Mass Spectrometry in Identification and Characterization of Biopolymers

P. Roepstorff
Department of Molecular Biology,
Odense University, DK 5230 Odense M, Denmark

May, 1999

Advances in DNA sequencing has changed the focus of analytical biochemistry

Advanced technology for determination of DNA sequences has become widely available in the last decade and have been used for sequencing cDNAs and entire genomes from a variety of organisms ranging from viruses to man. As a consequence, the amount of DNA sequence information entered in publicly accessible databases has increased exponentially in the last decade. This growth has by far exceeded the growth of sequences entered in databases based on *de novo* protein sequencing. In addition to the genomic sequencing, large scale partial cDNA's sequencing has resulted in another set of data, the so-called expressed sequence tags (EST's), containing stretches of sequence from a large number of genes from a variety of organisms. The genomic sequences only provides information of the potential of the selected microorganisms and cell types but does not reflect the actual situation at any given moment, i.e. which proteins are expressed and how are they modified. cDNA sequences or the incomplete EST's gives information on the actually expressed proteins, but no information on processing and secondary information. Therefore, studies of the proteins will never be obsolete, but the questions to address will be different. There is no doubt that the huge amount of sequence information will be one of the most important sources to new knowledge about biological systems. However, effective use of this information requires development of new analytical concepts by which data can be generated that matches the type of information in sequence databases.

Mass spectrometry the ideal analytical technique in the post genome era

Independently, but coinciding in time, mass spectrometric analysis has undergone an equally dramatic development. From being an analytical tool for analysis of small volatile molecules, new ionization methods, especially electrospray ionization (ESI) (Fenn et al. 1989) and matrix assisted laser desorption/ionization (MALDI) (Karas and Hillenkamp 1988), have increased the accessible mass range to include nearly all proteins. Mass accuracy and sensitivity have been improved to a routinely achievable level (Table 1) which allow analysis of proteins, nucleic acids and oligosaccharides isolated by the most sensitive separation methods presently available (Jensen et al. 1996).

Table 1. Mass accuracy and sensitivity routinely achieved by MS of peptides and proteins.

mass range		0.5 - 5 kDa	5 -20 kDa	> 20 kDa
Mass Accuracy	MALDI DE-RF-TOF	30 ppm	50-100 ppm	
	MALDI DE-lin-TOF	100 ppm	100-200 ppm	0.02-0.1ppm
	nanoESI all modes	0.1 Da	1 Da	0.01%
Sensitivity	MALDI all modes	0.1 - 1 fmoles	1 - 10 fmoles	0.1 - 1 pmoles
	nanoESI-quadrupole	10-50 fmoles	0.1 - 1 pmoles	1-10 pmoles
	nanoESI-QTOF	1-10 fmoles	50-100 fmoles	0.5-5 pmoles

Mass spectrometry (MS) has been proven ideal for analysis of peptide and protein mixtures and partial or complete sequence information can be generated from the single components in such mixtures by the so-called MS/MS techniques. In addition, MS is the ideal technique for analysis of post translational modifications in proteins, thus being the perfect complement to DNA sequencing. Recent progress in nucleic acid and oligosaccharide analysis indicate that concepts similar for those described below for protein studies will also be applicable in studies of these biopolymers.

Proteome analysis the next step after the genome

Once a genome is sequenced then the next natural step is analysis of the proteome which as defined by Wilkins et al. 1996 represent: The total protein complement expressed by an organism, a cell or a tissue type. Proteome analysis involves two essential steps. Firstly, separation and visualization of the proteins and, secondly, identification of the proteins relative to the genomic sequence, if known. 2- dimensional polyacrylamide gel electrophoresis (2D-PAGE) is the only technique presently available for separation of all or the majority of the proteins from a given cell type. Identification of the proteins is now routinely carried out by mass spectrometry after digestion of the proteins in the gel with appropriate proteolytic enzymes. This can be done either based on peptide maps produced by MALDI MS or partial sequences produced by ESI MS (For a complete strategy see: Shevchenko et al. 1996) Of these techniques peptide mapping by MALDI MS is the simplest and most sensitive whereas the sequence based techniques are more specific. Partial sequences also often allow identification in cases where only partial protein sequence information is available, e.g., in EST databases (Mann 1996). Upon positive identification, the corresponding cDNA can then be ordered and sequenced.

Characterization of secondary modifications is essential

The genomic sequence contains information about the amino acid sequence of the proteins. However, a majority of all proteins are modified after they have been synthesized based on the information in genetic code, so-called post translational modification. Once a protein is identified, then the next obvious questions are: Are the identified proteins post translationally modified and if so, how? If the purified protein is available then the strategy is to compare its molecular mass determined by MS with that calculated based on the DNA sequence. If these masses are different, then the modified sites and the types of modification are identified based on mass spectrometric peptide mapping if relevant supplemented with MS/MS of relevant peptide ions or degradation with appropriate enzymes, e.g., glycosidases or phosphatases (Burlingame 1996, Bean et al. 1995). If the proteins are only available as spots or bands in electrophoretic gels, then it is often not possible to determine the molecular mass of the intact protein and characterization of post translational modifications must rely on peptide mapping before and after enzymatic treatments and when appropriate supplemented with MS/MS of selected components in the mixture. In such cases it is essential to obtain complete or very high sequence coverage in the peptide maps (Moertz et al. 1996, Wilm et al. 1996a). The most frequently occurring secondary modifications are N- or C-terminal removal of part of the protein, acylation of the N-terminal amino group or the side chain amino group of Lysine, phosphorylation of serine, threonine and tyrosine hydroxyl groups and glycosylation of asparagine, serine or threonine. Characterization of all these modifications is now possible by mass spectrometry (Roepstorff 1997, Jensen et al. 1998)

When is "de novo" protein sequencing relevant?

If a protein under identification in proteome studies comes out as unknown in the data base search, then sufficient de novo sequence information must be generated to allow synthesis of nucleotide probes for subsequent isolation of the corresponding cDNA from cDNA libraries. This has been achieved from silver stained 2D-gels by ESI MS/MS (Wilm et al. 1996b). In cases where no genomic or cDNA information is available for the organism studied then complete sequencing on the protein level might be indicated as for example in our

studies of insect and crustacean cuticle proteins (Andersen et al. 1995). A number of strategies have been proposed for protein sequencing based on or supported by mass spectrometry (Fig. 1). However, attempts to perform de novo sequencing of entire proteins with mass spectrometry as the only sequencing method nearly always fails, because amino acid residues in some positions always remain unassigned or ambiguously identified. In our experience the best strategy for complete de novo protein sequencing involves a combination of Edman degradation and mass spectrometric analysis (Roepstorff and Hoejrup 1993).

Protein sequencing by mass spectrometry

MS-only based techniques

Collision induced dissociation (CID)

Post source decay (PSD)
(Spengler et al. 1992)

Combined methods with MS detection

Ladder sequencing using Edman degradation
(Chait et al. 1993)

Ladder sequencing using carboxypeptidases
(Patterson et al. 1995)

Partial acid hydrolysis
(Vorm and Roepstorff 1994)

MS supported strategy

Edman degradation combined with MS
(Roepstorff and Hoejrup 1993)

Figure 1. Different approaches for mass spectrometric de novo protein sequencing.

Conclusion

There is no doubt that mass spectrometry will play an increasingly important role as protein analytical tool in the post genome era. The information obtained from the mass spectra is perfectly complementary to the information derived from genome or cDNA sequencing. The combination of sensitivity and specificity is unsurpassed by any other technique for protein identification and for characterization of post translational modifications in proteins. Although not a fully reliable sequencing tool, then the sensitivity with which partial or full sequence information can be generated from peptides is extremely useful in many applications. The next major fields in protein studies where mass spectrometry must be expected to gain acceptance are those of protein surface topology and protein interaction studies. The first applications in these fields already have appeared and the future potential is very high (Roepstorff 1997). We can also expect to see an increasing number of applications of mass spectrometry to nucleic acid and oligosaccharide analysis.

Acknowledgments

The Danish Biotechnology Programme and the Danish National Research Foundation are acknowledged for financial support.

References

Andersen, S.O., Hoejrup P., and Roepstorff, P. Insect Biochem. Molec. Biol. **25**, 153 (1995).
Burlingame, A.L. Curr. Opin. Biotechnol. **7**, 4 (1996).

Bean, M.F., Annan, R.S., Hemling, M.E., Mentzer, M., Huddleston, M.J., and Carr, S.A., (1995) In *Techniques in Protein Chemistry VI.* Crabbe, J., (ed.) Academic Press, San Diego, pp 107-116.

Chait, B.T., Wang, R., Beavis,R.C., and Kent, S.B.H., Science **262**, 89 (1993).

Fenn, J.B., Mann, M., Meng, C.K., Wong, S.F., and Whitehouse, C.M., Science **246**, 64 (1989).

Jensen, O.N., Potelejnikov, A. and Mann, M., Rapid Comm. Mass Spectrom. **10**, 1371 (1996).

Jensen, O.N., Larsen, M.L., and Roepstorff, P. PROTEINS, Structure Function and Genetics Suppl. **2**, 74 (1998).

Karas, M., and Hillenkamp, F., Anal. Chem. **60**, 2299 (1988).

Mann, M., Trends Biol. Sci. **21**, 494 (1996).

Moertz, E., Saraneva, T., Haebel, S., Julkunen, I. and Roepstorff, P., Electrophoresis **17**, 1493 (1996).

Patterson, D.H., Tarr, G.E., Regnier, F.E., and Martin, S.A., Anal. Chem. **67**, 3971 (1995).

Roepstorff, P. and Hoejrup, P. *Methods in Protein Sequence Analysis.* K. Imahori & F. Sakiyama (eds.) Plenum Press, New York, pp 149 (1993).

Roepstorff, P., Curr. Opin. Biotechnology **8**, 6 (1997).

Shevchenko, A., Jensen, O.N., Podtelejnikov, A.V., Sagliocco, F., Wilm, M., Vorm, O., Mortensen, P., Shevchenko, A., Boucherie, H., and Mann, M., Proc. Nat. Acad. Sci. (USA) **93**, 14440 (1996).

Spengler, B., Kirsch, D., Kaufmann, R., and Jaeger, E., Rapid Commun Mass Spectrom **6,** 105-108 (1992).

Vorm O. and Roepstorff P. (1994) Biol Mass Spectrom **23**, 734 (1994).

Wilkins, M.R., Pasquali, C., Appel, R.D., Ou, K., Golaz, O., Sanchez, J.C., Yan, J.X., Gooley, A.A., Hughes, G., Humphery-Smith, I., Williams, K.L., and Hochstrasser, D.F., Bio/Technology **14**, 61 (1996).

Wilm, M., Neubauer, G., and Mann, M. Anal. Chem. **68**, 527 (1996a).

Wilm, M., Shevchenko, A., Houthaeve, T., Breit, S., Schweigerer, L., Fotis, T., and Mann, M. (1996b) Nature 379, 466 (1996b).