

Extracting Monoisotopic Single-Charge Peaks From Liquid Chromatography-Electrospray Ionization–Mass Spectrometry

Rune Matthiesen

Summary

Peak extraction from raw data is the first step in analysis of mass spectrometry (MS) data. The quality of this procedure is very important because it affects the quality of all subsequent analysis, such as database searches and peak quantitation. Many methods have been proposed in the literature, yet the number of practical solutions in terms of available software is rather limited. Virtual Expert Mass Spectrometrist (VEMS) v3.0 includes an algorithm for extracting mono-isotopic single-charged peaks and their corresponding retention time from liquid chromatography (LC)–MS data. The extracted peaks can subsequently be exported to other programs or used internally by VEMS to perform peptide mass fingerprinting searches or peptide quantitation. Additionally, VEMS interfaces the commercial program ProteinLynx Global server v2.0.5 for automatic peak extraction from MS/MS spectra obtained by LC–MS/MS.

Key Words: Noise filtering; peak extraction; deisotoping; decharging.

1. Introduction

Liquid chromatography-electrospray ionization–mass spectrometry (LC-ESI–MS) of tryptic peptides produces a wealth of information in the form of peptide masses and peptide retention time(s) on the LC column. In proteomics, the LC system is typically a single hydrophobic reverse-phase column (one-dimensional separation) or an anionic/cationic column followed by a hydrophobic reverse-phase column (multidimensional separation) (1). The electrospray ion source is responsible for production of charged peptides in the gas phase resulting in tryptic peptide charge states typically from +1 to +4, where the same peptide can appear with different charge states (2). The mass

spectrometer used for LC–MS in proteomics is most often a tandem mass spectrometer that produces MS or both MS and MS/MS data (*see* Chapter 1). The raw data obtained from these experiments contains, in general, transformed and distorted versions of the ideal physical quantity of interest, which is the masses of the intact peptide, the peptide fragments, and the retention time. The conversion of raw data to a peak list consists of the following three steps in Virtual Expert Mass Spectrometrists (VEMS): (1) the instrument-introduced noise in the spectra should be removed, (2) the monoisotopic single-charged mass should be extracted by decharging and deisotoping, and (3) the retention time(s) for the peptides should be extracted. How this is done in theory and practice with VEMS v3.0 is described in this chapter.

1.1. Noise Filtering

Two types of errors are present in experimental data: systematic and random error. Systematic error is often removed by calibration and will not be discussed further in this section. Random error is also called noise. Filtering out noise from the data ideally gives the true signal. The true difference between noise and signal is that noise is not reproducible, whereas signal is. The quality of signals is often expressed as the true signal divided by the standard deviation of the noise. There are many methods for noise removal, such as linear filters (3,4), penalized least square (5), Fourier transform filters (6), and wavelets (7). The presentation here concentrates on the linear filters, which are computationally fast and have satisfactory performance for proteomics data. Linear filters convert a time series to a new by a linear operation. Linear filters can in general be expressed as (4)

$$y_t = \sum_{r=-q}^{+s} a_r x_{t+r} \quad (1)$$

where y_t is the smoothed signal. x_t is the current data point, and r iterates over neighboring data points. The smooth width m is equal to $q+s+1$. a_r are weights and are dependent on the filter type. For example, for a simple, unweighed sliding, average smooth $a_r = 1/m$ for all r . A frequent filter used in MS is the Savitsky Golay filter (3), which has weights that result in a smoothed signal that corresponds to fitting a low-order polynomial to all smooth intervals (*see* Fig. 1). Savitsky Golay filters have been criticized for having end effect problems because it is a symmetrical filter (*see* Note 1). However, this is rarely a problem for MS data and can easily be circumvented by combining symmetrical filters together with asymmetrical filters. The quality of the smoothness can be evaluated by the lack of fit and by either the roughness of the data or by maximum entropy (*see* Note 2).

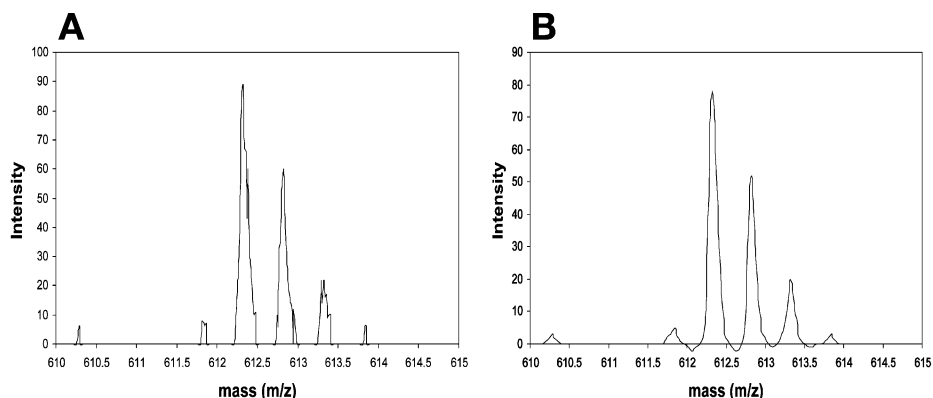


Fig. 1. Savitsky Golay noise filtering. (A) Raw mass spectrometry (MS) data. (B) MS data from (A) after three iterations with a nine-point Savitsky Golay filter.

Alternatively a geometric mean filter with window size $2m+1$ can be used in combination with a Savitsky Golay filter.

$$y_i = \left(\prod_i x(t + i\Delta t) \right)^{1/(2m+1)} \quad i = 0, \pm 1, \pm 2, \dots, \pm m$$

The geometric mean filter can remove spikes because neighboring data points need to be non-zero for a signal to be maintained, and has the additional advantage that the data remains on the same scale (8).

1.2. Deisotoping and Decharging LC-MS Data

The smoothed raw data is not practical to input into a search engine. Instead, a peak list containing what corresponds to the monoisotopic single-charged ion is often used as input for search engines. The first step is to extract all peaks (see **Note 3**) from the smoothed raw data. Peak extraction can be done by extracting peak tops, the centroid method, or by taking the first derivative of the signal (see **Fig. 2**). After the peak list is obtained, decharging and deisotoping is done simultaneously by the VEMS program. The algorithm described here for deisotoping and decharging has some similarities to earlier published methods (9,10). However, the method here is improved by considering information in all MS scan numbers, rather than only considering one scan number at a time. In addition, it considers all combinations of theoretical isotopic distributions of one to two compounds with charge state +1 to +4 to find the best fit to the observed isotopic distribution.

VEMS starts at the first MS scan number from the low mass end, and the program considers high-charge states first.

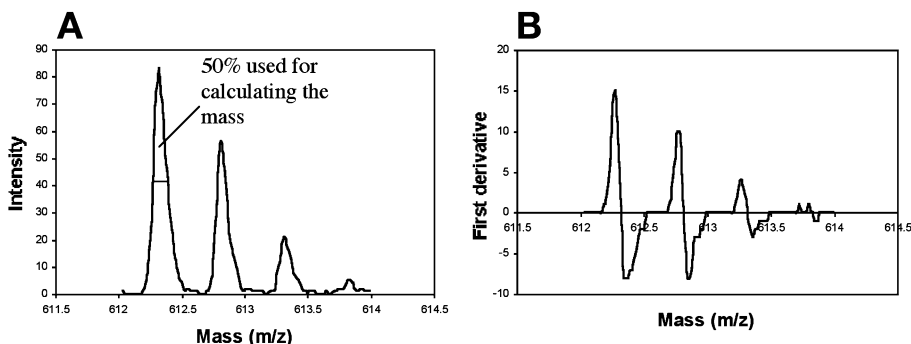


Fig. 2. Converting profile data to peak lists. **(A)** Fifty percent Centroid method. Fifty percent of the resolved part of the peak is used for determining the mass. The mass is calculated by an intensity-weighted average of the masses in the peak. This is equivalent to finding the vertical line passing through the center of gravity of the peak. **(B)** First derivative method. The first derivative of the signal in **(A)** is calculated and the peak masses are determined at the mass points where the first derivative is cutting the x-axis.

1. When a peak is encountered by scanning from the low mass end and with low scan numbers, VEMS scans the neighboring MS scans to find the intensity peak maximum in the elution profile.
2. It is likely that the interference from other compounds with similar m/z values and retention times is smallest at the scan number obtained in **step 1**. However, there can still be some overlapping peaks. VEMS, therefore, calculates approximate isotopic distributions (*see Note 4*) for all possible combinations of two compounds with charge states ranging from +1 to +4, and evaluates which combination fits the observed distribution best by calculating the lack of fit.
3. The best combination obtained in **step 2** is inserted in a new peak list as monoisotopic single-charged mass, intensity, and retention time. After insertion into a new peak list, the theoretical isotopic distribution at the determined charged state is used to remove peaks in the peak list obtained from the raw data corresponding to the observed isotopic distribution over the whole elution profile of the compound. If the best combination was found to contain two compounds, then only the compound corresponding to the peak found in **step 1** is inserted in the peak list and used for removing peaks in the elution profile.
4. **Steps 1–3** are continued until there are no more peaks in the peak list obtained from raw data.

2. Materials

2.1. Required Software

1. VEMS v3.0 (<http://yass.sdu.dk>). To follow this guide it is also necessary to download the raw data (<http://yass.sdu.dk/raw/my00234kr.raw.rar>).

2. Microsoft Windows. Currently VEMS is only fully tested on Windows XP and Windows 2000.

2.2. Optional Software

1. PLGS v2.05 and Masslynx v4.0 are commercial programs that can be obtained from Waters (Milford, MA). VEMS interfaces to some of the raw data processing tools of PLGS v2.05 and MassLynx v4.0. It is important that PLGS v2.05 and Masslynx v4.0 are installed in the default directory, otherwise the interfacing from VEMS will not work. If the commercial software is not available, then one can use ExrawNoPKX to convert mzData MS data to the VEMS MS data format. PLGS v2.05 and Masslynx v4.0 are only necessary for the methods described in **Subheading 3.2.**

3. Methods

This section describes how to extract monoisotopic single-charged peaks from raw LC–MS/MS files. In **Subheading 3.1.**, the extraction of the LC–MS data is presented that is accomplished by the VEMS algorithm described in **Subheading 1.** **Subheading 3.2.** shows how to extract monoisotopic single-charged peaks from all the MS/MS spectra in a number of LC–MS/MS runs.

3.1. Extract Monoisotopic Single-Charged Peaks From MS Scans

Download VEMS from (<http://yass.sdu.dk>) and uncompress the folder. In the VEMS directory folder there is a folder named “Exraw.” The files in this folder should be moved directly to “c:\data” folder. This folder contains the program “Exraw.exe” which is for format conversion. The program can extract the LC–MS part of LC–MS or LC–MS/MS runs to an indexed format that is used by the VEMS program. The program can, on the time of writing, convert mzXML files and Micromass raw files to the VEMS LC–MS format.

1. Start the “Exraw.exe” program (*see Fig. 3*).
2. Use the directory listbox in area 1 (*see Fig. 3*) to choose the folder where the raw data folder my00234kr.raw (can be obtain from <http://yass.sdu.dk/raw>) is located. Press the button “>>” to select the folder. It should now appear in the listbox in area 2.
3. The listbox in area 3 can now be used to specify the output directory.
4. Now press the button “Raw → VEMS” in area 4. The program will now convert all the specified raw data files to the VEMS LC–MS format.
5. In the output folder there should now be a directory named “my00234” containing the LC–MS data in the VEMS format.

The above steps converted a Micromass raw data file to the VEMS LC–MS format. The following steps describe how to convert mzXML files to the VEMS LC–MS format.

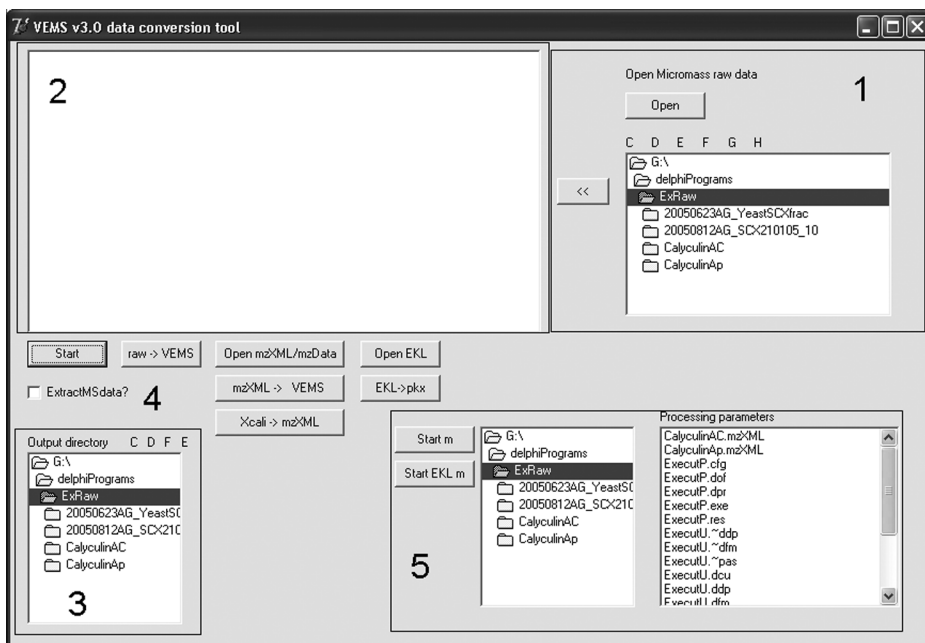


Fig. 3. Screen shot of the VEMS v3.0 data conversion tool. Area 1 is used to specify Micromass raw data files. Area 2 displays the chosen raw data files. Area 3 is used to specify the output directory. Area 4 activates different conversion functions. Area 5 is used to choose files containing different data processing parameters. This is used for optimization of processing parameters.

1. Click the button “Open mzXML” in area 4 to open an mzXML file containing LC–MS or LC–MS/MS data.
2. Choose an output directory in area 3.
3. Click on the button “mzXML → VEMS” in area 4 to create the VEMS LC–MS format in the output directory.

The operations performed so far did not do any data processing, they only extracted the LC–MS data to a more efficient format both in terms of size and data access. The VEMS LC–MS format just created can be used in the VEMS program to extract monoisotopic single-charged peaks from MS scans. VEMS can also use this format for peptide quantitation (*see* Chapter 8). The following describes how to use VEMS to extract peaks from the format. The nomenclature used to describe the user interface is presented in **Appendix E**.

1. Start VEMS_3.exe. Open the data import window from the file menu (File → Open data → Open multiple spectra or press sequentially “Alt”+“F”+“O”+“P”).
2. Select the VEMS LC–MS raw data files and close the data import window.

3. Now click on “File → Save → Extract MS peaklist.” This will automatically extract peaks from all the specified LC–MS data files and save them in the same folder.

The created peak list(s) can now be specified in the data import window and can be used for peptide mass fingerprinting searches. This is useful when working with a simple protein mixture and higher sequence coverage than achieved by the MS/MS spectra gives is important. Please note that the function activated by **step 1–3** is currently being improved.

3.2. Extract Monoisotopic Single-Charged Peaks From MS/MS Spectra

The VEMS program currently accepts MS/MS peak lists in mgf, pkl, dta, bsc, and pxx. All these formats are ASCII formats containing mass and intensity of parent ion and fragment ions. The pxx format is a VEMS format. The critical reader would probably ask why a new format was made when there are so many already. The reason is that the other formats do not contain all the necessary information for a proper data analysis. For example the pxx format contains the retention time and the original charge state of the peptide fragments before decharging. This section will describe how to make the pxx format from the raw data file “my00234kr.raw.”

1. Start the “Exraw.exe” program (*see Fig. 3*).
2. Use the directory listbox in area 1 (*see Fig. 3*) to choose the folder where the raw data folder my00234kr.raw (can be obtain from <http://yass.sdu.dk>) is located. Press the button “>>” to select the folder. It should now appear in the listbox in area 2.
3. The listbox in area 3 can now be used to specify the output directory.
4. Now press the button “Start” to create pxx files in the specified output directory.

Alternatively one can check the checkbox “Extract MS data?” then both the VEMS LC–MS format and the pxx formatted files are created in the output window when the “Start” button is pressed.

4. Notes

1. For symmetrical filters $q = s$ in (Eq. 1). Symmetrical filters have the drawback that they cannot be evaluated in the start and end of the spectrum that is the q first data points and the s last data points in the spectrum. However, asymmetrical filters where q or s equals zero can be evaluated (4).
2. The quality-of-function fitting is often evaluated by the lack of fit, which is given by $E_{lof} = \sum_{it} (x_i - y_i)^2$. It is not only the lack of fit that is important for the quality of a fit. For example, the roughness of spectrum, which is given by $R = \sum_t (y_t - y_{t-1})^2$, is also important and a best fit can be found by minimizing a weighted sum of E_{lof} and R . Alternatively, the E_{lof} could be evaluated together with maximum entropy of

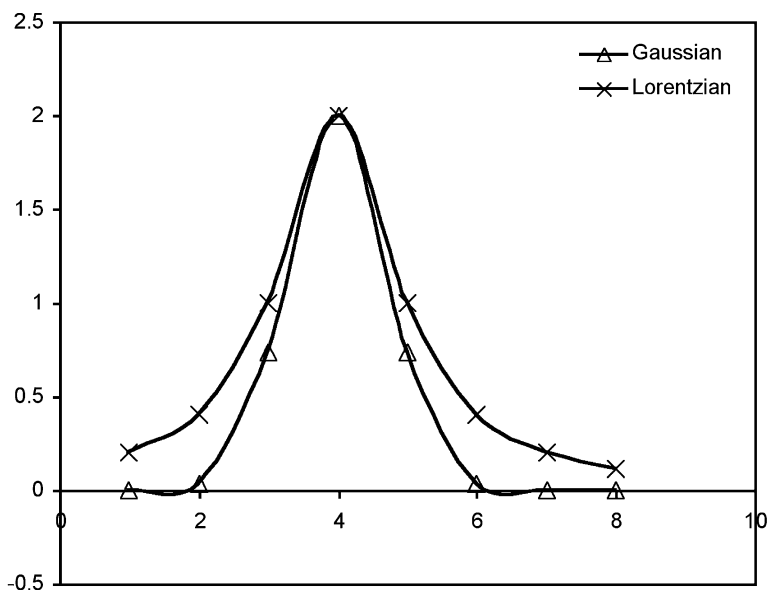


Fig. 4. The peak shape defined by a Gaussian or Lorentzian equation.

residuals, which is given by $S = -\sum_i p_i \log(p_i)$, where p is residuals at different time-points. Maximum entropy is very useful for choosing between different models that give the same E_{lof} .

- Peaks in spectroscopy can have several different shapes that need different mathematical functions for fitting. Peaks can often be approximated by a Gaussian (see Fig. 4), Lorentzian (see Fig. 4), or a mixture of the two functions (see Figs. 5–7). The equation for the Gaussian function is based on the normal distributions and can be formulated as $f(m_i) = A \cdot \exp(-(m_i - m_0)^2/s^2)$. Where m_0 is at the center, A is the maximum height at x_0 , and s defines the peak width. The width at half-height of a Gaussian peak is given by $s(4 \cdot \ln 2)^{1/2}$ and the area is $As(\pi)^{1/2}$. The equation for a Lorentzian function is given by $f(x_i) = A/(1+(m_i - m_0)^2/s^2)$, where m_0 is at the midpoint of the peak, and A is the height at the midpoint. The width at half-height of a Lorentzian peak is given by $2s$ and the area is $As\pi$ (II). In MS the mass of such peaks are often determined by calculating the centroid mass, which is more accurate than just taking the mass at the peak maximum. The centroid mass m_c and the corresponding intensity I_c can be calculated by the following expressions:

$$m_c = \frac{\sum_{y_i > y_{i,\max}^x} m_i I_i}{I_c}$$

$$I_c = \sum_{y_i > y_{i,\max}^x} I_i$$

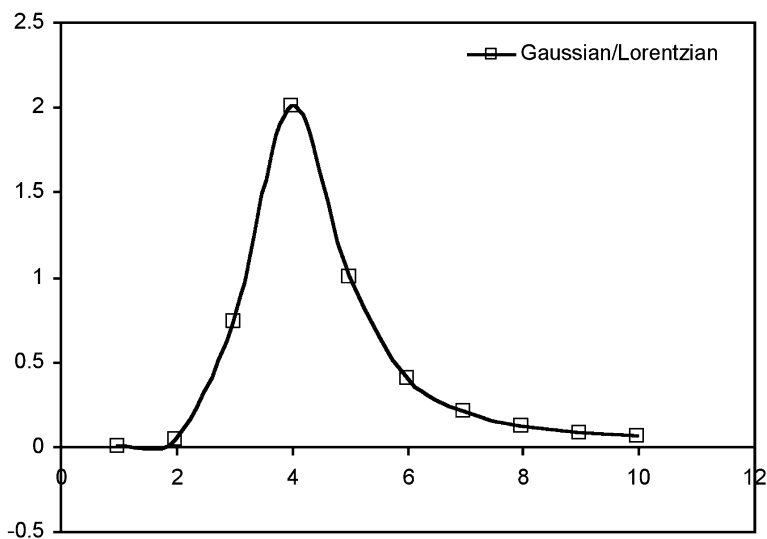


Fig. 5. A mixed model where the peak shape is defined by a Gaussian equation up to the midpoint, and by a Lorentzian function after the midpoint.

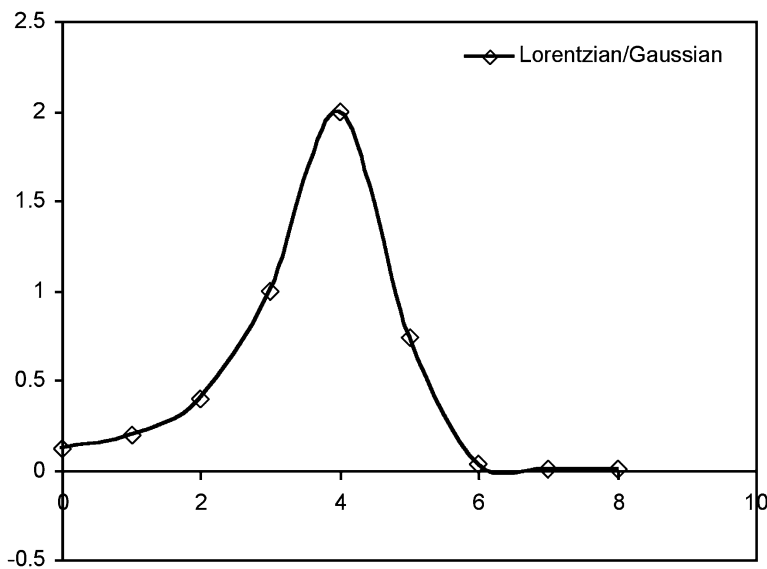


Fig. 6. A mixed model where the peak shape is defined by a Lorentzian function up to the midpoint, and by a Gaussian equation after the midpoint.

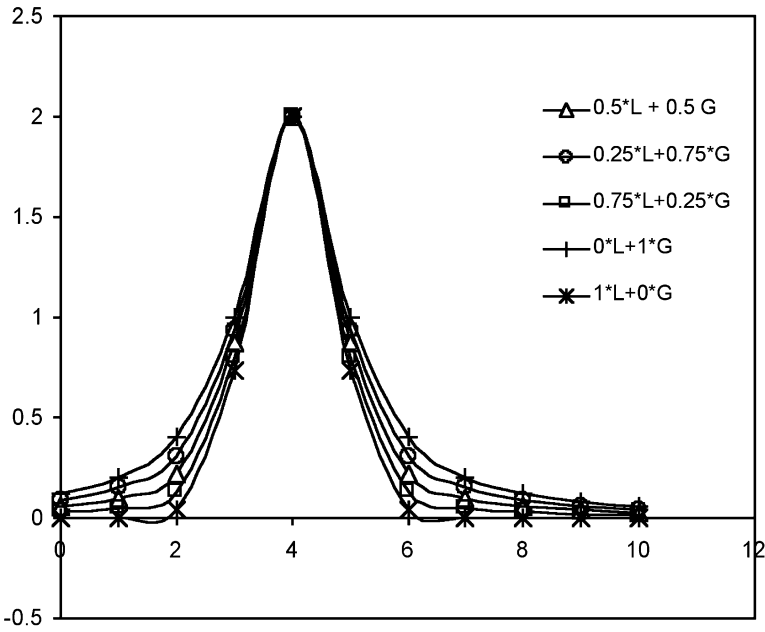


Fig. 7. Examples of Lorentzian and Gaussian mixed models. L is the Lorentzian function and G is the Gaussian.

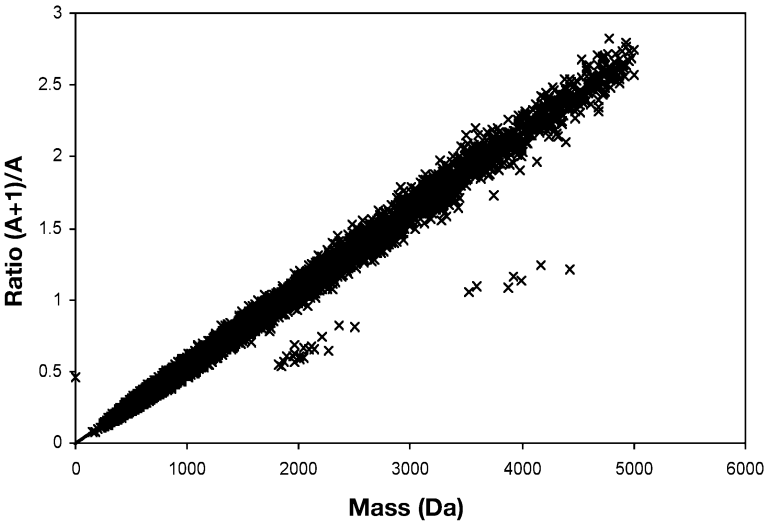


Fig. 8. The ratio between the theoretical abundance of the monoisotopic plus one and the monoisotopic peak plotted as a function of the monoisotopic mass.

where m_i is the mass at a certain mass bin and I_i is the corresponding intensity. x is a specified percentage of the maximum intensity.

4. Approximate isotopic distributions are calculated based on theoretical isotopic distribution of 20,000 standard tryptic peptides. Given the intensity of the monoisotopic peak, the intensity of the following isotopic peaks can be approximated by a linear equation (12). The ratio R between the intensity of the monoisotopic peak and the monoisotopic plus is approximated by $R = 0.0005412 * m - 0.01033$, where m is the mass of the monoisotopic peak (see Fig. 8). Similar approximation can be made for the higher masses in the isotopic distribution. The approximate isotopic distributions are used to generate all possible combinations of two overlapping isotopic distributions. The combination used is the one that gives the best fit on the neighboring peaks. The deisotoping problem can also be solved by linear algebra (13) instead of checking all reasonable possibilities.

Acknowledgments

R. M was supported by grants from EU TEMBLOR and by Carlsberg Foundation Fellowships.

References

1. Washburn, M. P., Wolters, D., and Yates, J. R., 3rd. (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **19**, 242–247.
2. Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F., and Whitehouse, C. M. (1989) Electrospray ionization for mass spectrometry of large biomolecules. *Science* **246**, 64–71.
3. Savitzky, A. and Golay, J. E. M. (1964) Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.* **36**, 1627–1639.
4. Chatfield, C. (ed.) (1989) *The Analysis of Time Series: An introduction*. Chapman and Hall, New York, pp. 1–8.
5. Eilers, P. H. (2003) A perfect smoother. *Anal. Chem.* **75**, 3631–3636.
6. Kast, J., Gentzel, M., Wilm, M., and Richardson, K. (2003) Noise filtering techniques for electrospray quadrupole time of flight mass spectra. *J. Am. Soc. Mass Spectrom.* **14**, 766–776.
7. Morris, J. S., Coombes, K. R., Koomen, J., Baggerly, K. A., and Kobayashi, R. (2005) Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. *Bioinformatics* **21**, 1764–1775.
8. Bylund, D. (2001) *Chemometrics Tools for Enhanced Performance in Liquid Chromatography-Mass Spectrometry*. Uppsala University, Uppsala, Sweden.
9. Wehofskey, M. and Hoffmann, R. (2002) Automated deconvolution and deisotoping of electrospray mass spectra. *J. Mass Spectrom.* **37**, 223–229.
10. Zhang, Z. and Marshall, A. G. (1998) A universal algorithm for fast and automated charge state deconvolution of electrospray mass-to-charge ratio spectra. *J. Am. Soc. Mass Spectrom.* **9**, 225–233.

11. Brereton, R. G. (2003) *Data Analysis for the Laboratory and Chemical Plant*. John Wiley and Sons, New York, pp. 119–168.
12. Wehofsky, M., Hoffmann, R., Hubert, M., and Spengler, B. (2001) Isotopic deconvolution of matrix-assisted laser desorption/ionization mass spectra for substances-class specific analysis of complex samples. *Eur. J. Mass Spectrom.* **7**, 39–46.
13. Meija, J. and Caruso, J. A. (2004) Deconvolution of isobaric interferences in mass spectra. *J. Am. Soc. Mass Spectrom.* **15**, 654–658.



<http://www.springer.com/978-1-58829-563-7>

Mass Spectrometry Data Analysis in Proteomics

Matthiesen, R. (Ed.)

2007, X, 320 p. 87 illus., Hardcover

ISBN: 978-1-58829-563-7

A product of Humana Press