Trawling for proteins in the post-genome era

Brian T. Chait

Genome sequencing projects are producing an unprecedented information resource for biologists. Efficient use of this remarkable treasure trove demands the development of new tools for rapidly analyzing mature proteins and correlating them with their genes and ultimately their functions. One particularly powerful new set of tools for rapidly identifying and characterizing proteins uses modern mass spectrometric techniques such as matrix-assisted laser desorption/ionization and electrospray ionization mass spectrometry[1] in combination with genome database searching strategies[2-5]. In this issue, Figeys et al.[6] describe a new mass spectrometric system for identifying proteins that promises to have a major impact on biological research. Significantly, their system allows the analysis of nanomolar amounts of sample—levels previously too low for detection using mass spectrometric systems.

The main general features of these rapid new mass spectrometric approaches to protein identification are illustrated in Figure 1. The proteins of interest are separated, for example, by gel electrophoresis (Fig. 1A), individually subjected to proteolysis with an enzyme of known specificity such as trypsin (Fig. 1B), and the molecular masses of the resulting peptides accurately and rapidly determined by mass spectrometry (Fig. 1C), either with or without previous chromatographic or electrophoretic separation. In the simplest approach to identification, these experimentally determined masses are compared to the calculated masses of all tryptic peptides that can be theoretically produced from sequences corresponding to all of the proteins in the genomic database of the organism under study. The protein yielding the best match between the experimental and theoretical peptides is identified[2,5]. Although this peptide mapping approach is fast and simple, its success can be compromised by the presence of more than one protein in the gel spot or by extensive posttranslational modifications of the protein of interest and errors in the database sequence.

A second approach has been developed that circumvents these difficulties[3,4]. Here, a particular tryptic peptide is selected in the mass spectrometer and dissociated to produce a fragmentation mass spectrum that is characteristic of the sequence of the peptide (Fig. 1D)—a protocol that is referred to as tandem mass spectrometry because there are two stages of mass analysis. In this case, the database search for the protein uses the molecular mass of the tryptic peptide together with its fragmentation spectrum[3-5]. Although the molecular mass of the peptide by itself only moderately constrains the search—and leads to hundreds or even thousands of possible proteins—a good match of the fragmentation spectrum usually identifies a unique protein. To further increase the confidence of the call, fragmentation mass spectra of one or more additional tryptic peptides are obtained. Finally, the identification can be verified by checking how many of the tryptic peptides (Fig. 1C) have measured masses that are in accord with hypothetical tryptic peptides from the putative protein.

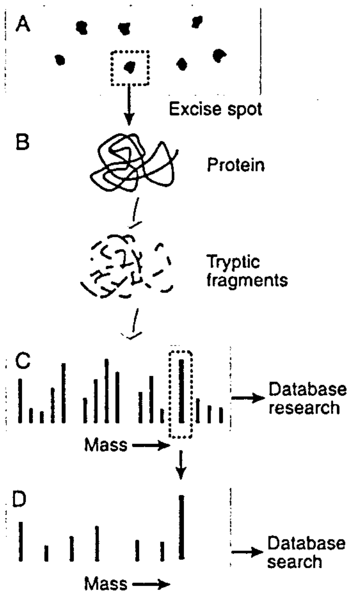

Figure 1. Identifying proteins by mass spectrometry and database searching. A. Separate proteins on two-dimensional gel. B. Excise protein and digest with trypsin. C. Obtain mass spectrum of tryptic peptides (with or without prior separation) and compare with database. D. Obtain fragmentation mass spectrum of a chosen tryptic peptide and compare with database.

Because protein identification incorporating mass spectrometric fragmentation and database searching uses single (or at most a few) tryptic peptides from any given protein, this strategy confidently identifies multiple proteins in mixtures and is highly tolerant of posttranslational modifications or errors in the database. In addition, it is orders of magnitude faster than more conventional approaches for protein identification based on Edman sequencing—speed that should be attractive for a host of applications such as the definition of components of complex molecular machines, the elucidation of kinase substrates during cell cycle progression, and the identification of proteins associated with disease states. It may even be feasible to identify the majority of proteins associated with whole cellular organelles.

For many applications, it is also critical that the method be highly sensitive. Although modern mass spectrometric methods have high intrinsic sensitivities, losses during sample handling severely limit the sensitivity that can be achieved in practice. Figeys et al.[6] describe a system for identifying proteins that successfully addresses the problem of handling low abundance samples (e.g., low nanogram amounts of yeast proteins separated by high-resolution two-dimensional gel electrophoresis). The system consists of a miniature solid-phase microextraction column (for capturing tryptic peptides generated from the protein of interest) integrated with a capillary zone electrophoresis separation/concentration system, which is connected through a microspray ion source to an electrospray ionization tandem mass spectrometer. Manual manipulations are kept to a minimum during the sample concentration, separation, and analysis steps, and the use of a flowing liquid system allows automation of much of the process.

Commercialization of robust systems of the type described by Figeys et al. is the next step. Once this has been achieved, they are likely to become "essential tools for the comprehensive analysis of biological systems in which substantial or complete DNA sequence databases have been established."

Brian T. Chait is Camille and Henry Dreyfus Professor and Head of the Laboratory for Mass Spectrometry and Gaseous Ion Chemistry at The Rockefeller University, 1230 York Ave., New York, NY 10021, and Director of the NIH National Resource for the Analysis of Biological Macromolecules (chait@rockvax.rockefeller.edu).

1. Chait, B.T. and Kent, S.B.H. 1992. *Science* 257:1885–1894.
2. Patterson, S.D. and Aebersold, R. 1995. *Electrophoresis* 16:1791–1814.
3. Mann, M. and Wilm, M. 1994. *Anal. Chem.* 66:4390–4399.
4. Yates, J.R., III, Eng, J.K., McCormack, A.L., and Schieltz, D. 1995. *Anal. Chem.* 1995. 67:1426–1436.
5. Mass spectrometric protein identification computer algorithms are publicly available over the World Wide Web, e.g., at URL http://chait-sgi.rockefeller.edu.
6. Figeys, D., Ducret A., Yates, J.R. and Aebersold, R. 1996. *Nature Biotechnology* 14:1579–1583.