

Mass Spectrometry in Biology

John R Yates, *University of Washington, Seattle, Washington, USA*

Applications of mass spectrometry in biological studies range from identifying expressed proteins using two-dimensional gel electrophoresis to identifying those proteins involved in protein–protein interactions.

Introduction

Since the 1990s there has been rapid growth in the field of biological mass spectrometry. Ionization methods such as electrospray ionization (ESI) and matrix-assisted laser desorption have revolutionized the creation of gas-phase ions from high-molecular weight compounds. Developments in mass spectrometry have simultaneously decreased the cost and increased the resolution and flexibility of these instruments. Advances in data processing have enabled the high-throughput analysis of mass spectra. Taken together, these innovations have made mass spectrometry an integral component of biological research.

The study of proteins through mass spectrometry has been aided by information obtained from the genome sequencing projects. Sequences for *Haemophilus influenzae*, *Saccharomyces cerevisiae*, *Escherichia coli* and many other genomes have now been completed; the human genome should soon follow. Understanding the relationship between primary sequence information and its biological context (processing events, protein–protein interactions and cellular localization) will be the next task for biologists. Mass spectrometry is well suited to provide a powerful link between proteomics and genomics (Figure 1).

Ionization of Biomolecules

A critical requirement for analysing molecules by mass spectrometry is the ability to create gas-phase ions. Two methods are commonly used for peptide and protein analysis, and are equally suitable for small fragments of DNA, carbohydrates and lipids. Matrix-assisted laser desorption (MALDI) ionizes molecules cocrystallized with an organic matrix. A laser is focused on the crystals, converting molecules in the solid state to gas-phase ions. The photon energy deposited into the matrix converts to thermal energy, causing rapid evaporation of matrix and analyte. Photoionization processes also occur and may contribute to the ionization of analyte. Ions are produced primarily as singly protonated molecules, and in combination with a time-of-flight (TOF) mass spectrometer, the sensitivity is in the low femtomole range.

Secondary article

Article Contents

- Introduction
- Ionization of Biomolecules
- Peptide and Protein Analysis by Mass Spectrometry
- Computer Algorithms for Database Searches with Mass Spectrometry Data
- Mass Spectrometry Applications in Biology
- Quantitative Mass Spectrometry
- Shotgun Identification of Peptides and Proteins from Complex Mixtures
- Future Directions

Electrospray ionization (ESI) is a process that introduces solution-bound molecules into a mass spectrometer as gas-phase ions. To form the electrospray, a solution is sprayed from the tip of a capillary at a high voltage (0.5–4 kV). A fine mist is created and directed to an opening in the mass spectrometer. Small droplets pass through a heated capillary or a heated countercurrent gas to desolvate (remove solvent from the ions) the droplets and create ions. Ions are created as multiple protonated molecules, lowering the mass-to-charge ratio (m/z). Mass spectrometers measure m/z ratios, consequently the addition of more than one charge to the molecule lowers this ratio. For example, a molecule with a molecular mass of 10 000 Da and 10 protons added will produce an m/z value of 1001 ($10\,000 + 10$ divided by 10).

Excellent sensitivity with minute quantities of material (low femtomoles) is achieved when ESI is performed at low flow rates ($20\text{--}200\text{ nL min}^{-1}$). One of the biggest benefits of ESI is the creation of a robust interface between liquid separation techniques and the mass spectrometer. Complex mixtures of peptides can be separated and ionized in combination with liquid chromatography or capillary electrophoresis. These methods effect a temporal separation based on the chemical properties of the molecules and a physical separation of the m/z ratios is obtained in the mass spectrometer.

Peptide and Protein Analysis by Mass Spectrometry

Two types of mass spectrometry experiments are in common use for biological studies. The first experiment measures molecular masses of intact molecules such as peptides and proteins. Such molecular measurements are often performed using a TOF mass spectrometer. The m/z ratios of ions are measured by determining the time required for an ion to move from the ion source to a detector. Ions typically drift through a region free of fields

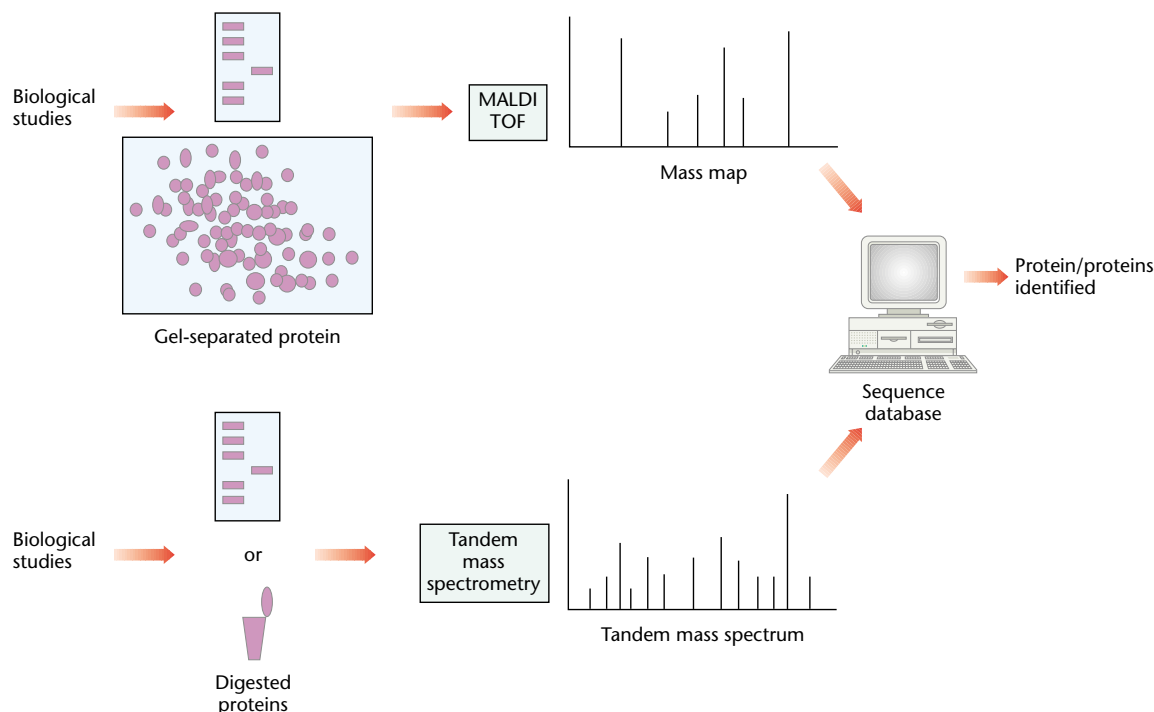


Figure 1 Identification of proteins obtained in biological studies. Proteins from a study can be separated by one- or two-dimensional gel electrophoresis and proteins identified using MALDI-TOF and peptide mass mapping. Alternatively, peptides obtained by site-specific digestion of proteins isolated from polyacrylamide gels can be analysed using tandem mass spectrometry. The tandem mass spectrometry data can then be used to search a database. A third method is to digest a protein mixture proteolytically and use microcapillary high-performance liquid chromatography and tandem mass spectrometry to generate fragmentation for individual peptides contained in the mixture. Tandem mass spectra, representing the amino acid sequences of peptides from the protein(s) present, are then used to search sequence databases. Proteins are identified when tandem mass spectra are matched to individual protein sequences.

(electric or magnetic) over a distance of 1 to 3 m. Smaller ions arrive at the detector sooner than larger ones. For peptides in the range 1000 to 4000 Da, the accuracy with which molecular mass can be measured is in the low parts per million (5–50 ppm) range. A molecular mass measurement can be useful for structural elucidation of molecules when combined with other types of data (e.g. NMR, protein sequencing).

Structural information can be obtained directly in mass spectrometry experiments by using a second type of instrument – a tandem mass spectrometer. This instrument couples together ion isolation and ion activation (excitation of vibrational states of the ion) with mass analysis to provide detailed structural information. A key component of analysis using the tandem mass spectrometer is ion activation. This process is usually performed using gas-phase collisions with neutral and inert target gases such as argon. Ion activation by photoexcitation or surface collisions has also been used. During the ion activation process the vibrational modes of the ion are excited. This causes bonds within the ion to dissociate, producing structural diagnostic fragments. The dissociation products provide the ‘structural fingerprint’ of a particular ion and thereby provide detailed structural information (e.g.

amino acid sequence) on the species of interest. Fragmentation pathways for peptides are generally predictable and allow sequences to be derived for peptides. Several different types of mass spectrometers are capable of this experiment including triple quadrupoles, TOFs, ion traps (quadrupole and Fourier transform mass spectrometers), and quadrupole-TOFs.

The process of acquiring tandem mass spectra consists of two steps. The first step is to determine the m/z values of the ions currently entering the mass spectrometer. Typically scanning the desired mass range performs this step (e.g. 400–1200). Ions are then chosen for dissociation and the first mass analyser is set to pass only that ion into the collision cell. The scan range of the second mass analyser is adjusted to scan over the mass range expected for fragments (e.g. 50–1400 for a peptide of relative molecular mass 1390). Computer control of the operating parameters of the tandem mass spectrometer permits automated acquisition of tandem mass spectra. A benefit is the ability to acquire tandem mass spectra continuously over the course of a liquid-based separation such as liquid chromatography. The high-throughput nature of this type of data acquisition, however, requires computer-assisted analysis of the data.

Computer Algorithms for Database Searches with Mass Spectrometry Data

An unanticipated benefit from the genome sequencing projects is the use of sequence databases to analyse mass spectrometry data. As genome sequences are completed, the context of protein biochemistry changes from sequence discovery to sequence identification. As protein biochemical experiments are performed that generate collections of proteins, there is a need to determine the identity of the proteins. Two types of mass spectrometry experiments are effective for protein identification. The first method uses the measured molecular masses of peptides created by site-specific proteolysis of proteins. A database is then searched, comparing the collection of peptide molecular masses with those predicted for every protein in the database. When a majority of the molecular masses match to a particular protein, then it is highly likely the protein match is correct. By using highly accurate mass measurements, a fingerprint or 'mass map' forms a unique signature for a protein and is sufficient to identify proteins in large databases. Limitations of the accuracy of identifications can result when heavily modified or mixtures of proteins are analysed. The digestion of heavily modified proteins results in peptide molecular masses that are difficult to assign *a priori* to a particular protein, and protein mixtures create mass maps that match reasonably well to several proteins and thus provide an ambiguous match. In some cases, the individual identities of proteins can be resolved through iterative search procedures.

Tandem mass spectrometry fragmentation data also can be used to match directly sequences in the database. Collisional induced bond dissociation in peptide ions creates a fragmentation pattern that reflects the amino acid sequence of the peptide. The molecular mass of the peptide and the fragmentation pattern is then used to search the database to identify stretches of amino acid sequence corresponding to the molecular mass of the peptide represented in the tandem mass spectrum. Once a sequence is identified, it is compared with the fragmentation pattern. A score is generated that determines the closeness-of-fit of the sequences to the tandem mass spectrum. Two criteria must be met for a protein to be uniquely identified. First, the amino acid sequence must be unique to a protein. If a sequence is conserved in several proteins a definitive identification cannot be made. Second, the fragmentation pattern must clearly match the sequence. The method of identifying proteins using tandem mass spectra is accurate and can be performed using as few as one or two peptides. The sites of covalent modification within an amino acid sequence can be determined as well. Tandem mass spectra can also be searched against nucleotide sequence databases by using a six-reading frame translation to convert the nucleotide sequences to amino acid sequences. A primary advantage of using tandem mass spectrometry data is the

informational independence of each spectrum. The practical result is the ability to identify proteins present in mixtures.

Mass Spectrometry Applications in Biology

A powerful method of protein separation is sodium dodecylsulfate (SDS)–polyacrylamide gel electrophoresis (SDS-PAGE). This method is widely used by biologists as a method to separate proteins of interest in their experiments. Once proteins are separated the next step is to identify the protein or proteins. One common method is to use Western blotting, a technique requiring an antibody specific for a protein's structure. The presence of a signal indicates that protein is present. Western blotting is not useful for the discovery of new proteins, but it is a powerful method to confirm the existence of a known protein in an experiment. Naturally, significant effort has been expended to use mass spectrometry to identify gel-separated proteins for two reasons. First, a method applicable to any protein, irrespective of the existence of an antibody reactive to it, would be general. Second, the ability to process proteins quickly with high sensitivity would make the analysis high-throughput. Two general approaches have been developed employing either ESI-tandem mass spectrometry or MALDI-TOF. A band containing protein is excised from the polyacrylamide matrix and digested *in situ* with a protease to create peptides. Typically a protease such as trypsin is used because it creates peptides in the molecular mass range of 800–4000 Da by cleaving after arginine or lysine residues. Repeated extraction removes peptides from the gel matrix for analysis. Peptides are then analysed by MALDI-TOF to generate a mass map or by tandem mass spectrometry to obtain fragmentation data specific to individual peptides. This approach has been effective in identifying proteins from a variety of biological studies.

The ability to identify accurately proteins purified by gel electrophoresis has been very useful in the study of biological processes. A typical example is a study of chromosome stability that depends on efficient repair of damage to DNA. Several enzymatic repair systems exist to repair DNA damage. Nijmegen breakage syndrome (NBS) is a disorder similar in clinical feature to ataxia telangiectasia, which shows increased incidence in cancer and radiation sensitivity. Both disorders exhibit evidence of defects in the DNA damage repair systems. Two proteins, Mre11p and Rad50p, are involved in the repair of double-stranded DNA damage. An immunoprecipitation of a protein complex consisting of Mre11p and Rad50p revealed several additional proteins when separated by SDS-PAGE. Mass spectrometric analysis of a coprecipitating protein of 95 kDa led to the identification of a new gene. The gene mapped at the NBS locus and the protein

was subsequently shown to be a good candidate as a sensor of double-stranded DNA damage. This example is but one of many that have appeared in the literature where mass spectrometry has been used to help solve important biological problems through protein identification.

Polyacrylamide gel electrophoresis is an effective method for protein separation. For expression studies of complex protein mixtures such as total cell lysates, two-dimensional SDS-PAGE is an effective tool. A 2D SDS-PAGE (2DE) separation resolves proteins by both molecular mass and isoelectric point. An additional step involving protein analysis is required to identify the separated proteins. As with single dimension polyacrylamide gels, proteins must be digested and extracted from the matrix. Several mass spectrometry methods have evolved as high-throughput and accurate approaches to identify proteins separated by 2DE. In general, protein identification using mass spectrometry must be sensitive and thus methods employing very low flow rate separations (ESI is a concentration-dependent method) such as microcolumn liquid chromatography and capillary electrophoresis are often employed. Other methods have included nanoelectrospray infusion, a low flow rate electrospray method for injecting a mixture of peptides without separation into the mass spectrometer, and MALDI-TOF. Studies identifying a large numbers of proteins separated by 2DE have been performed. These experiments complement the expression profile obtained upon staining of 2DE separations by identifying the observed proteins. Studies have been performed in *S. cerevisiae*, *E. coli*, and many other organisms. By using 2D SDS-PAGE in conjunction with mass spectrometry protein identification methods, proteins expressed in response to a change in cellular conditions can be observed and identified. The numbers of proteins that can be observed has been increased through the use of narrower pH gradients (e.g. pH 5–6 or 6–7) and then stitching the images of the gels together. The principal disadvantage to 2D SDS-PAGE gels is that several additional steps are required to identify the proteins. The 'spots' must be excised, the proteins proteolytically digested in the polyacrylamide matrix, and then digested proteins introduced into the mass spectrometer. Robotics are now available to perform many of these steps. The other factor is a limited dynamic range that allows only the most abundant proteins to be observed.

The ability to separate large numbers of proteins in complex mixtures allows the study of protein expression and modification. The most abundant proteins of *H. influenzae* (NTCC 8143) have been analysed from a 2-DE experiment. Two hundred and sixty-three proteins were identified as the most abundant proteins of the *H. influenzae* proteome, representing 12 different functional categories. Nineteen different proteins were identified as unannotated open reading frames with no significant sequence similarity to other proteins in the database. One

protein was found with no match to a sequence in the *H. influenzae* Rd strain genome sequence. This protein was subsequently identified as tryptophanase by using the *E. coli* sequence database to analyse the tandem mass spectra. On the Coomassie-stained gel approximately 400 stained proteins were observed, representing 23.5% of the predicted proteome. An additional 200 proteins were observed when the more sensitive silver stain was employed. 2-DE is capable of resolving proteins with different structural features or modifications. Twenty-two per cent of the proteins identified had the same identities but different 2-DE mobilities. Protein expression studies coupled with protein identification are useful for studying differences among cell types and states, as well as differences between organisms.

Quantitative Mass Spectrometry

Mass spectrometers can be used to measure relative amounts of molecules present in experiments. By using stable isotope-labelled molecules, the relative quantities of proteins between two different states can be determined. In these experiments, cells are grown in the presence of ammonium sulfate containing a ^{15}N ammonium sulfate, or the proteins from one state are covalently tagged with an isotopically heavy reagent and the other state is tagged with the isotopically light reagent. The proteins from the two different states are combined and the mass spectrometer is used to measure the relative heights of the same peptide or protein ion in the mass spectrometer. The difference in intensity of the two different ions reflects the relative change in abundance. These measurements have been performed on intact proteins and on the digested products – peptides. By measuring changes in the relative abundance for proteins in different cellular states, the dynamics of processes can be better understood.

Shotgun Identification of Peptides and Proteins from Complex Mixtures

Heterogeneous complex protein mixtures are most often separated by one- and two-dimensional SDS-PAGE. These methods are frequently used to separate the components of coprecipitation or coenrichment experiments and to observe proteins localized to specific subcellular compartments. The ability of tandem mass spectrometry to identify proteins contained in a mixture has prompted the use of this method to circumvent gel electrophoresis as a sample preparation method for protein identification. To identify the components of mixtures, proteins are purposely treated with a protease to generate a mixture of peptides. By using a combination of high-

performance liquid chromatography and tandem mass spectrometry (LC-MS-MS), peptides are fractionated and tandem mass spectra acquired. The identities of the proteins are determined by searching a database with each tandem mass spectrum. This approach is amenable to a variety of biochemical experiments such as immunoprecipitation and protein-affinity interaction chromatography. Both methods are useful to identify protein-protein interactions. Proteins can also be identified in large protein complexes (> 50 proteins). Surveys of the proteins localized to specific cellular compartments are also possible. By using this approach, the identities of proteins in the *E. coli* periplasmic space, the area between the cell membrane and periplasmic membrane, were quickly surveyed. Protein fragments, present in the endosomal and lysosomal compartments of murine antigen-presenting cells, have been identified to aid understanding of the mechanisms of formation of MHC class II complexes and their trafficking from the endoplasmic reticulum to the cell surface. Direct analysis of proteins holds promise for fast and automated identification of proteins and is particularly powerful in organisms with a completed genome. This method has enabled the identification of the components of large protein complexes containing 100 or more components.

The ability to identify proteins rapidly and accurately by using mass spectrometry has been an unexpected benefit of the genome projects. Mass spectrometry, and in particular tandem mass spectrometry, will be an important tool for high-throughput functional studies of the gene products. Protein identification using tandem mass spectra has several important advantages. First, identification is possible based on one peptide tandem mass spectrum. Second, each tandem mass spectrum represents an independent piece of information, and additional spectra that match to the same protein add considerable strength to the identification. Third, the ability to identify proteins based on a single tandem mass spectrum allows the identification of proteins present in complex mixtures. Finally, posttranslational modifications do not appear to complicate the identification and they can be localized to the specific amino acid residue with the aid of computer programs.

Future Directions

Genome sequencing will result in the identification of many new genes and gene products of unknown function. The roles of new and existing gene products in organisms and between organisms will be dissected by using biochemical and genetic studies. The ability to identify proteins rapidly and automatically, directly from mass spectrometry data, is a powerful capability. Thus, mass spectrometry will garner a broader role in biological studies as the mass accuracy and resolution of instruments improves. These improvements will allow mass spectrometers to be used without the need for gel electrophoresis to profile protein expression and to quantitate relative protein levels between different states of cells. Improvements will come through refinement of existing instruments and the development of new ones. New types of ion trap and TOF mass spectrometers will be developed as well as new variations of tandem mass spectrometers based on hybrid technologies (combining different types of mass analysers).

Further Reading

- Carney JP, Maser RS, Olivares H *et al.* (1998) The hMre11/hRad50 protein complex and Nijmegen breakage syndrome: linkage of double-strand break repair to the cellular DNA damage response. *Cell* **93**: 477–486.
- James P (1997) Of genomes and proteomes. *Biochemical and Biophysical Research Communications* **231**: 1–6.
- Link AJ, Hays LG, Carmack EB and Yates IJR (1997) Identifying the major proteome components of *Haemophilus influenzae* type-strain NCTC 8143. *Electrophoresis* **18**: 1314–1334.
- Pappin DJC, Bleasby AJ, Sutton CW and Cottrell JS (1993) Rapid identification of proteins by database matching of proteolytic peptide masses. *Protein Science* **2** (supplement 1): 90.
- Shevchenko A, Jensen ON, Podtelejnikov AV *et al.* (1996) Linking genome and proteome by mass spectrometry: large-scale identification of yeast proteins from two dimensional gels. *Proceedings of the National Academy of Sciences of the USA* **93**: 14440–14445.
- Wilkins MR, Williams KL, Appel RD and Hochstrasser DF (1997) *Proteome Research: New Frontiers in Functional Genomics*. Heidelberg: Springer-Verlag.
- Yates JR III (1998) Mass spectrometry and the age of the proteome. *Journal of Mass Spectrometry* **33**: 1–19.