7

OPTIMIZING SENSITIVITY AND SPECIFICITY IN MASS SPECTROMETRIC PROTEOME ANALYSIS

Jan Eriksson and David Fenyö

Proteomics studies aim at answering questions about a biological system by characterizing all its proteins (see also Chapter 10). The proteins are typically characterized by analyzing carefully chosen samples from the biological system by mass spectrometry [1]. The mass spectrometric information should ideally be sufficient to answer two questions about each sample: what does it contain and how much? In order to answer these questions appropriately a researcher has to face the three central problems of mass spectrometry based proteomic research: (i) the design of the experiment to allow for detection of proteins that are present in low abundance in the biological system [2]; (ii) the optimal use of the experimental information to allow for statistically significant identification [3] and quantitation [4] of the proteins detected; and (iii) the accurate assignment of the significance levels of the results [5].

The success in solving these three central problems will depend on many factors in a given experiment. We will here use different terms to describe how the approach in a given experiment can handle the central problems: *Success rate* and *relative dynamic range* [2], which are specific to proteomics experiments and will be defined stringently below, and the terms *sensitivity* and *selectivity*, which originate from mathematical statistics. The *sensitivity* is a measure of how good the method employed is at identifying a protein that is actually present in the sample. The *specificity* is a measure of how good the method is at not reporting a result when a protein is absent from the sample. The focus in this chapter is on the question of what is in the sample, that is, identification of proteins, but

Mass Spectrometry. Edited by Ekman, Silberring, Westman-Brinkmalm, and Kraj

we will first describe briefly what should be considered with respect to the information obtained from experiments aiming at answering the question of how much there is of different components in the sample using MS-based quantitation.

7.1. QUANTITATION

In proteomics, quantitation is typically a *comparison of two biological systems*, for example, cells in a normal state versus cells in a transformed state. MS-based quantitation utilizes analyses of digested extracted proteins (Fig. 7.1a). The comparison

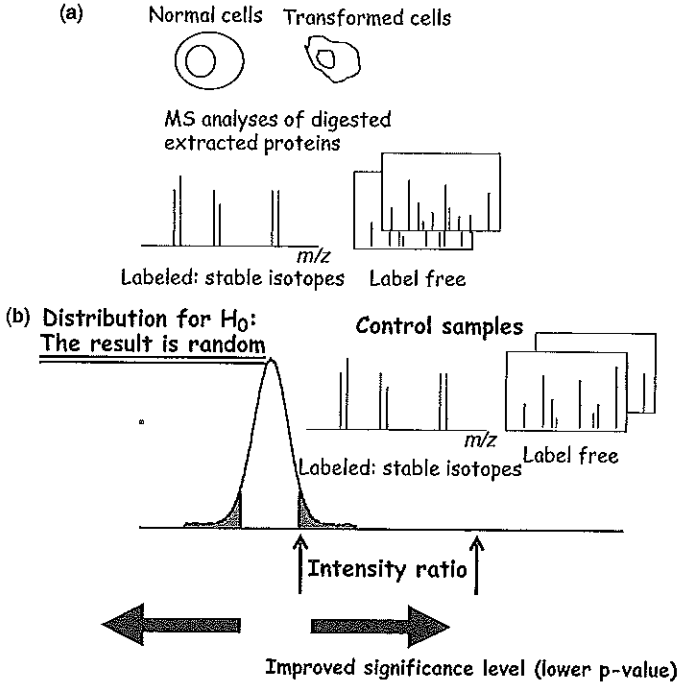

Figure 7.1. (a) Quantitative comparison of MS-signals from two different cell systems can be done by either of two basic principles: (i) stable isotope labeling of one system followed by mixing the systems prior to the MS-analysis and comparison of the intensities of pairs of signals from labeled and unlabeled (or differently labeled) ions; (ii) by label-free analysis where mass spectra are acquired separately from the systems and comparisons of signal intensities of specific m/z values are done between the spectra: (b) The statistical significance of ratios between signal intensities from the system can be judged once the distribution of intensity ratios for control samples with no known systematic difference have been obtained. The significance level of a quantitation measurement is better the smaller the overlap with the distribution for the control samples. The significance level is given by the red area under the distribution curve.

between the two systems is done either using stable isotope labeling of one of the systems and mixing of the proteolytic peptides from the two systems prior to MS analysis, or by so-called label-free analysis in which spectra are acquired separately from each system [6]. It is an advantage to introduce the stable isotope label and mix the samples as early in the experimental protocol as possible, because experimental variation is then minimized. In all of these approaches the intensity ratios between MS signals from individual peptides must be determined and are employed as a measure. It is important to determine whether it is plausible that the ratio represents a true difference between the two systems, that is, if the result can be discerned from a result corresponding with no difference between the systems. In order to answer that question control samples with no biological difference between the systems must be analyzed and all intensity ratios computed. This set of intensity ratios yields a distribution that represents the hypothesis that a given result is random. Hence, from this distribution the p-value (significance level) of a result (intensity ratio) from a real quantitation experiment can be computed (Fig. 7.1b).

7.2. PEPTIDE AND PROTEIN IDENTIFICATION

The identification of peptides and proteins using MS information can be done in three different fashions: (i) de novo sequencing, (ii) library searching, and (iii) sequence collection searching.

De novo sequencing utilizes the information from an MS-MS spectrum of a peptide isolated and fragmented in the mass spectrometer (see also Chapter 2). The spectra are analyzed with respect to mass differences that correspond to mass values of individual amino acids or stretches of peptide sequences. This information is employed for proposing the most likely sequence of the peptide analyzed [7, 8]. Advantages of this approach include no need for a sequence collection, allowing the sequencing of proteins from organisms that have not yet been sequenced. The main disadvantage is the need for excellent data quality.

Library searching compares MS-MS spectra of a peptide isolated and fragmented in the mass spectrometer to a library of peptide MS-MS spectra [9–11]. The analysis aims at identifying a peptide by finding the best similarity between an MS-MS spectrum and a member of the spectrum library. This approach is very fast and sensitive, since the comparisons involve real spectra of observed peptides using the intensity information in the comparison. It is, however, important that only high-quality spectra are included in the libraries. This approach does not work for analysis of peptides not already detected, but there is a rapidly growing number of peptide MS-MS spectra in the public domain [12–14], allowing the construction of spectrum libraries with good coverage for many mammals, fungi, and bacteria.

Sequence collection searching aims at identifying proteins or peptides from mass spectrometric information and information from protein sequence collections generated from genome sequencing. Sequence collection searching is the major approach for protein identification and exists in two different versions: (i) peptide mass fingerprinting, which utilizes a mass spectrum of proteolytic peptides from an individual protein digested with a specific enzyme and assumes that the proteolytic peptide masses yield

a fingerprint of the protein [15, 16]. It is also assumed that the fingerprint can be recognized when searching a set of theoretical mass fingerprints derived by computing the mass values resulting from in silico digestion of each sequence in a sequence collection. (ii) Sequence searching using MS-MS information, where a set of mass values detected from a proteolytic peptide ion isolated and fragmented in the mass spectrometer is compared with theoretical proteolytic peptide fragment mass values generated in silico for each proteolytic peptide in a protein sequence collection (Fig. 7.2) [17–20].

Significance testing is important for minimizing false results. Identification using any of the methods mentioned above involves the scoring of each comparison between the experimental data and the model, followed by ranking of the models. Unfortunately, there is a risk of obtaining false results, since mass values measured are not unique for an individual peptide or peptide mass fragment (Fig. 7.3) [21]. In analogy with what is described above for quantitation there is a need to evaluate an identification result with respect to its statistical significance. The significance level (p-value) of a result can be determined once the distribution of scores for false (random) results is known (Fig. 7.4). Such distributions are specific for each algorithm employed in the scoring procedure. There are three different ways of generating the score distribution for false results: simulation [5], collecting statistics during the search [22–27], and direct computation [3].

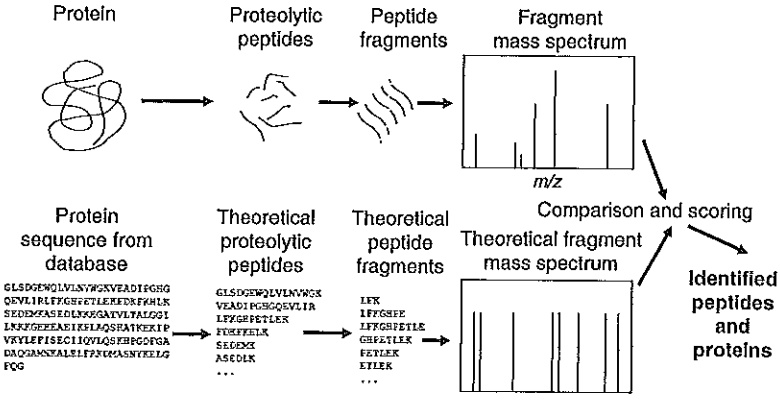

Figure 7.2. Protein identification using sequence collection searching and MS/MS data of proteolytic peptides fragmented in the mass spectrometer. Search conditions such as fragmentation pathways and mass accuracy are specified prior to the search and in the search procedure a computational algorithm compares the mass information from the experiment with theoretical mass information obtained by in silico fragmentation of each proteolytic peptide in a sequence collection. The peptides in the sequence collection are given a score that measures the degree of matching with the experimental MS/MS information and the peptide in the sequence collection that displays the best score is given the highest rank and is assumed as the identification result.

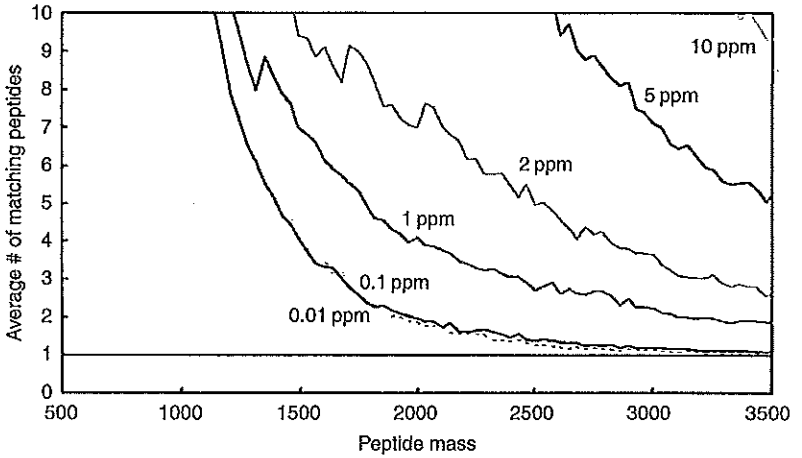

Figure 7.3. The average number of peptides matching within various mass windows (ppm) as a function of the peptide mass (Da) for proteins from *H. sapiens* completely digested with trypsin. Note that there is negligible increase in the information value (no reduction in the number of matches) below 0.1 ppm.

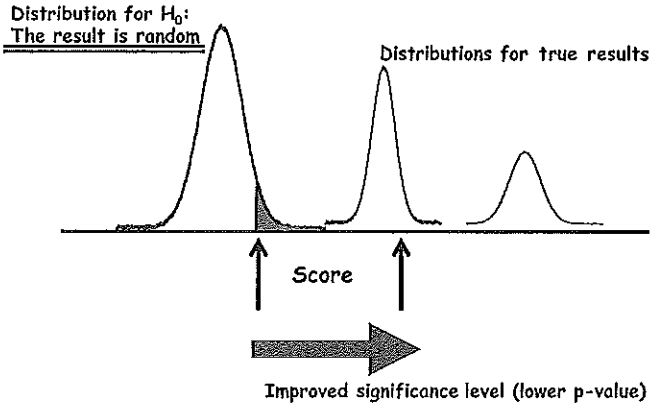

Figure 7.4. The significance level of an identification result can be determined once the distribution of scores for false identification results is known. Score distributions for true results can vary between experiments and are typically unknown, in contrast with the distribution of scores for false identification results, which can be derived by various methods (see text for details). A score that is in a region with little overlap with the distribution for false results yields a good significance level (the gray area is small).

It is critical to optimize the experimental design and data analysis to maximize the resulting information. Score distributions for true results vary between experiments and typically these distributions are unknown, since it is difficult to prove that a result is true unless the data used is synthetic or the data is from a control sample characterized with an independent and reliable method. It is desirable that the score distributions for true and false results are well separated so that the score itself can be employed as a means for minimizing the number of false results not rejected and to minimize the number of true results rejected (Fig. 7.4). An indirect view of the separation between these distributions is given by a so-called ROC-curve. In the simplest form a ROC-curve is plotted with the frequency of true results as a function of the frequency of false results and with the data points organized so that the score becomes worse with increasing distance from the origin of the graph. This can be slightly modified into plotting the *sensitivity* versus 1 − *selectivity*, where

$$\text{Sensitivity} = \frac{\#\text{ of true results not rejected}}{\text{total }\#\text{ of true results}}$$

and

$$1 - \text{Selectivity} = \frac{\#\text{ of false results not rejected}}{\text{total }\#\text{ of false results}}$$

The sensitivity and the selectivity depend on the choice of algorithm. A simple way to examine what influences the sensitivity and the selectivity is to employ synthetic data and simulate protein or peptide identification. Figure 7.5a displays a ROC-curve comparing peptide mass fingerprint-based identification of *Saccharomyces cerevisiae* proteins using a set of PMFs generated in silico where in each PMF four mass values were correlated with a single protein and 16 mass values were chosen randomly. The same data set was employed for searching the *S. cerevisiae* sequence collection using two different search algorithms: algorithm 1, Probity and algorithm 2, which ranks simply based on the number of matching mass values in each PMF. It is seen in Fig. 7.5a that for algorithm 1 there is a region along the y-axis where good scores yield only true results, whereas for algorithm 2, the score more or less arbitrarily indicates a true or a false result also for the best scores. From this simulation example we learn that the sensitivity and the selectivity depend on the choice of algorithm.

The sensitivity and the specificity depend on the search conditions. Fig. 7.5b indicates results from a simulation using synthetic MS-MS spectra generated in silico from the *S. cerevisiae* sequence collection. Each spectrum contained 25 peptide fragment mass values, but only seven mass values corresponding to an individual peptide. These spectra were employed for searching the *S. cerevisiae* sequence collection using the algorithm X! Tandem in two sessions employing different search conditions. In one session the windows for accepted mass errors of both the peptide itself and its fragments were ten times larger than in the other session. Based on the distinct different in the outcome from the two sessions displayed in Fig. 7.5b we conclude that the sensitivity and the specificity depend on the search conditions.

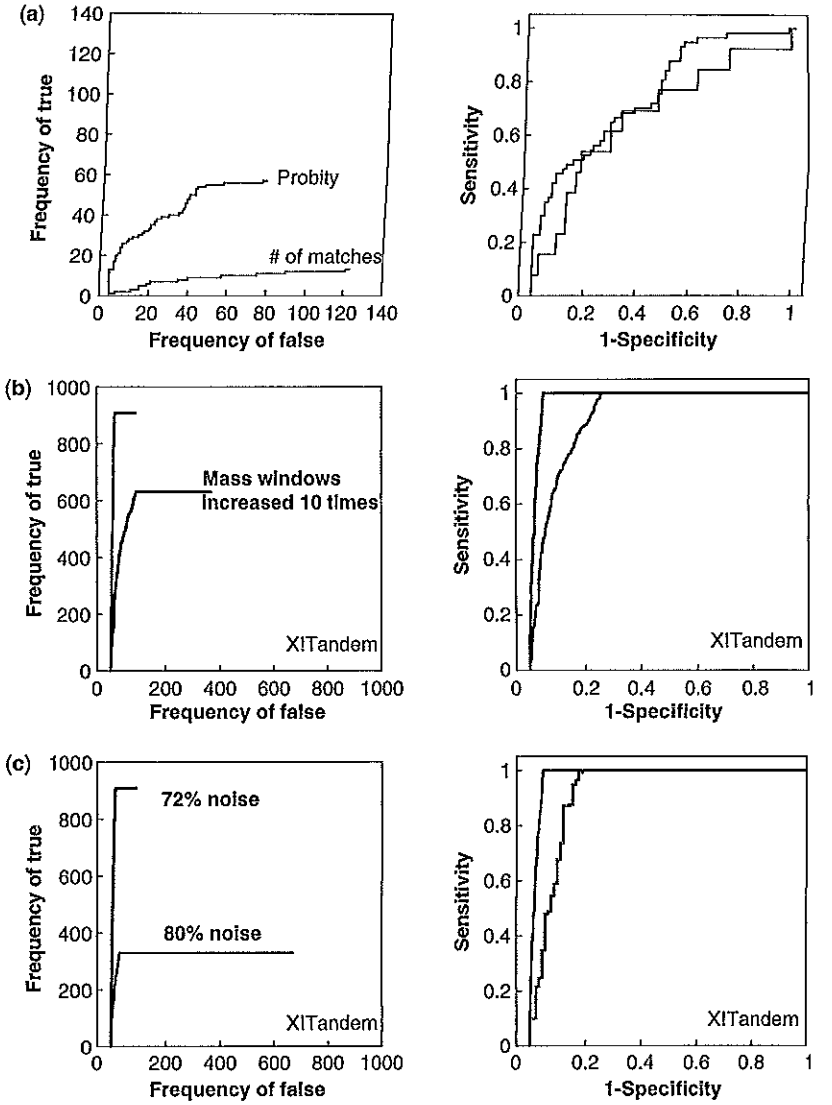

Figure 7.5. Simulation results that elucidate how the sensitivity and the selectivity of a proteomics experiment depend on various features: (a) The choice of algorithm. The probity algorithm displays better sensitivity and selectivity than an algorithm that ranks strictly based on the number of matches. (b) The search conditions. Increasing the mass window of a search 10 times when searching with data that display small mass errors yields worse sensitivity and selectivitry. (c) The quality of the data. Data with less noise yields better sensitivity and selectivity.

The sensitivity and the selectivity depend on the data quality. Figure 7.5c displays results from a simulation employing the same data set as was used in Fig. 7.5b, together with a simulation in which the MS-MS spectra have only five mass values corresponding to an individual peptide (out of 25). Hence we see that the sensitivity and the selectivity depend on the data quality.

7.3. SUCCESS RATE AND RELATIVE DYNAMIC RANGE

We have already concluded that the data quality is an issue for the sensitivity and the selectivity for protein identification in MS-based proteomics experiments. A related issue is that we do not acquire data for all the proteins actually present in the sample. The reason for this is that there is a discrepancy between the experimental dynamic range and the range of protein abundances in the proteome. The bell-shaped curve shown in Fig. 7.6a is an approximation of the protein amount distribution measured for yeast (*S. cerevisiae*) using immunodetection methods [28]. The range of protein abundances in yeast spans six orders of magnitude. It is believed that for human body fluids the range of protein abundances is at least 10^{10}. The dynamic range of a mass spectrometer can be as low as 10^2 (for generating signals from two substances present in the sample at a given point in time). Proteomics researchers have realized that the complexity and the range of protein abundance of a proteome make it necessary to

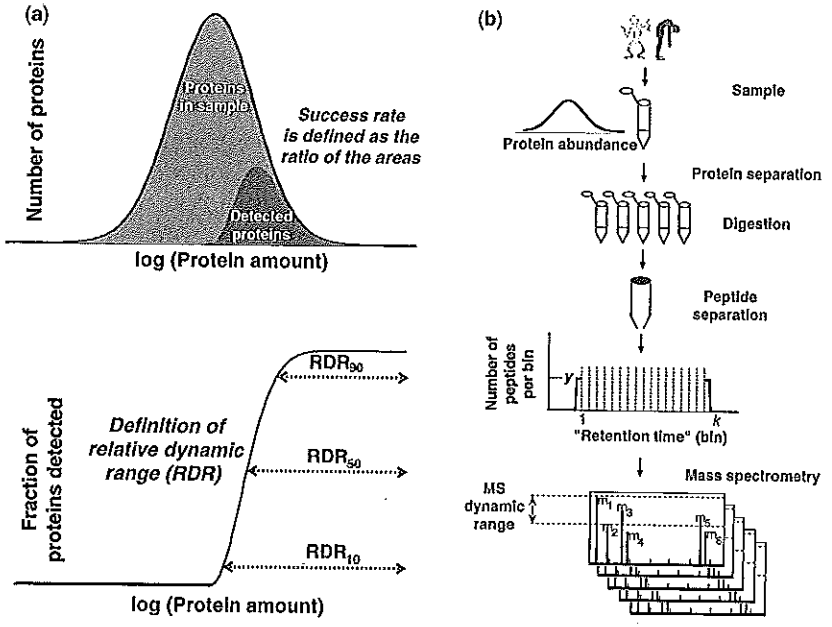

Figure 7.6. (a) Definitions of success rate and relative dynamic range. (b) Model of a proteomics experiment. (See color insert.)

apply various separation protocols prior to the MS analysis. There are many options to choose from in this respect and it is impossible to examine the merits of all combinations experimentally. By constructing a model of a proteomics experiment and a model of the protein abundance distribution of a proteome it is possible to use computer simulations to examine how good a particular experimental design would be for detecting the proteins of that proteome. In such simulations the quantities studied are the success rate and the relative dynamic range (RDR), where the success rate indicates how many proteins are detected divided by the total number of proteins in the proteome and the RDR indicates how deep down into the low abundance proteins an experimental design can manage to detect proteins (see Fig. 7.6a). The experimental design can be described by a set of parameters (Fig. 7.6b) and we will here give an example of how one feature of the sample preparation and two features of the mass spectrometer influence the success rate and theRDR: the degree of protein separation, the MS detection limit,

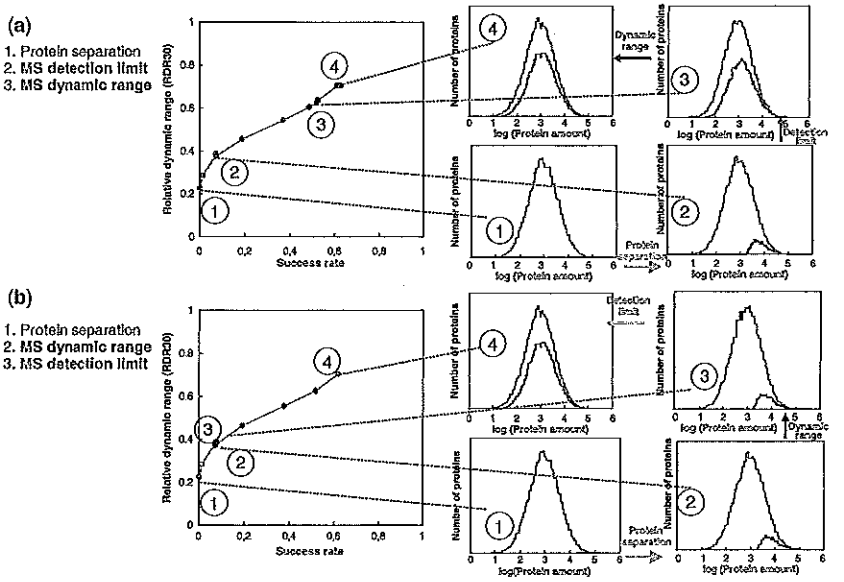

Figure 7.7. Results from model simulations showing the effect of protein separation and the effect of MS detection limit and MS dynamic range on the success rate and the relative dynamic range (RDR) for detection of proteins from *H. sapiens* tissue samples. (a) *Left:* RDR as a function of success rate when first improving the protein separation going from 30,000 proteins (1) to 300 proteins (2), then enhancing the sensitivity of the mass spectrometer from 1 fmol to 1 amol (3), and finally improving the MS dynamic range from 10^2 to 10^4 (4). *Right:* The protein abundance distribution assumed for human tissue together with the distribution of the proteins detected for the experimental designs 1 to 4. (b) Same as in (a), but with the MS dynamic range improved prior to improving the MS detection sensitivity. Note that the effect of improving the dynamic range is negligible compared with the effect of improving the detection sensitivity. (See color insert.)

and the MS dynamic range. The top left panel of Fig. 7.7a indicates how the success rate and the RDR vary when first improving the protein separation, then improving the MS detection limit and finally improving the MS dynamic range. The right panel of Fig. 7.7a shows the protein abundance distribution model employed in the simulation together with the distribution of the proteins detected for the initial design (1), the design with improved protein separation (2), after improving the detection limit (3), and after enhancing the MS dynamic range (4). It is evident that all these three features of the experimental design can influence strongly the outcome of an experiment. The way in which design parameters are changed can, however, have a strong influence on the result. For example, if instead of improving the protein separation, the MS dynamic range is improved, there is no improvement of the success rate and the RDR until the MS detection limit also is improved (Fig. 7.7b, 1–4).

7.4. SUMMARY

Computations and simulations are important tools for examining the performance of mass spectrometry-based proteomic research. Computations are necessary for deriving distributions for results corresponding with "no difference between the systems" for quantitation experiments and for results corresponding with "a false result" for identification experiments. We have demonstrated using simulations that the sensitivity, that is, the ability to identify a protein present in the sample, and the selectivity, that is, the ability to not report proteins absent from the sample, depend on three factors: (i) the choice of protein identification algorithm (including search conditions), (ii) the data quality, and (iii) the experimental design.

REFERENCES

1. R. Aebersold and M. Mann. Mass Spectrometry-Based Proteomics. *Nature*, **422**(2003): 198–207.
2. J. Eriksson and D. Fenyo. Improving the Success Rate of Proteome Analysis by Modeling Protein-Abundance Distributions and Experimental Designs. *Nat. Biotechnol.*, **25**, no. 6 (2007): 651–655.
3. J. Eriksson and D. Fenyo. Probity: A Protein Identification Algorithm with Accurate Assignment of the Statistical Significance of the Results. *J. Proteome Res.*, **3**, no. 1 (2004): 32–36.
4. R. Aebersold. Quantitative Proteome Analysis: Methods and Applications. *J. Infect. Dis.*, **187**, Suppl 2(2003): S315–20.
5. J. Eriksson, B. T. Chait, and D. Fenyo. A Statistical Basis for Testing the Significance of Mass Spectrometric Protein Identification Results. *Anal. Chem.*, **72**, no. 5 (2000): 999–1005.
6. C. Fenselau. A Review of Quantitative Methods for Proteomic Studies. *J. Chromatogr. B, Analyt. Technol. Biomed. Life Sci.*, **855**, no. 1 (2007): 14–20.
7. J. A. Taylor and R. S. Johnson. Sequence Database Searches via de novo Peptide Sequencing by Tandem Mass Spectrometry. *Rapid Commun. Mass Spectrom.*, **11**, no. 9 (1997): 1067–1075.

8. A. Frank and P. Pevzner. PepNovo: De novo Peptide Sequencing via Probabilistic Network Modeling. *Anal. Chem.*, **77**, no. 4 (2005): 964–973.
9. R. Craig, et al., Using Annotated Peptide Mass Spectrum Libraries for Protein Identification. *J. Proteome Res.*, **5**, no. 8 (2006): 1843–1849.
10. B. E. Frewen, et al., Analysis of Peptide MS/MS Spectra from Large-Scale Proteomics Experiments Using Spectrum Libraries. *Anal. Chem.*, **78**, no. 16 (2006): 5678–5684.
11. H. Lam, et al., Development and Validation of a Spectral Library Searching Method for Peptide Identification from MS/MS. *Proteomics*, **7**, no. 5 (2007): 655–667.
12. R. Craig, J. P. Cortens, and R. C. Beavis. Open Source System for Analyzing, Validating, and Storing Protein Identification Data. *J. Proteome Res.*, **3**, no. 6 (2004): 1234–1242.
13. L. Martens, et al., PRIDE: The Proteomics Identifications Database. *Proteomics*, **5**, no. 13 (2005): 3537–3545.
14. F. Desiere, et al., The PeptideAtlas Project. *Nucleic Acids Res.*, **34**, Database issue (2006): D655–D658.
15. W. J. Henzel, et al., 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**, no. 11 (1993): 5011–5015.
16. W. Zhang and B. T. Chait. ProFound: An Expert System for Protein Identification Using Mass Spectrometric Peptide Mapping Information. *Anal. Chem.*, **72**, no. 11 (2000): 2482–2489.
17. M. Mann and M. Wilm. Error-Tolerant Identification of Peptides in Sequence Databases by Peptide Sequence Tags. *Anal. Chem.*, **66**, no. 24 (1994): 4390–4399.
18. J. Eng, A. L. McCormack, and J. R. Yates, III, *J. Am. Soc. Mass Spectrom.*, **5**(1994): 976–989.
19. D. N. Perkins, et al., Probability-Based Protein Identification by Searching Sequence Databases Using Mass Spectrometry Data. *Electrophoresis*, **20**, no. 18 (1999): 3551–3567.
20. R. Craig and R. C. Beavis. TANDEM: Matching Proteins with Tandem Mass Spectra. *Bioinformatics*, **20**, no. 9 (2004): 1466–1467.
21. D. Fenyo, J. Qin, and B. T. Chait. Protein Identification Using Mass Spectrometric Information. *Electrophoresis*, **19**, no. 6 (1998): 998–1005.
22. H. I. Field, D. Fenyo, and R. C. Beavis. RADARS, a Bioinformatics Solution that Automates Proteome Mass Spectral Analysis, Optimises Protein Identification, and Archives Data in a Relational Database. *Proteomics*, **2**, no. 1 (2002): 36–47.
23. A. Keller, et al., Empirical Statistical Model to Estimate the Accuracy of Peptide Identifications made by MS/MS and Database Search. *Anal. Chem.*, **74**, no. 1 (2002): 5383–5392.
24. R. E. Moore, M. K. Young, and T. D. Lee. Qscore: An Algorithm for Evaluating SEQUEST Database Search Results. *J. Am. Soc. Mass Spectrom.*, **13**, no. 4 (2002): 378–386.
25. D. Fenyo and R. C. Beavis. A Method for Assessing the Statistical Significance of Mass Spectrometry-Based Protein Identifications using General Scoring Schemes. *Anal. Chem.*, **75**, no. 4 (2003): 768–774.
26. A. I. Nesvizhskii, et al., A Statistical Model for Identifying Proteins by Tandem Mass Spectrometry. *Anal. Chem.*, **75**, no. 17 (2003): 4646–4658.
27. J. Peng, et al., Evaluation of Multidimensional Chromatography Coupled with Tandem Mass Spectrometry (LC/LC-MS/MS) for Large-Scale Protein Analysis: The Yeast Proteome. *J. Proteome Res.*, **2**, no. 1 (2003): 43–50.
28. S. Ghaemmaghami, et al., Global Analysis of Protein Expression in Yeast. *Nature*, **425**, no. 6959 (2003): 737–741.