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

Article *in* International Journal for Computational Biology · March 2015

DOI: 10.34040/IJCB.4.1.2015.46

CITATIONS
0

READS
232

4 authors, including:

Badir Hassan
Abdelmalek Essaâdi University
123 PUBLICATIONS **325** CITATIONS

SEE PROFILE

Ahmed Moussa
National School of Applied Scienes, Tangier, Morocco
88 PUBLICATIONS **461** CITATIONS

SEE PROFILE

International Journal for Computational Biology (IJCB)
Vol.4, No.1, April 2015, pp. 05~12
ISSN: 2278-8115 ❒5

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

Sarra HAMZAOUI[1], Smail BOUZERGANE[1], Tiratha Raj SINGH[2], Hassan BADIR[1], Ahmed MOUSSA[1]

[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.

Article Info

Article history:

Received Nov 22nd, 2014
Revised Dec 20th, 2014
Accepted Jan 4th, 2015

Keyword:

Mass Spectrometry data
Differential analysis
SAM
Comparison of features selection

ABSTRACT

Mass Spectrometry (MS) has arguably become thecore technology in proteomics. MALDI and SELDI-TOFtechniques enable the study biological fluids, e.g. human blood.Analysis of these samples can lead to discover new biomarkerswhich can ease the diagnostic and prognostic of several diseases,e.g. various cancers. In this work, we focus on MS data fromSELDI-TOF experiments. We begin with a preprocessing step inorder 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) methodimplemented in Matlab. Results using the SAM method arecompared with those obtained by the conventional *t*-test andAnalysis Of Variance (ANOVA) in order to evaluate its efficacyand its performance. As a result, we demonstrate that the SAMmethod can be adapted for effective significance analysis ofSELDI-MS data. It is deemed powerful and provides betterresults that totes. An easy-to-use application is developed withMatlab for mass spectrometry data analysis from raw spectra todifferential analysis, including the SAM method.

Corresponding Author:

Ahmed MOUSSA,
LabTIC laboratory, ENSA,
University Abdelmalek Essaadi,
Tangier, Morocco.
Email: amoussa@uae.ac.ma

How to Cite:

Sarra HAMZAOUI *et. al.*Analysis of Mass Spectrometry Data: Significance Analysis of Microarrays for SELDI-MS Data in Proteomics. IJCB. 2015; Volume 4 (Issue 1): Page 05-12.

1. INTRODUCTION

The Proteomics is a rapidly developing field among other omics disciplines, focusing on large biological datasets. Thislevel 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 onbiological samples containing large numbers of proteins.Plasma represents a challenging sample: its analysis can leadto discovering new biomarkers, thereby offering new diagnostic or prognostic tools.

Cancer research has been an active field of research usingfluid proteomics: researchers are interested in identifying aprotein signature of a specific cancer in serum or plasmasamples, easily accessible in contract to most biopsies. Apioneering study[1] brought to light the use of the SurfaceEnhanced Laser Desorption – Time of Flight (SELDI-TOF)technology as a powerful method to detect ovarian cancers.Although this study later was criticized[2], numerous otherstudies followed, extending MS use to a wide range ofdiseases. It appears that MS methods in this field still requireimprovement and validation [3, 4, 5] but offer a powerfulapproach for cancer research as well as other domains inmedicine and biology. Another very similar technology,Matrix Assisted Laser Desorption Ionization (MALDI-TOF)showed the same promises and led to a significant numberofstudies. We focus on the use of blood MS (SELDI) in the fieldof cancer diagnosis and prognosis, essentially as a biomarkerdiscovery tool but also as predictive tool.

A mass spectrometer generates multiple ions from thesample under investigation; it then separates them accordingto their specific mass-to-charge ratio (m/z), and then recordsthe relative abundance of each ion

type. The spectra producedare noisy and functional;they must undergo *pre-processingsteps*, described in *section 2*, to produce a coherent structureof information so that it can be exploited by statistical analysismethods.

Analysis of proteomic data involves testing simultaneouslythe expression of a large number of proteins between two ormore conditions. To this end, we explore the use of *the SAM*method[6] for analysis of MS data. We used two SELDIdataset from[1] which includes samples of a female populationwith ovarian cancer and control samples of unaffected women.We report the results of quantifying the protein relativeabundance between the controls (normal) and the ovariancancers, with an emphasis on significance analysis of proteindifferential expression using the *SAM* method in comparisonwith *t-test*[6] *and ANOVA* test[7], as described in *section 3*.In *sub-section 3.3*, we present very briefly the mainfunctionalities of the application "SAM_MSDA" for analysisof MS data, developed with Matlab, which comes to meet theneeds of bioinformatics researchers. In fact, three main stepsare identified with specific sub-steps in this application: i)importing raw spectra; ii) preprocessing MS data; iii)Differential analysis, using *SAM* method implemented inMatlab.

2. MS DATA PREPROCESSING

All steps described in this section aim at removing all forms of noise and artifacts introduced in the data by specificproperties 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 theirdesign, rapid in execution to be able to use them on large-scaleand 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 ofpublications [8, 9]and a good review of the methods ofpreprocessing for mass spectra is found in[10]. All these stepsare related to the following equation:

$$f(t) = B(t) + \mathrm{N} * S(t) + \epsilon(t) \qquad (1)$$

- ***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).*
- ***ϵ(t) :*** is a random noise.

The aim of pre-processing step is to isolate *S(t)* to be ableto make valid comparison between samples and ultimatelyidentify biomarkers. Denoising filters out the *ϵ(t)*componentwhile baselinecorrection(*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 wherethere peaks and peak alignment (*section 2.4*) is concernedwith setting up a common time scale for all spectra.

2.1. Noise Filtering and Baseline Correction

The random noise component of the observed signal, *ϵ(t)*, is mainly of electronicorigin. A simple way of reducing thenoise is to perform smoothing of the spectrumby using asliding window and replacing the intensity values in thewindow by a single value based on all of the values in thewindow, for example their weighted average. Fouriertransform, smoothing splines, and wavelets are among moresophisticated 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 attributablemainly to chemical noise generated by the molecules of theenergy absorbing matrix. For each spectrum, this offsetline 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 decreaseswith the increase in m/z values. Popular methods of the baseline approximation fitare polynomial or exponential functions to the local minima of the spectrum. Otherapproaches may be based on fast Fourier transform or wavelets. The authors proposed a method that combines baselinecorrection 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 peakis calculated simply as the differencebetween the localmaximum 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, theestimated signal is:

$$f(t) = \mathrm{N} * S(t) \qquad (2)$$

2.2. Normalization

To remove the normalization factor N in *equation (2)*, we simply divide each denoised and baseline corrected spectra by its area under the curve (AUC) which is a standard normalizing choice in MS spectra analysis.

2.3. Peak Detection

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

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

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

2.4. Peak Alignment

Due to measurement errors, peaks corresponding to the same protein may, in differentspectra, be associated with different m/z values. The m/z errors are usually estimatedasnot 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 ortheirsignal-to-noise ratios. Then, starting from the mostprominent peaks, we may matchpeaks from different spectraif their m/z values differ less than an appropriate m/zerrorinterval. Peak alignment based on hierarchicalclustering ofpeaks from all considered spectra has also been done. Peaksare clustered by their m/z values, with constraints based on them/z measurement error rate, **merr**. The distance between twom/z values (or two clusters ofm/z values) is calculated inrelation to their mean, so it can be directly compared to therelative measure of the m/z error. Although the centroidlinkage distance is used by the authors to identify clusters thatare candidates for merging, two clusters may be merged onlyif their complete linkage distance is below 2xmerr, thedoubledmass 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 representproteins or peptides. While it would be more precise to callsuch data the *peak expression matrix*, it has the same form asthe more general *protein expression matrix* and can beanalyzed in exactly the same way. Only after an optimal biomarker is identified, we have to remember the necessity ofmatching its peaks to proteins, before looking for biologicalinterpretation of the biomarker.

The protein expression matrix has the same form as the gene expression matrix. Furthermore, the goals of proteinexpression studies are basically the same as the goals of geneexpression 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 ProgramDatabank* 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 featureselection methods may be utilized for proteomics data[15].

Usually, a *t-test*[6] or an *ANOVA* test[7] are used to identifydifferentially 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 Falsediscovery rate (FDR) method [16, 17].

In general, a *t-test* is used to evaluate whether the means ofcontrol 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 beused to identify differentially expressed proteins in two-classexperiments. Depending on whether we can or cannot assumeequal variances of protein expression in both differentiatedpopulations, we will use either the *t*-test for equal variances orthe *t*-test for unequal variances.

ANOVA can be used in a univariate way to test whetherthe 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 classesand 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 suchanalysis 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 changein protein expression relative to the standard deviation ofrepeated measurements. For proteins with scores greater thanan adjustable threshold, SAM uses permutations of therepeated 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 manipulationand 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.

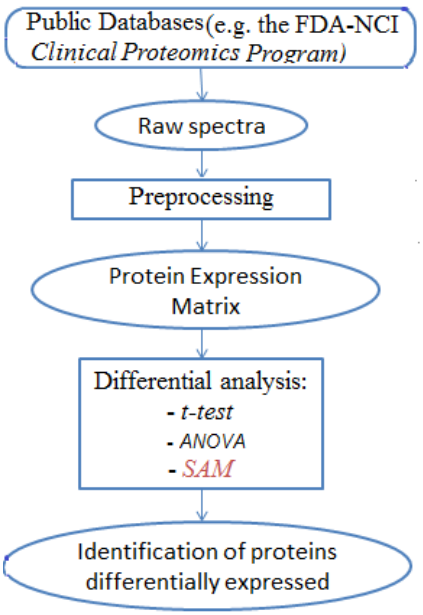

Figure 1. Workflow for MS data analysis from raw spectra to differential analysis.Therectangular boxesrepresenta processing step, oval boxes describe the type of data obtained when changes.

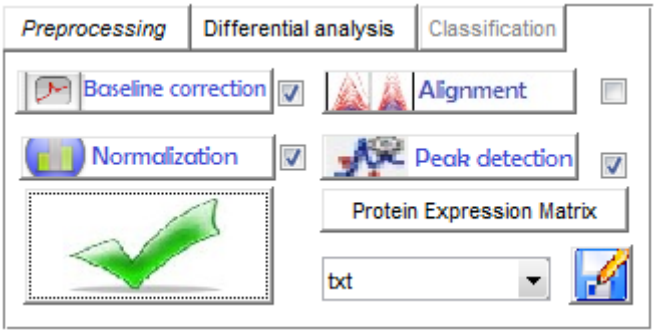

Figure 2. "Preprocessing" tab which allows performing the main methods of MS data preprocessing.

The user cannot proceed to the next step until thevalidation of the previous step according to the sequence of workflow described in Figure 1I 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 3shows the "Differential analysis" tab which allowsidentifying proteins differentially expressed using the *t*-test ANOVA test and SAM method we have developed to make it suitable for proteomic data.

Figure 6shows the results from the differential analysis, usinga 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.

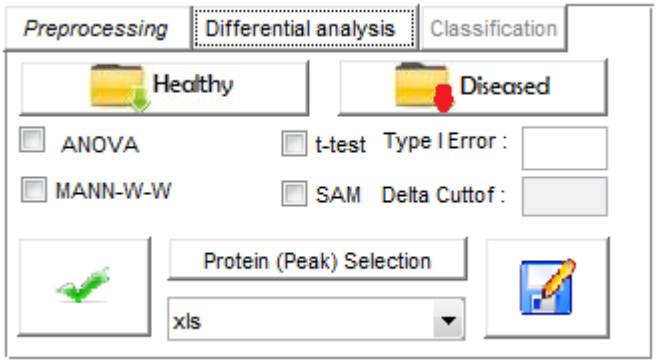

Figure 3. "Differential analysis" tab , including the Significance Analysis of Microarrays (SAM) method.

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 comparedwith the *t*-test is determined by sensitivity and specificity.These two are components that measure the inherent validityof a test. Receiver Operating Characteristics (ROC) graphs area useful technique for organizing classifiers and visualizingtheir 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 theROC curve represents a sensitivity/specificity paircorresponding to a particular decision threshold[24]. A testwith perfect discrimination (no overlap in the twodistributions) has a ROC curve that passes through the upperleft corner (100% sensitivity, 100% specificity). Therefore thecloser the ROC curve is to the upper left corner, the higher theoverall accuracy of the test. Figure 4shows ROC curves for SAMmethod, conventional *t*-test and ANOVA test, using a SELDIdata set (4-3-02) [1].

Considering the area under the ROC curve that iscomputed using cross-validation[25], SAM test is better than *t*-test and ANOVA and has good validity as the curve appearsmore 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 ourapplication developed with Matlab.

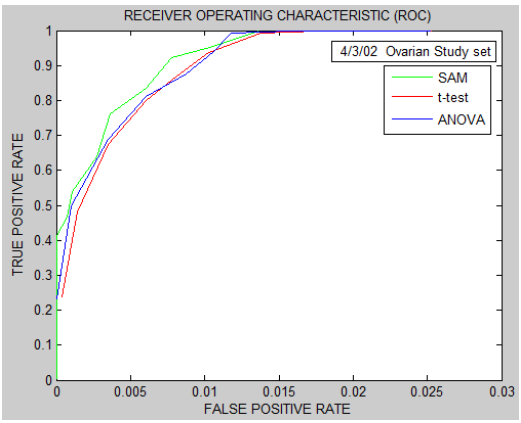

Figure 4. ROC curves for SAM method, conventional t-test and ANOVA test (SELDI data set: Petricoin and al."4-3-02").

Figure 5represents 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. Thisfigure shows thatwe obtainedthe same results asthe previous. In fact, theSAM method appears better than *t*-test and ANOVA. We can say that the SAM test is deemed powerful and can be adapted for effective significanceanalysis of proteomic data.

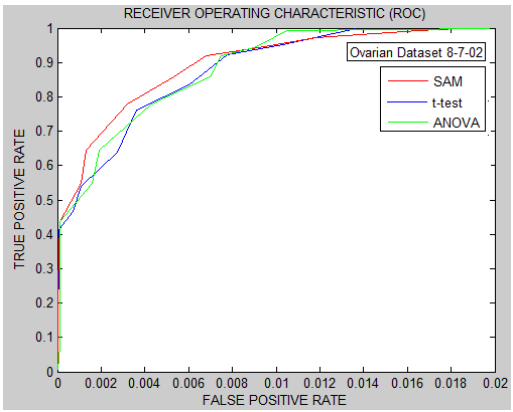

Figure 5.ROC curves for SAM method, conventional t-test and ANOVA test (SELDI data set: Petricoin et al."8-7-02").

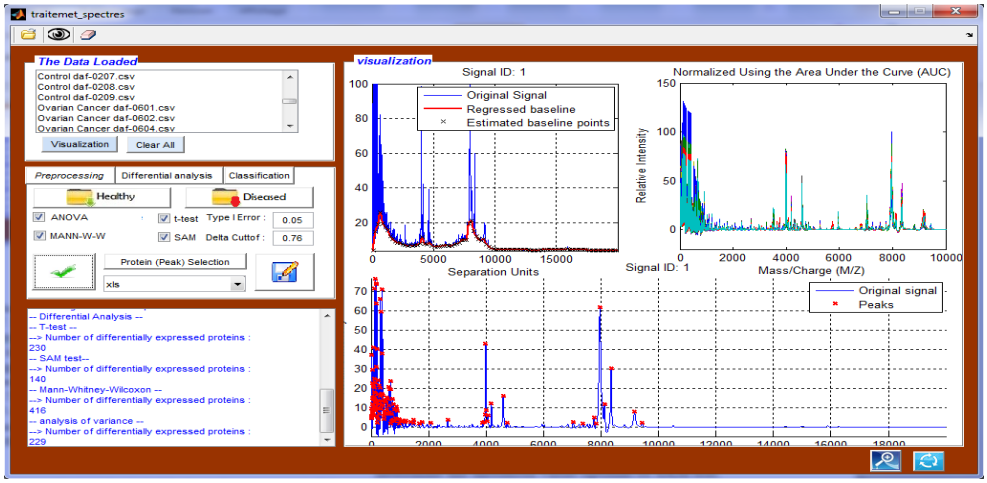

Figure 6. Results obtained by applying t-test, ANOVA and SAM method using the *Ovarian Dataset 4-3-02.* The value of SAM testto use, having best sensitivity for a number of false positives as low as possible is equal to approximately 0.78 cut-off. Of 15154 peptides, 261were significant in differential expressionby the *t*-test and ANOVA test ($p < 0.05$) and 143 were significant in differential expression by SAMwith $\Delta = 0.76$ cut-off, usinga

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. Thevariablesof this matrix can represent proteins either directly (as in the case of antibody microarrays) or indirectly (for instance, SELDI-TOF m/z variables).If thegoal 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 significanceanalysis of proteomic data (Especially, SELDI data sets). It provides much richer information about the protein differentialprofiles. This result is obtained using ROC curve, it is amethod of choice forthe studyoftheclinical efficacy of a bioassay.Indeed, comparison of the areas under the curve of the tree tests (t-test, ANOVA and SAM) allowsusto assessand classifythe diagnostic performanceof these three tests. The ROC curvehas alsoallowed us to determinethe threshold valueoptimal of SAMtest.

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

[1] Emanuel F Petricoin, Ali M Ardekani, Ben A Hitt, Peter J Levine,Vincent A Fusaro, Seth M Steinberg, Cordon B Mills, Charles Simone, David A Fichman, Elise C Kohn, and Lance A Liotta. "*Use ofproteomic patterns in serum to identify ovarian cancer*". Lancet, 359(9306):572-577, 2002.

[2] Keith A Baggerly, Jeffrey S morris, and Kevin RCoombes.*"Reproducibility of seldi-tof protein patterns in serum:comparing datasets from different experiments".* Bioinformatics, 20(5):777-785, 2004.

[3] Kevin R Coombes, Jeffrey S Morris, Jianhua Hu, Sarah R Edmonson, andKeith A Baggerly. *"Serum proteomics profiling-a young technologybegins to mature"*. Nat Biotechnol, 23(3): 391-292, 2005.

[4]Eleftherios P Diamandis. "*Mass spectrometry as a diagnostic and acancer biomarker discovery tool: opportunities and potentiallimitations"*. Mol Cell Proteomics, 3(4): 376-378, 2004.

[5] Jianhua Hu, Kevin R Coombes, Jeffrey S Morris, and Keith ABaggerly.*"The importance of experimental design in proteomic massspectrometry experiments: some cautionary tales"*. Brief FunctGenomic Proteomic, 3(4): 322-331, 2005.

[6] Bryan AP Roxas and Qingbo Li. *"Significance analysis of microarray for relative quantitation of LC/MS data in proteomics".* BMCBioinformatics, 9:187, 2008.

[7] Ann L. Oberg,Douglas W. Mahoney, Jeanette E. Eckel-Passow,Christopher J. Malone, Russell D. Wolfinger, Elizabeth G. Hill, LeslieT. Cooper, Oyere K. Onuma, Craig Spiro, Terry M. Therneau, and H.Robert Bergen, III. *"Statistical analysis of relative labeled massspectrometry data from complex samples using ANOVA".* J ProteomeRes, 7(1): 225–233, 2008.

[8] J. Prados, A. Kalousis and M. Hilario. *"On preprocessing of seldi-ms data and its evaluation*".In Proceedings of the 19th IEEE Symposium onComputer-Based Medical Systems, pages 953-958, 2006.

[9] J. Prados, A. Kalousis, L. Allard, O. Carrette, J. C. Sanchez, and M. Hilario. *" Mining mass-spectra for diagnosis and biomarker discoveryof cerebral accidents".* Proteomics, 4:2320–2332, 2004.

[10] M. Hilario, A. Kalousis, C. Pellegrini, and M. Muller.*"Processing and classification of protein mass spectra".* Mass Spectrometry Reviews,25:409 – 449, 2006.

[11] Kevin R. Coombes, SpiridonTsavachidis, Jeffrey S. Morris, Keith A. Baggerly, Mien-Chie Hung and Henry M. Kuerer. *"Improved PeakDetection and Quantification of Mass Spectrometry Data Acquiredfrom Surface-Enhanced Laser Desorption and Ionization by DenoisingSpectra with the Undecimated Discrete Wavelet Transform".*
Proteomics, 5(16):4107-17, 2005.

[12] ThomasJouve, DelphineMaucort-Boulch, Patrick Ducoroy and Pascal Roy.*"Local features based methods in mass spectrometry proteomics:a review"* .Bioinformatics. *2009.*

[13] Malyarenko DI, Cooke WE, Adam BL, Malik G, Chen H, Tracy ER, Trosset MW, Sasinowski M, Semmes OJ, Manos DM. *"Enhancementof sensitivity and resolution of surface-enhanced laserdesorption/ionization time-of-flight mass spectrometric records forserum peptides using time-series analysis techniques"*. ClinicalChemistry, 51:65–74, 2005.

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

[15] Christin C, Hoefsloot HC, Smilde AK, Hoekman B, Suits F, Bischoff R, Horvatovich P. *"A Critical Assessment of Feature Selection Methodsfor Biomarker Discovery in Clinical Proteomics".* Mol CellProteomics. 12(1):263-76, 2013.
[16] William Stafford Noble. *"How does multiple testing correction work? ".* Nat Biotechnol. 27(12): 1135–1137, 2009.
[17] Angel P. Diz, Antonio Carvajal-Rodríguez, and David O. F. Skibinski.*"Multiple Hypothesis Testing in Proteomics: A Strategy forExperimental Work".*Mol Cell Proteomics.10(3): M110.004374, 2011.
[18] Tusher VG, Tibshirani R, Chu G. *"Significance analysis of microarray applied to the ionizing radiation response".* ProcNatlAcadSci U S A,98(9):5116-5121, 2001.
[19] Tusher VG, Tibshirani R and Chu G. *" Significance analysis of microarrays applied to the ionizing radiation response".*ProcNatlAcadSci U S. 28;98(18):10515, 2001.
[20] John T. Prince and Edward M. Marcotte. *"mspire: mass spectrometry proteomics in Ruby".*Bioinformatics. 24(23): 2796–2797, 2008.
[21] Colin A. Smith , Elizabeth J. Want , Grace O'Maille, Ruben Abagyan and Gary Siuzdak. *" XCMS: Processing Mass Spectrometry Data forMetabolite Profiling Using Nonlinear Peak Alignment, Matching, andIdentification".*Anal. Chem. , *78 (3), 2006.*
[22] Vagisha Sharma, Jimmy K. Eng,Michael J. MacCoss and Michael Riffle.*"A Mass Spectrometry Proteomics Data Management Platform".*MolCell Proteomics. 11(9): 824–831, 2012.
[23] Tom Fawcett. "*ROC Graphs: Notes and Practical Considerations for Data Mining Researchers".* HP Laboratories Palo Alto.CA 94304,2004.
[24] Faraggi D, Reiser B. *"Estimating of area under the ROC curve"*. Stat Med, 21:3093-3106, 2002.
[25] Richard M. Simon, Jyothi Subramanian, Ming-Chung Li and SupriyaMenezes. *"Using cross-validation to evaluate predictive accuracy ofsurvival risk classifiers based on high-dimensional data".*Brief Bioinform. 12(3): 203–214, 2011.