

Preprocessing of Mass Spectrometry Data in the field of Proteomics

Sabine Bachmayer
Proteomics and Bioinformatics
Masters Degree Program for Bioinformatics
University of Helsinki, Finland

May 25, 2007

Contents

1	Introduction	3
1.1	Mass Spectrometry	3
1.1.1	General Workflow for Proteomics Analysis	3
1.1.2	Type of Data	7
1.1.3	Problems with the raw data	7
2	Preprocessing of Mass Spectrometry Data	8
2.1	Data Reduction by Binning	8
2.2	Normalization	9
2.3	Peak Identification and Extraction	10
2.3.1	Peak Detection	10
2.4	Peak Alignment	11
2.5	Noise Reduction and Smoothing	12
3	Discussion / Conclusion	14

List of Figures

1.1	Workflow in Proteomics Analysis [5], where MS = Mass Spectrometry	3
1.2	Example 2D-Gel page, where the proteins are separated on their Immobilized pH gradients (horizontal axis / 1^{st} dimension) and on their molecular weights (vertical axis / 2^{st} dimension). Referenz: http://gelmatching.inf.fu-berlin.de/	4
1.3	A schematic and generalized illustration of the steps of a mass spectrometer [5]	4
1.4	Time of Flight analyzer Tube [5]	5
1.5	MALDI-TOF-Spectra for Polystyrene [7], where each peak corresponds to the exact mass-to-charge ratio of a peptide ion.	5
1.6	Electro Spray Ionization Process [14]	6
1.7	Matrix consisting of m/z and intensity value pairs and its resulting mass-to-charge spectra [7]	7
2.1	The above graphic shows the raw spectrum, the below graphic shows the reduced spectrum (window size: 100) [7]	9
2.2	The top panel shows the original graph. The bottom panel shows the transformed protein intensity measures [20].	11
2.3	The Calibration of a set of SELDI (which is a variant of MALDI) outputs for different samples, where the peak marked with an x is calibrated. The right side shows a partial enlargement of the calibrated peak. [20]	12
2.4	Baseline Removal flattens the baseline of a spectrum [7].	13

Chapter 1

Introduction

As the title already declare, this paper gives some introduction into the preprocessing of mass spectrometry data and an overview of common algorithms in use.

Caused by the circumstance that mass spectrometry is used in a lot of different applications and contexts, it is necessary to define the surrounding and some basic definitions which are used in this paper before start talking about the core topic of data-preprocessing.

1.1 Mass Spectrometry

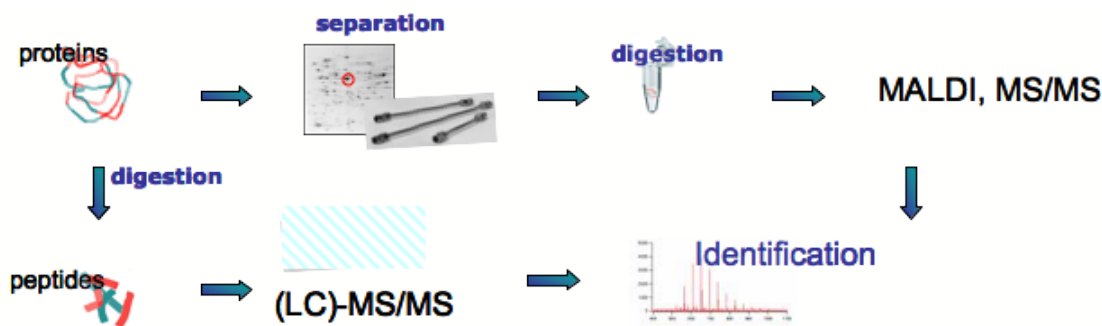


Figure 1.1: Workflow in Proteomics Analysis [5], where MS = Mass Spectrometry

As already mentioned Mass Spectrometry is used in a lot of applications and contexts like Protein Identification and Structure, Atom Probe, Pharmacokinetics and so on. In this paper we take the Mass Spectrometry as one part in the workflow of proteomics analysis which is shown in figure 1.1.

1.1.1 General Workflow for Proteomics Analysis

To understand in which context the Mass Spectrometry is taken in this paper we have to take a quick look to the General Workflow of Proteomics Analysis.

Protein Separation with 2D-Gel Analysis

This method separates proteins based on their isoelectric point (PI) or charge (1^{st} dimension) and on their molecular weight (2^{nd} dimension). The first part is the isoelectric focusing (IEF) where the mixed protein sample runs on a PH gradient and some electric-ity is applied. The proteins in the sample will now be positively charged at the PH level

below their PI and negatively charged at the PH level above their PI. The protein will stop moving, when the PH of the surrounding is equal to its PI because there will be no more charging.

The next step is to separate them concerning to their molecular weight. An electric current is applied again and the proteins move vertically - the heavier proteins move faster than the lighter ones which means that the heavier proteins are beneath in the 2D-Gel. Picture 1.2 shows an example 2D-Gel.[1]

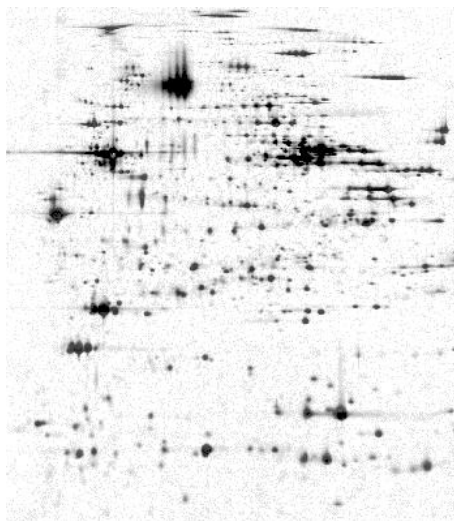


Figure 1.2: Example 2D-Gel page, where the proteins are separated on their Immobilized pH gradients (horizontal axis / 1^{st} dimension) and on their molecular weights (vertical axis / 2^{nd} dimension).

Referenz: <http://gelmatching.inf.fu-berlin.de/>

Digestion

To digest a protein sequence into smaller peptides, different digest enzymes like pepsin, trypsin and peptidases can be used [16]. Each of these enzymes cut the protein sequence at specific amino acids. By using 2D-Gels, the spots (see 1.2) are cut out of the gel and then digested into smaller peptides. If we go directly to the digestion, we directly cut the protein sequence/s from the sample (without separation).

Mass Spectrometry

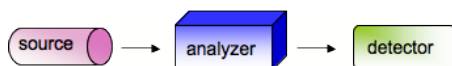


Figure 1.3: A schematic and generalized illustration of the steps of a mass spectrometer [5]

Figure 1.3 shows a schematic illustration of the procedure of a mass spectrometer and in general we can say that the mass spectrometry measures the mass-to-charge ratio of the particles. The following paragraphs describe two common mass spectrometers, the Matrix assisted laser disorbtion / ionization mass spectrometry (MALDI) and the Liquid chromatography mass spectrometry (LC) briefly.

Matrix assisted laser disorbtion / ionization - Time of Flight mass spectrometry (MALDI-TOF MS)

- **Source**

For the mass spectrometry, an ion source is needed which ionizes our protein sample we got from the digestion. The ionization is triggered by a laser beam. To protect the fragile biomolecules from being destroyed by the laser on the one hand, and to facilitate vaporization and ionization on the other hand, a matrix which consists of crystallized molecules is used. The laser is fired at the MALDI spot which absorbs the laser energy, so in this first step only the matrix is ionized. To charge the molecules of our sample, the matrix transfers a part of its charge to it which protect our sample from the disruptive energy of the laser. The most common analyzer used with MALDI is the Time of Flight (TOF) analyzer (see next paragraph) [5] [3].

- **Time of Flight Analyzer**

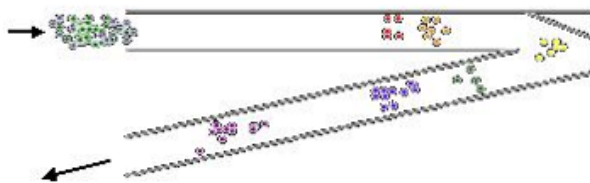


Figure 1.4: Time of Flight analyzer Tube [5]

The ions we got from the source are now separated by the analyzer according to their mass-to-charge ratio. The estimation of the mass-charge ratio happens by measuring the time of flight (TOF) of the ions. For that, the ions are accelerated by an electric field and the ions fly through a so called analyzer tube (see figure 1.4. Caused by the relation $E = \frac{1}{2}mv^2$, the time of flight of the ions is proportional to the ratio mass / charge values, so $tof \propto \sqrt{\frac{m}{q}}$ which means the greater the mass-charge ratio, the faster they fly [5] [3] [13].

- **Detector**

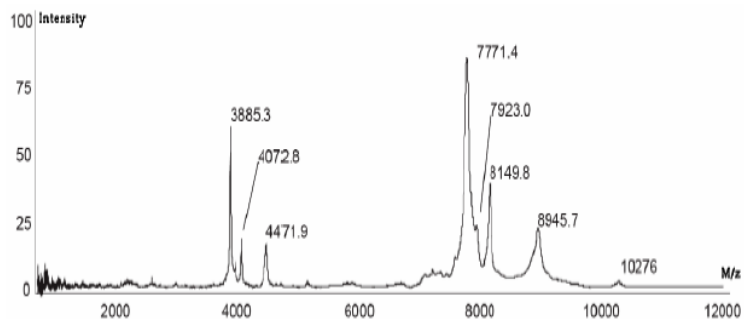


Figure 1.5: MALDI-TOF-Spectra for Polystyrene [7], where each peak corresponds to the exact mass-to-charge ratio of a peptide ion.

The separated ions are now detected and their signal is sent to a computer which stores this mass-to-charge ratios together with their relative abundance. In order to get a survey to the results, they are stored in the format of a mass-to-charge spectrum (see figure 1.5). The list of the peaks shown in figure 1.5 are the list of masses and this is the so called peptide-mass-fingerprint which identifies a single peptide [5].

Liquid Chromatography Mass Spectrometry (LC MS)

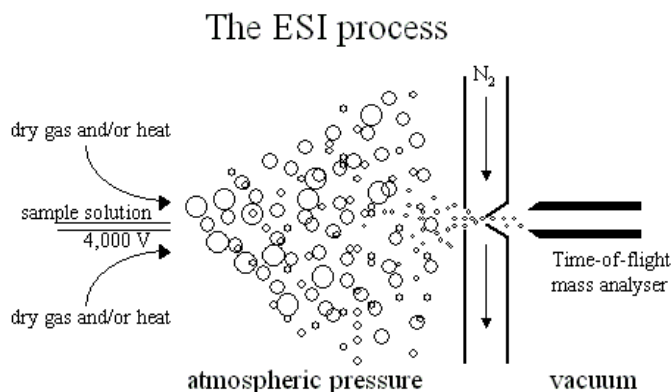


Figure 1.6: Electro Spray Ionization Process [14]

- **Source**

In this method, the ionization process is done by atmospheric pressure (API). For that a liquid, which contains our sample, is sprayed out through a small charged needle which produces small and also charged droplets. The solvent is evaporated leaving the sample molecule in the gas and then ionized (see Figure 1.6). In the next step, a time of flight analyzer is used [14].

- **Time of Flight Analyzer**

See 1.1.1.

- **Detector**

See 1.1.1.

1.1.2 Type of Data

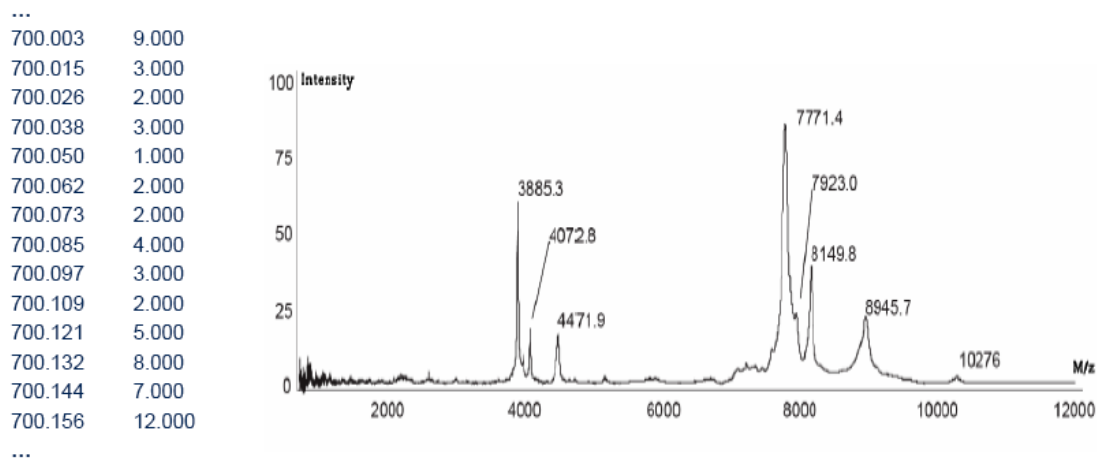


Figure 1.7: Matrix consisting of m/z and intensity value pairs and its resulting mass-to-charge spectra [7]

For the raw data, we must not take into account which type of mass spectrometry was used. As figure 1.7 shows, our mass-to-charge spectrum consists of value pairs containing the mass-to-charge ratio (m/z) and the intensity. One matrix can contain thousands of values which depends on the resolution of the spectra. In particular, data produced by mass spectrometer are affected by errors and noise due to sample preparation, sample insertion into the instrument and the instrument itself [4].

1.1.3 Problems with the raw data

Using the raw / original data for the analyzation is not the optimum because we have to consider the contaminations [4] (I) noise, (II) peak broadening, (III) instrument distortion and saturation, (IV) isotopes, (V) wrong calibration, (VI) different contaminants and problems (I) size of the matrix (II) m/z measurement errors.

In the following chapter algorithms, methods and software tools, to treat with the above listed contaminations and problems, are described and discussed.

Chapter 2

Preprocessing of Mass Spectrometry Data

2.1 Data Reduction by Binning

Binning is one of the most used preprocessing technique in the mass spectrometry analysis to reduce the huge amount of data.

The basic idea is to scan the original spectra and group adjacent values of the data into so called bins which causes a dimensionally reduction of the data (if we think of the matrix). Furthermore a representative member of each group is selected which represents the mass-to-charge ratio and the intensive value for its whole group.

The tricky part is to scan the spectra by using a so called sliding window. Tricky because one has to estimate the width of this sliding window - if it is too large, the reduced data can be inexact or incorrect. If it is too small, the effect of the data reduction is not given. This scanning operation creates the bins. One bin then consists of N mass-to-charge / intensity value pairs (= peaks) in the form of $[(I_1, m/z_1), (I_2, m/z_2), (I_3, m/z_3), \dots, (I_N, m/z_N)]$ which is combined to one value pair (or peak) $(I, m/z)$. The intensity value of this bin is calculated by using an aggregate function (like the sum) on all N original intensity values and its m/z value is determined among the N original m/z values by taking the median, corresponding value to the maximum intensity, average value or something similar.

This binning offers a subset of the original data which are much more easier to handle and to analyze as figure 2.1 shows. However it can cause a loss of accuracy and because of that it is absolutely necessary to take the characteristics of the spectrometers (for example MALDI-TOF or LC-MS) into account and also the experiment parameters (sample, needed resolution, etc) [10].

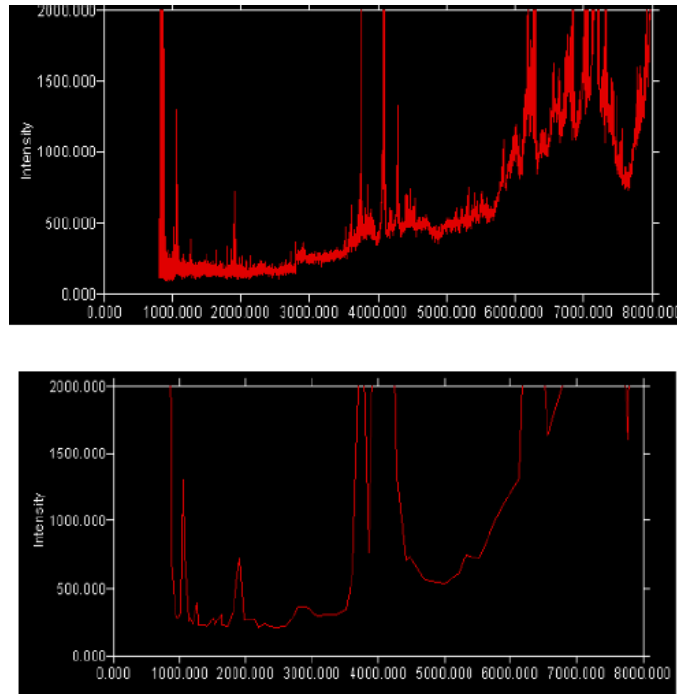


Figure 2.1: The above graphic shows the raw spectrum, the below graphic shows the reduced spectrum (window size: 100) [7]

2.2 Normalization

Normalization is done to make the data independent of experimental variations (like varying amounts of protein, degradation in the sample or variations in the detector sensitivity) and make different spectra comparable. To make this possible, the relative intensities of the spectrograms are normalized. The following paragraph describes the common used global normalization in different variations.

This method normalizes the signal intensities across several samples to identify and remove variations between these samples as described above. The global normalization assumes first that the sample intensities are related by a constant factor. Second, one assume that the number of proteins that are overrated is almost equal to those that are underrated. And third, that the number of proteins, whose expression level changed, is small relative to the total number of proteins. This re-scaling factor for the intensity can be a fixed scaled value like 100 (which is not recommended) or the median or mean of the spectrum [9], which is justified by assumption two [17]. The problem with these re-scaling factors, if they are used globally, is the non-random missing of peptides which is caused by ions that intensity level undercut a certain threshold. This phenomenon is caused by the limited sensitivity of the detector and instrumental noise.

So Wang et al. [17] recommend the top L order statistics of feature intensities in each sample. The parameter L is chosen by the user. An example calculation could be:

Given are two samples X and Y and their intensity values $X = (x_1, x_2, x_3, \dots, x_n)$ and $Y = (y_1, y_2, y_3, \dots, y_m)$.

The order statistic of the intensity values is denoted as $x(1) > x(2) > \dots > x(n)$ and $y(1) > y(2) > \dots > y(m)$.

For a chosen $L < \min(n, m)$, the re-scaling factor is estimated as $L = \frac{\text{median}(x(1), \dots, x(L))}{\text{median}(y(1), \dots, y(L))}$. Sauve et al [9] recommend the average area under curve (AUC) of all spectra for re-scaling because in the MS, the concentration of a protein is defined by the AUC of its peak. On the other hand Cannataro et al [7] [4] developed the following four different methods for calculating the re-scaled intensity value:

Direct normalization:
$$I_{j_{norm}} = \frac{I_j - I_{min}}{I_{max} - I_{min}}$$

Inverse normalization:
$$I_{j_{norm}} = 1 - \frac{I_j - I_{min}}{I_{max} - I_{min}}$$

Canonical normalization:
$$I_{j_{norm}} = \frac{I_j}{\sum I_j}$$

Logarithmic normalization: Is used if a skewed distribution is give. There the intensity value for a spectra x , at m/z value j is transformed logarithmically to fix this skewed distribution.

For more detailed descriptions of the above presented approaches, see [17], [9], [7] and [4].

2.3 Peak Identification and Extraction

2.3.1 Peak Detection

As already mentioned in 1.1.1, one peak corresponds to the intensity value and mass-to-charge ratio of one particle (for example a peptide). For further machine analysis we have to detect and extract these peaks automatically - of course it would be possible to do that manually but if we take the quantity and the necessary accuracy into account it is recommended to do it automatically. For a human being, the detection of (most) peaks is not a great deal, our eye is able to cover the largest peaks within a few seconds but with less accuracy concerning the smaller peaks. To do the peak detection automatically, different algorithms exist. The easiest method is the nearest neighbor analyzation presented in [20]. Yasui et al. define the size of a nearest neighbor set N and charge then for each m/z point whether the protein intensity at that point is the highest among its nearest neighbors $+/- N$ points in direction of the m/z - axis. If we get a positive result for that peak which means that it is the highest in this region, we mark it as a peak. Yasui et al [20] started with $N = 10$ by trial and error. Of course the N has to be chosen carefully because its an error source.

The results presented in [20] shows that this easy method is not that optimal. Figure 2.3 shows on its upper plot the original spectra and below the peaks detected and one can see the differences which can falsify the results.

Another approach is presented by Pan Du et al. [8] which assumes that each peak has a certain pattern and width in contrary to noise. This algorithm uses the Signal Noise Ratio (SNR) and Continuous Wavelet Transforms which are widely in use in the field of pattern recognition. This algorithm is quoted for the sake of completeness and to list a reference which achieves useful results but is not further discussed because it would go beyond the scope of this paper. For detailed information take a look to

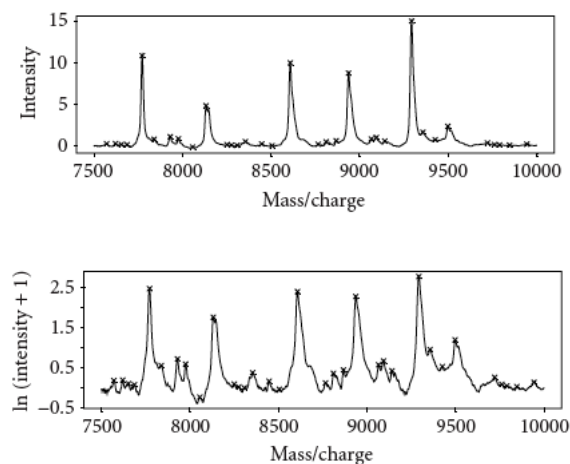


Figure 2.2: The top panel shows the original graph. The bottom panel shows the transformed protein intensity measures [20].

2.4 Peak Alignment

Peak alignment is used to find out which peaks among different spectra correspond to the same peak (protein, ...). The problem here is the machine measurement error of 0.03% – 0.06% which has to be considered by the decision if two peaks from different spectra are the same or not. This preprocessing is absolutely necessary because there exist a significant variability between samples in intensity, background and location of the m/z peaks which must be resolved before comparisons can be made across samples.

One can see that urgent requirements exist because of the big amount of solutions and algorithms which came up during the past five years:

For example Yasui et al. [20], Randolph and Yasui [15] and Wong et al. [19] developed solutions to align the m/z values across replicate mass spectra due to the variability in instruments, operators and processing centers. Ball et al. [2], Li et al. [12] and again Yasui et al. [20] reported a need for a variability in m/z ranging from 0.1% in the lower m/z range up to 0.3% in the greater m/z range. In the following paragraph, the algorithm reported by Yasui et al. [20] is presented in detail.

The general idea of the algorithm, developed by Yasui et al., is to replace the original m/z values with a set of calibrated (or aligned) m/z values. For that they define a so called "window of potential shift" which is the range of the potential m/z shifting from a m/z point which contains this m/z point itself. Normally one can assume the "window of potential shift" for a m/z point P is about $\pm 0.1 - 0.2\%$ of the m/z value of that point. This value is based on the quality control experiments of the manufacturer of mass spectrometers. The algorithm itself only calculates the number of peaks in each sample at each m/z point P which are within the "window of potential shift" for P . The m/z point P which has the highest number of peaks over all samples within the "window of potential shift" is taken into the new m/z set as a calibrated m/z value - the other m/z points are not included in the algorithms subroutine anymore. This procedure is repeated until all peaks are exhausted from every sample. As a result we get the new calibrated m/z set.

Finally, the algorithm calculates a calibrated data-set which consists of protein intensity measures of each sample that correspond to the points in the new calibrated m/z set.

So for each sample i and for each point j in the new calibrated m/z set, the maximum protein intensity measure of the sample i is taken, among the protein intensity measures corresponding to the "window of potential shift" for point j . The final calibrated data-set is Y_{ij} whose elements represent protein intensity measures indexed by the sample i and the calibrated m/z value j .

Figure ?? shows the results for the algorithm described in [20].

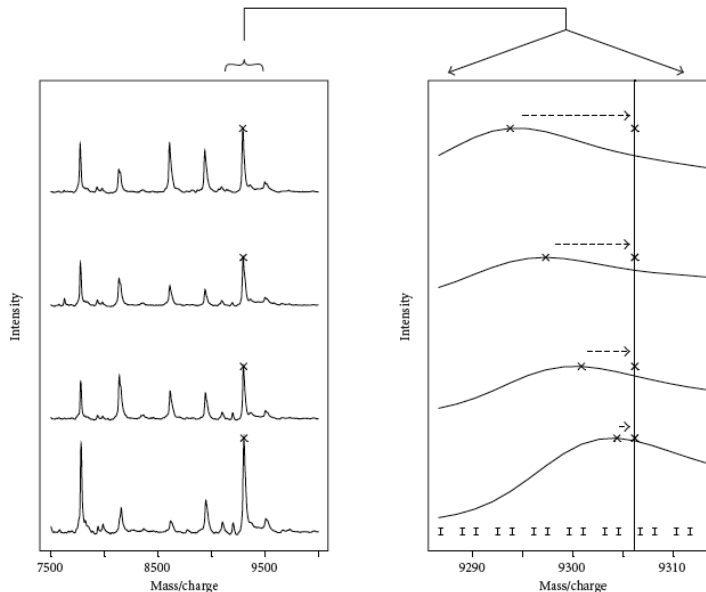


Figure 2.3: The Calibration of a set of SELDI (which is a variant of MALDI) outputs for different samples, where the peak marked with an x is calibrated. The right side shows a partial enlargement of the calibrated peak. [20]

2.5 Noise Reduction and Smoothing

Noise is defined as any unwanted signal interfering with the clarity and intelligibility of desired signals [6]. In the case of doing mass spectrometry, we have two types of noise:

1. Electrical Noise

One can say that this type of noise is caused by physics, namely by the instruments that are used and change the intensity randomly. To reduce this kind of noise we have to do a so called smoothing. Smoothing is a kind of low-pass filter which reduces the noise level in the whole spectrum. The following paragraphs present the smoothing algorithms which are implemented in the tool IGOR which is a graphing, data analysis and image processing tool [18].

Igors Smooth operation performs box and binomial smoothing. The different smoothing algorithms convolve the input data with different coefficients.

- *Box Smoothing* The principle behind box smoothing is quite simple and replaces each value in the spectrum with the average of its neighboring values. It is recommended to average the same number of values before and after the calculated average to avoid shifting of the data. The equation to that explanation is the following: $\bar{x}[i] = \frac{1}{(2*M+1)} * \sum_{j=-M}^{+M} x[i + j]$, where $\bar{x}[i]$ is the replaced value on

position i in our spectrum, $x[i]$ is the original value on position i in the spectrum and M is the number of neighbors.

- *Binomial Smoothing* Binomial smoothing is a Gaussian filter. In the case of mass spectrometry, a one dimensional Gaussian filter with the form

$G(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}$ is adequate, where σ is the standard deviation of the distribution. For a further description see [18] and their references.

2. Chemical Noise

is caused by contaminants in the sample or matrix molecules and is manifested in the baseline of a signal. The baseline is an offset of the intensities of masses as figure 2.4 shows. This happens only at low masses and is caused by the molecules of the energy absorbing matrix (see 1.1.1) and varies in general between the spectra. To reduce this kind of noise we have to identify and remove this baseline which flattens the base profile of a spectrum. The following paragraph presents the baseline subtraction (described in [11]), a common used method to reduce this kind of noise. Baseline Subtraction

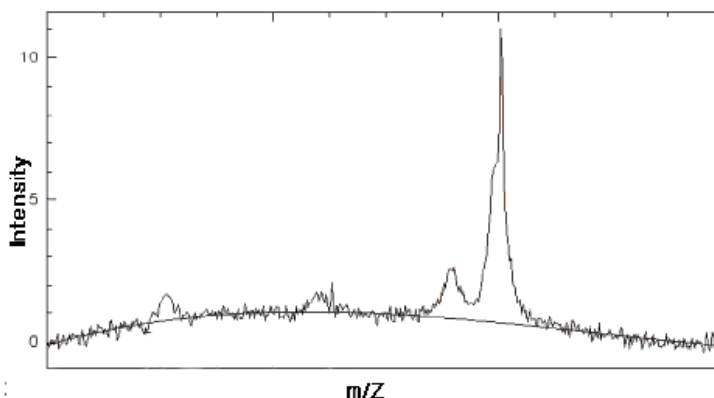


Figure 2.4: Baseline Removal flattens the baseline of a spectrum [7].

[11] extract first the local minima from the spectrum and do then a local weighted quadric fitting on these extracted values. On these new fitted values, a new search for local minima is done which smooth out small variations. The new local minima are used to split then the signal into constant parts and finally the baseline is computed by doing the initial local weighted quadric fitting on this pieces of constant signals calculated before. The baseline subtraction is useful when the spectrum consists of a large number of peaks. When measuring a large number of peaks, it is often more effective to subtract an estimated baseline from the data.

Smoothing and baseline removal are two common methods for data reduction 2.1 and image processing.

Chapter 3

Discussion / Conclusion

The ability to identify biological molecules, if it is for the science in general or in the field of the medicine to identify cancer cells or to identify molecules that might serve as novel therapeutic goals, could have recondite benefits.

The core topic of this paper was the preprocessing of mass spectrometry data and an overview of a range of common used techniques in this field, were shown. This procedure of data preprocessing seems to be quite complex and time consuming but is absolutely necessary to treat with the following problems: (I) noise, (II) peak broadening, (III) instrument distortion and saturation, (IV) isotopes, (V) wrong calibration, (VI) different contaminants, (VII) size of the matrix, (VIII) m/z measurement errors [4]. To minimize these irritations a huge amount of techniques, algorithms and corresponding software tools exist. The ones described above are just a few picked out to show the possibilities but also the problems and pitfalls.

Bibliography

- [1] Christine Antler. Information on protein identification [available on <http://www.scq.ubc.ca/investigating-the-cellular-machinery-protein-identification/>].
- [2] G. Ball, S. Milan, F. Holding, R. O. Allibone, J. Lowe, S. Ali, G. Li, S. McCardle, I.O. Ellis, C. Creaser, and R. C. Rees. An integrated approach utilizing artificial neural networks and seldi mass spectrometry for the classification of human tumours and rapid identification of potential biomarkers. *Bioinformatics*, 18(3):395–404, 2002.
- [3] Gerardo Brucker. Information on MALDI-TOF Mass Spectrometry [available on <http://www.chemistry.wustl.edu/~msf/damon/>].
- [4] Mario Cannataro, Pietro Hiram Guzzi, Tommaso Mazza, and Pierangelo Veltri. Pre-processing, management and analysis of mass spectrometry proteomics data.
- [5] Definition of noise [available on [http://de.wikipedia.org/wiki/Rauschen_\(Physik\)](http://de.wikipedia.org/wiki/Rauschen_(Physik))].
- [6] Pan Du. Improved peak detection in mass spectrometry spectrum by incorporating continuous wavelet transform-based pattern matching, 2006.
- [7] Pan Du, Warren A. Kibbe, and Simon M. Lin. Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics*, 22(17):2059–2065, 2006.
- [8] P. H. Guzzi, T. Mazza, and G. Tradigo. Normalization, baseline correction and alignment of high-throughput mass spectrometry. In *Data Proceedings Gensips 2004*. GENSIPS, 2004.
- [9] P. H. Guzzi, T. Mazza, and G. Tradigo. Preprocessing of mass spectrometry proteomics data on the grid. In *CBMS '05: Proceedings of the 18th IEEE Symposium on Computer-Based Medical Systems*, pages 549–554, Washington, DC, USA, 2005. IEEE Computer Society.
- [10] Alexandros Kalousis. General information on mass spectrometry [available on <http://cui.unige.ch/AI-group/research/massspectrometry/massspectrometryframe.htm>].
- [11] Sophia Kossida. Proteomics and bioinformatics. Lecture Slides, University of Helsinki, 2007.
- [12] J. Li, Z. Zhang, J. Rosenzweig, Y. Y. Wang, and D. W. Chan. Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer. *Clinical Chemistry*, 48(8):1296–1304, 2002.

- [13] Information on MALDI-TOF Mass Spectrometry [available on <http://de.wikipedia.org/wiki/TOFMS>].
- [14] Charly Morgan. Information on LC-ESI Mass Spectrometry [available on <http://qbab.aber.ac.uk/roy/mss/lct.htm>].
- [15] Timothy Randolph and Yutaka Yasui. Multiscale processing of mass spectrometry data. UW Biostatistics Working Paper Series 1063, Berkeley Electronic Press, July 2004. available at <http://ideas.repec.org/p/bep/uwabio/1063.html>.
- [16] safe2Use. Information on digest enzymes [available on <http://www.safe2use.com/data/enzymes.htm>].
- [17] Pei Wang, Hua Tang, Heidi Zhang, Jeffrey Whiteaker, Amanda G. Paulovich, and Martin McIntosh. Normalization regarding non-random missing values in high-throughput mass spectrometry data. In *Pacific Symposium on Biocomputing 2006*. World Scientific Publishing, 2006.
- [18] wavemetrics. Smoothing algorithms implemented in IGOR [available on <http://www.wavemetrics.com/products/IGORPro/dataanalysis/signalprocessing/smoothing.htm>].
- [19] Jason W.H. Wong, Gerard Cagney, and Hugh M. Cartwright. Specalign—processing and alignment of mass spectra datasets. *Bioinformatics*, 21(9):2088–2090, 2005.
- [20] Yutaka Yasui, Dale McLerran, Bao-Ling Adam, Marcy Winget, Mark Thornquist, and Ziding Feng. An automated peak identification/calibration procedure for high-dimensional protein measures from mass spectrometers. *Journal of Biomedicine and Biotechnology*, 2003(4):242–248, 2003. doi:10.1155/S111072430320927X.

All links were validated in Mai 2007