Laboratory for Computational Proteomics

www.FenyoLab.org

E-mail: Info@FenyoLab.org

Facebook: *NYUMC Computational Proteomics Laboratory*

Twitter: *@CompProteomics*

Chapter 8

Modeling Mass Spectrometry-Based Protein Analysis

Jan Eriksson and David Fenyö

Abstract

The success of mass spectrometry based proteomics depends on efficient methods for data analysis. These methods require a detailed understanding of the information value of the data. Here, we describe how the information value can be elucidated by performing simulations using synthetic data.

Key words: Protein identification, Simulations, Synthetic mass spectra, Significance testing, Value of information, Peptide mass fingerprinting, Tandem mass spectrometry

1. Introduction

Mass spectrometry based proteomics is a method of choice for identifying, characterizing, and quantifying proteins. Proteomics samples are often complex and the range of protein amounts is typically large ($>10^6$), whereas the dynamic range of mass spectrometers is limited ($<10^3$) (1). Because of this mismatch, it is necessary to process the protein samples so that the protein mixture that reaches the mass spectrometer at any given time is much less complex. This is often achieved by first separating the proteins, followed by digestion, and separation of the peptides. The peptides are subsequently analyzed in the mass spectrometer.

With mass spectrometry, it is possible to measure the mass and the intensity of peptide ions and their fragments. To identify proteins and to characterize their posttranslational modifications, the mass measurements are used (2–4) and sometimes to lesser degree the intensity measurements can also be used (5, 6). For quantification, the intensity measurements can be used, but only if the intensity scale is calibrated for each peptide, because the intensity of a peptide ion signal depends strongly on its sequence.

Cathy H. Wu and Chuming Chen (eds.), *Bioinformatics for Comparative Proteomics*, Methods in Molecular Biology, vol. 694,
DOI 10.1007/978-1-60761-977-2_8, © Springer Science+Business Media, LLC 2011

The two most common types of analysis are peptide mass fingerprinting and tandem mass spectrometry. In both these approaches, the proteins are digested with an enzyme having high digestion specificity (usually trypsin) prior to the mass spectrometric analysis. The digestion results in mixtures of proteolytic peptides. In peptide mass fingerprinting the mass spectrometer detects ions of the proteolytic peptides and measures their respective mass. The mass of a proteolytic peptide is typically not unique (7) and therefore observation of several proteolytic peptides from a single protein is needed to generate a peptide mass fingerprint that is useful for protein identification. The peptide mass fingerprinting approach is usually used for samples where the protein of interest can be purified quite well, because peptide ion signals from different proteins can interfere with each other in an individual mass spectrum and the inclusion of mass values of peptides from more than one protein reduces the specificity of the peptide mass fingerprint. In tandem mass spectrometry, individual proteolytic peptide ion species are isolated in the mass spectrometer and are subjected to fragmentation. The masses of the proteolytic peptides and their fragments are measured, making it more applicable to complex mixtures, because a large amount of information is obtained for each peptide and the interference from peptides originating from other proteins is reduced.

Here we describe a few methods for generating synthetic mass spectra, including peptide mass fingerprints and tandem mass spectra. We also give a few examples of how these synthetic mass spectra can be used to better understand the dependence of the value of information in mass spectra on the nature and accuracy of the measurements.

2. Methods

2.1. Peptide Mass Fingerprinting

In peptide mass fingerprinting, protein identification is achieved by comparing the experimentally obtained peptide mass fingerprint to masses calculated from theoretical proteolytic digests of protein sequences from a sequence collection. Each sequence in the collection that has some extent of matching with the experimental peptide mass fingerprint is given a score, the statistical significance of the high scoring matches is tested, and the statistically significant proteins are reported. The statistical significance is tested by generating a distribution of scores for false and random matches. The score of the high-scoring proteins are then compared to the distribution of scores for false and random matches, and the significance level of the match is calculated. The distribution of scores for false and random matches can be obtained by direct calculations (8), by collecting statistics during

the search (9, 10), or by simulations using random synthetic peptide mass fingerprints (11). Here we describe a method for generation of synthetic random peptide mass fingerprints to obtain a distribution of scores for false and random identification that can be used to test the significance of protein identification results (11) (Fig. 1):

1. Analyze the experimental data to obtain information about the parameter space that the synthetic random peptide mass fingerprints should cover, including number of peaks, intensity distribution, mass distribution, and mass accuracy.
2. Select a protein sequence collection, digest it with the enzyme used in the experiment, and calculate the masses of the proteolytic peptides.
3. Randomly pick a set of masses from the proteolytic peptide masses of the sequence collection according to the distributions obtained from the analysis of experimental data, and making sure that no more than one peptide is picked from each protein (see Note 1).
4. Add a mass error sampled from the expected error distribution.
5. Assign intensities to each mass (see Note 2).
6. Search the protein sequence collection and record the highest score.
7. Repeat steps 3–6 until sufficient statistics are obtained, and construct a distribution of scores for false and random identifications.

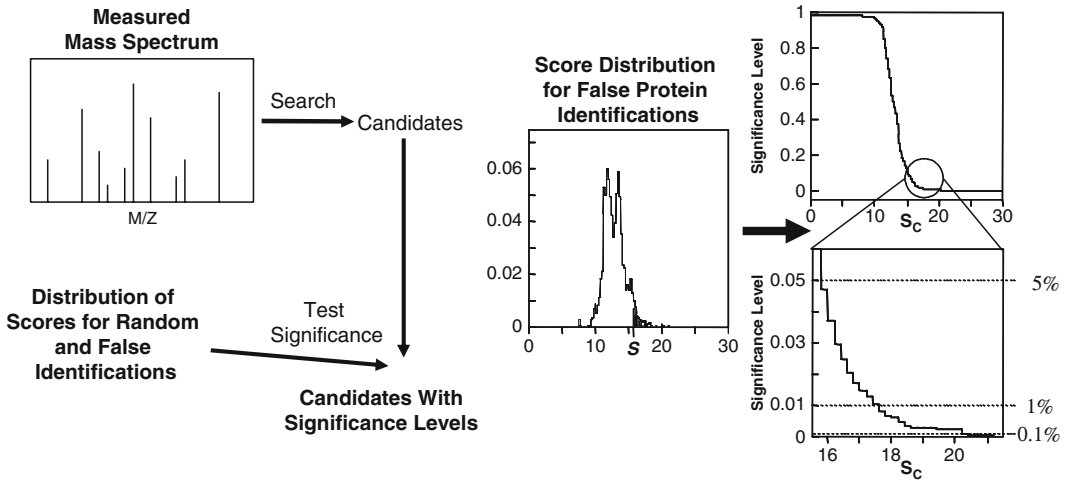

Fig. 1. *Left panel*: The principle of significance testing utilizing the distribution of scores for random and false identifications. *Right panel*: Detailed view of a simulated score distribution for random and false identifications (adapted from (11)).

8. Use the score distribution generated in step 7 to convert the scores from the search with the experimental data to a significance level.

For investigating other aspects of protein identification, it is useful to construct nonrandom peptide mass fingerprints. This can be achieved by modifying step 3:

3a. Select one or more proteins.

3b. For each of the selected proteins, pick a few peptides (see Note 3).

3c. Add background peaks by randomly picking a set of masses from the entire set of proteolytic peptide masses of the sequence collection according to the distributions obtained from the analysis of experimental data, and making sure that no more than one peptide is picked from each protein.

These nonrandom synthetic peptide mass fingerprints can be used to for example improve or compare algorithms, and investigate the effect of search parameters including mass accuracy, enzyme specificity, number missed cleavage sites, and size of sequence collection searched (8, 12). Nonrandom synthetic peptide mass fingerprints have also been used to investigate the potential of identifying complex mixtures of proteins by peptide mass fingerprinting (13). It was concluded that mass fingerprinting could be applied to complex mixtures of a few hundred proteins, if the mass accuracy and the dynamic range of the measurement are sufficient (Fig. 2).

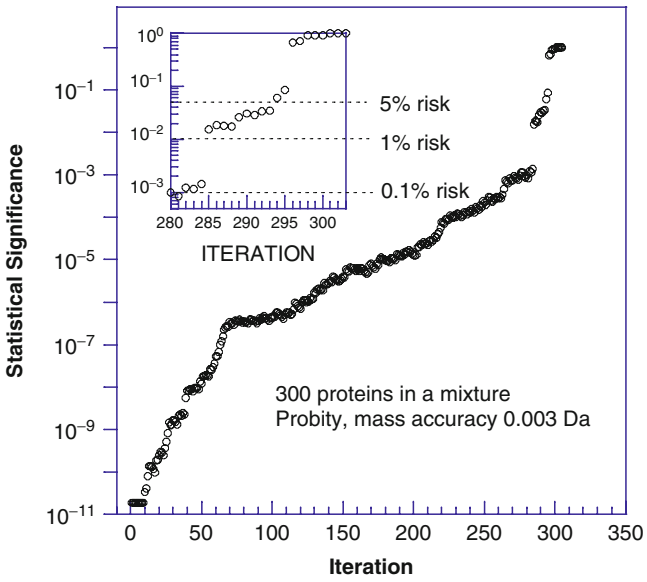

Fig. 2. The statistical significance of proteins identified by peptide mass fingerprinting in a mixture of 300 proteins using an iterative method. The inset displays a magnified portion of the graph for the 280–300th protein identified (Source: ref. 13).

In most practical cases, however, the dynamic range of the measurement is severely limiting and only a few proteins can be identified by peptide mass fingerprinting (14).

2.2. Tandem Mass Spectrometry

The method of choice for complex protein mixtures is to search sequence collections using the observed mass of an intact individual peptide ion species together with the masses of the fragment ions observed upon inducing fragmentation of the peptide in the mass spectrometer. This method requires much lower sequence coverage, and in some cases, even one peptide can be sufficient to identify a protein. Synthetic peptide tandem mass spectra can be generated by the following method:

1. Analyze the experimental data to obtain information about the parameter space of interest (see Note 4 and Fig. 3).
2. Select a protein sequence collection and digest it with the enzyme used in the experiment.
3. Randomly pick a peptide and calculate the peptide mass.
4. Add to the peptide mass an error sampled from the expected error distribution.
5. Calculate the mass of all expected fragment ions.
6. Randomly pick a set of fragment ion masses (Fig. 3a, b).
7. Add to the fragment ion masses an error sampled from the expected error distribution.
8. Assign intensities to each fragment ion mass sampled from the expected error distribution (Fig. 3e).
9. Add background ions by randomly picking peptides that have similar mass as the peptide in step 3, and randomly picking one fragment ion mass from each (Fig. 3c, d).
10. Add to the background masses an error sampled from the expected error distribution.
11. Assign intensities to background fragment ions sampled from the expected intensity distribution (Fig. 3f).
12. Search the protein sequence collection and record the highest score.
13. Repeat steps 6–12 until sufficient statistics are obtained.
14. Repeat steps 3–13 to cover the desired parameter space.

Random synthetic tandem mass spectra can be constructed by skipping steps 3–8 above. These random synthetic tandem mass spectra can be used for significance testing in a similar way as for peptide mass fingerprinting (15).

Nonrandom synthetic tandem mass spectra can, for example, be used to answer the question: How many fragment ions are needed for identification? By generating nonrandom synthetic

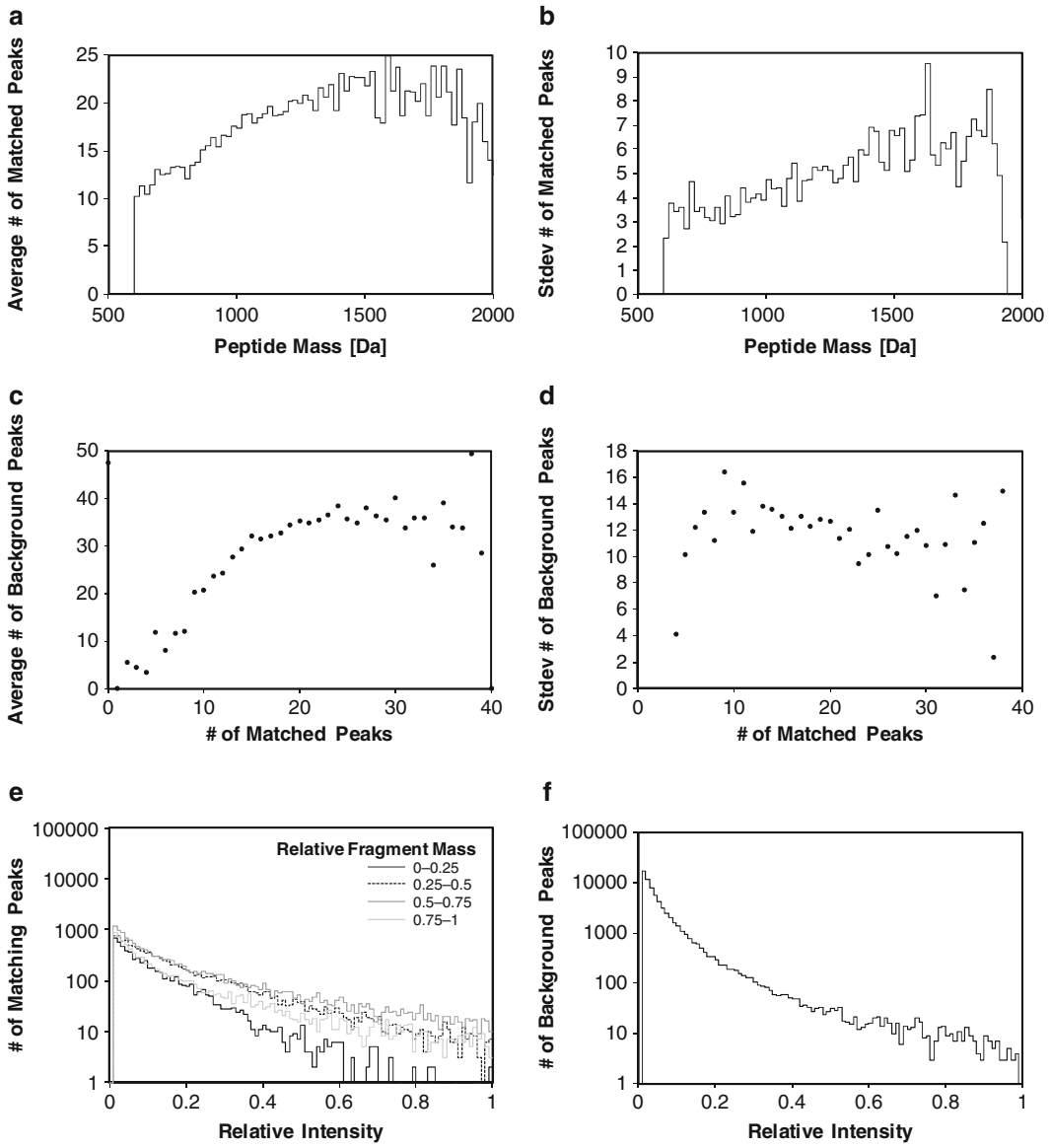

Fig. 3. Properties of tandem mass spectra with significant matches to a dataset acquired with an LTQ-Orbitrap (Thermo Fisher, San Jose, CA): (**a**) the average number and (**b**) the standard deviation of peaks matching the sequence as a function of peptide mass; (**c**) the average number and (**d**) the standard deviation of background peaks as a function of the number of peaks matching the sequence; (**e**) the intensity distribution of matching peaks; and (**f**) the intensity distribution of background peaks.

tandem mass spectra containing varying amounts of sequence information the number of matching fragments needed for identification can be determined (see Note 5 and Fig. 4). In this way it is possible to investigate how many fragment ions are needed for identification depending on the precursor mass, precursor and fragment mass errors, background levels, and modification states (16).

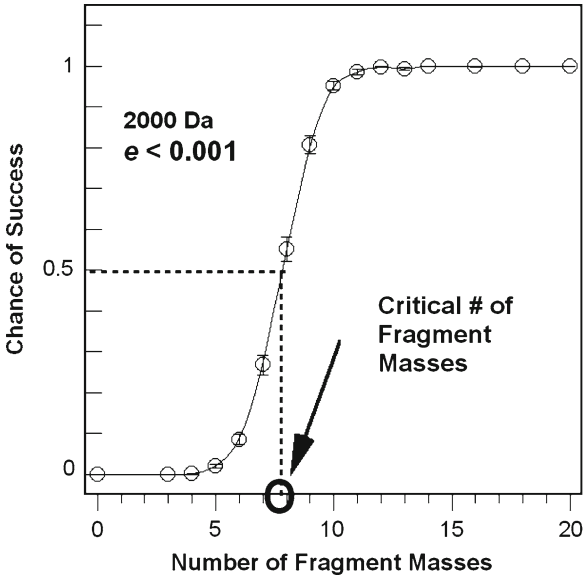

Fig. 4. The chance of success of identification, i.e., the fraction of the spectra that yield a true result and an *e*-value below a desired threshold, as a function of the number of fragment masses in the spectra. Each data point represents the mean value with standard error of the results for 50 randomly selected peptides and with 20 different randomly generated spectra from each peptide. The chance of success is low for few matching fragment and high for many matching fragments. The critical number of fragment masses is defined as the number of fragment masses that yield a 50% chance of success.

3. Notes

1. The distribution of peptide masses is far from uniform, because peptides contain only a few different types of atoms, and it is, therefore, important to use actual peptide masses in simulations. The distribution of peptide masses consists of peaks with centroids approximately 1 Da apart, and regions in between the peaks that are devoid of peptide masses. Using a uniform mass distribution would therefore result in unrealistic synthetic peptide mass fingerprints.
2. The intensities are often set to the same value for all masses. Alternatively, an intensity distribution derived from experimental data can be used.
3. The number of peptides to pick can for example be determined by selecting a target coverage for the proteins, and then randomly picking peptides until that coverage is reached.
4. An example of the kind of information that can be extracted from experiments is shown in Fig. 3. First the data acquired

on an LTQ-Orbitrap was searched using X! Tandem and all peptides with expectation value $<10^{-3}$ were used to characterize the data set. The average and the standard deviation of the number of ions that match the peptide sequence first increases with mass, and at masses above 1,500 Da the average saturates (Fig. 3a, b). The average number of background peaks increases with the number of matching peaks up to about 15 matching peaks, and then saturates (Fig. 3c). The standard deviation of the number of background peaks is constant within the uncertainty of the measurement (Fig. 3d). The matching peaks dominate at high intensity, but even though the majority of peaks with low relative intensity are background (<20% of the base peak), there are still a considerable number of low-intensity peaks that match the sequence (Fig. 3e, f).

5. Tryptic peptides were randomly selected from a proteome, and a set of fragment mass spectra was generated for each selected peptide assuming that they were unmodified or phosphorylated. These fragment mass spectra were constructed by randomly selecting fragment ions, and the number of fragments selected was varied over a wide range. The fragment mass spectra were searched against the proteome using X! Tandem and the probability of successful peptide identification was obtained as a function of the number of fragment ions in the spectra. From these curves, the critical number of fragment masses was derived for a given experimental condition, i.e., the number of fragment masses needed for successfully identifying half of the peptides.

Acknowledgments

This work was supported by funding provided by the National Institutes of Health Grants CA126485, DE018385, NS050276, RR00862 and RR022220, the Carl Trygger foundation, and the Swedish research council.

References

1. Eriksson J, Fenyo D. (2007) Improving the success rate of proteome analysis by modeling protein-abundance distributions and experimental designs. *Nat Biotechnol* **25**, 651–655.
2. Henzel WJ, Billeci TM, Stults JT, Wong SC, Grimley C, Watanabe C. (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.
3. Mann M, Wilm M. (1994) Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal Chem* **66**, 4390–4399.
4. Eng JK, McCormack AL, Yates JR. (1994) An approach to correlate mass spectral data with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* **5**, 976.

5. Craig R, Cortens JC, Fenyo D, Beavis RC. (2006) Using annotated peptide mass spectrum libraries for protein identification. *J Proteome Res* **5**, 1843–1849.
6. Lam H, Deutsch EW, Eddes JS, Eng JK, King N, Stein SE, Aebersold R. (2007) Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* **7**, 655–667.
7. Fenyo D, Qin J, Chait BT. (1998) Protein identification using mass spectrometric information. *Electrophoresis* **19**, 998–1005.
8. Eriksson J, Fenyo D. (2004) Probity, a protein identification algorithm with accurate assignment of the statistical significance of the results. *J Proteome Res* **3**, 32–36.
9. Field HI, Fenyo D, Beavis RC. (2002) RADARS, a bioinformatics solution that automates proteome mass spectral analysis, optimises protein identification, and archives data in a relational database. *Proteomics* **2**, 36–47.
10. Fenyo D, Beavis RC. (2003) A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal Chem* **75**, 768–774.
11. Eriksson J, Chait BT, Fenyo D. (2000) A statistical basis for testing the significance of mass spectrometric protein identification results. *Anal Chem* **72**, 999–1005.
12. Eriksson J, Fenyo D. (2004) The statistical significance of protein identification results as a function of the number of protein sequences searched. *J Proteome Res* **3**, 979–982.
13. Eriksson J, Fenyo D. (2005) Protein identification in complex mixtures. *J Proteome Res* **4**, 387–393.
14. Jensen ON, Podtelejnikov AV, Mann M. (1997) Identification of the components of simple protein mixtures by high accuracy peptide mass mapping and database searching. *Anal Chem* **69**, 4741–4750.
15. Eriksson J, Fenyo D. (2009) Peptide identification with direct computation of the significance level of the results. *Proceedings of the 57th ASMS Conference on Mass Spectrometry.*
16. Fenyo D, Ossipova E, Eriksson J. (2008) The peptide fragment mass information required to identify peptides and their post-translational modifications. *Proceedings of the 56th ASMS Conference on Mass Spectrometry.*