers consist of three essential parts (Fig. 1). The first, an ionization source, converts molecules into gas-phase ions. Once ions are created, individual mass-to-charge ratios (m/z) are separated by a second device, a mass analyzer, and transferred to the third, an ion detector. A mass analyzer uses a physical property [e.g. electric or magnetic fields, or time-of-flight (TOF)] to separate ions of a particular m/z value that then strike the ion detector. The magnitude of the current that is produced at the detector as a function of time (i.e. the physical field in the mass analyzer is changed as a function of time) is used to determine the m/z value of the ion. Although mass analyzers are an important (and continually improving) component of mass spectrometers and determine critical performance characteristics, an important innovation for proteomics has been the development of two robust techniques to create ions of large molecules.

Matrix-assisted laser desorption ionization (MALDI) creates ions by excitation of molecules that are isolated from the energy of the laser by an energy absorbing matrix. The laser energy strikes the crystalline matrix to cause rapid excitation of the matrix and subsequent ejection of matrix and analyte ions into the gas-phase.

Electrospray ionization (ESI) creates ions by application of a potential to a flowing liquid causing the liquid to charge and subsequently spray. The electrospray creates very small droplets of solvent-containing analyte. Solvent is removed as the droplets enter the mass spectrometer by heat or some other form of energy (e.g. energetic collisions with a gas), and multiply-charged ions are formed in the process. The detection limits that can be achieved with ESI have improved with a reduction in the flow rates. These ionization techniques have stimulated development of tandem mass spectrometers (MS/MS), which are diagnostic of amino acid sequence (Fig. 1b). Many types of MS/MS analyzers have been developed, and new innovations allow greater automation and efficiency in data acquisition. Data can be generated in a data-dependent manner through interaction of the m/z data in each scan with a computer program to control the type of experiment performed. For example, a scan of the mass range can reveal the presence of several ions above a preset ion-abundance threshold. The computer can signal the instrument to perform tandem mass spectrometry on each of the ions, thus improving the efficiency of data acquisition, particularly during separations when ions appear for only a brief period of time.

Identifying proteins using mass spectrometry data and database searching

Mass spectrometers are capable of generating data quickly and thus have a great potential for high-throughput analysis. An essential component to achieving greater throughput is simplifying data analysis. There is a direct relationship between mass spectrometry data and amino acid sequences. Peptide molecular weight measurements are predictive of amino acid composition, and peptide fragmentation information (as described in the glossary) relates to amino acid sequence. Both types of information can be correlated to protein sequences in the database. A single peptide sequence can provide critical performance characteristics and determine critical performance characteristics, an important innovation for proteomics has been the development of two robust techniques to create ions of large molecules.
molecular weight, however, is not generally unique to a specific protein, thus a collection of peptides (‡) that are derived from the same protein must be used to find a unique match. The identity of an ‘unknown’ protein is determined by comparing the molecular-weight map of the ‘unknown’ protein with the theoretical molecular weights of peptides that are produced by digestion of each of the proteins in a database2,4. Proteins that contain peptide molecular weights that match a preponderance of the m/z values in the mass spectrum are then considered a match. An ability to acquire highly accurate m/z values has helped this method of protein identification a great deal. As the accuracy of molecular weight measurement increases, the number of peptides that will match that weight in the database will decrease5.

A second method employs amino acid fragmentation data that are generated by MS/MS (Refs 6, 7). In this method, data that are specific to an individual peptide are collected. These data contain information that is specific to and diagnostic of the amino acid sequence of peptides. In the collision-induced dissociation (CID) process, peptides fragment in a predictable manner, thus sequences from the database can be used to predict an expected fragmentation pattern and match the expected pattern to that observed in the spectrum. An advantage of this approach is that each peptide tandem mass spectrum represents a unique piece of information; consequently, matching one or more tandem mass spectra to sequences in the same protein provides a high level of confidence in the identification6,8. The identification process is not adversely affected by the presence of peptides from other proteins and is amenable to searching expressed sequence tag (EST) databases9. Thus, a collection of peptides that originate from a mixture of proteins allows the identification of the proteins that are present.

**Protein expression mapping**

The ability to identify proteins rapidly using mass spectrometry data has catalyzed the development of methods for large-scale protein analysis as well as the development of new approaches to analyze protein mixtures. A natural application of mass spectrometry is to identify the individual proteins that have been separated by gel electrophoresis. Two gel-separation methods are used to separate complicated protein mixtures. For simple protein mixtures (<100 components), single dimension (1-D) sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) is often used. Complex protein mixtures, such as total cell lysates, require the use of the highly resolving two-dimensional (2-D) SDS-PAGE. In this technique, proteins are separated by isoelectric point (pI) in the first dimension and then by...
molecular weight in the second. Unlike a cDNA array, in which the location of every gene is known, a 2-D gel separates proteins and the identity of each ‘spot’ is therefore unknown. Mass spectrometry provides a rapid and sensitive method to ‘address’ the ‘spots’ with a protein identity. 

Link et al.13 and Shevchenko et al.14 have performed large-scale identification studies on proteins that have been separated by 2-D SDS-PAGE from Haemophilus influenzae NCTC 8143 and Saccharomyces cerevisiae, respectively. Data from MS/MS or a combination of MALDI–TOF and MS/MS have been used to identify the proteins. This approach requires efficient methods to digest proteins in the gel, extract the resulting peptides and transfer them to the mass spectrometer. By employing robotics to excise spots from the gel or to slice the gel into a series of small cubes and to digest the proteins, whole-gel analysis might be possible. Identification of proteins that have been purified by 1-D or 2-D SDS-PAGE can now be performed with low nanogram quantities of protein.

An interesting question that is raised by the ability to identify and quantify protein production and turnover is the correlation of these measurements to gene expression. Anderson and Seilhamer13 have compared levels of gene expression as measured by the number of human liver cDNAs that have been sequenced to known quantities of protein, which have been determined through 2-D SDS-PAGE studies. Gygi et al.14 have attempted a more rigorous comparison of the levels of radioactively labelled S. cerevisiae proteins with those of mRNA through serial analysis of gene expression (SAGE analysis and with gene codon bias). In both studies, the mRNA and protein levels correlated poorly overall. Both studies compared levels of mRNA with those of abundant proteins; consequently, further study will be required to determine the correlation between less-abundant proteins and their corresponding mRNA. Because it is difficult to observe minor protein constituents in a total cell lysate using 2-D SDS-PAGE, some form of enrichment is generally required.

**Identification of proteins in complexes**

To observe proteins involved in specific biological processes it is possible to specifically enrich for these proteins. However, this requires knowledge of activity or at least one protein in the biological process. Under non-denaturing conditions, interacting proteins can be co-enriched in these methods, which include: chromatography, co-immunoprecipitation (using antibodies against one of the components), co-precipitation (using affinity-tagged proteins) or protein affinity-interaction chromatography (Fig. 2). The final step of the analysis employs SDS-PAGE to separate the proteins for analysis. The use of mass spectrometry to either sequence or identify proteins from 1-D or 2-D SDS-PAGE separations is growing because of the relative ease of data analysis and increased levels of sensitivity, as well as the potential for comprehensive analysis15,16. This situation is particularly true for studies that are conducted in organisms with completed genomes or a large collection of ESTs. By using this approach, a linkage between the S. cerevisiae SAGA complex and TATA-binding-protein-associated factors (TAFiIs) has been shown convincingly17. Other examples include linking the S. cerevisiae ataxia telangiectasia mutated (ATM)-related cofactor Tra1 with the SAGA complex18, identification of proteins in the S. cerevisiae spindle-pole body19, and identification of the proteins in the yeast and human spliceosome20,21. This list is by no means exhaustive, but illustrates a clear trend in the

**FIGURE 2. Identifying protein-complex components by mass spectrometry**

The components of protein complexes can be determined using mass spectrometry from co-precipitation reactions, protein affinity-interaction chromatography or isolation. One of two approaches can be used to isolate and identify the components. (a) The collections of proteins can be separated using sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) and then bands can be excised for proteolysis. Each protein from the gel can be subjected to mass spectrometry. Data is then searched through databases to identify the protein. A second approach (b) subjects the collection of proteins to proteolysis directly. The complicated collection of peptides is then analysed by high performance liquid chromatography (HPLC), combined with tandem mass spectrometry (LC/MS/MS), and the resulting spectra searched through a database. Software is used to assign peptides to their respective proteins, and thus identify the proteins present.
use of mass spectrometry to identify the components of functionally important protein complexes.

Direct identification of proteins in complexes

The use of MSMS to identify proteins provides a unique analytical capability. Because MSMS can separate an ion from its precursors and collect other ions, they provide a powerful tool to analyze protein mixtures. When numerous peptide ions enter an MSMS, one peptide m/z value can be isolated and then dissociated to obtain a fragmentation pattern that is indicative of the amino acid sequence. The ability to search a database of sequences to match a tandem mass spectrum uniquely to a sequence allows proteins in mixtures to be identified46. Thus, an approach that is based on the proteolytic digestion of protein mixtures, which is followed by reversed-phase liquid chromatography to separate or partially fracate the complex peptide mixture and direct introduction into a tandem mass spectrometer, has been developed. Several advantages flow from this strategy. This approach reduces the reliance on SDS-PAGE to separate proteins for analysis, because mass spectrometry has emerged as the method of choice for separation, digestion, and manipulation, and can take advantage of the sensitivity of mass spectrometry. Direct identification of proteins in mixtures has been used in several types of experiments. McCormack et al. have used the method with immunoaffinity precipitation, proteins affinity-interaction chromatography, and to identify proteins that interact with large protein complexes47 (Fig. 2). This approach has also been used to identify components of the proteolytic digestion using 2-D liquid and MS/MS. A total of 80 proteins were identified in a single experiment, at least ten of which were not observed by 2-D gel electrophoresis48. Coupling the direct-identification approach with quantitative methods to measure relative protein expression will greatly increase the value of the data that are produced49.

Conclusions

Integration of the information that is produced through structural genomics has significantly improved protein discovery by proteomics. It is now possible to be more thorough and rigorous in the analysis of proteins that have been obtained through molecular biology experiments and to do so with higher throughput. As more genomes are completed, the task of reconstructing metabolic and regulatory networks, pathways and subsequently the functions of all proteins, will be more straightforward. Much of this information will be derived from the study of protein–protein interactions and protein complexes, experiments that are simplified through the availability of better tools for protein analysis. Future prospects are bright the sensitivity of protein analysis by mass spectrometry will continue to improve, as will the sophistication of data-dependent acquisition methods and data analysis software.

Acknowledgements

The author is supported by NSF BIR-921482, NIH RR11823-03, and NCI R33CA81665-01.