Analysis of Mass Spectrometry data: Significance Analysis of Microarrays for SELDI-MS Data in Proteomics

Sarra HAMZAOU1, Smail BOUZERGANE1, Tiratha Raj SINGH2, Hassan BADIR1, Ahmed MOUSSA1

1 LabTIC laboratory, ENSA, University Abdelmalek Essaadi, Tangier, Morocco.
2 Department of Biotechnology and Bioinformatics, Jaypee University of Information and Technology, Solan, H.P., India.

ABSTRACT

Mass Spectrometry (MS) has arguably become the core technology in proteomics. MALDI and SELDI-TOF techniques enable the study of biological fluids, e.g., human blood. Analysis of these samples can lead to discovering new biomarkers which can ease the diagnostic and prognostic of several diseases, e.g., various cancers. In this work, we focus on MS data from SELDI-TOF experiments. We begin with a preprocessing step in order to remove noises due to the acquisition process of the data. Then, we apply the differential analysis to a SELDI-MS data using the Significance Analysis of Microarray (SAM) method implemented in Matlab. Results using the SAM method are compared with those obtained by the conventional t-test and Analysis Of Variance (ANOVA) in order to evaluate its efficacy and its performance. As a result, we demonstrate that the SAM method can be adapted for effective significance analysis of SELDI-MS data. It is deemed powerful and provides better results that to the others. An easy-to-use application is developed with Matlab for mass spectrometry data analysis from raw spectra to differential analysis, including the SAM method.

1. INTRODUCTION

The Proteomics is a rapidly developing field among other omics disciplines, focusing on large biological datasets. This level of study requires the use of appropriate high throughput methods, as is also the case in transcriptomic study. As such, Mass Spectrometry (MS) offers an interesting insight on biological samples containing large numbers of proteins. Plasma represents a challenging sample: its analysis can lead to discovering new biomarkers, thereby offering new diagnostic or prognostic tools.

Cancer research has been an active field of research using fluid proteomics: Researchers are interested in identifying a protein signature of a specific cancer in serum or plasma samples, easily accessible in contrast to most biopsies. Apionering study[1] brought to light the use of the Surface Enhanced Laser Desorption – Time of Flight (SELDI-TOF) technology as a powerful method to detect ovarian cancers. Although this study later was criticized[2], numerous other studies followed, extending MS use to a wide range of diseases. It appears that MS methods in this field still require improvement and validation[3, 4, 5] but offer a powerful approach for cancer research as well as other domains in medicine and biology. Another very similar technology, Matrix Assisted Laser Desorption Ionization (MALDI-TOF) showed the same promises and led to a significant number of studies. We focus on the use of blood MS (SELDI) in the field of cancer diagnosis and prognosis, essentially as a biomarker discovery tool but also as predictive tool.

A mass spectrometer generates multiple ions from the sample under investigation; it then separates them according to their specific mass-to-charge ratio (m/z), and then records the relative abundance of each ion.
MS DATA PREPROCESSING

All steps described in this section aim at removing all forms of noise and artifacts introduced in the data by specific properties of the method. Note that in our preprocessing mass spectra approach, we looked for certain simplicity, limiting the number of steps and focusing on procedures simple in their design, rapid in execution to be able to use them on large-scale and which do not destroy the signal of the mass spectra. Tables and Figures are presented center, as shown below and cited in the manuscript.

The essential of methods have made subject of publications [8, 9] and a good review of the methods of preprocessing for mass spectra is found in [10]. All these steps are related to the following equation:

\[ f(t) = B(t) + N \cdot S(t) + \epsilon(t) \]  (1)

- \( f(t) \): corresponds to the observed signal.
- \( S(t) \): is the true signal.
- \( B(t) \): a baseline term.
- \( t \): refers to time of flight values (which can be easily converted to m/z).
- \( N \): normalization factor, it is very important to correct the intensities of the peaks of \( S(t) \).
- \( \epsilon(t) \): is a random noise.

The aim of pre-processing step is to isolate \( S(t) \) to be able to make valid comparison between samples and ultimately identify biomarkers. Denoising filters out the \( \epsilon(t) \) component while baseline correction (section 2.1) aims at removing the \( B(t) \) component from the signal. Normalization (section 2.2) deals with the intensity scale. Peak detection (section 2.3) aims at extracting of the spectrum the list of positions where there peaks and peak alignment (section 2.4) is concerned with setting up a common time scale for all spectra.

2.1. Noise Filtering and Baseline Correction

The random noise component of the observed signal, \( \epsilon(t) \), is mainly of electronic origin. A simple way of reducing the noise is to perform smoothing of the spectrum by using a sliding window and replacing the intensity values in the window by a single value based on all of the values in the window, for example their weighted average. Fourier transform, smoothing splines, and wavelets are among more sophisticated approaches to noise reduction. For example, Coombes et al. use the Undecimated Discrete Wavelet Transform (UDWT) for denoising SELDI spectra[11].

The baseline offset of the spectrum, \( B(t) \), is attributable mainly to chemical noise generated by the molecules of the energy absorbing matrix. For each spectrum, this offset line can be approximated and subtracted from the raw spectrum intensities. Usually, the baseline is highest at the low range of m/z values and exponentially decreases with the increase in m/z values. Popular methods of the baseline approximation fit a polynomial or exponential function to the local minima of the spectrum. Other approaches may be based on fast Fourier transform or wavelets. The authors proposed a method that combines baseline correction with the peak detection step. Instead of explicit fitting of the baseline for the entire spectrum, they defined the baseline locally, for each identified peak, as the local minimum in the fixed-width window containing the peak. The baseline adjusted height of the peak is calculated simply as the difference between the local maximum and local minimum. However, in situations when peaks overlap, the local minimum may be significantly higher than the real base line and the height of the peak may be underestimated[12].

After denoising and suppression of background noise, the estimated signal is:
Press the done. Peaks are clustered by their m/zs appear as local maxima (i.e., peak-licability of SAM method, it has the same form as the more general dies. For example, -relative, we simply divide each denoised and baseline such data the irss /ins, before looking for biological interpretation of intensity at each molecular ma-

Different feature selection methods may be utilized for proteomics data[15].

3.2. Peak Detection

Usually, peptide signals appear as local maxima (i.e., peaks) in MS spectra. However, detecting these signals still remains challenging due to the following reasons:

(1) Some peptides with low abundance may be buried by noise, causing high false positive rate of peak detection.

(2) The chemical, ionization, and electronic noise often result in a decreasing curve in the background of MALDI/SELDIMS data, which is referred to as baseline[13]. The existence of baseline produces strong bias in peak detection. It is desirable to remove baseline before peak detection.

2.4. Peak Alignment

Due to measurement errors, peaks corresponding to the same protein may, in different spectra, be associated with different m/z values. The m/z errors are usually estimated as not greater than 0.3 percent of the m/z values. Peaks with their m/z values within such m/z error intervals should be aligned across spectra and treated as the same peak. For example, the identified peaks may be first sorted by their intensity values or their signal-to-noise ratios. Then, starting from the most prominent peaks, we may match peaks from different spectra if their m/z values differ less than an appropriate m/z error interval. Peak alignment based on hierarchical clustering of peaks from all considered spectra has also been done. Peaks are clustered by their m/z values, with constraints based on them/z measurement error rate, merr. The distance between two m/z values (or two clusters of m/z values) is calculated in relation to their mean, so it can be directly compared to the relative measure of the m/z error. Although the centroid linkage distance is used by the authors to identify clusters that are candidates for merging, two clusters may be merged only if their complete linkage distance is below 2xmerr, the doubled mass measurement error[12].

3. DIFFERENTIAL ANALYSIS

For unbiased technologies, such as SELDI-TOF or MALDI-TOF mass spectrometry, the variables represent the identified spectra peaks, which hypothetically represent proteins or peptides. While it would be more precise to call such data the peak expression matrix, it has the same form as the more general protein expression matrix and can be analyzed in exactly the same way. Only after an optimal biomarker is identified, we have to remember the necessity of matching its peaks to proteins, before looking for biological interpretation of the biomarker.

The protein expression matrix has the same form as the gene expression matrix. Furthermore, the goals of protein expression studies are basically the same as the goals of gene expression studies. For example, biomarker discovery aims at the identification of small sets of proteins (or m/z peaks) whose joint expression pattern can significantly separate differentiated classes.

3.1. MS Datasets

In this paper, we used two Low Resolution SELDI-TOF Datasets downloaded from the Clinical Proteomics Program Databank website[14]. The first data, Ovarian Dataset 4-3-02, includes 100 unaffected women (controls) and 100 patients who later developed ovarian cancer. The second data set, Ovarian Dataset 8-7-02 consists of serum profiles of 162 subjects with ovarian cancer and 91 non-cancer control subjects. The raw spectral data of each sample contains the relative amplitude of the intensity at each molecular mass / charge (M/Z) identity. There are total 15154 M/Z identities. In the follow, we explore the applicability of SAM method to our two datasets to identify proteins differentially expressed.

3.2. Methods Used for Identifying Differentially Expressed Proteins

The univariate exploratory analysis is the common first step in analyzing protein expression data. Different feature selection methods may be utilized for proteomics data[15].

Usually, a t-test[6] or an ANOVA test[7] are used to identify differentially expressed variables. The variables may be ordered by p-values representing the significance of their differential expression. Due to a large number of simultaneous univariate tests (equal to the number of variables), the p-values have to be corrected for multiple testing by the False discovery rate (FDR) method [16, 17].
In general, a \( t \)-test is used to evaluate whether the means of control and experiment groups are statistically different. The \( p \)-value is the ratio between the difference of group means and the variability of groups. One of the classical \( t \)-tests may be used to identify differentially expressed proteins in two-class experiments. Depending on whether we can or cannot assume equal variances of protein expression in both differentiated populations, we will use either the \( t \)-test for equal variances or the \( t \)-test for unequal variances.

ANOVA can be used in a univariate way to test whether the mean expression levels of a particular protein differ significantly between the \( J \) populations, where \( J > 2 \). It is based on the ratio of the variance between classes to the variance within classes and is used to decide whether we can reject the null hypothesis of no difference between the \( J \) population means.

SAM is one of the widely accepted methods for such analysis in DNA microarray [6]. In the following, we explore the applicability of the SAM method to SELDI proteomics data analysis. SAM was originally developed for microarray analysis by Tusher et al. [18]. In this study, we developed the SAM method with Matlab to be adapted for effective significance analysis of proteomic data. SAM assigns a score to each protein on the basis of change in protein expression relative to the standard deviation of repeated measurements. For proteins with scores greater than an adjustable threshold, SAM uses permutations of repeated measurements to estimate the percentage of proteins identified by chance, the false discovery rate (FDR) [19].

3.3. “SAM_MSDA”: Workflow for MS Data Analysis

Many software applications have been developed to analyze mass spectrometry data such as mspire [20], XCMS [21], and MSDaPI [22]. In this work, we developed an easy-to-use application entitled “SAM_MSDA” with Matlab.

This application provides a set of tools for the manipulation and analysis of proteomic data. It is very intuitive to use making it an ideal tool for the biologist. Indeed, “SAM_MSDA” is a platform that offers a “constellation” of tools to analyze, manipulate and visualize proteomic data, without the need for programming knowledge. The user can perform four types of operations:

- Importing Mass Spectrometry data (raw data),
- Preprocessing MS data to remove all forms of noise and artifacts introduced in the data, Differential analysis: using \( t \)-test, ANOVA and SAM method.
- Visualization of data and results.

![Workflow for MS data analysis from raw spectra to differential analysis.](image)

Figure 1. Workflow for MS data analysis from raw spectra to differential analysis. The rectangular boxes represent a processing step, oval boxes describe the type of data obtained when changes.
The user cannot proceed to the next step until the validation of the previous step according to the sequence of workflow described in Figure 1 is being performed. In fact, before importing raw data the “Preprocessing” tab is useless so the user cannot proceed to the preprocessing step after importing the data successfully. Similarly, the “Differential analysis” tab is useless until validation of preprocessing MS data and extraction of expression matrix. Thus, the user can easily follow the sequence of treatments without requiring knowledge of data analysis.

Figure 2 shows the “Preprocessing” tab which allows performing the main methods of MS data preprocessing.

Figure 3 shows the “Differential analysis” tab which allows identifying proteins differentially expressed using the t-test ANOVA test and SAM method we have developed to make it suitable for proteomic data.

Figure 6 shows the results from the differential analysis, using a SELDI data set 4–3–02 (Petricoin et al. 2002a). Of 15154 peptides, 230 and 229 were significant in differential expression by the t-test and ANOVA test ($p < 0.05$), respectively, and 140 were significant in differential expression by SAM with $\Delta = 0.76$ cut-off.

Figure 3. “Differential analysis” tab, including the Significance Analysis of Microarrays (SAM) method.

3.4. Results and Discussion

Our goal in this study is to evaluate the efficacy and the performance of the SAM method in comparison with the t-test and ANOVA test. The validity of the SAM method compared with the t-test is determined by sensitivity and specificity. These two are components that measure the inherent validity of a test. Receiver Operating Characteristic (ROC) graphs are a useful technique for organizing classifiers and visualizing their performance[23]. In a ROC curve the true positive rate (Sensitivity) is plotted in function of the false positive rate (100-Specificity) for different cut-off points. Each point on the ROC curve represents a sensitivity/specificity pair corresponding to a particular decision threshold[24]. A test with perfect discrimination (no overlap in the distributions) has a ROC curve that passes through the upper left corner (100% sensitivity, 100% specificity). Therefore the closer the ROC curve is to the upper left corner, the higher the overall accuracy of the test. Figure 4 shows ROC curves for SAM method, conventional t-test and ANOVA test, using a SELDI data set (4-3-02) [1].

Considering the area under the ROC curve that is computed using cross-validation[25], SAM test is better than t-test and ANOVA and has good validity as the curve appears importantly. Thus, the SAM test is deemed powerful and the most adapted for identify proteins differentially expressed. The value of this test providing the best sensitivity for a number of false positives as low as possible is equal to approximately 0.76 (the closest to the upper left corner points): it is 140 proteins differentially expressed among 15154 peptides. Figure 6 shows these results using our application developed with Matlab.
Figure 4. ROC curves for SAM method, conventional t-test and ANOVA test (SELDI data set: Petricoin and al. “4-3-02”).

Figure 5 represents ROC curves for SAM method, conventional t-test and ANOVA test, using a SELDI data set (8-7-02) [1] that consists of 162 samples from ovarian cancer patients and 91 samples from individuals without cancer. This figure shows that we obtained the same results as the previous. In fact, the SAM method appears better than t-test and ANOVA. We can say that the SAM test is deemed powerful and can be adapted for effective significance analysis of proteomic data.

Figure 5. ROC curves for SAM method, conventional t-test and ANOVA test (SELDI data set: Petricoin et al. "8-7-02").

Figure 6. Results obtained by applying t-test, ANOVA and SAM method using the Ovarian Dataset 4-3-02. The value of SAM test to use, having best sensitivity for a number of false positives as low as possible is equal to approximately 0.78 cut-off. Of 15154 peptides, 261 were significant in differential expression by the t-test and ANOVA test ($p < 0.05$) and 143 were significant in differential expression by SAM with $\Delta = 0.76$ cut-off, using
SELDI data set: Petricoin et al. (2002) (8-7-02). Therefore, we can say that the SAM method gives the same results whatever the data used and appears better than the t-test and ANOVA test.

4. CONCLUSION

Protein expression analysis is likely to become one of the main sources of new biomarkers for personalized medicine, which may include early medical diagnosis, tailoring therapy selection to the prediction of individual response to available treatment modalities, and assessing treatment progression and drug efficacy. Multivariate approaches to feature selection coupled with large and good quality training data sets will lead to the identification of parsimonious proteomic biomarkers representing multi-protein expression patterns characteristic for the differentiated classes. Preprocessing of raw proteomic data depends on the technology that generated the data. Nevertheless, after low-level preprocessing we can represent any protein expression data in the form of a protein expression matrix. The variables of this matrix can represent proteins either directly (as in the case of antibody microarrays) or indirectly (for instance, SELDI-TOF m/z variables). If the goal of our analysis is biomarker discovery, we try to identify a small set of variables whose joint expression pattern can significantly separate the differentiated classes.

We demonstrate that the SAM method can be adapted for effective significance analysis of proteomic data (Especially, SELDI data sets). It provides much richer information about the protein differential profiles. This result is obtained using ROC curve, it is amethod of choice for the study of the clinical efficacy of a bioassay. Indeed, comparison of the areas under the curve of the tree tests (t-test, ANOVA and SAM) allows us to assess and classify the diagnostic performance of these three tests. The ROC curves also allowed us to determine the threshold value of optimal SAM test.

The development of “SAM_MSDA” application makes statistical analysis of mass spectrometry data simpler and it is anticipated that the developed method will provide efficient contribution to the analysis of protein expression data.

In perspectives, we want to develop “SAM_MSDA” to make it a web application allowing users/proteomists to realize the entire workflow of mass spectrometry data analysis from the importation of proteomic data to the differential analysis of mass spectrometry.

REFERENCES


