Chapter 11

Mass Spectrometric Protein Identification Using the Global Proteome Machine

David Fenyö, Jan Eriksson, and Ronald Beavis

Abstract

Protein identification by mass spectrometry is widely used in biological research. Here, we describe how the global proteome machine (GPM) can be used for protein identification and for validation of the results. We cover identification by searching protein sequence collections and spectral libraries as well as validation of the results using expectation values, rho-diagrams, and spectrum databases.

Key words: Proteomics, Mass spectrometry, Protein identification, Spectrum libraries, Validation

1. Introduction

Mass spectrometry-based protein identification has become an invaluable tool for elucidating protein function, and several methods have been developed for protein identification, including sequence collection searching with masses of peptides or their fragments, spectral library searching, and de novo sequencing (Fig. 1).

The first step in protein identification is to find peaks in the mass spectra that correspond to peptides and their fragments. It is important to find all the relevant peaks and at the same time minimizing the number of background peaks. This can be achieved by scanning the spectra for peaks of the expected width and selecting peaks above a signal to noise threshold (see Note 1), and then picking the monoisotopic peak for each isotope cluster (see Note 2). After picking the peaks, spectra with low information content that could not produce any meaningful results can be removed to increase the speed of subsequent analysis (1).

David Fenyö (ed.), *Computational Biology*, Methods in Molecular Biology, vol. 673,
DOI 10.1007/978-1-60761-842-3_11, © Springer Science+Business Media, LLC 2010

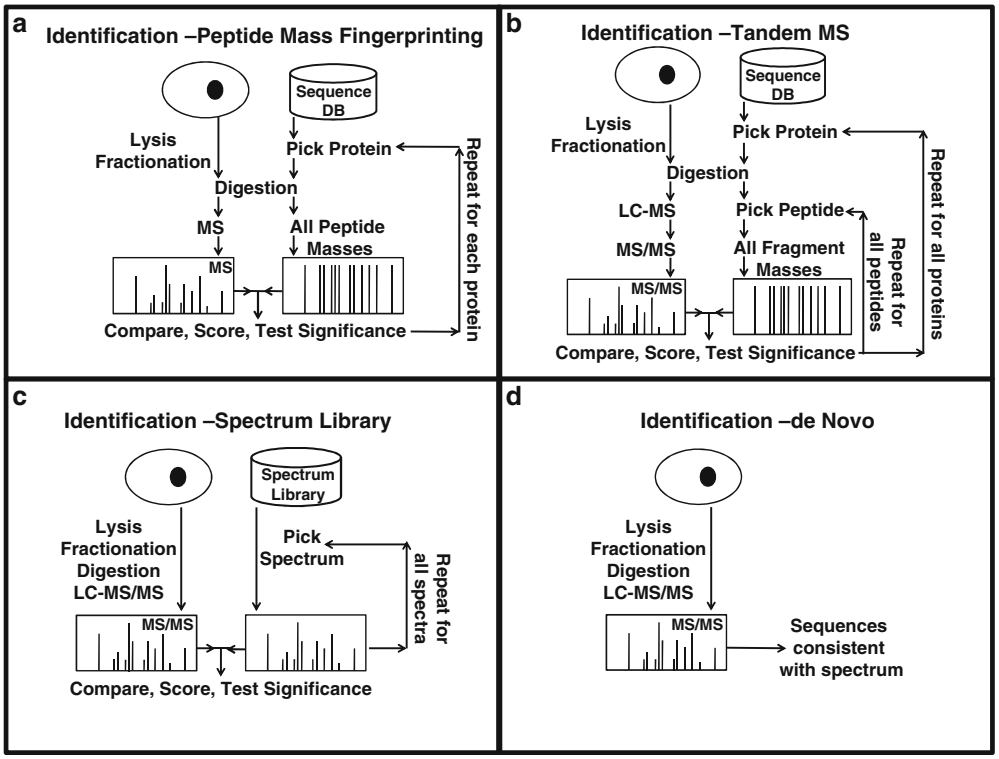

Fig. 1. *Mass spectrometry based workflows for protein identification*: (**a**) searching a protein sequence collection with peptide mass information; (**b**) searching a protein sequence collection with peptide fragment mass information; (**c**) searching a spectrum library with peptide fragment mass information; (**d**) de novo sequencing.

The first method for protein identification developed was peptide mass fingerprinting, PMF (2), i.e., matching measured proteolytic peptide masses to the theoretical proteolytic peptide masses of proteins in a sequence collection and calculating a score based on the matching peptides (see Note 3 and Fig. 1a). A basis of peptide mass fingerprinting is that the mass measurement of a single proteolytic peptide matches the masses of only a few different proteolytic peptide sequences (3). For example, a mass around 2,000 Da measured with an accuracy of 1 ppm matches on the average 4 and 1.5 unmodified tryptic peptides in the entire proteome of human and yeast, respectively (Fig. 2). A single peptide mass measurement is typically not matched uniquely with a single protein species and is therefore not sufficient to identify a protein (the probability for more than one protein identified = 1). But, a set of measured peptide masses from a single digested protein is useful for identification, since the probability is <<1 of randomly matching these mass values to a protein sequence in the collection searched. In theory, not only single proteins but also a large portion of the proteins in a complex protein mixture can be identified by the PMF approach (4).

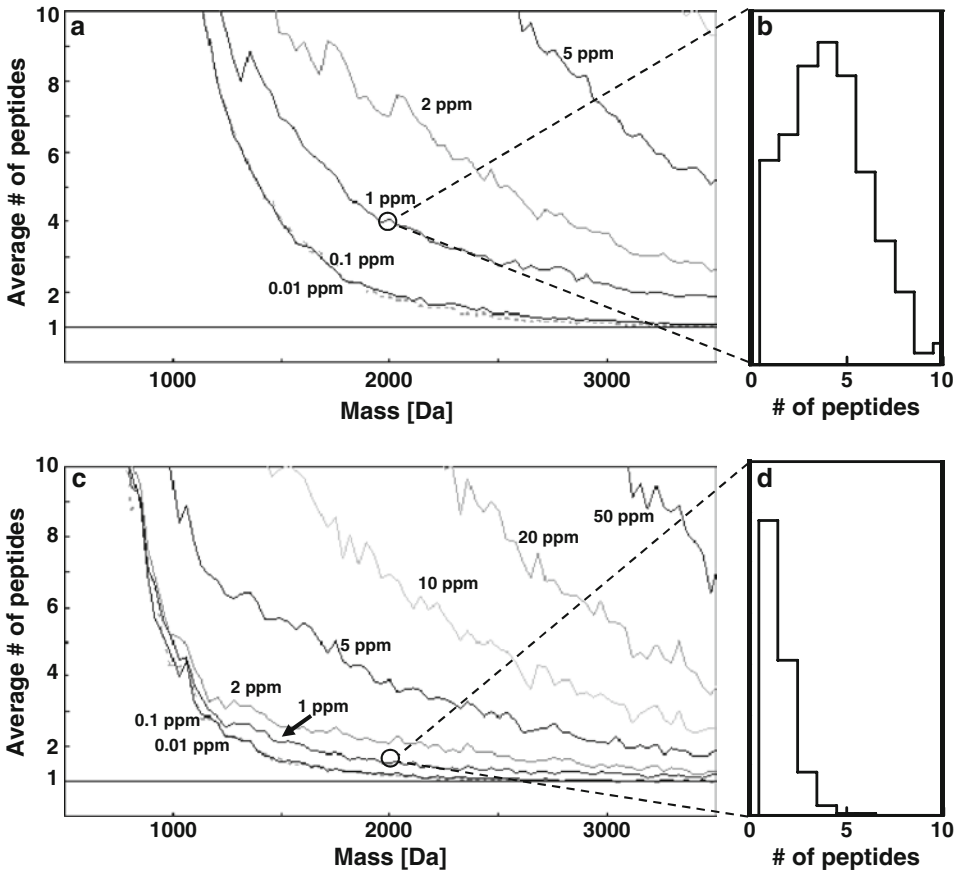

Fig. 2. *The information value of a mass measurement.* The number of unmodified tryptic peptides as a function of peptide mass for different mass accuracies for (**a**) human and (**c**) yeast. The distribution of number of matching unmodified tryptic peptides at mass 2,000 Da and mass accuracy of 1 ppm for (**b**) human and (**d**) yeast.

However, in practice, mass spectrometers fail to detect simultaneously peptides originating from different sample proteins that differ significantly in abundance (5). Hence, a prerequisite for PMF-based protein identification is that the samples analyzed are reasonably pure and only contain a few different proteins (6).

A more robust method 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 (Fig. 1b). This method requires only one or a few identified peptides to identify a gene. Peptides are fragmented by increasing their internal energy, usually through collisions. When their internal energy is increased, peptides fragment along their backbone, and ions characteristic of the amino acid sequence and the activation method are produced. The masses of these ions are compared with the theoretical fragment masses of the peptides in the sequence collection that match the mass of the

intact peptide, and a score is calculated based on the matching fragments (7, 8). This method is based on the method developed for identifying organic molecules from their fragment mass spectra (9–11). The advantage of using a sequence collection is that it is not necessary to observe fragmentation next to every amino acid in the peptide; a few fragment ions are usually sufficient because the sequence collection can be used to fill in the missing information (see Note 4). The drawback is, however, that if the sequence is not in the sequence collection, it cannot be found using this method, but as more and more complete genome sequences are becoming available, this becomes less of an issue. The probability of fragmentation between a pair of adjacent amino acids is dependent on their chemical properties and to a lesser degree on the amino acids further away from the fragmentation site; therefore, the intensity of fragment ions is highly sequence dependent. The information in the peak intensities cannot fully be utilized when searching protein sequence collections because most implementations use the same intensity for all theoretical fragments owing to the difficulty in accurately predicting their relative intensities from the amino-acid sequence.

One way of utilizing the sequence-specific fragment ion intensities and thereby improving the sensitivity is to instead search spectrum libraries (Fig. 1c), i.e., large collections of experimentally acquired fragment mass spectra that have been annotated. This is currently the predominant method for identification of small organic molecules (12) and has during the last few years been applied to peptide identification (13, 14). In this method, the intensity information is fully utilized (see Note 5) because the matching is between two experimentally acquired fragment mass spectra, and therefore, this is the most sensitive of the identification methods. The challenge is, however, to collect large high-quality sets of spectra that have sufficient coverage of the proteome.

In cases, when the genome has not been sequenced and there are no spectrum libraries available, the only possibility is to use de novo sequencing (Fig. 1d), i.e., use only the information in the fragment mass spectra and the mass of the intact peptide to obtain the peptide sequences (15–18). This requires much higher quality data because the entire space of all possible sequences is the search space (see Note 6). To search the entire space of potential sequences is impractical even for short peptides, but several algorithms have been developed that attempt at searching the relevant part of the search space in a reasonable time frame (15–18).

In all mass spectrometry-based identification methods, a score is calculated to quantify the match between the observed mass spectrum and the collection of possible sequences. These scores are highly dependent on the details of the algorithm used, and they are not always easy to interpret because the interpretation of

the score depends on properties of the data and the search results. Therefore, it is desirable to convert the score to a measure that is easy to interpret, such as the probability that the result is random and false. For this conversion, the distribution of random and false scores is needed (Fig. 3). Estimates of this distribution can be generated using either simulations (19, 20), collecting statistics during the search (21–23), or direct calculations (24).

Here, we describe how the different components of the global proteome machine (GPM) can be used for protein and peptide identification and validation.

2. Methods

2.1. Searching Protein Sequence Collections

X! Tandem (25–27) is a search engine for identifying proteins by searching sequence collections. X! Tandem scores the match between an observed tandem mass spectrum and a peptide sequence, by calculating a score that is based on the intensities of the fragment ions and the number of matching b- and y-ions (see Note 7). This score is converted to an expectation value using the distribution of the scores of randomly matching peptides (Fig. 3). Before the search, the user needs to specify a set of parameters including which sequence collection to search, the mass accuracy of peptides and their fragments, and modifications of the peptide sequence (see Note 8). The search is done iteratively; only proteins that have at least one peptide identified in an iteration are

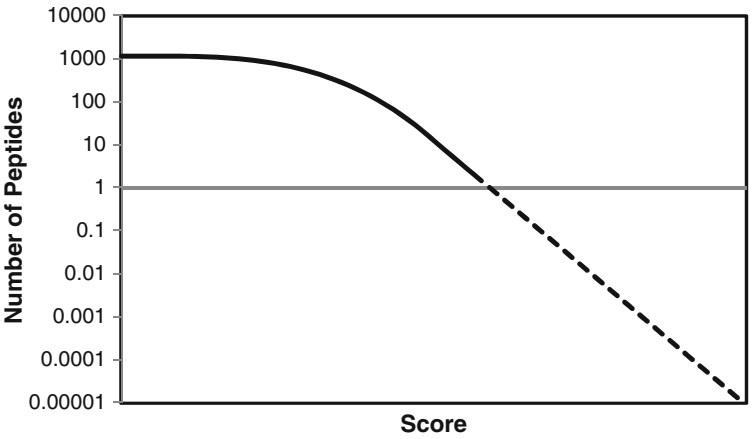

Fig. 3. *Expectation values.* The score can be transformed to an expectation value, i.e., the number of peptides that through random matching generate the score, if the distribution of random scores is known. This random distribution can be obtained for expectation values >1 by collecting statistics during the search because most peptides in a sequence collection match a given mass spectrum purely through random matching. Estimating expectation values <1 can be done by fitting the tail of the distribution to a Gumbel distribution and extrapolating.

searched in subsequent iterations (25). This iterative search can be used to speed up and increase the sensitivity of the identification of modifications, nonspecific enzymatic cleavage, and point mutations by restricting the search to unmodified tryptic peptides in the first iteration, and then widening the search in subsequent iterations. Another way to speed up the searches and make them more sensitive is to restrict the search to proteotypic peptides using X! P3 (27), which searches only peptides that have been previously identified and deposited in the GPM DataBase (GPMDB) (28).

2.2. Searching Spectrum Libraries

X! Hunter (13) is a search engine for searching annotated spectrum libraries. X! Hunter uses the same scoring as X! Tandem, except for that it compares the observed mass spectrum to libraries of spectra derived from experiments. Therefore, the peptide sequence-dependent intensity information can be fully utilized, and the sensitivity of the search is increased. It is, however, critical that the spectrum libraries are constructed carefully. The libraries for X! Hunter are constructed by taking the fragment mass spectra from GPMDB and grouping them so that one library spectrum is constructed for each peptide modification and charge state. The selection criteria are that (1) the spectrum matches to a peptide with an expectation value less than 0.001 and (2) at least 40% of the ion intensity in a spectrum is assignable as y- or b-ions or their corresponding neutral loss products. For the selected spectra, the m/z values of the matching peaks are substituted with the exact theoretical values. The ten spectra with lowest expectation value are selected for each peptide modification and charge state, and a composite spectrum is created and added to the library. These annotated spectrum libraries can also be extended to modification that do not affect the fragmentation pattern (e.g., some types of stable isotope labeling), by using the ion intensities of the fragmented unmodified peptide and reassigning the m/z values to correspond to the modified peptide.

2.3. Validation of Results

The search results for all GPM search engines are displayed in a unified interface that allows the user to get an overview of the results as well as inspect the details of the results when needed. In the basic display, proteins for which there is evidence for their presence in the sample are listed. The strength of the evidence is quantified with an expectation value (see Note 9) (23), and the proteins are listed in the order of increasing expectation value, i.e., in the order of decreasing strength of the evidence. Other information that can be used to assess the identified proteins are also shown, including the sum of the intensity of the matching fragment ions for all peptides, the number of matching peptides, and the fraction of the protein sequence covered by the observed peptides. Details of the evidence for a protein can be displayed,

listing all matching peptides sequences, modifications and charge state together with the peptide expectation values, error in the mass measurement, and the sum of the intensity of the fragment ions matching to the peptide sequences. For an individual peptide, the annotated fragment mass spectrum can be displayed showing the peak assignments. There are also alternative ways to display the list of identified protein, including their distribution among gene ontology categories, pathways, and protein interaction networks. In these displays, a *p*-value is calculated to asses which gene ontology categories, pathways, or interactions are enriched or depleted in the dataset.

Comparison of identification results to the large set of search results collected in GPMDB is an effective way to validate the results. One way to use GPMDB is to visually compare the peptides observed for a protein with observations in other experiments in GPMDB (Fig. 4). Commonly, the same peptides are observed for a given protein in most proteomics experiments, and therefore, an observation of a peptide that has not been observed in other experiments should be investigated manually.

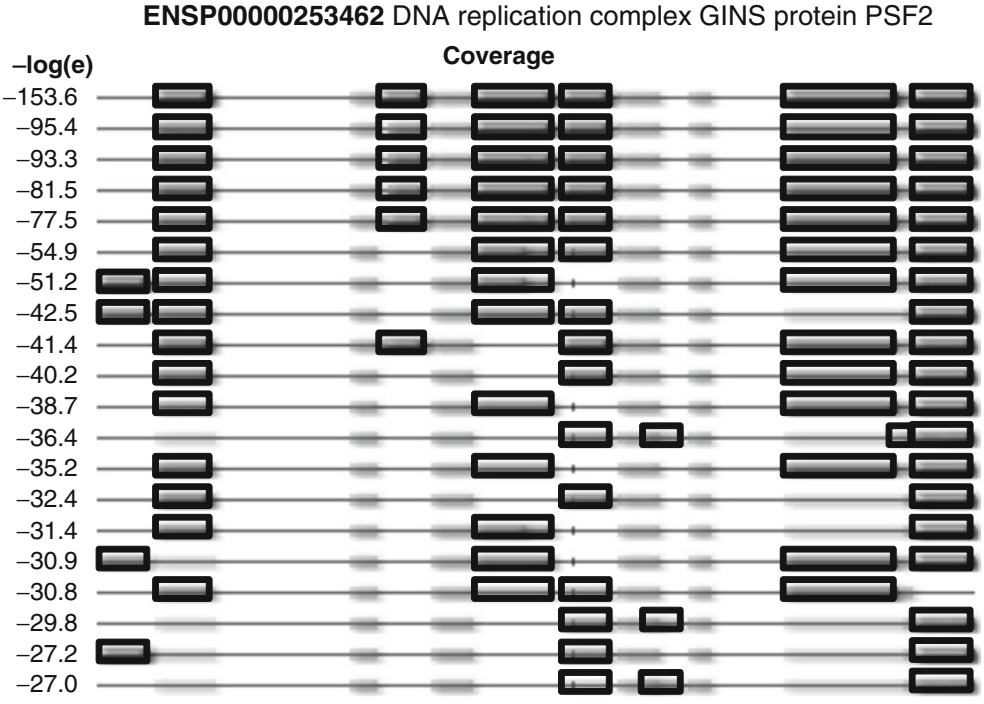

Fig. 4. *Using proteotypic peptides for validation of identification results.* The peptides identified for a protein can be compared with observations in other experiments in GPMDB. Commonly, the same peptides are observed for a given protein in proteomics experiments, and therefore, an observation of a peptide that has not been observed in other experiments should be investigated manually. The peptides observed for PSF2, a protein associated with the replication fork, are shown with black borders and regions of the protein that are difficult to observe in proteomics experiments are shown without borders. In a majority of the 20 experiments shown, the same 5 peptides are observed.

Another way of validating search results is to compare the sequence dependent ion intensity distribution of tandem mass spectra with spectra in GPMDB to evaluate if the fragmentation pattern is similar (Fig. 5). Several frequency measures from GPMDB for proteins and peptides are also reported together with the search results. For peptides, the number of times it has been observed in GPMDB and the fraction of the peptide identifications that are in a specific charge state (ω) are used. For proteins, Ω, a measure of peptide coverage with respect to charge state is used. Ω is a list of ratios denoting what fraction of the peptides in a particular charge state for a given protein was seen in a single protein identification. Proteins expectation values are also compared with other identifications of the protein in GPMDB, and the rank is reported, allowing the user to judge how their result compares with other results. All these measures are shown to make the validation of the results easier by allowing detailed comparison with the large set of experimental results that are available in GPMDB.

The information in GPMDB can also be used to design experiments. It is advisable to start planning an experiment by inspecting the information associated with proteins of interest to find out what has been observed in other proteomics experiments. For example, GPMDB supports the design of experiments targeted to investigate a group of proteins (multiple reaction monitoring (MRM)). Through the MRM module, the information in GPMDB is used to aid in the selection of peptides and their fragment ions that produce a strong signal and are specific to the protein.

The quality of the overall match between the whole dataset and the sequence collection can be evaluated using ρ-diagrams and ρ-scores (29). A ρ-diagram is a comparison between the distribution of peptide expectation values for a dataset and the predicted distribution for random matching (see Note 10). For a dataset that only has random matches to a sequence collection, the data points in the ρ-diagram will fall on the diagonal, $\rho = \log(e)$, i.e., the expectation values for the peptides are distributed as expected from random matching (Fig. 6a). In contrast, for datasets that are of high quality, typically many peptides match well with the sequence collection, and the data points in the ρ-diagram deviate from the diagonal and are closer to $\log(e) = 0$ (Fig. 6b). The ρ-score corresponding to a ρ-diagram is defined as the area between the data points and the diagonal [$\rho = \log(e)$] normalized to a value between 0 and 100, where ρ-score of 0 corresponds to purely random matching and ρ-score of 100 corresponds to no random matching. The ρ-score, being a measure of the quality of a match between an entire dataset and a sequence collection, can be used for optimizing search parameters, for evaluating algorithms, and for controlling the quality of datasets.

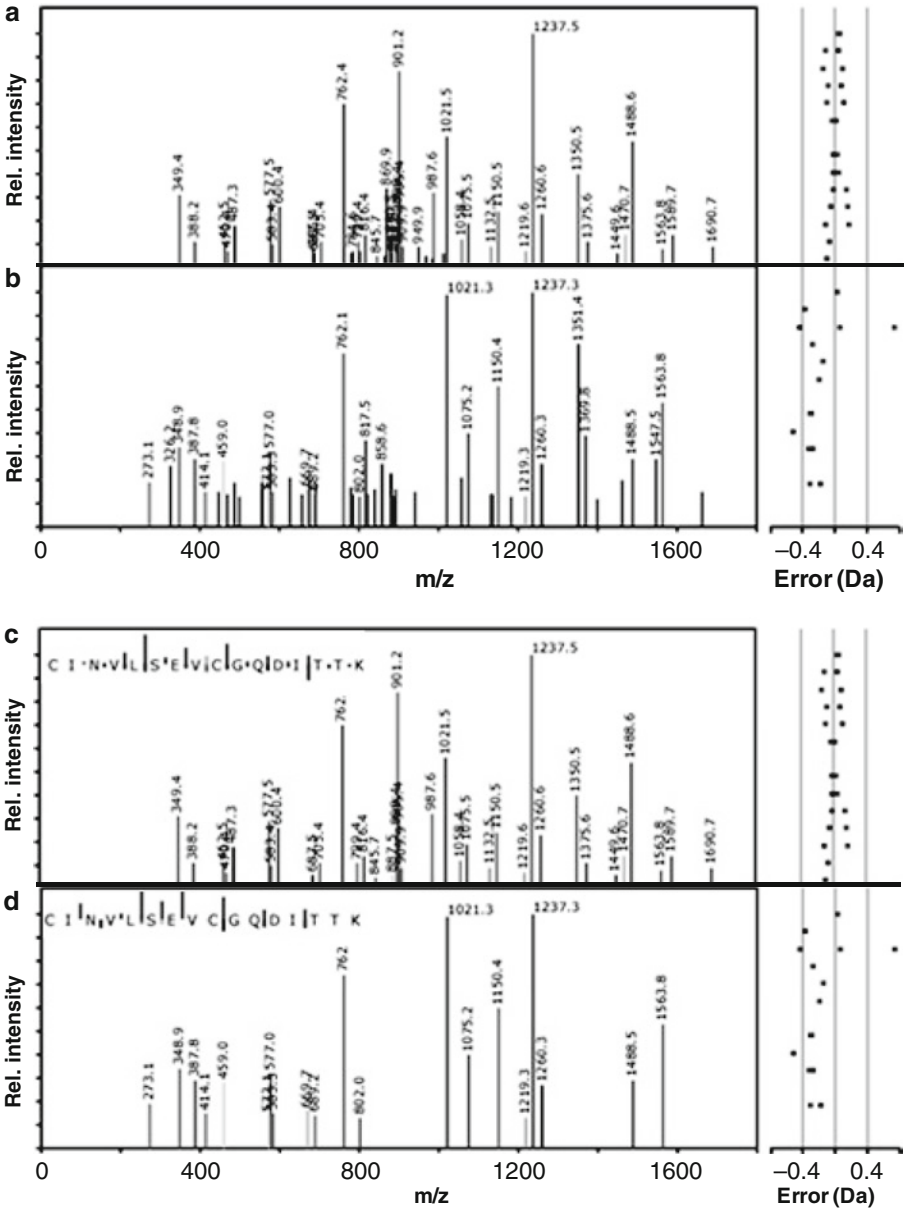

Fig. 5. *Using tandem mass spectra for validation of identification results.* The intensity distribution of tandem mass spectra is mainly dependent on the peptide sequence. Therefore, comparing a fragment mass spectrum with spectra in GPMDB can be used for validation of the results. (**a**, **c**) A stronger [log(*e*) = –12.8] and (**b**, **d**) a weaker [log(*e*) = –3.6] spectrum matching to the sequence CINVLSEVCGQDITTK are shown [(**a**, **b**) – all peaks (**c**, **d**) – peaks matching the sequence]. The stronger spectrum has many peaks matching the peptide sequence and little background, while the weaker spectrum has fewer matching peaks and more background peaks, but the intensity profile of the matching peaks is similar.

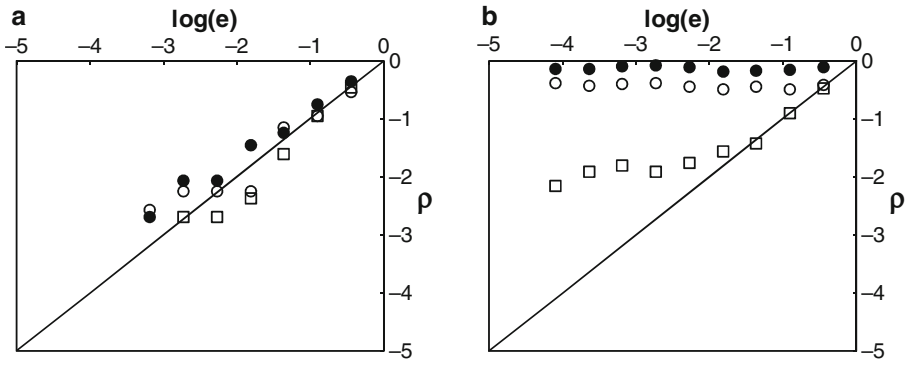

Fig. 6. *ρ-diagram.* A *ρ*-diagram shows the quality of the match between a dataset and a proteome. (**a**) The data points are close to the line $\rho = \log(e)$ when the results are dominated by random matching between the data and the proteome. The three datasets shown were obtained by searching against a collection of reversed sequences. (**b**) Three datasets of different quality are shown (*ρ*-scores are 95, 87, 57, respectively). The highest quality dataset (*filled circles*) is closest to the line $\log(e) = 0$ and the lowest quality dataset (*open squares*) is closest to the line $\rho = \log(e)$.

3. Notes

1. Peaks in mass spectra are detected by finding local maxima in $S(l) = \sum_{|k-l|<w_l/2} I(k)$ over the expected peak width w_l for each point, l, in the spectrum, where $I(k)$ is the measured intensity at a point $k, 0 \le k \le N, 0 \le l \le N$ and N is the total number of points in the mass spectrum. The signal to noise ratio (the ratio of the root mean square deviation of the peak and of the background) is usually used to decide if the peak should be used for identification. The mass of an analyte can be determined using the centroid;

$$C(l) = \frac{\sum_{|k-l|<w'/2} I(k) \cdot \frac{m}{z}(k)}{\sum_{|k-l|<w'/2} I(k)}$$

(where $\frac{m}{z}(k)$ is the mass to charge ratio at a point k) of the corresponding peak in the mass spectrum, where w' is the width of the centroid calculation.

2. Because peptides naturally contain heavy isotopes of atoms (e.g., 1.11% ^{13}C and 0.366% ^{15}N), they are observed as clusters of peaks. The relative intensities of these isotope clusters are dependent on the mass of the peptide because the number of atoms increases with mass, and therefore, the probability of the peptide containing one or more heavy isotopes increases. The largest effect comes from ^{13}C and a first order estimate of the peak intensities is given by, $T_m = \binom{n}{m} p^m (1-p)^{n-m}$

where T_m is the intensity of peak m in the distribution, m is the number of ^{13}C, n the total number of carbon atoms in the peptide, and p is the probability for ^{13}C (i.e., 1.11%).

3. The simplest method for peptide mass fingerprinting is to count the number of peptides in the mass spectrum that match to each protein in the sequence collection. This count can then be used as a score to rank the proteins. This simple scoring scheme works well when the data are of high-quality, but with low-quality data, typically, a large protein will get the highest score due to random matching. This is because the probability for random matching increases with the size of the protein simply because there are more peptides to match. More sophisticated scoring methods have been developed as an attempt to compensate for this effect (24, 30–32).
4. The sequence collections used for protein identification are based on the genes predicted from the genome sequence, and are therefore a very small subset of all possible sequences. For example, there are ~2.5×10^{14} unique tryptic peptides of length 15 in the human proteome compared with $20^{15} = 3.3 \times 10^{19}$ possible unmodified peptides of length 15. Because a vast majority of possible peptides are not used in an organism, the distance between real peptides in sequence space is typically large, and therefore, missing information can be filled in using the sequence collection.
5. Typically, the normalized inner product of the two spectra is used to score how well their intensities match. If the spectra are represented as vectors with the number of dimension equal to the number of matching peaks, n, and the length of the vector in each dimension equal to the intensity of the corresponding ion, the dot product is given by, $\frac{\mathbf{I}\cdot\mathbf{L}}{|\mathbf{I}||\mathbf{L}|} = \sum_{k=1}^{k=n} I_k L_k \Big/ \sqrt{\sum_{k=1}^{n} I_k^{\,2} \sum_{k=1}^{n} L_k^{\,2}}$, where $\mathbf{I} = (I_1, I_2, \ldots, I_n)$ is the observed spectrum, and $\mathbf{L} = (L_1, L_2, \ldots, L_n)$ is the library spectrum. The range of the normalized dot product is from –1 to 1. If the observed and library spectra are identical, the resulting dot product is 1, and any differences between them will result in lower values of the dot product.
6. The search space for de novo sequencing of unmodified peptides is 20^N where N is the length of the peptide. If there are m types of potential modifications, then search space increases to $(20+m)^N$.
7. The score, called hyperscore, is based on the assumption of a hypergeometric distribution and is given by $S_H = S_I \cdot n_b! \cdot n_y!$, where n_y is the number of matching y-ions, n_b the number

of matching b-ions, and S_I is the dot product between the observed spectrum and the spectrum predicted from the peptide sequence. The intensities for the spectrum predicted from the peptide sequence are usually set to 1 for each expected fragment mass and 0 for all other masses. However, X! Tandem also supports using intensities that are dependent on the two amino acids on each side of the fragmented bond.

8. A complete description of the input parameters for X! Tandem, X! P3, and X! Hunter can be found at http://thegpm.org/TANDEM/api/.

9. Protein expectation values can be estimated from the expectation values of its matching peptides. If more than one peptide has been found for a protein, the expectation values for the peptides are combined with a simple Bayesian model for the probability of having two peptides from the same protein having the best score in different spectra: $e_{pro} = \left(\frac{\beta^n(1-\beta^{s-n})}{sN^{n-1}}\right) \times (\Pi_{j=1}^{n} e_j) \times \left(\Pi_{i=0}^{n-1} \frac{s-i}{n-i}\right)$ where n is the number of unique peptide sequences matching the protein, e_j is the expectation value of the jth peptide, N is the total number of peptides scored to find the n unique peptides, s is the number of mass spectra in dataset, and β is $N/$(the total number of peptides in the proteome considered). If only one peptide is matching the protein, then the protein expectation value is set to the peptide expectation value, $e_{pro} = e_1$.

10. ρ is defined as $\rho(i) = \log\left(\frac{E_i}{E_0}\right)$ where i is an integer, $i = \log(e)$, e is the expectation value, and E_i is the number of peptides with expectation values between $\exp(i)$ and $\exp(i-1)$. For purely random matching, $E_i = \int_{\exp(i-1)}^{\exp(i)} N\,de = N\left[\exp(i) - \exp(i-1)\right] = N \exp(i)[1-\exp(-1)]$ where N is the total number of peptides that have been assigned to spectra, and therefore $\rho(i) = \log\left(\frac{E_i}{E_0}\right) = i = \log(e)$ for random matching.

Acknowledgments

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

References

1. K. Flikka, L. Martens, J. Vandekerckhove, K. Gevaert, and I. Eidhammer (2006) Improving the reliability and throughput of mass spectrometry-based proteomics by spectrum quality filtering, *Proteomics*, **6**, 2086–94.
2. W.J. Henzel, T.M. Billeci, J.T. Stults, S.C. Wong, C. Grimley, and C. Watanabe (1993) Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases, *Proc Natl Acad Sci USA*, **90**, 5011–5.
3. D. Fenyo, J. Qin, and B.T. Chait (1998) Protein identification using mass spectrometric information, *Electrophoresis*, **19**, 998–1005.
4. J. Eriksson and D. Fenyo (2005) Protein identification in complex mixtures, *J Proteome Res*, **4**, 387–93.
5. J. Eriksson and D. Fenyo (2007) Improving the success rate of proteome analysis by modeling protein-abundance distributions and experimental designs, *Nat Biotechnol*, **25**, 651–5.
6. O.N. Jensen, A.V. Podtelejnikov, and M. Mann (1997) Identification of the components of simple protein mixtures by high-accuracy peptide mass mapping and database searching, *Anal Chem*, **69**, 4741–50.
7. J.K. Eng, A.L. McCormack, and J.R. Yates (1994) An approach to correlate mass spectral data with amino acid sequences in a protein database, *J Am Soc Mass Spectrom*, **5**, 976.
8. M. Mann and M. Wilm (1994) Error-tolerant identification of peptides in sequence databases by peptide sequence tags, *Anal Chem*, **66**, 4390–9.
9. A.M. Duffield, A.V. Robertson, C. Djerassi, B.G. Buchanan, G.L. Sutherland, E.A. Feigenbaum, and J. Lederberg (1969) Applications of artificial intelligence for chemical inference. II. Interpretation of low-resolution mass spectra of ketones, *J Am Chem Soc*, **91**, 2977–81.
10. J. Lederberg, G.L. Sutherland, B.G. Buchanan, E.A. Feigenbaum, A.V. Robertson, A.M. Duffield, and C. Djerassi (1969) Applications of artificial intelligence for chemical inference. I. The number of possible organic compounds. Acyclic structures containing C, H, O, and N, *J Am Chem Soc*, **91**, 2973–6.
11. G. Schroll (1969) Applications of artificial intelligence for chemical inference. III. Aliphatic ethers diagnosed by their low-resolution mass spectra and nuclear magnetic resonance data, *J Am Chem Soc*, **91**, 2977–81.
12. S. Heller (1999) The history of the NIST/EPA/NIH mass spectral database, *Today's Chemist at Work*, **8**, 45–50.
13. R. Craig, J.C. Cortens, D. Fenyo, and R.C. Beavis (2006) Using annotated peptide mass spectrum libraries for protein identification, *J Proteome Res*, **5**, 1843–9.
14. H. Lam, E.W. Deutsch, J.S. Eddes, J.K. Eng, N. King, S.E. Stein, and R. Aebersold (2007) Development and validation of a spectral library searching method for peptide identification from MS/MS, *Proteomics*, **7**, 655–67.
15. J.A. Taylor and R.S. Johnson (1997) Sequence database searches via de novo peptide sequencing by tandem mass spectrometry, *Rapid Commun Mass Spectrom*, **11**, 1067–75.
16. V. Dancik, T.A. Addona, K.R. Clauser, J.E. Vath, and P.A. Pevzner (1999) De novo peptide sequencing via tandem mass spectrometry, *J Comput Biol*, **6**, 327–42.
17. B. Ma, K. Zhang, C. Hendrie, C. Liang, M. Li, A. Doherty-Kirby, and G. Lajoie (2003) PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry, *Rapid Commun Mass Spectrom*, **17**, 2337–42.
18. B. Spengler (2004) De novo sequencing, peptide composition analysis, and composition-based sequencing: a new strategy employing accurate mass determination by fourier transform ion cyclotron resonance mass spectrometry, *J Am Soc Mass Spectrom*, **15**, 703–14.
19. J. Eriksson, B.T. Chait, and D. Fenyo (2000) A statistical basis for testing the significance of mass spectrometric protein identification results, *Anal Chem*, **72**, 999–1005.
20. J.E. Elias and S.P. Gygi (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry, *Nat Methods*, **4**, 207–14.
21. H.I. Field, D. Fenyo, and R.C. Beavis (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.
22. A. Keller, A.I. Nesvizhskii, E. Kolker, and R. Aebersold (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search, *Anal Chem*, **74**, 5383–92.
23. D. Fenyo and R.C. Beavis (2003) A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes, *Anal Chem*, **75**, 768–74.

24. J. Eriksson and D. Fenyo (2004) Probity, a protein identification algorithm with accurate assignment of the statistical significance of the results, *J Proteome Res*, **3**, 32–6.
25. R. Craig and R.C. Beavis (2003) A method for reducing the time required to match protein sequences with tandem mass spectra, *Rapid Commun Mass Spectrom*, **17**, 2310–6.
26. R. Craig and R.C. Beavis (2004) TANDEM: matching proteins with tandem mass spectra, *Bioinformatics*, **20**, 1466–7.
27. R. Craig, J.P. Cortens, and R.C. Beavis (2005) The use of proteotypic peptide libraries for protein identification, *Rapid Commun Mass Spectrom*, **19**, 1844–50.
28. R. Craig, J.P. Cortens, and R.C. Beavis (2004) Open source system for analyzing, validating, and storing protein identification data, *J Proteome Res*, **3**, 1234–42.
29. D. Fenyo, B.S. Phinney, and R.C. Beavis (2007) Determining the overall merit of protein identification data sets: rho-diagrams and rho-scores, *J Proteome Res*, **6**, 1997–2004.
30. D.N. Perkins, D.J. Pappin, D.M. Creasy, and J.S. Cottrell (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data, *Electrophoresis*, **20**, 3551–67.
31. W. Zhang and B.T. Chait (2000) ProFound: an expert system for protein identification using mass spectrometric peptide mapping information, *Anal Chem*, **72**, 2482–9.
32. J. Magnin, A. Masselot, C. Menzel, and J. Colinge (2004) OLAV-PMF: a novel scoring scheme for high-throughput peptide mass fingerprinting, *J Proteome Res*, **3**, 55–60.