Protein Sequencing in the Post-Genomic Era

Indira Rajagopal and Kevin Ahern

Any high school biology student of today can tell you that the information coded in the sequences of DNA directs, via messenger RNA, the synthesis of the thousands of proteins that give cells and organisms their characteristic phenotypes. Whole genomes have been sequenced and many millions of base pairs of nucleic acid sequence have been deposited into vast databases. The success of these initiatives and the speed with which they were carried out were made possible by the availability of DNA sequence determination methods that are efficient, rapid, and inexpensive. Microarray technologies, developed in the wake of the sequencing projects, have made it possible to monitor system-wide changes in patterns of gene expression at the mRNA level.

To further understand cellular functioning, the next logical level of analysis is proteomics, the study of global protein expression patterns that define cells in specific biological states. Until recently, such analyses were frustratingly slow and labor intensive. Protein mixtures from cells had to be resolved first by two-dimensional (2D) electrophoresis, where proteins are separated by charge in the first dimension and then by size in the second dimension, yielding spots on a polyacrylamide gel. Then, each protein spot had to be individually recovered and cleaved into short peptide fragments. Each of these fragments, in turn, had to be sequenced using the strategy originally developed by Pehr Edman (*1*) (or modifications of it) in which a protein is systematically broken down and the released amino acids are identified by chromatographic analysis. The Edman process has several limitations. First, the chemical reactions required in the process alter some of the amino acids in the peptide chain. Second, covalent modifications of amino acids, such as phosphorylation, may interfere with the analysis. Last, the process takes 30 to 50 min per amino acid identified, making it inefficient for analyzing long peptides. Despite these drawbacks, the Edman method, together with techniques for protein transfer to synthetic membranes, allowed scientists to obtain sequences from small amounts of protein. These sequences, painstakingly worked out, formed the basis for the first protein sequence databases.

Though the need to determine individual protein sequences remains, the availability of genome-wide nucleotide sequences, in recent years, has radically reshaped approaches to protein sequencing. When an organism's entire genomic sequence is known, identification of all possible open reading frames is a relatively simple matter, as is the deduction of the sequences of the proteins they encode. The problem, then, becomes one of determining which of these protein sequences corresponds to a given protein spot on a 2D gel of, say, a total cell lysate.

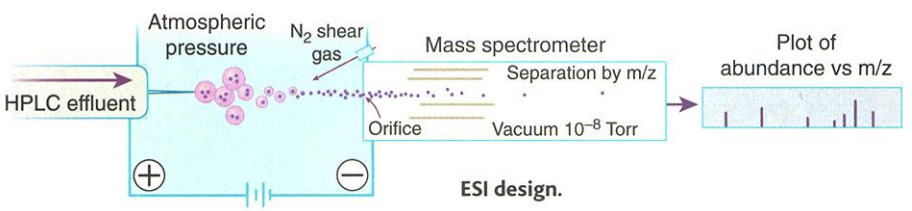
But improved mass spectrometric methods, coupled with robotic sampling techniques, can help the researcher tackle this problem.

The authors are at the Department of Biochemistry and Biophysics, Oregon State University, Corvallis, OR 97331, USA. E-mail: rajagopi@ucs.orst.edu (I.R.) and ahearnk@ucs.orst.edu (K.A.)

Because mass spectrometry can determine the mass of a peptide with such high accuracy, this method allows for the calculation of a protein spot's probable amino acid composition. Further fragmentation and analysis of the peptide can yield data of its mass, which permit the determination of its amino acid sequence with a high level of certainty.

Mass spectrometers use an electrical field to accelerate an ionized molecule of interest toward a detector. If all molecules in a sample are given the same kinetic energy, smaller pieces will travel faster than larger pieces. Charge, too, affects the speed of movement. Molecules with a double-positive charge will move faster toward a negatively charged detector than singly charged molecules of the same mass. The time a molecule requires to move from the point of ionization to the detector is a function of the mass-to-charge ratio (m/z) of a particle, and is termed time of flight (TOF).

Mass spectrometry of biomolecules became feasible with the development of methods to produce ionic forms of relatively large molecules such as peptides. The two most widely used ion-producing methods are called ElectroSpray Ionization (ESI) and Matrix Assisted Laser Desorption Ionization (MALDI). ESI, developed by John Fenn and co-workers in the late 1980s (*2*), is readily interfaced to a high-performance liquid chromatograph (HPLC) and allows peptides to be introduced into the mass spectrometer in solution, as they are eluted from a chromatographic column. Mass spectral techniques are most effective on the products of proteolytic cleavage rather than on intact proteins. Microscopic droplets of the protonated peptide are released directly from the HPLC system using a potential difference between the end of an HPLC needle and the entrance to the mass spectrometer to accelerate the droplets (see figure above). During this process, a shear gas is introduced to desolvate the ionized peptide cluster, so they are free of solvent by the time they reach the mass spectrometer's orifice. Without the solvent, each ion's mass-to-charge ratio can readily be determined by the instrument. The advantage of ESI lies in the ease with which the isolation of peptides is linked to the determination of their amino acid sequences. However, this method is not ideal for the high-throughput sequence determination needed for proteomics.

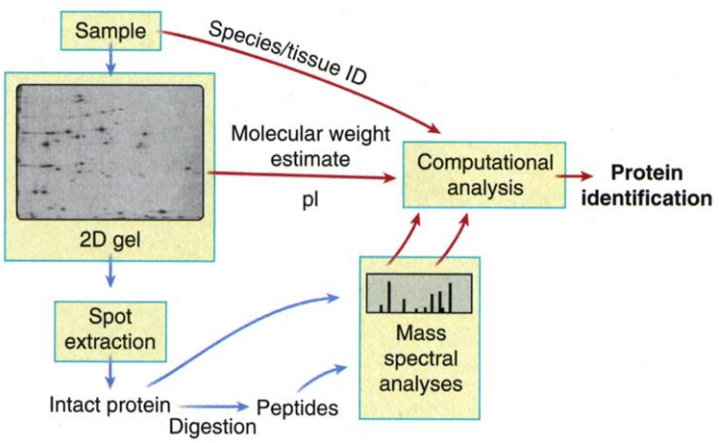


ESI design.

The method of choice for rapid, high-volume sequence analysis is MALDI, developed in 1987 by Hillenkamp and Karas (*3*). In this process, peptides are first made soluble in a solvent containing an organic acid, such as nicotinic acid, and are then deposited onto a metal stage as the solvent is evaporated. The organic acid plays a dual role—it is capable of absorbing light energy and it serves as a matrix that holds the protein molecules in place. Energy from a short laser pulse is absorbed by the matrix, which vaporizes, releasing stable precharged peptide ions into the ion chamber. Ionized molecules released in this manner may remain intact or break down into smaller pieces en route to the detector.

MALDI's tendency to deliver intact protein masses to the detector is very useful for determining the full molecular weight of a protein, but provides little information about the amino acid sequence that

composes it. This limitation can be overcome by enzymatically or chemically cleaving the protein to obtain a mixture of peptides that is then analyzed by mass spectrometry. The peptide mass pattern obtained in this manner is characteristic of the original protein and constitutes a sort of "fingerprint" that can aid in its identification. Each peptide in the fingerprint can be further analyzed by fragmentation to yield complex patterns of ion mass that may be used to identify the molecule. Two methods for fragmentation are generally used: (i) peptide cleavage(s) before instrumental analysis (by a carboxypeptidase, for example, which sequentially removes amino acids from the carboxyl end of the peptide). In this method, fragments of the peptide that differ from each other by a single amino acid are obtained. By comparing the masses of two such fragments, one can identify the amino acid at the terminus of the larger fragment. (ii) Fragmentation of unmodified peptides within the mass spectrometer. Fracturing a peptide within a machine requires energy. Though a small amount of breakage occurs with the initial laser excitation of a protein matrix in



Protein sequencing. Computational analysis uses tissue- or species-specific databases for each sequence determination. The isoelectric points (pIs) and the mass measurements of the intact protein, its peptide fragments and their degradation products are used as screening criteria for the identification of the corresponding open reading frame in the genome of the species. Mass spectral patterns may also be compared against a peptide fingerprint database to identify unknown samples. Blue arrows indicate movement of sample. Red arrows show transfer of information.

MALDI, these processes are inefficient and cannot produce all the pieces necessary for a sequence determination or protein identification. A modification of MALDI-based mass spectrometry causes the accelerated ions to collide with a neutral gas, such as argon, which fragments the peptide in a predictable manner (*4*, *5*). Peptide mass fingerprinting and mass analysis of fragments from individual peptides are carried out sequentially to obtain information that can be used to unambiguously identify a protein.

Peptides can have multiple positive charges that depend on the number of protonated amino groups in amino acid side chains. Because analyzers on mass spectrometers can easily determine *m/z*, the instrument's optimization varies with the amount of charge; the greater the charge of an ion, the smaller the range of the *m/z* the instrument must be able to detect. The mass resolution of a mass spectrometer is defined as the ratio of the mass of an ion divided by the smallest difference that can be measured. Thus, an instrument that can detect a 1-dalton difference in an ion of 10,000 daltons has a resolution of 10,000. The larger the mass of an ion, the greater is the difficulty in accurately measuring its *m/z*. Other than increasing the peptide ion's charge (as one would do when using ESI), two simple solutions to optimizing mass determinations are available. First, analyses can be calibrated with internal standards of precisely known mass, such as cytochrome c. Second, small peptide fragments can be analyzed instead of intact proteins, thus reducing the size of molecules being analyzed.

Once a suitable mass spectrum is obtained, it is compared against a protein mass spectral database to see if the sample pattern matches the fingerprint of any known peptides (see figure, this page). Useful online resources for this purpose include Mascot (www.matrixscience.com/), Prowl (http://prowl.rockefeller.edu/contents/resource.htm), and Protein Prospector (http://prospector.ucsf.edu/). The experimentally obtained peptide mass fingerprint may also be compared with patterns generated by "virtual proteolytic digests" of protein sequences in databases, such as PIR (www-nbrf.georgetown.edu/pir/) and Swiss-Prot (www.expasy.ch/sprot/sprot-top.html).

Highly accurate mass estimates for peptides are essential for determination of the amino acids composing them. (*6*, *7*). For example, the amino acid sequence of all *Escherichia coli* proteins is known because of the knowledge already delineated about the sequence of its genome. Thus, the amino acid composition of every *E. coli* protein (and proteolytic fragment) can readily be determined. Mass spectral analysis of a peptide that reveals its molecular weight as 1011.04 defines the amino acid composition of that peptide. By scanning the genomic sequence of the organism from which the original peptide came and tallying molecular weights of all possible peptides, one can determine the set of peptides with a possible molecular weight of 1011.04. This information may be insufficient to uniquely identify the protein from which the peptide originated. Determination of the amino acid composition of the remaining peptide fragments from the same protein allows the computer to successively eliminate candidate proteins, until only one remains. Thus, the original protein can be identified as the only one made of the several distinct peptide mass fragments observed experimentally.

Post-translational modifications, such as glycosylation or phosphorylation, can complicate the identification of proteins. However, such modifications can be flagged in several ways (*8*). For example, all glycopeptides contain a hexose, *N*-acetylhexosamine, and/or neuraminic acid isomers, all of which produce distinctive fragment ions that help to tag such molecules. Alternatively, matching of an unknown peptide fragment's fingerprint to that in a previously characterized fingerprint in a database may also help clarify otherwise confusing peptide patterns. As a last resort, researchers can resort to traditional methods, such as attempting to reconstruct unknown molecules from the pattern of fragment sizes in the mass spectral fragmentation pattern produced by the peptide as it passes through the instrument.

Thanks to mass spectrometry and the explosion of information from genomic sequencing, the field of proteomics is progressing rapidly. With the use of robotics, as many as 8000 protein spots per day can be analyzed. Further advances in instrumentation, chromatographic separation techniques, and data analysis, as well as improvements in the quality of data stored in the databases, will speed genome-wide analyses and ultimately enhance our understanding of the roles of proteins in cells.

References and Notes

1. P. Edman, *Arch. Biochem.* **22**, 475 (1949).
2. J. B. Fenn *et al.*, *Science* **246**, 64 (1989).
3. F. Hillenkamp, M. Karas, *Methods Enzymol.* 193, 280 (1990).
4. B. Spengler, D. Kirsch, R. Kaufmann, *J. Phys. Chem.* **96**, 9678 (1990).
5. ———, E. Jaeger, *Rapid Commun. Mass Spectrom.* **2**, 105 (1992).
6. M. Mann, P. Højrup, P. Roepstorff, *Biol. Mass Spectrom.* **22**, 338 (1993).
7. K. K. R. Clauser, P. Baker, A. L. Burlingame, *Anal. Chem.* **71**, 2871 (1999).
8. For further details, see: www.abrf.org/ABRF/ResearchCommittees/deltamass/deltamass.html
9. We thank M. Deinzer and L. Wheeler.