Laboratory for Computational Proteomics

www.FenyoLab.org
E-mail: Info@FenyoLab.org
Facebook: NYUMC Computational Proteomics Laboratory
Twitter: @CompProteomics
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 © 2000 Elsevier Science Ltd. All rights reserved.

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.
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

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.

### Mascot

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


This paper discusses ProFound — an expert system for protein identification employing MS or MS/MS and database searching. ProFound takes into account detailed information about each individual protein sequences in the protein database and allows for incorporation of additional experimental information.

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

Identifying the proteome: software tools