Identifying the proteome: software tools

David Fenyö

The interest in proteomics has recently increased dramatically and proteomic methods are now applied to many problems in cell biology. The method of choice in proteomics for identifying and characterizing proteins is mass spectrometry combined with database searching. Software tools have been improved to increase the sensitivity of protein identification and methods for evaluating the search results have been incorporated

Address
ProteoMetrics, LLC, 7 West 36th Street, New York, NY 10018, USA

Current Opinion in Biotechnology 2000, **11**:391–395

0958-1669/00/$ – see front matter

Introduction

Proteomics is the study of all the expressed proteins of an organism. The information that proteomics studies can provide includes expression levels, post-translational modifications, subcellular localizations, protein–protein interactions, and protein–nucleic acid interactions.

The first step in proteomic analysis is usually sub-fractionation of the cells of interest followed by separation of the proteins by 1- or 2-dimensional gel electrophoresis. The next step is to cut out and enzymatically digest the proteins of interest from the gel for mass spectrometric analysis [1–14]. An alternative method is to digest the protein before separation by gel electrophoresis and instead separate the peptides by liquid chromatography and analyze the peptides by tandem mass spectrometry [15–18].

The mass spectrometric analysis usually begins with peptide mapping, that is, the separated proteins are digested with an enzyme and the masses of the proteolytic peptides are measured with mass spectrometry. The masses of the measured proteolytic peptides are compared to predicted proteolytic peptides from protein sequence databases [19–23]. Each protein sequence in the database is digested according to the specificity of the enzyme and the masses of the resulting peptides are calculated and a theoretical mass spectrum is constructed (Figure 1). The measured mass spectrum is compared with the theoretical mass spectrum and a score qualifying the comparison is calculated. The protein sequences in the database are sorted according to the score and the protein sequence with the best score is selected.

The success of protein identification by peptide mapping is a result of certain characteristics of proteins, including the limited number of proteins for each organism, the large differences in amino acid sequence, and the large mass difference between different amino acids. Figure 2 shows the number of proteins in different organisms that match the mass of a single tryptic peptide [24], indicating that a measurement of a few tryptic peptides is sufficient for identification of a protein when the genome sequence is available. Recent improvements in instrumentation have made it possible to determine peptide masses with a higher mass accuracy, which has improved the success rate for protein identification by peptide mapping [24–26]. Other information that can be used to improve the quality of identifications includes amino acid composition, number of exchangeable hydrogens [27] and partial amino acid sequence [3,7,28]. The searches are usually restricted with

Figure 1

Protein identification using peptide mapping information. **(a)** In the experiment, the proteins are digested with an enzyme and the masses of the proteolytic peptides are measured with mass spectrometry. **(b)** In the database search, each protein sequence in the database is digested according to the specificity of the enzyme. The masses of the resulting peptides are calculated and a theoretical mass spectrum is constructed. The measured mass spectrum is compared with the theoretical mass spectrum.

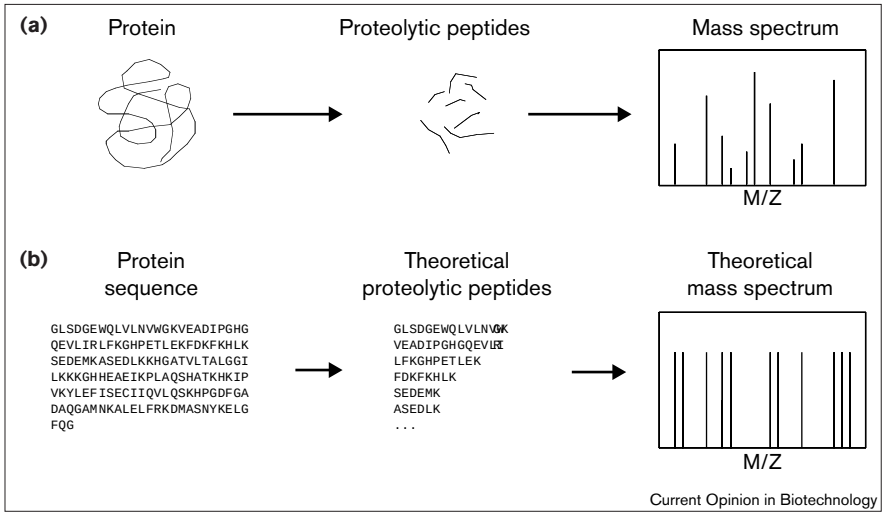
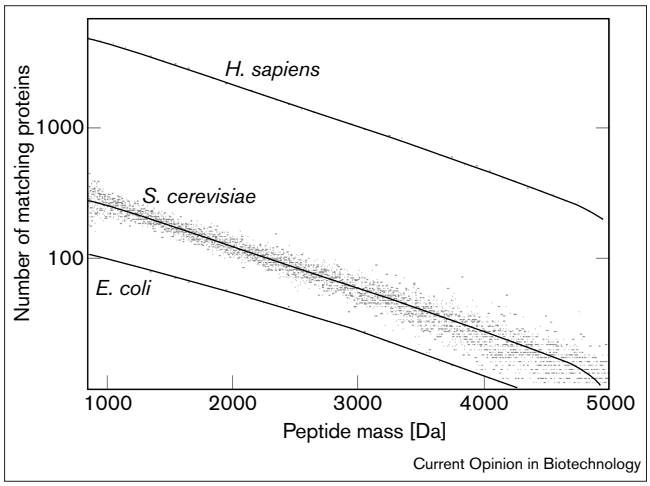


Figure 2



Information content in the mass of a single tryptic peptide for *Escherichia coli* (~4000 open reading frames [ORFs]), *Saccharomyces cerevisiae* (~6000 ORFs), and *Homo sapiens* (~100,000 ORFs), at a mass accuracy of 0.5 Da. For *S. cerevisiae,* the number of proteins at every mass unit is shown together with a smooth curve fitted to the data. For *E. coli* and *H. sapiens*, only the smooth fits are shown for clarity [24].

additional information, such as species or taxonomic category, protein mass, and protein isoelectric point. Although peptide mapping is usually applied to pure proteins, the constituents of simple protein mixtures can also be identified by peptide mapping [29••,30].

Peptide mapping has a high success rate for identifying simple protein mixtures from microorganisms with fully sequenced genomes; however, when studying mammals the success rate is presently considerably lower. The success rate of peptide mapping will increase in the near future when the human and, soon after, the mouse genomes will be completed. In the cases where peptide mapping does not provide sufficient information for confident identification, it is necessary to obtain more information. The most common method is to isolate ions corresponding to a proteolytic peptide in the mass spectrometer, fragment them by collisional excitation, and measure the masses of the fragment ions to obtain partial sequence information. The measured fragment mass spectrum is compared to theoretical mass spectra calculated from the protein sequences in the database [31,32].

In this article, we will discuss different software tools that are available for searching protein sequence databases with mass spectrometric information and how the quality of the results can be assessed.

Software for protein identification

Peptide mapping

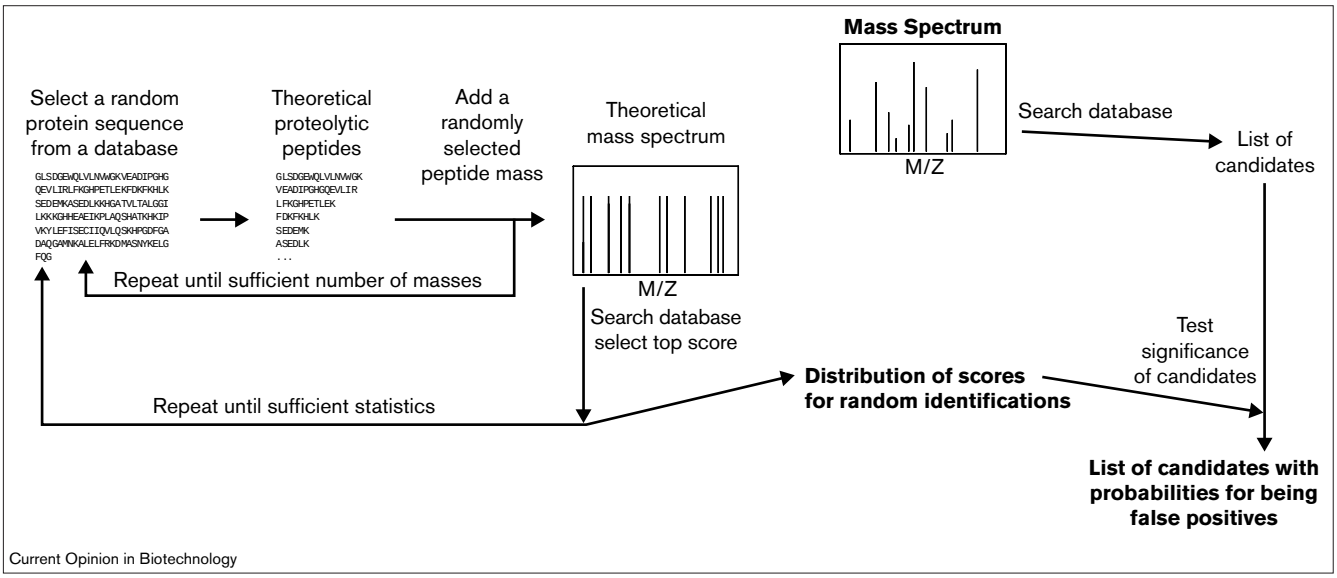
The simplest and most obvious scoring method for peptide mapping is to count the number of measured peptide masses that correspond to calculated peptide masses in the theoretical mass spectrum of each protein in the database. Several software tools are available on the Internet that use this method of ranking the proteins in the database according to the number of matching peptides, for example, PepSea [20] (http://pepsea.protana.com/PA_PepSeaForm.html), PeptIdent/MultiIdent [33,34] (http://www.expasy.ch/tools/peptident.html), and MS-Fit [26] (http://prospector.ucsf.edu/ucsfhtml3.2/msfit.htm). This simple scoring method works well for high-quality experimental data, but has the disadvantage that it usually gives higher scores to larger proteins because the probability of random matching is higher. More sophisticated methods for identifying proteins are all based on counting the number of measured peptide masses that correspond to calculated peptide masses but they attempt to make better use of the mass spectrometric information compensating, for example, for effects of protein size [21,29••,34–36]. This usually leads to methods that are more selective and sensitive.

MOWSE [21] (http://srs.hgmp.mrc.ac.uk/cgi-bin/mowse and also implemented in MS-Fit at http://prospector.ucsf.edu/ucsfhtml3.2/msfit.htm) uses average properties of the proteins in the database to improve the sensitivity and selectivity of the identification. It takes into account the relative abundance of the peptides in the database (see Figure 2) when calculating the score, that is, the chance of getting a random match to a larger peptide is lower and therefore it will contribute to a higher degree to the score. Also the protein size effect is compensated for.

ProFound [29••] (http://prowl.rockefeller.edu/cgi-bin/ProFound or http://www.proteometrics.com/prowl-cgi/ProFound.exe) is an expert system for protein identification using Bayesian theory to rank the protein sequences in the database by their probability of occurrence. It takes into account detailed information about each individual protein sequence in the database and allows for incorporation of additional experimental information (e.g. amino acid composition or sequence information) when available. In addition, empirical information about patterns observed for the distribution of proteolytic peptides along the protein sequence is included in the algorithm. One advantage of the Bayesian approach is that different types of information can be included in a natural way and therefore it is possible to make optimal use of all available information and increase the sensitivity and selectivity of the algorithm. ProFound can also be used to identify simple protein mixtures. A two-step approach is used where first the proteins in the database are ranked according to how well they match the experimental data assuming a single protein is present. In the second step, the top ranking proetins are fused together pairwise, in groups of three, and so on. These fusion proteins are then ranked according to how well they match the experimental data.

Mascot [35] (http://www.matrixscience.com/cgi/searchform.pl?SEARCH=PMF) is based on the MOWSE

Figure 3



Simulations provide a method for determining the quality of the search results [41••].

algorithm [21] but in addition it uses probability-based scoring. The probability that the observed match between experimental data and a protein sequence is a random event is approximately calculated for each protein sequence in the database. The proteins are then ranked with decreasing probability of being a random match to the experimental data.

PeptIdent2 [37] is an algorithm that has been optimized using a genetic algorithm. PeptIdent2 is a generic algorithm with many coefficients and does not incorporate any knowledge about protein properties. The coefficients are optimized using a training set of protein mass spectra. This is a very different approach than that of ProFound, MOWSE, and Mascot, where the algorithms are based on either our knowledge of the properties of individual proteins or database averages.

Peptide fragmentation

In contrast to mass spectra of peptide maps, which contain global information about a protein, peptide fragmentation mass spectra contain rich information on a small section of a protein. The information on the sequence of each peptide enables the identification of a protein from a single peptide. This allows searching of databases that contain incomplete gene information, for example, expressed sequence tags (ESTs). The use of peptide fragmentation mass spectra is also the method of choice for identifying complex protein mixtures. There are several approaches to using peptide fragment information for protein identification.

PepSea [31] (http://pepsea.protana.com/PA_PeptidePatternForm.html) uses information from fragmented proteolytic peptides. First, a peptide sequence tag has to be extracted. A peptide sequence tag is a short partial amino acid sequence of a proteolytic peptide together with information of the mass of the peptide and the masses of the parts of the peptide that have not been sequenced. This approach is very fast but requires extraction of the peptide sequence tag prior to searching.

SEQUEST [32,38–40] uses data from un-interpreted peptide fragment mass spectra (i.e. the information from the whole mass spectrum is used). A cross-correlation function is calculated between the measured fragment mass spectrum and the protein sequences in the database. The cross-correlation function is used to score the proteins in the database. SEQUEST supports the use of information from several fragment mass spectra in the database search. This approach does not require extraction of any information from the mass spectra but the searches are time consuming.

PepFrag [24] (http://www.proteometrics.com/prowl/pepfragch.html) and MS-Tag [26] (http://prospector.ucsf.edu/ucsfhtml3.2/mstagfd.htm) use peptide fragment mass information in combination with other mass spectrometric information, such as amino acid composition, to identify proteins.

Mascot [35] (http://www.matrixscience.com/cgi/search_form.pl?SEARCH=MIS) uses the same probability-based scoring algorithm for fragment information as for peptide maps. It also supports the use of information from several fragment mass spectra in the database search.

Quality of search results

The software tools for protein identification using mass spectrometric information will give a top-ranking candidate

even if all the matching peptides are random matches. It is important to determine the quality of the identification, that is, what the probability is that the identified protein is a false positive [35,36,41••].

One method for assessing this is by using simulations [40] (Figure 4). In the simulations, protein sequences were randomly selected from a protein sequence database, digested according to the specificity of an enzyme, a single peptide was randomly chosen, and its mass calculated and stored. This procedure was repeated and a theoretical mass spectrum was constructed. This theoretical mass spectrum was then used in a database search and the top score was saved. The protein sequence with the highest score was in nearly all cases a false positive, that is, the peptide matches were random. These searches were repeated with different theoretical mass spectra and a distribution of scores for random identification was obtained. Subsequently, the distribution of scores for random identification can be used to assess the quality of the results when experimental data is used in a database search, that is, each protein candidate in the list can be associated with a probability for it being a false positive. Other methods are attempts at directly calculating the probability that the masses observed in a mass spectrum would correspond to proteolytic peptides from a protein sequence [35,36]. The direct calculations are, however, less reliable than the simulation because it is necessary to make approximations because of the complexity of the process.

Objective methods for assessing the quality of search results have become more important as high-throughput proteome analysis is becoming more widespread [36,42–44].

Conclusions

The software tools for protein identification have matured and the algorithms have been refined to give higher selectivity and sensitivity. High-throughput analysis has become increasingly common in proteome projects and requires automatic analysis of the mass spectrometric data. An important part of automation is quality control and therefore development of methods to determine the quality of the search results has become a focus.

Acknowledgements

This work was funded in part by a grant from the National Institutes of Health (2R44RR13503-02). We thank RC Beavis, BT Chait, J Eriksson, C Tang, and W Zhang for fruitful discussions.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Fey SJ, Nawrocki A, Larsen MR, Gorg A, Roepstorff P, Skews GN, Williams R, Larsen PM: **Proteome analysis of *Saccharomyces cerevisiae*: a methodological outline.** *Electrophoresis* 1997, **18**:1361-1372.

2. Mortz E, Vorm O, Mann M, Roepstorff P: **Identification of proteins in polyacrylamide gels by mass spectrometric peptide mapping combined with database search.** *Biol Mass Spectrom* 1994, **23**:249-261.

3. Rasmussen HH, Mortz E, Mann M, Roepstorff P, Celis JE: **Identification of transformation sensitive proteins recorded in human two-dimensional gel protein databases by mass spectrometric peptide mapping alone and in combination with microsequencing.** *Electrophoresis* 1994, **15**:406-416.

4. Clauser KR, Hall SC, Smith DM, Webb JW, Andrews LE, Tran HM, Epstein LB, Burlingame AL: **Rapid mass spectrometric peptide sequencing and mass matching for characterization of human melanoma proteins isolated by two-dimensional PAGE.** *Proc Natl Acad Sci USA* 1995, **92**:5072-5076.

5. Shevchenko A, Jensen ON, Podtelejnikov AV, Sagliocco F, Wilm M, Vorm O, Mortensen P, Boucherie H, Mann M: **Linking genome and proteome by mass spectrometry: large-scale identification of yeast proteins from two dimensional gels.** *Proc Natl Acad Sci USA* 1996, **93**:14440-14445.

6. Wilkins MR, Pasquali C, Appel RD, Ou K, Golaz O, Sanchez JC, Yan JX, Gooley AA, Hughes G, Humphery-Smith I *et al.*: **From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis.** *Biotechnology* 1996, **14**:61-65.

7. Patterson SD, Thomas D, Bradshaw RA: **Application of combined mass spectrometry and partial amino acid sequence to the identification of gel-separated proteins.** *Electrophoresis* 1996, **17**:877-891.

8. Qin J, Fenyö D, Zhao Y, Hall WW, Chao DM, Wilson CJ, Young RA, Chait BT: **A strategy for rapid, high-confidence protein identification.** *Anal Chem* 1997, **69**:3995-4001.

9. Rout MP, Aitchison JD, Suprapto A, Hjertaas K, Zhao Y, Chait BT: **The yeast nuclear pore complex. Composition, architecture, and transport mechanism [In Process Citation].** *J Cell Biol* 2000, **148**:635-652.

10. Jensen ON, Houthaeve T, Shevchenko A, Cudmore S, Ashford T, Mann M, Griffiths G, Krijnse Locker J: **Identification of the major membrane and core proteins of vaccinia virus by two-dimensional electrophoresis.** *J Virol* 1996, **70**:7485-7497.

11. Shevchenko A, Wilm M, Vorm O, Jensen ON, Podtelejnikov AV, Neubauer G, Mortensen P, Mann M: **A strategy for identifying gel-separated proteins in sequence databases by MS alone.** *Biochem Soc Trans* 1996, **24**:893-896.

12. Patterson SD: **Identification of low to subpicomolar quantities of electrophoretically separated proteins: towards protein chemistry in the post-genome era.** *Biochem Soc Trans* 1997, **25**:255-262.

13. Link AJ, Hays LG, Carmack EB, Yates JR III: **Identifying the major proteome components of *Haemophilus influenzae* type-strain NCTC 8143.** *Electrophoresis* 1997, **18**:1314-1334.

14. Wigge PA, Jensen ON, Holmes S, Soues S, Mann M, Kilmartin JV: **Analysis of the *Saccharomyces* spindle pole by matrix-assisted laser desorption/ionization (MALDI) mass spectrometry.** *J Cell Biol* 1998, **141**:967-977.

15. Yates JR III, McCormack AL, Schieltz D, Carmack E, Link A: **Direct analysis of protein mixtures by tandem mass spectrometry.** *J Protein Chem* 1997, **16**:495-497.

16. Courchesne PL, Jones MD, Robinson JH, Spahr CS, McCracken S, Bentley DL, Luethy R, Patterson SD: **Optimization of capillary chromatography ion trap-mass spectrometry for identification of gel-separated proteins.** *Electrophoresis* 1998, **19**:956-967.

17. Yates JR III, Carmack E, Hays L, Link AJ, Eng JK: **Automated protein identification using microcolumn liquid chromatography-tandem mass spectrometry.** *Methods Mol Biol* 1999, **112**:553-569.

18. Figeys D, Corthals GL, Gallis B, Goodlett DR, Ducret A, Corson MA, Aebersold R: **Data-dependent modulation of solid-phase extraction capillary electrophoresis for the analysis of complex peptide and phosphopeptide mixtures by tandem mass spectrometry: application to endothelial nitric oxide synthase.** *Anal Chem* 1999, **71**:2279-2287.

19. Henzel WJ, Billeci TM, Stults JT, Wong SC, Grimley C, Watanabe C: **Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases.** *Proc Natl Acad Sci USA* 1993, **90**:5011-5015.

20. Mann M, Hojrup P, Roepstorff P: **Use of mass spectrometric molecular weight information to identify proteins in sequence databases.** *Biol Mass Spectrom* 1993, **22**:338-345.

21. Pappin DDJ, Højrup P, Bleasby AJ: **Rapid identification of proteins by peptide-mass finger printing.** *Curr Biol* 1993, **3**:327-332.

22. Yates JRD, Speicher S, Griffin PR, Hunkapiller T: **Peptide mass maps: a highly informative approach to protein identification.** *Anal Biochem* 1993, **214**:397-408.

23. James P, Quadroni M, Carafoli E, Gonnet G: **Protein identification by mass profile fingerprinting.** *Biochem Biophys Res Commun* 1993, **195**:58-64.

24. Fenyö D, Qin J, Chait BT: **Protein identification using mass spectrometric information.** *Electrophoresis* 1998, **19**:998-1005.

25. Jensen ON, Podtelejnikov A, Mann M: **Delayed extraction improves specificity in database searches by matrix-assisted laser desorption/ionization peptide maps.** *Rapid Commun Mass Spectrom* 1996, **10**:1371-1378.

26. Clauser KR, Baker P, Burlingame AL: **Role of accurate mass measurement (± 10 ppm) in protein identification strategies employing MS or MS/MS and database searching.** *Anal Chem* 1999, **71**:2871-2882.

27. James P, Quadroni M, Carafoli E, Gonnet G: **Protein identification in DNA databases by peptide mass fingerprinting.** *Protein Sci* 1994, **3**:1347-1350.

28. Jensen ON, Vorm O, Mann M: **Sequence patterns produced by incomplete enzymatic digestion or one-step Edman degradation of peptide mixtures as probes for protein database searches.** *Electrophoresis* 1996, **17**:938-944.

29. •• Zhang W, Chait BT: **ProFound — an expert system for protein identification using mass spectrometric peptide mapping information.** *Anal Chem* 2000, **72**:2482-2489.

This paper discusses ProFound – an expert system for protein identification of simple protein mixtures using Bayesian theory. ProFound takes into account detailed information about each individual protein sequences in the database and allows for incorporation of additional experimental information.

30. Jensen ON, Podtelejnikov AV, Mann M: **Identification of the components of simple protein mixtures by high-accuracy peptide mass mapping and database searching.** *Anal Chem* 1997, **69**:4741-4750.

31. Mann M, Wilm M: **Error-tolerant identification of peptides in sequence databases by peptide sequence tags.** *Anal Chem* 1994, **66**:4390-4399.

32. Eng JK, McCormack AL, Yates JR: **An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database.** *J Am Soc Mass Spec* 1994, **5**:976.

33. Wilkins MR, Gasteiger E, Bairoch A, Sanchez JC, Williams KL, Appel RD, Hochstrasser DF: **Protein identification and analysis tools in the ExPASy server.** *Methods Mol Biol* 1999, **112**:531-552.

34. Wilkins MR, Gasteiger E, Wheeler CH, Lindskog I, Sanchez JC, Bairoch A, Appel RD, Dunn MJ, Hochstrasser DF: **Multiple parameter cross-species protein identification using MultiIdent — a world-wide web accessible tool.** *Electrophoresis* 1998, **19**:3199-3206.

35. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS: **Probability-based protein identification by searching sequence databases using mass spectrometry data.** *Electrophoresis* 1999, **20**:3551-3567.

36. Berndt P, Hobohm U, Langen H: **Reliable automatic protein identification from matrix-assisted laser desorption/ionization mass spectrometric peptide fingerprints.** *Electrophoresis* 1999, **20**:3521-3526.

37. Gras PR, Muller M, Gasteiger E, Gay S, Binz PA, Bienvenut W, Hoogland C, Sanchez JC, Bairoch A, Hochstrasser DF *et al.*: **Improving protein identification from peptide mass fingerprinting through a parametized multi-level scoring algorithm and an optimized peak detection.** Electrophoresis 1999, **20**:3535-3550.

38. Griffin PR, MacCoss MJ, Eng JK, Blevins RA, Aaronson JS, Yates JR III: **Direct database searching with MALDI-PSD spectra of peptides.** *Rapid Commun Mass Spectrom* 1995, **9**:1546-1551.

39. Yates JR III, Eng JK, McCormack AL: **Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases.** *Anal Chem* 1995, **67**:3202-3210.

40. Yates JRd, Eng JK, McCormack AL, Schieltz D: **Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database.** *Anal Chem* 1995, **67**:1426-1436.

41. •• Eriksson J, Chait BT, Fenyö D: **A statistical basis for testing the significance of mass spectrometric protein identification results.** *Anal Chem* 2000, **72**:999-1005.

This paper introduces significance testing of protein identification results using simulations. It shows how the frequency function for false and random identification can be obtained and how significance testing can be automated and integrated with database search programs.

42. Traini M, Gooley AA, Ou K, Wilkins MR, Tonella L, Sanchez JC, Hochstrasser DF, Williams KL: **Towards an automated approach for protein identification in proteome projects.** *Electrophoresis* 1998, **19**:1941-1949.

43. Binz PA, Muller M, Walther D, Bienvenut WV, Gras R, Hoogland C, Bouchet G, Gasteiger E, Fabbretti R, Gay S *et al.*: **A molecular scanner to automate proteomic research and to display proteome images.** *Anal Chem* 1999, **71**:4981-4988.

44. Quadroni M, James P: **Proteomics and automation.** *Electrophoresis* 1999, **20**:664-677.