Preprocessing, Management, and Analysis of Mass Spectrometry Proteomics Data

M. Cannataro, P. H. Guzzi, T. Mazza, and P. Veltri

Università Magna Græcia di Catanzaro, Italy

1 Introduction

Mass Spectrometry (MS) based proteomics is becoming a powerful, widely used technique in order to identify different molecular targets in different pathological conditions [1]. Proteomics experiments involve different and heterogeneous technological platforms so a clear understanding of the function and errors related to each one has to be taken into account. In particular, data produced by mass spectrometer are affected by errors and noise due to sample preparation, sample insertion into the instrument (different operators can lead to different results using the same sample) and instrument itself. Mass spectrometry-based proteomics experiments usually comprise a data generation phase, a data preprocessing phase and a data analysis phase (usually data mining, pattern extraction or peptide/protein identification). Mass spectrometry produces a huge volume of data, said spectra, that are represented as a very large set of measures (*intensity, m/Z*), representing the abundance (intensity) of biomolecules having certain mass to charge ratio (m/Z) values.

In this paper, after introducing Mass Spectrometry, we survey different techniques for spectra preprocessing and we present a first design of a software tool that allows to manage efficient storing and preprocessing of mass spectrometry data. A first performance evaluation of MS-Analyzer is also presented.

2 Mass spectrometry proteomics data

Mass Spectrometry is a technique more and more used to identify macromolecules in a compound. The mass spectrometer is an instrument designed to separate gas phase ions according to their *m/Z* (mass to charge ratio) values. Matrix-Assisted Laser Desorption / Ionization - Time Of Flight Mass Spectrometry (MALDI-TOF MS) is a relatively novel technique that is used for detection and characterization of biomolecules, such as proteins, peptides, oligosaccharides and oligonucleotides, with molecular masses between 400 and 350000 Da [2]. The Mass Spectrometry process [1] can be decomposed in three subphases: (i) Sample Preparation (e.g. Cell Culture, Tissue, Serum); (ii) Proteins Extractions; and (iii) Mass Spectrometry processing. Mass Spectrometry output is represented, at a first stage, as a (large) sequence of value pairs, where each pair contains a measured *intensity*, which depends on the quantity of the detected biomolecules and a mass to charge ratio (*m/Z*), which depends on the molecular mass of detected biomolecules.

When obtaining a spectrum we have to consider some imperfection causes: (i) noise, (ii) peak broadening, (iii) instrument distortion and saturation, (iv) isotopes, (v) miscalibration, (vi) contaminants of various kinds. Data cleaning is performed in different phases by using: (i) best-practices sample preparation; (ii) mass spectrometer software; (iii) further data pre-processing algorithms. In the rest of the paper we focus on pre-processing techniques conducted after data have been produced and eventually cleaned by the spectrometer.

3 Preprocessing and analysis of mass spectrometry proteomics data

Each point of a spectrum is the result of two measurements, m/Z and intensity, that are corrupted by noise. Preprocessing is the process that consists of spectrum noise and contaminants *cleaning up*. Moreover, preprocessing can also be used to reduce dimensional complexity of the spectra, but it is important to use efficient and biologically consistent algorithms. Currently this is an open problem. In summary, preprocessing (see [5] for a survey) aims to correct intensity and m/Z values in order to: (i) reduce noise, (ii) reduce amount of data, and (iii) make spectra comparable.

3.1 Noise reduction and normalization

Noise reduction and normalization are conducted in part by the spectrometer and in part by external preprocessing tools. In the following we describe some approaches to noise reduction and normalization.

Base line subtraction and smoothing. Each of these techniques aims to reduce the noise. Base line subtraction flattens the base profile of a spectrum while smoothing reduces the noise level in the whole spectrum. Each mass spectrum exhibits a base intensity level (a baseline) which varies from fraction to fraction and consequently needs to be identified and subtracted. This noise varies across the m/Z axis, and it generally varies across different fractions, so that a one-value-fits-all strategy cannot be applied.

Base line subtraction uses an iterative algorithm to attempt to remove the baseline slope and offset from a spectrum by iteratively calculating the best fit straight line through a set of estimated baseline points. *Smoothing* is a process by which data points are averaged with their neighbors as in a time-series of data. The main reason for smoothing is to increase signal to noise ratio.

Normalization of intensities. Normalization enables the comparison of different samples since the absolutes peak values of different fraction of spectrum could be incomparable. The purpose of spectrum normalization is to identify and remove sources of systematic variation between spectra due for instance to varying amounts of sample or degradation over time in the sample or even variation in the instrument detector sensitivity. We have analyzed and implemented four normalization methods not described here due to space limits: the *Canonical Normalization*, the *Inverse Normalization*, cited in [3] and used in [4], the *Direct Normalization*, and the *Logarithmic Normalization*, described by B. Wu [6] and Y. Yasui [7].

3.2 Data Reduction

Binning. Binning is one of the most used preprocessing technique in MS data analysis. Its aim is to preserve raw data information while performing a dimensional reduction for subsequent processing and mining phases. Binning performs data dimensionality reduction by grouping measured data into bins. This process involves grouping adjacent values and electing for each group a representative member.

3.3 Identification and extraction of peaks

Algorithms that do not require human intervention are needed for rapid and repeatable quantitative processing of spectra that often contain hundreds of discrete peaks. *Peaks extraction* consists of separating real peaks (e.g. corresponding to peptides) from peaks representing noise. Although sometimes such task can be performed by using the data-processing embedded in mass spectrometer, custom identification methods fitting both informatics and biological considerations are more effective.

3.4 Peaks alignment

A point in a spectrum represents a measurement of mass to charge ratio and electrical intensity. Each of these measurements is affected by an error. Correction of error in m/Z measurement is also known as data-calibration or *alignment* of correspondent peaks across samples. Without alignment, the same peak (e.g. the same peptide) can have different values of m/Z across samples. To allow an easy and effective comparison of different spectra, peaks alignment methods find a common set of peak locations (i.e. m/Z values) in a set of spectra, in such a way that all spectra have common m/Z values for the same biological entities. Each detected m/Z value is afflicted by noise causing the presence of a window in which mass/charge ratio can be shifted. In [7] this window is defined as *window of potential shift* indicating the range of potential mass/charge shifting for each m/Z point. Characteristics of shift are strictly lied to the mass spectrometer used. Peaks alignment consists of shifting m/Z values in such window such that peaks in all spectra will have the same m/Z.

4 MS-Analyzer

MS-Analyzer is a Grid-based Problem Solving Environment for proteomics applications, that uses domain ontologies to model basic software tools and data sources and workflow techniques to design complex in silico experiments. MS-Analyzer sits in the middle of proteomic facilities and data mining software tools, so its main requirements are: interfacing with proteomics facility; storing and managing MS proteomics data; interfacing with off-the-shelf data mining and visualization software tools (e.g. WEKA, IBM Intelligent Miner, etc.). In particular, MS-Analyzer provides the following functions:

1. **MS proteomic data acquisition** loads MS raw spectra produced by different kind of Mass Spectrometers.

2. **MS proteomic data pre-processing** loads MS raw spectra and applies the pre-processing techniques described before.

3. **MS proteomic data preparation** loads pre-processed spectra and prepare them to be given in input to different kind of data mining tools.

4. **Data Mining analysis** allows to select and execute different data mining tasks (e.g. classification, clustering, pattern analysis), and the corresponding data mining algorithms and tools (e.g. Q5, C5, K-means, etc.), producing knowledge models.

5. **Data Visualization and/or Visual Data Mining**.

5 Performance evaluation

The general scheme of a mass spectrometry-based data mining experiment comprises the following activities: (i) loading of the raw spectra produced by mass spectrometer; (ii) preprocessing of the raw spectra data, (iii) preparation of the data mining input file (e.g. the Weka ARFF file), and (iv) data mining analysis (e.g. classification) of of mass spectra.

We considered three different mass spectra datasets publicly available on Internet: (i) Pancreatic Cancer dataset (it contains 142 spectra each one with 6,772 (m/Z, intensity) measurements); (ii) Prostate Cancer dataset (it contains 322 spectra each one with 15,154 measurements); and (iii) Ovarian Cancer dataset (it contains 49 spectra each one with 59,386 measurements). The Pancreatic and Prostate datasets are available at [1] while the Ovarian datases is available at [2].

For each dataset, we conducted a set of classification experiments and we measured: (i) *execution times* of, respectively, loading and preparation, preprocessing, and classification activities ; (ii) *memory occupancy* of, respectively, raw spectra, preprocessed spectra and ARFF files; (iii) quality indexes of classification under different preprocessing conditions. [3]

As an example of such measurements, figure 1 shows the sizes of different datasets when applying none or different preprocessing techniques.

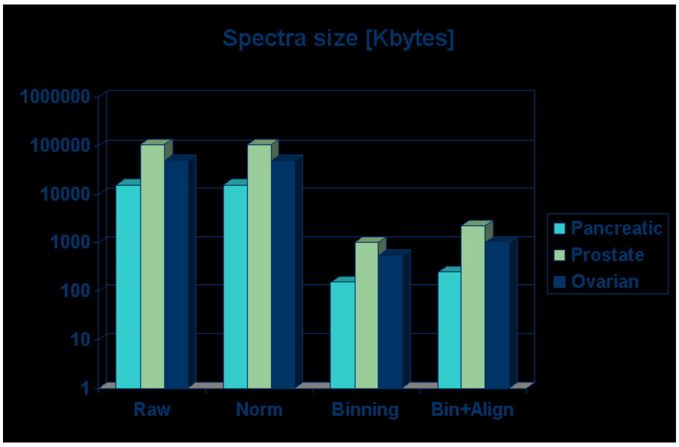

Figure 1. Datasets sizes and preprocessing

Figure 2 shows the execution time of a classification experiment when applying none or different preprocessing techniques.

6 Conclusion

In this paper we presented a survey of preprocessing techniques for MS proteomics data and the design and a first prototype of MS-Analyzer, a tool for the management, preprocessing and analysis of proteomics mass spectra. We currently implemented all the preprocessing tools and the data preparation for the WEKA data mining platform. The performance evaluation of MS-Analyzer using three public

[1] http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp

[2] http://www-stat.stanford.edu/%7Etibs/PPC/Rdist/index.html

[3] Execution times are expressed in milliseconds, and memory occupancy is expressed in KiloBytes.

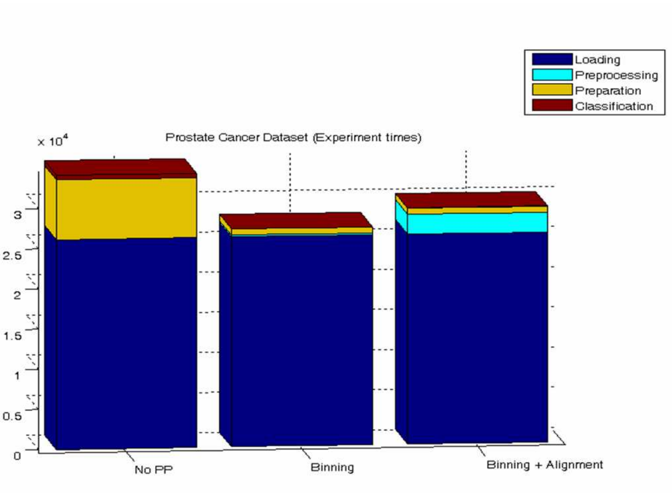

Figure 2. Execution times and reprocessing

available mass spectra datasets showed how the selective use of preprocessing techniques improve execution times, memory occupancy and quality of data mining (measures not reported here due to space limitations).

References

[1] R. Aebersold and M. Mann. Mass spectrometry-based proteomics. *Nature*, 422:198–207, 13 March 2003 2003.

[2] G. L. Glish and R. W. Vachet. The basic of mass spectrometry in the twenty-first century. *Nature Reviews*, 2:140–150, February 2003 2003.

[3] V. Gopalakrishnan, E. William, S. Ranganathan, R. Bowser, M. E. Cudkowic, M. Novelli, W. Lattazi, A. Gambotto, and B. W. Day. Proteomic data mining challenges in identification of disease-specific biomarkers from variable resolution mass spectra. In *Proceedings of SIAM Bioinformatics Workshop 2004*, pages 1–10, Lake Buena Vista, FL, April 2004.

[4] E. I. Petricoin, A. Ardekani, B. Hitt, P. Levine, V. Fusaro, and S. S. et al. Use of proteomic patterns in serum to identify ovarian cancer. *Lancet*, 9306(359):572–577, 2002.

[5] M. Wagner, D. Naik, and A. Pothen. Protocols for disease classification from mass spectrometry data. *Proteomics*, 3(9):1692–8, 2003.

[6] B. Wu, T. Abbott, D. Fishman, W. McMurray, G. Mor, K. Stone, D. Ward, K. Williams, and H. Zhao. Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics*, 1(19):1636–43, September 2003.

[7] Y. Yasui, D. McLerran, B. Adam, M. Winget, M. Thornquist, and Z. Feng. An automated peak identification/calibration procedure for high-dimensional protein measures from mass spectrometers. *Journal of Biomedicine and Biotechnology*, (4):242–248, 2003.