Database searching with mass-spectrometric information

Database searching using molecular-mass information has become a popular method for determining the sequence of a protein, a key step in proteomics. This article discusses the various database-search algorithms that are available for protein identification and the issues involved in interpreting the results.

Ronald C. Beavis
beavis@proteometrics.com

David Fenyö
fenyo@proteometrics.com

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

Mass spectrometry (MS) combined with database searching is the method of choice for identifying proteins during proteome projects. In a typical proteomics experiment, the proteins of interest are first enriched and then separated by one- or two-dimensional gel electrophoresis[1–7]. The separated proteins are digested with an enzyme and the proteolytic peptides are analysed by mass spectrometry [or tandem mass spectrometry (MS–MS)]. The protein-separation step is sometimes left out and instead the complex mixture of proteins is digested and the resultant peptides are separated by liquid chromatography before mass analysis[8–10].

The results of using mass spectrometry and database search engines currently depend on the user to a significant extent. The process requires several disparate pieces of software as well as manual intervention by a skilled operator to achieve optimum results. Determining the mass of peaks in a mass spectrum frequently requires a user to look at the values, and fix mistakes made by the signal-processing algorithms. The software that has been developed assumes a highly motivated, knowledgeable user who can quickly evaluate the results and make their own decisions.

In this article, we assess the different database-search algorithms that are available for protein identification and discuss methods for determining the quality of the search results.

Peptide-mapping experiments

The starting point for MS analysis is usually peptide mapping: proteins are digested with an enzyme and the molecular masses of these peptides measured. The database search engines mimic the experiment by calculating the possible peptide masses using the specificity of the enzyme for each protein sequence in a database. The measured masses are then compared to the calculated masses and a score is calculated[11–15]. There are currently two ways to identify proteins from mixtures by peptide mapping. In one, the peptides matching the top-ranking proteins in the first search are removed and a second search is performed[12]. In the other, fusion proteins are formed between the top-ranking proteins and the score is calculated for these fusion proteins[16].

The simplest scoring method for peptide mapping is to count the number of measured peptide masses that match calculated peptide masses within the accuracy of the measurement. This scoring method works well for high-quality experimental data but has the weakness that it usually gives higher scores to larger proteins, for which more possible peptides can be calculated and thus have a higher probability of random matching. More-sophisticated methods for identifying proteins are also based on counting the number of matching peptide masses but they make better use of the experimental information by using our knowledge of proteins to increase the selectivity and sensitivity of the identification[13,16–19].

Extending peptide mapping: MS–MS fragmentation

Proteins from organisms with fully sequenced genomes can often be successfully identified using peptide-mapping information. The success rate for peptide mapping is much lower when applied to organisms with incompletely known genomes or to complex protein mixtures. More experimental information can be obtained by tandem mass spectrometry (MS–MS). Ions corresponding to a single peptide sequence are isolated by a mass spectrometer and fragmented by excitation, which results in unimolecular dissociation reactions, and the masses of the fragment ions measured.

The database search engines mimic the experiment by calculating the possible peptide-fragment masses using the specificity of the enzyme for each protein sequence in a database. The measured fragment masses are compared

 1471-1931/00/$ – see front matter PII: S1471-1931(00)00009-4

with the calculated fragment masses and a score is calculated for the comparison[20,21]. Alternatively, search engines can extract partial sequence information from the spectrum based on 'ion series'.

In contrast to peptide maps, which can contain global information about a protein, tandem mass spectra contain a lot of information on small sections of a protein. There is often enough information available that databases with incomplete gene sequences can be searched, such as the expressed sequence tag (EST) database, which contains partial information in particular on human and mouse genes. MS–MS has also been used to identify complex protein mixtures[22].

Available database-searching algorithms

After the peak masses have been extracted from the mass spectrum, the next step in the analysis is to use this information to identify proteins by database searching. There are many search tools available (Box 1), which are based on different algorithms. Table 1 shows these search tools and the various parameters that can be used to define the experiment and to restrict the search to sections of the protein-sequence database. All the search programs work by comparing the measured masses with masses calculated using protein sequences in a database. Software tools that rank the proteins in the database according to the number of matching peptides include PepSea for peptide mapping[12], MS-Fit[23], MS-Tag[23], PepFrag[24] and PeptIdent/MultiIdent[25].

The MOWSE algorithm[13] has a higher selectivity and sensitivity than these algorithms, which simply count the number of peptide matches. MOWSE takes into account the relative abundances of the peptides of a given mass in the database and also compensates for protein size. MASCOT[17] is based on the MOWSE algorithm but also calculates an approximate probability that the observed match between experimental data and a protein sequence is a random event. MASCOT can use information from both peptide maps and tandem mass spectra to identify proteins.

ProFound[16] uses Bayesian statistics to rank the protein sequences in the database by their probability of having generated the experimental data. The algorithm uses detailed information about each individual protein sequence in the database and can incorporate additional experimental information (e.g. peptide-fragment-mass information, amino acid composition or sequence information) when available. Systematic information that is experimentally obtained is also included in the algorithm (e.g. information about the distribution patterns of proteolytic peptides). ProFound uses a two-step approach to identify protein mixtures. First, the proteins in the database are ranked according to how well they match the experimental data, assuming that a single protein is present; second, fusion proteins are constructed from the top-ranking proteins and ranked. The advantage of the Bayesian approach is that different types of information can be included in a natural way and so it is possible to make optimal use of all of the available information and to increase the sensitivity and selectivity of the algorithm.

PeptIdent2[19] uses a general algorithm that does not incorporate any knowledge about protein properties. PeptIdent has been optimized with a genetic algorithm – using a training set of protein mass spectra to 'evolve' the search parameters to find values that give the best results. This approach is different from that of the other algorithms (ProFound, MOWSE and Mascot), which are based on either our knowledge of the properties of individual proteins or of database averages.

PepSea for MS–MS[21] uses information from peptide sequence tags – short partial amino acid sequences of proteolytic peptides, the mass of the peptide and the masses of the parts of the peptide that have not been sequenced. These searches are fast but require the extraction of the peptide sequence tag before searching (although this can be automatic).

Box 1. URLs for the primary sites associated with database-search algorithms

ProFound
http://www.proteometrics.com/prowl-cgi/ProFound.exe

Mascot
http://www.matrixscience.com/cgi/search_form.pl?SEARCH=PMF

PepSea
http://pepsea.protana.com/PA_PepSeaForm.html
http://pepsea.protana.com/PA_PeptidePatternForm.html

MS-Fit
http://prospector.ucsf.edu/ucsfhtml3.2/msfit.htm

MOWSE
http://srs.hgmp.mrc.ac.uk/cgi-bin/mowse

PeptIdent
http://www.expasy.ch/tools/peptident.html

MultiIdent
http://www.expasy.ch/tools/multiident/

SEQUEST
http://thompson.mbt.washington.edu/sequest/

Mascot
http://www.matrixscience.com/cgi/search_form.pl?SEARCH=MIS

PepFrag
http://www.proteometrics.com/prowl/PepFragch.html

MS-Tag
http://prospector.ucsf.edu/ucsfhtml3.2/mstagfd.htm

Table 1. A comparison of database search algorithm characteristics for MS-related protein identification

Name	WWW[a]	MS type	Taxonomy data	Enzymes	Sequence modifications	Protein properties	Masses	Other inputs
ProFound	Yes	MS and MS–MS	Yes	8 + user defined	User defined, partial and complete	Mass + *p*I	m_0 and A	AA
Mascot	Yes	MS and MS–MS	Yes	10	Predefined partial and complete	Mass	m_0 or A	–
PepSea	Yes	MS and MS–MS	No	8	Cys blocking and Met oxidation	Mass	m_0 or A	Sequence tags
MS-Fit	Yes	MS	Yes	11 and 12 mixtures	Predefined: partial and complete	Mass + *p*I	m_0 or A	AA
MOWSE	Yes	MS	Yes	8	None	Mass	m_0 and A	AA + sequence tags
PeptIdent	Yes	MS	Yes	1	Cys blocking and Met oxidation	Mass + *p*I	m_0 or A	–
MultiIdent	Yes	MS	Yes	9	Cys blocking and Met oxidation	Mass + *p*I	m_0 or A	AA + sequence tags
SEQUEST	No	MS–MS	Yes	User defined	User defined partial and complete	Mass + *p*I	m_0 or A	–
PepFrag	Yes	MS–MS	Yes	5	Cys blocking and phosphorylation	Mass + *p*I	m_0 or A	AA
MS-Tag	Yes	MS–MS	Yes	11 and 12 mixtures	Predefined partial and complete	Mass + *p*I	m_0 or A	AA

Abbreviations: A, average chemical mass; AA, amino acid composition; m_0, monoisotopic mass; MS, mass spectrometry; MS–MS, tandem MS.
[a]Database availability on the world-wide web

SEQUEST[20] calculates a cross-correlation function between the measured tandem mass spectrum and the protein sequences in the database and this 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 the extraction of any information from the mass spectra, but the searches are time consuming.

Evaluating the results generated by automated identifications

If protein identification moves out of specialized laboratories and becomes a widely practiced, high-throughput technique, the manual intervention and decision making prevalent today must disappear. Projections for the throughput of a large-scale proteomics facility are as high as 10^5–10^6 protein identifications per week using current mass spectrometers and computers[26]. Mass spectra must be processed completely automatically to achieve this rate of analysis, using peak-finding techniques that extract the necessary information from the spectra.

Mass spectra obtained from low-resolution instruments ($M/\Delta M < 2000$, where M is the mass assigned to the peak and ΔM is the full width of the peak at half maximum) are already routinely processed in this manner (e.g. the spectra obtained from quadrupole MS and MS–MS devices). The information content of medium-resolution spectra ($2000 < M/\Delta M < 20\,000$) has made fully automated processing more difficult because of partially resolved isotopes and overlapping isotopic distributions. However, there are now algorithms available to perform this task. These algorithms range from simple heuristics (such as deisotoping) to more-reliable data-modeling procedures that calculate isotope-distribution parameters rather than simply counting peaks. The goal of these algorithms is not to assign every peak but to identify proteins automatically and confidently. However, most of the commonly used algorithms are proprietary and will probably never be described in the literature.

The advent of sophisticated peak-finding schemes will remove the current bottleneck in applying easy-to-use, medium-resolution mass spectrometers to protein identification, allowing them to be used for high-throughput

operations. Automated search engines can then be built with more weight placed on mass accuracy (i.e. accurate mass data can be used more intelligently). This level of confidence in mass measurement will allow the search engines to detect and to correct the processed data, to remove operator intervention even more. Data faults such as minor mass-calibration errors and the occasional choice of inappropriate isotope peaks will be detectable and appropriate remediation could occur automatically. These developments in the front end of the process will allow peptide maps to be processed and identifications to be performed and checked by very-limited MS–MS experiments in a matter of seconds.

Even with the best mass measurements, random matching between the MS data and the protein sequences in the database can lead to false identifications. Most search algorithms will return a protein sequence with a highest score, even if the matches are random. It is thus very important to be able to tell what the chance is that a result with a certain score is random. The probability of a false identification can be calculated if the score-frequency function for random identifications is known. One approach to obtaining the score-frequency function for random identifications is by computer simulations[27]. An alternative approach is directly to calculate the probability that a certain protein sequence matches the experimental data[17,18]. The currently available direct calculations are, however, less reliable than the simulation because the process is complex and so it is necessary to make approximations.

The score-frequency function for random identifications can be obtained by computer simulations using the following method[27].

- Select a random protein sequence from a protein-sequence database.
- 'Digest' the selected protein according to the specificity of the enzyme.
- Select a random peptide and calculate its mass.
- Repeat this procedure and construct a synthetic peptide map with the calculated peptide masses, all from different proteins.
- Search the protein-sequence database using the information from the synthetic peptide map and save the top score (corresponding to random matching between a protein sequence in the database and the synthetic peptide map).
- Repeat the searches with different synthetic peptide maps to obtain a distribution of scores for random identification.

The quality of protein-identification results can subsequently be assessed when the score-frequency function for random identifications is known. The score is compared to the frequency function to give the probability that the score is obtained from random matching (i.e. the probability that the protein identified is a false positive). An objective method to assess the quality of the search results is a prerequisite for the automation of protein identification.

Future developments

The future of database search engines that use experimentally determined masses to identify proteins will be shaped by the following developments:

- improvements in the signal-processing algorithms used to generate mass spectra;
- the automation of protein-identification searches and the rational storage of results in large databases;
- the development of data-dependent search engines that can guide the data-gathering process in real time; and
- post-processing using additional statistical data from compiled databases of theoretical and experimentally determined sequence properties.

Simple approaches to increasing the speed of the algorithms, such as increasing processor power to make up for the limitations of existing algorithms (e.g. using clusters of inexpensive off-the-shelf computers), will allow processing to be done on the same time-scale as the measurement process. The invention and adoption of reliable, reproducible front-end algorithms will allow database search engines to develop into much-more-useful discovery tools. The current search engines require complete sequence data in order to make identifications. Peptide-map-based engines require nearly complete, translated, contiguous sequence information and MS–MS-based engines require much shorter regions of translated contiguous information (e.g. EST sequences); however, most techniques are currently limited to fully translated information[28].

The next generation of search engines must deal with gene-level information. A straightforward extension of the use of MS-based information is its use in finding open reading frames and exon–intron combinations in chromosomal DNA. Armed with the knowledge that a gene product exists and that it produces a given set of MS and MS–MS results, it should be possible to assign hypothetical exon–intron patterns to genome sequences using very-simple iterative schemes of data-dependent analysis (Fig. 1). Information made available in this way will be complementary to data obtained from mRNA sequencing and from sequence-interpretation software. The state machine shown in Fig. 1 can be applied to a variety of data-dependent experiments, as long as a hypothesis can be formulated and tested in an iterative fashion. Examples of other problems that are amenable to this approach are the assembly of expressed sequence tags into complete contiguous sequences and the experimental assignment of mRNA alternate-splicing sites.

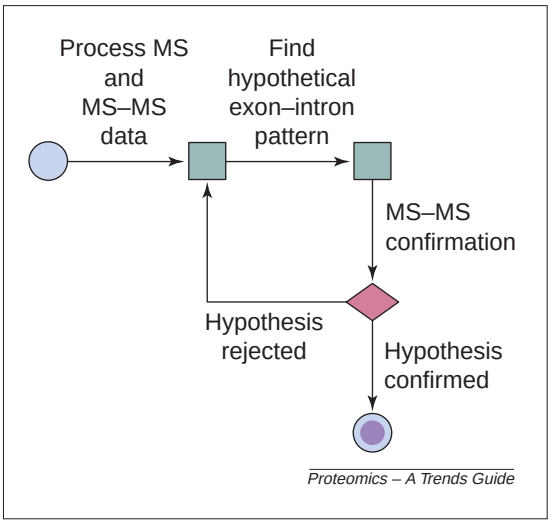

Figure 1. Exon–intron-pattern discovery

A simplified state-machine representation of an exon–intron-pattern discovery search engine. This type of state machine can be used for a wide variety of experiments simply by altering the conditions used to find hypothetical patterns in sequence databases. Abbreviations: MS, mass spectrometry; MS–MS, tandem MS.

Improvements in the speed and error tolerance of protein-identification experiments will allow new types of biological question to be asked and answered by the incorporation of post-processing heuristics into search engines. A large part of interpreting the results of a protein-identification experiment is trying to separate the wheat from the chaff. The investigator must manually perform the following steps:

- reject trivial identifications (e.g. cytoskeletal components, housekeeping and stress-related proteins);
- find identifications confirmatory of their primary hypothesis; and
- carefully watch for unexpected but possibly important sequences that might correlate with their experimental system.

A simple type of post-processing heuristic that could replace the current practice would be a set of appropriate questions that could be asked of large sets of data using either fuzzy logic or Bayesian methods. Results from a large set of experiments could be 'sliced' with a collection of conditions. Compiled databases created using existing sequence-analysis methods, such as BLAST or secondary-structure-prediction algorithms, would be used to formulate a group of fuzzy sets corresponding to these questions in real time. Seeing the inverse degree of participation of all identified proteins in each set would allow the investigator to assess the results: relevant data would be located near the origin. The resulting data 'slices' could then be displayed visually using similar visual metaphors to those currently used commercially to display demographic or customer-preference information. An investigator could then concentrate on repeating these higher-level questions to answer specific biological questions more accurately or completely based on a data set, rather than slogging through the mechanical process of running various types of search engines through databases and manually assembling the results.

References

1 Mortz, E. *et al.* (1994) Identification of proteins in polyacrylamide gels by mass spectrometric peptide mapping combined with database search. *Biol. Mass Spectrom.* 23, 249–261
2 Clauser, K.R. *et al.* (1995) Rapid mass spectrometric peptide sequencing and mass matching for characterization of human melanoma proteins isolated by two-dimensional PAGE. *Proc. Natl. Acad. Sci. U. S. A.* 92, 5072–5076
3 Fey, S.J. *et al.* (1997) Proteome analysis of *Saccharomyces cerevisiae*: a methodological outline. *Electrophoresis* 18, 1361–1372
4 Patterson, S.D. (1997) Identification of low to subpicomolar quantities of electrophoretically separated proteins: towards protein chemistry in the post-genome era. *Biochem. Soc. Trans.* 25, 255–262
5 Link, A.J. *et al.* (1997) Identifying the major proteome components of *Haemophilus influenzae* type-strain NCTC 8143. *Electrophoresis* 18, 1314–1334
6 Wigge, P.A. *et al.* (1998) Analysis of the *Saccharomyces* spindle pole by matrix-assisted laser desorption/ionization (MALDI) mass spectrometry. *J. Cell Biol.* 141, 967–977
7 Rout, M.P. *et al.* (2000) The yeast nuclear pore complex: composition, architecture, and transport mechanism. *J. Cell Biol.* 148, 635–652
8 Yates, J.R., III *et al.* (1997) Direct analysis of protein mixtures by tandem mass spectrometry. *J. Protein Chem.* 16, 495–497
9 Courchesne, P.L. *et al.* (1998) Optimization of capillary chromatography ion trap–mass spectrometry for identification of gel-separated proteins. *Electrophoresis* 19, 956–967
10 Figeys, D. *et al.* (1999) 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.* 71, 2279–2287
11 Henzel, W.J. *et al.* (1993) Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. *Proc. Natl. Acad. Sci. U. S. A.* 90, 5011–5015
12 Mann, M. *et al.* (1993) Use of mass spectrometric molecular weight information to identify proteins in sequence databases. *Biol. Mass Spectrom.* 22, 338–345
13 Pappin, D.D.J. *et al.* (1993) Rapid identification of proteins by peptide-mass fingerprinting. *Curr. Biol.* 3, 327–332
14 Yates, J.R., III *et al.* (1993) Peptide mass maps: a highly informative approach to protein identification. *Anal. Biochem.* 214, 397–408
15 James, P. *et al.* (1993) Protein identification by mass profile fingerprinting. *Biochem. Biophys. Res. Commun.* 195, 58–64
16 Zhang, W. and Chait, B.T. (2000) ProFound – an expert system for protein identification using mass spectrometric peptide mapping information. *Anal. Chem.* 72, 2482–2489
17 Perkins, D.N. *et al.* (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 20, 3551–3567
18 Berndt, P. *et al.* (1999) Reliable automatic protein identification from matrix-assisted laser desorption/ionization mass spectrometric peptide fingerprints. *Electrophoresis* 20, 3521–3526
19 Gras, R. *et al.* (1999) Improving protein identification from peptide mass fingerprinting through a parameterized multi-level scoring algorithm and an optimized peak detection. *Electrophoresis* 20, 3535–3550
20 Eng, J.K. *et al.* (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* 5, 976
21 Mann, M. and Wilm, M. (1994) Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.* 66, 4390–4399
22 Washburn, M.P. and Yates, J.R., III (2000) New methods of proteome analysis: multidimensional chromatography and mass spectrometry. In *Proteomics: A Trends Guide* (Blackstock, W. and Mann, M., eds), pp. 27–30, Elsevier
23 Clauser, K.R. *et al.* (1999) Role of accurate mass measurement (± 10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal. Chem.* 71, 2871–2882

24 Fenyö, D. *et al.* (1998) Protein identification using mass spectrometric information. *Electrophoresis* 19, 998–1005
25 Wilkins, M.R. *et al.* (1998) Multiple parameter cross-species protein identification using MultiIdent – a world-wide web accessible tool. *Electrophoresis* 19, 3199–3206
26 Service, R.F. (2000) Proteomics: can Celera do it again? *Science* 287, 2136–2138
27 Eriksson, J. *et al.* (2000) A statistical basis for testing the significance of mass spectrometric protein identification results. *Anal. Chem.* 72, 999–1005
28 Küster, B. *et al.* (1999) Identifying proteins in genome databases using mass spectrometry. In *Proceedings of the 47th ASMS Conference on Mass Spectrometry and Allied Topics*, pp. 1897–1898, American Society for Mass Spectrometry

New methods of proteome analysis: multidimensional chromatography and mass spectrometry

Shortcomings in two-dimensional gel electrophoresis have encouraged the search for new methods of high-throughput proteome analysis. A variety of chromatographic methods can be coupled directly to a mass spectrometer to accomplish this task. However, multidimensional chromatography must be used to achieve the resolving power of two-dimensional gel electrophoresis. Current systems still fall short of the resolving power of two-dimensional gel electrophoresis but they have the potential to improve.

The analysis of a proteome involves the resolution of the proteins in a sample followed by the identification of the resolved proteins. Two-dimensional polyacrylamide-gel electrophoresis (2D PAGE) followed by mass spectrometry (MS) is the most widely used method of protein resolution and identification. In 2D PAGE, proteins are separated in one dimension by isoelectric point (pI) and in the other dimension by molecular weight. As a result, a single 2D-PAGE system can resolve more than 1500 proteins.

In spite of the contributions of 2D PAGE to proteomics, there are shortcomings to this technology. High-throughput analysis of proteomes is challenging because each spot from 2D PAGE must be individually extracted, digested and analysed – a time-consuming process. In addition, owing to the limited loading capacity of 2D-PAGE gels and the detection limit of staining methods, 2D PAGE currently has an insufficient dynamic range for complete proteome analysis. The dynamic range of a proteome system can be defined as the number of copies per cell of the most-abundant protein identified by a system divided by the number of copies per cell of the least-abundant protein identified by a system. These shortcomings of 2D PAGE as a separation medium for proteomics have encouraged the development of alternative methods for protein or peptide separation.

A true substitute to 2D-PAGE–MS for proteome analysis should resolve proteins as well as 2D PAGE does and also allow the rapid identification of resolved proteins. This article describes some alternative methods that combine separation and identification of proteins, including one- and two-dimensional (1D and 2D, respectively) chromatography methods using high-performance liquid chromatography (HPLC), capillary isoelectric focusing (CIEF), capillary electrophoresis (CE) or microcapillary chromatography. As with 2D PAGE, MS is the method of choice for identifying proteins resolved by liquid separation methods. By eliminating the steps required to transfer proteins from the separation device to the mass spectrometer, several of the methods described might be better suited to high-throughput analysis.

Michael P. Washburn*
mwashburn@u.washington.edu

John R. Yates, III†
jyates@scripps.edu

*Department of Molecular Biotechnology, Box 357730, University of Washington, Seattle, WA 98195-7730, USA.
†Department of Cell Biology, SR11, The Scripps Research Institute, La Jolla, CA 92037, USA.

 1471-1931/00/$ – see front matter PII: S1471-1931(00)00005-7