Preprocessing of Mass Spectrometry Proteomics Data on the Grid

M. Cannataro¹, P. H. Guzzi¹, T. Mazza¹, G. Tradigo², P. Veltri¹

¹Università Magna Græcia di Catanzaro, Italy, ²ICAR-CNR, Italy
{cannataro, hguzzi, t.mazza, veltri}@unicz.it
gtradigo@si.deis.unical.it

Abstract

The combined use of Mass Spectrometry and Data Mining is a novel approach in proteomic pattern analysis for discovering novel biomarkers or identifying patterns and associations in proteomic profiles. Data produced by mass spectrometers are affected by errors and noise due to sample preparation and instrument approximation, so different preprocessing techniques need to be applied before analysis is conducted. We survey different techniques for spectra preprocessing, and we present a first design of a software tool that allows the preprocessing, management and analysis of mass spectrometry data on the Grid.

1. Introduction

Mass Spectrometry (MS) 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 / Ionisation - 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 [4].

The Mass Spectrometry process can be decomposed in three sub-phases: (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. Mass spectra are usually represented in a graphical form as in Figure 1, that shows the spectrum of a real sample containing insulin and mioglobin. The peak (\( m/Z=5736.85 \)) denotes the insulin, whereas the peak (\( m/Z=8688.14 \)) denotes the mioglobin.

Mass spectrometry-based proteomics is becoming a powerful, widely used technique in order to identify different molecular targets in different pathological conditions [6] [7]. 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. Moreover, other imperfection causes are: (i) noise, (ii) peak broadening, (iii) instrument

¹This work has been partially supported by projects GRID.IT, grant n. RBNE01KNFP, funded by CNR, and COFIN 2003, grant n. 2003060917, funded by MIUR.
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 preprocessing algorithms. In Mass spectrometry-based proteomics, the Mass Spectrometry process is only one of the phases, and proteomics experiments usually comprise: (i) a data generation phase (i.e. the Mass Spectrometry processing cited before), (ii) a data preprocessing phase, and a (iii) data analysis phase (usually data mining, pattern extraction or peptide/protein identification).

In the rest of the paper we focus on spectra preprocessing, surveying different techniques conducted after data have been produced and eventually cleaned by the spectrometer. Since such techniques can be applied alone or in combination, we present a first design of a software tool that allows to manage efficient storing and preprocessing of mass spectrometry data. An application of the proposed approach in a Grid scenario is also depicted. The rest of the paper is organized as follows. Section 2 surveys main preprocessing techniques. Section 3 describes the architecture and a first prototype of MS-Analyzer, a software tool for the management of mass spectra on the Grid. Finally, Section 4 concludes the paper and outlines future work.

2. Preprocessing 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 [7] 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.

2.1. Noise reduction and normalization

Noise reduction and normalization are conducted in part by the spectrometer software 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, as depicted in Figure 2. 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.
**Figure 2. Base line subtraction**

*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. Smoothing techniques are frequently used in Mass Spectrometry as reported in [7]. The basic process of smoothing is very simple. The simpler software technique for smoothing signals consisting of equidistant points is named moving average. An array of raw (noisy) data \([y_1, y_2, ..., y_N]\) is converted to a new array of smoothed data, where each “smoothed point” \((y_k)\) is the average of an odd number of consecutive \(2n + 1\) points of the raw data, \((n = 1, 2, 3, ...)\), where the odd number \(2n + 1\) is usually named filter width.

**Normalization of intensities.** Normalization enables the comparison of different samples since the absolutes peak values of different fractions 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 [5] and used in [6], the Direct Normalization, and the Logarithmic Normalization, described in [9] and [10].

### 2.2. 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. The binning algorithm takes a subset of \(N\) peaks from a spectra, represented by the couples \([ (I_1, m/Z_1), (I_2, m/Z_1), ..., (I_N, m/Z_N) ]\), and substitutes all of them with a unique peak \((I, m/Z)\), whose intensity \(I\) is an aggregate function of the \(N\) original intensities (e.g. their sum), and the mass \(m/Z\) is usually chosen among the original mass values (e.g. the median value, or the value corresponding to the maximum intensity). Such basic operation is conducted by scanning all the spectrum by using a sort of sliding window. So the main parameters of the method are: (i) the width of the sliding window (it should be noted that a constant window can contain a variable number of peaks along the sample), (ii) the function used to calculate the aggregate intensity \(I\), (iii) the function used to choose the representative \(m/Z\). Moreover, binning algorithms should take into account the characteristics of spectrometers (e.g. MALDI-TOF or LC-MS), and experiment parameters (resolution of mass spectrometer used). We have analyzed and implemented two binning
2.3. 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 or thousands 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. We are implementing an ad-hoc algorithm based on Spectral Polygonalization that locates peaks and calculates their associated area, first discussed in [8]. This algorithm is based on a time-series segmentation routine that reduces the data set to groups of three strategic points where each group defines the beginning, the medium and the ending of each located peak. Peaks with statistically insignificant height or area are then discarded [3], [8].

2.4. 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 [10] 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 same peaks in all spectra will have the same m/Z.

3. MS-Analyzer

In recent works [2] we developed PROTEUS, a Grid-based Problem Solving Environment for bioinformatics applications that uses domain ontologies to model basic software tools and data sources and workflow techniques to design complex in silico experiments. MS-Analyzer is a specialization of PROTEUS for the integrated preprocessing, management and data mining analysis of MS proteomics data. 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 (see Figure 3).

1. **MS proteomic data acquisition** loads MS raw spectra produced by different kind of Mass Spectrometers (e.g. MALDI-TOF, LC) and stores them in a Raw Spectra Repository (RSR).

2. **MS proteomic data preprocessing** loads MS raw spectra (directly from mass spectrometer or from the RSR), and applies the preprocessing techniques described before, some of which can be optional, and stores data in an efficient manner into a Preprocessed Spectra Repository (PSR). Preprocessing can be applied to one spectrum or contemporarily to many spectra. Preprocessing techniques take into account semantics of data.
3. **MS proteomic data preparation** loads preprocessed spectra and prepare them to be given in input to different kind of data mining tools. E.g. Weka (www.cs.waikato.ac.nz/ml/index.html) requires that MS spectra being organized in a unique file containing a metadata header named ARFF. Preparation techniques take into account semantics of data mining software, but in general do not take into account semantics of MS spectra, that has been considered in the preprocessing phase. Data ready to be analyzed by data mining software is also stored in a file-based repository, the *Prepared Preprocessed Spectra Repository* (PPSR), for further analysis.

4. **Data Mining analysis** allows to select different data mining tasks (e.g. classification, clustering, pattern analysis), and the corresponding data mining tools (e.g. Q5, C5, K-means, etc.), then it allows to load preprocessed and prepared MS spectra and to give them as input to the chosen data mining tools, producing knowledge models.

5. **Data Visualization and/or Visual Data Mining**: this phase, that can be complementary or alternative to the Data Mining one, allows advanced analysis techniques based on interactive visualization of data.

A main function of MS-Analyzer is the management of spectra. The *MS proteomics data storage* implements basic data management functions. It comprises the RSR, PSR and PPSR repositories, containing respectively raw spectra as text files, preprocessed spectra as binary files and prepared spectra as text files. For prepared spectra we currently support the ARFF format, required by Weka.

### 3.1. Using MS-Analyzer in a Grid scenario

Mass spectrometry-based proteomics may require the collection and analysis of data produced in remote laboratories. The collection, storage and analysis of huge mass spectra can leverage the computational power of Grids, that offer efficient data transfer primitives (e.g. GridFTP), effective management of large data stores (e.g. replica management), and the computational power required by data mining algorithms. Our university has reported the identification of a recurrent mutation (5083del19) in the Breast Cancer (BRCA1) gene in a patient population with breast cancer from the Calabria region [1]. So we consider a mass spectrometry-based proteomics application for the characterization of inherited breast cancer as case study. The scenario we envision consists of a set of distributed proteomics facilities that produce spectra either belonging to the same experiment...
(e.g. the known population affected by the 5083del19 BRCA1-gene mutation), or to different experiments, and leverage the specialized services of MS-Analyzer through the Grid. In fact, often university hospitals conducting genomics and proteomics research do not have the storage and computational capability to face the huge amount of produced data, moreover often data produced by different centers need to be collectively analyzed. In such a scenario, the produced raw spectra (some basic preprocessing can directly be done at the proteomics facilities) are collected by using the data transfer capabilities of Grid and stored in a central Grid node hosting MS-Analyzer (see Figure 3). Here, more advanced preprocessing and spectra preparation are conducted on all spectra to be analyzed. At this point comparative analysis can be conducted by using different data mining tools and approaches in parallel, using the allocation and scheduling functions of Grid. Built knowledge models are then presented to the requesting scientists.

We currently designed the overall architecture and fully implemented the preprocessing tools described so far. In particular, MS-Analyzer provides a simple workflow editor where user can choose the preprocessing tools, can compose them by using simple drag&drop functions, and finally can execute the designed preprocessing functions. Using MS-Analyzer, the user can design different analysis experiments where preprocessing is a new design variable, whose effectiveness can be evaluated not only in terms of memory occupations and performance (e.g. binned vs not-binned spectra), but also in terms of data analysis quality (e.g. predictivity of data mining classification with respect to preprocessing type).

4. Conclusion and future works

In this paper we presented a survey of preprocessing techniques for MS proteomics data 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. Future work will regard the implementation of MS-Analyzer functions as Grid services, and the evaluation of the platform by considering different combination of preprocessing and data mining techniques.

References


