Journal of
Computer Science & Systems Biology

Belmonte et al., J Comput Sci Syst Biol 2013, 6:5
DOI: 10.4172/jcsb.1000125

Research Article Open Access

SpADS: An R Script for Mass Spectrometry Data Preprocessing before Data Mining

Luca Belmonte[1], Rosanna Spera[1] and Claudio Nicolini[1,2,3,4]*

[1]*Laboratories of Biophysics and Nanobiotechnology, Department of Experimental Medicine, University of Genova, Italy*
[2]*Nanoworld Institute Fondazione EL.B.A. Nicolini, Largo Piero Redaelli 7, Pradalunga (Bg), Italy*
[3]*Virginia G. Piper Center for Personalized Diagnostics, Biodesign Institute, Arizona State University, Tempe, AZ, USA*
[4]*New England Biolabs, Inc, 240 County Road, Ipswich, MA 01938, USA*

Abstract

The recent application of Mass Spectrometry (MS) to Nucleic Acid Programmable Protein Array (NAPPA) technique for proteins identification by non-classical methods leads to the needs of more sophisticated algorithm for peak recognition. NAPPA technique allows for functional proteins to be synthesized *in situ* directly from printed cDNAs but faces the difficulty generated by the presence of master mix and lysate molecules peaks appearing as background in the overall spectra. A wide range of tools are available to analyze proteins conventional mass spectra corresponding to few molecular species. None of them is optimized for background subtraction. Moreover, peak identification is performed by statistical analysis on characteristics peaks and thus background subtraction can alter outcome by erasing characteristic peaks. A first attempt to overcome the so far discussed problem is here discussed. The result of this effort is the development of SpADS: Spectrum Analyzer and Data Set manager-an R script for MS data preprocessing-therein discussed. SpADS provides useful preprocessing functions such binning and peak extractions, as available tools, and provides functions of spectra background subtraction and dataset managing. It is entirely developed in R, thus free of charge. A cluster k means implementation is here used to improve results of SpADS preprocessing on test datasets and on NAPPA expressed proteins.

Keywords: Computational biology; Qualitative data processing; Preprocessing of MS data; Statistical analysis; Validation; Clustering

Introduction

Data generated by mass spectrometer and usually by all kinds of spectrometers are plagued by noise and errors due to different factors that are, roughly, compound preparation, distortion and noise introduced by analysis tools. Other imperfections are due to distortion, peak broadening, saturation, distortion, incorrect calibration and various kinds of contamination. Data cleaning is performed, mainly, in two different steps: better preparation of the compound and application of preprocessing algorithms on the acquired data. In this manuscript we will focus only on the stage of the preprocessing algorithms implementation. The development of an in-house R script - SpADS (Spectrum Analyzer and Data Set manager) - for non classical mass spectrometry data preprocessing and its usage is here discussed. In addition to classical methods of differential protein gel or blot staining mass spectrometry became much more famous in the last years [1]. In classical techniques mass tags can be introduced in proteins for an indirect quantification of protein in compound while, in non-classical techniques, such label-free quantification approaches, a correlation of the mass spectrometric signal of intact proteolytic peptides (or the number of peptide sequencing events) can be quantified directly [1]. The recent application of MS for compound identification with label free quantification on expression systems like NAPPA/SNAP lead to the needs of more sophisticated and ad hoc algorithms to be used in combination with spectrometer software. Several reasons brought authors of this manuscript to develop a new script and are discussed in more details in the following. The proposed software was developed in a wider framework whose final goal is to implement a standardized analysis procedure, able to analyze the protein-protein interactions occurred on protein array in a label free manner by means of MS. To this aim we employed a MALDI-TOF mass spectrometer [2]. To analyze the protein interactions occurred on an innovative kind of protein array named NAPPA (Nucleic Acid Programmable Protein Array) [3]. Explosion of label-free techniques for NAPPA microarrays [4] allows for functional proteins to be synthesized *in situ* directly from printed cDNAs and utilized for personalized medicine; with the proteins being recently translated also using a reconstituted *E. coli* coupled cell-free expression system [2,5]. One of the challenges in evaluating the mass spectra obtained from NAPPA was the extra-biological material present on the NAPPA together with the target proteins, such as the master mix and lysate molecules. The same trouble is faced with the usage of NAPPA/SNAP tag that require spectra interpretation of synthesized proteins by additional peptide chain, the SNAP tag, and the anti-SNAP antibody [5]. These molecules are in all features of the array, and represent a "common background". These latter "background" molecules represent the main obstacle to the data interpretation. The proposed software would discriminate between protein and background peaks, allowing protein identification.

Some attempts have been done for mixture analyses, e.g. Label free LC-MS profiling and can be found in [6,7], these latter proteomic methods for peptide abundances detection do not admit identification while a comparison can be done. On the software side a wide range of tools are available to identify protein mass spectra, but the original samples have to be composed of few molecular species, i.e., 3 or 4.

***Corresponding author:** Claudio Nicolini, Laboratories of Biophysics and Nanobiotechnology, Department of Experimental Medicine, University of Genova, Italy, Tel: +3901035338217; Fax: +3901035338215; E-mail: claudio.nicolini@unige.it

Received July 17, 2013; **Accepted** September 16, 2013; **Published** September 23, 2013

Citation: Belmonte L, Spera R, Nicolini C (2013) SpADS: An R Script for Mass Spectrometry Data Preprocessing before Data Mining. J Comput Sci Syst Biol 6: 298-304. doi:10.4172/jcsb.1000125

Moreover, some of them are freely available. PEAKS is a software package that starting from raw mass spectra convert data and perform peptide and protein identification, mutation characterization, result validation and reporting.

Other software packages are available in the web; some of them are freely available and free of charge (mMass, DB Tool kit, Mass Kinetics, dante) [7-11]. Some of them are able to provide a theoretical digestion such Sherpa optimized for LC-ESI. While a lots of them provide functions for preprocessing, such smoothing, baseline subtraction and normalization, only few are able to perform clustering and data mining. Only Dante R is written in R and provides clustering functionality. Anyway, none of this discussed software is able to perform protein peaks extraction when a background molecule is provided with the spectra, i.e., master mix and lysate. Finally, peak identification is performed by statistical analysis of characteristics peaks [12] and thus, erasing some peaks, by subtracting background spectra, can alter outcome. A way to overcome this latter trouble is data mining on spectra datasets. SpADS provides preprocessing functionalities such as smoothing, peak extraction and normalization, it is able to perform peak alignment, to apply a threshold and finally provides functions for background subtraction in order to keep peak identification easier even in presence of noise. Finally, it can be used coupled to simple data mining algorithms such k means clustering in order to identify proteins when a data base search in MASCOTT is no possible.

Materials and Methods

NAPPA/SNAP mass spectrometry data acquisition

For MS analysis the array printing was realized in a special geometry getting protein samples with higher density, in order to obtain an amount of protein appropriate for MS analysis. The spots of 300 microns were printed in 12 boxes of 10×10 (spaced of 350 microns, centre to centre). The spots in a box were of the same gene, and in particular one box apiece was reserved to the sample genes (p53, CDK2, Src-SH2 and PTPN11-SH2), two boxes were printed with master mix (MM) as negative control and reference samples, and six boxes, labeled with the letters from A to F, were printed with the sample genes in an order blinded to the MS user.

The analysis was performed using a MALDI-TOF Ultraflex III (Bruker Daltonics, Leipzig, Germany). Digested trypsin was synthesized on the NAPPA for MS analysis. At the end of the digestion the solvent was let evaporating at RT and the slides were stored at 4°C for Ultraflex III MALDI-TOF MS analysis. The MALDI-TOF measures were performed in reflectron mode; the resulting mass accuracy for protein was < 50 ppm. MALDI TOF mass spectra were acquired with a pulsed nitrogen laser (337 nm) in positive ion mode. For each sample we acquired 8 spectra.

Protein mass spectrometry datasets

Results of three different tests are showed in this manuscript. The first dataset is composed of 20 spectra, acquired by standard MS technique. Since tests were blind to the user of this software in this manuscript we will refer to "a" and "b" spectra. Data set is available for download as "test dataset" at SpADS URL and it is composed of 20 samples, 10 for "a" group and 10 for "b". Some tests were performed using the same dataset duplicating entries in order to check data consistency, this latter dataset was indeed composed of 40 spectra and results are showed in this manuscript.

Afterwards, a second test was performed, in which SpADS was run with a dataset composed of four different NAPPA/SNAP expressed proteins, (p53, CDK2, Src-SH2 and PTPN11-SH2). The second dataset was thus composed of 32 spectra. Finally, a third dataset was used. This latter dataset was composed of four known protein spectra (p53, CDK2, Src-SH2 and PTPN11-SH2) and of six blind spectra (A, B, C, D, E, and F). Each of these samples was acquired 8 times, thus the final dataset was composed of 80 spectra plus 2 spectra of master mix.

Input and output

Both input and output SpADS files are in ASCII format. ASCII format have the advantage of being platform independent, easy to use and to read, light and it can be obtained regardless of the device used for data acquisition. The input file is composed of two columns representing mass/charge ratio (mZ) and Intensity. The output file, instead, is composed of a single mZ row and as many rows for intensities as many spectra files were inputted by the user, e.g. eleven rows for ten spectra. Spectra must be calibrated in the same range before preprocessing.

Software for MS data

Pre-processing: In order to reduce noise, preprocessing functions are useful for noise filtering and data reduction. Moreover, since MS spectra often contain hundreds of thousands peaks, semi automatic algorithms are needed for an easy, quick and reproducible spectra processing. Mainly there are three preprocessing functions provided by SpADS: normalization, smoothing and binning. Each of these techniques aims to MS spectra noise reduction [13].

Smoothing: Smoothing processes points averaging each point with its neighbor in a data time series. The purpose of this technique is to increase signal noise ratio. The simplest way to perform smoothing is the moving average method based on equidistant points. An array of raw data is converted into an array of smoothed points, in which, each of these point, is the average of 2n +1 consecutive unpaired raw data. The number 2n +1 unpaired is generally called filtering window.

Normalization: Normalization process allows different samples comparison when different values of absolute peak are not feasible. The purpose of normalization is to identify and delete sources of systematic variations between the spectra. Three types of normalization are provided by SpADS:

1. Direct normalization: $V_{NORM} = 1 - (\frac{V_{max} - V_{orig}}{V_{max} - V_{min}})$ where V_{NORM} is the normalized intensity value, V_{max} and V_{min} are respectively the maximum and minimum intensity value, and V_{orig} is the intensity value to be normalized;
2. Inverse normalization: Normalized intensity of each spectrum is obtained according to the formula $V_{NORM} = (\frac{V_{max} - V_{orig}}{V_{max} - V_{min}})$ where V_{NORM} is the normalized intensity value, V_{max} and V_{min} are respectively the maximum and minimum intensity value, and V_{orig} is the intensity value to be normalized;
3. Canonical normalization: Normalized intensity of each spectrum is obtained according to the formula where $V_{NORM} = \frac{V_i}{\sum_{i=1}^{n} V_i}$ V_{NORM} is the normalized intensity value and V_i is the intensity value to be normalized.

Binning: Binning procedure is one of the most used techniques in mass spectrometry data pre-processing [14]. It aims to preserve information from raw data performing while a dimensional reduction is performed. Binning groups adjacent values by electing a representative value on the basis of an aggregate function (Cannataro et al.). Therefore, main parameters of binning function are: (1) Binning window width; (2) Aggregate function used to calculate the I value (maximum, minimum or average); (3) Function used to select the representative mZ. Dimensional data reduction make data processing steps easy, allowing a data size reduction optimizing the overall performance.

Summarizing, this latter function groups adjacent values by electing a representative one inputting a pairs dataset such $[(I_1, mZ_1), (I_2, mZ_2), ..., (I_n, mZ_n)]$ and outputting it with a single point of the type $[(I, mZ)]$, where I is given by an aggregate function of N intensity values of the data and the charged mass mZ is usually chosen from among the values original mZ. This basic operation is conducted using a sliding window across the mZ spectrum axis.

Peak extraction and related functions

Peak extraction identifies the most significant peak value choosing real peaks among a huge amount of noisy peaks. A binning procedure is applied and then SpADS search for a local maximum in each binning window. Once those real peaks are extracted, three other functions can be used and provided by SpADS:

1. Peak alignment: align main peak of the two compared spectra, i.e. spectrum and noise. In its current version SpADS is able to align only main peaks of a binned window. It searches for a maximum peak in a binning region and use it as referral point. Then, all other main peaks, and spectra, in the same section are shifted and aligned to this latter, in both protein spectra and noise spectra.
2. Threshold: flat intensity to zero if it is found minor than a desired value. Threshold is not determined in automated way and it can be inserted depending on the user practice and needs.
3. Flat all peaks to one: it is used coupled with threshold, and it flats peaks to one. It can be useful to identify at which mZ peaks values, since a dataset of 0 and 1 values is provided where the value 1 corresponds to a real peak

At the end of the preprocessing stage a table format dataset in an ASCII file is given. The table have m*n dimensions where m is the columns number and n is the rows number with m<n. The first row represents the m/Z values. The first column represents the sample, e.g. CDK2, p53 etc. Every row of the column is filled in with the intensity value corresponding to the mZ preprocessed intensity value.

K means Clustering: Clustering through kmeans algorithm [15] was performed using an R implementation by [16]. The input of this section was given by a SpADS preprocessed dataset in ASCII format. Cluster charts are given at the end of this stage. For further details on k means implementation please refer to [16,17].

Results and Discussion

Running SpADS script

After invocation of the "main R" file, SpADS will start giving some guidelines for data preprocessing. First of all, it will ask for the usage of an existing dataset. SpADS will ask for some details, such us: number of files to be preprocessed, labels to apply for each item, and path on your local systems on which it will find the spectra files.

Parameters needed at this stage is also the binning window for advanced users could be useful to insert a threshold, a region of interest and finally a noise signal that will be subtracted to all the spectra that will compose the final dataset. Moreover, SpADS provide a flatten function that admit user to flat all peaks to 1. This function is really useful in case of comparison of mZ peaks. Finally, if a ROI is selected SpADS will ask for peak alignment. After preprocessing SpADS will ask to launch k means algorithm as showed in [16].

Preprocessing and cluster results

A single spectrum viewer tool provided by SpADS show that selection of ROI in the provided spectra can be easily performed by giving a lower and an upper bound of the ROI. This selection is done even on the noise signal in case this latter is provided. Then, all the preprocessing functions are performed over the spectrum, showing as results what have been done in the single viewer latest framework. In Figure 1, a threshold value of 1000 in intensity was set. In addition, once you have selected a threshold you can choose to keep all peaks value of 1, figure not shown. This function is very useful in case you need to quickly and easily identify the main peaks. Figure 2 shows noise subtraction over single spectra. In this last case the file coincides with the noise signal, thus a flat line is obtained as result. When user choose to noise subtraction function the main peak alignment can be performed, this function is useful only in case of ROI very small and only in the case in which due to the operations of preprocessing in the main peak is in a not appropriate position.

All the operations, so far discussed, can be performed on the whole specimen in order to obtain a homogeneous dataset. The result, which is then input to data mining algorithms, it is saved in ASCII format output as discussed previously.

As can be seen from Figures 3 and 4, clustering through k means algorithm on classical MS acquired spectra has produced excellent results. In both cases, in order to test data consistency number of samples was duplicated. Thus, the points related to the same occurrence and are shown in the same cluster, thus the algorithm is perfectly capable of clustering in the right way the preprocessed spectra dataset.

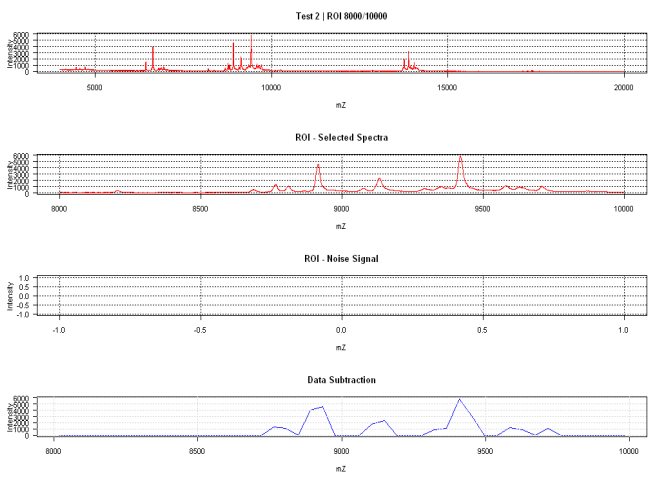

Figure 1: SpADS single spectrum viewer. From top to bottom: given spectra, selected ROI on the given spectra, selected Roi on given noise signal, preprocessed spectrum. If a noise spectrum it is not given then SpADS simply won't show it.

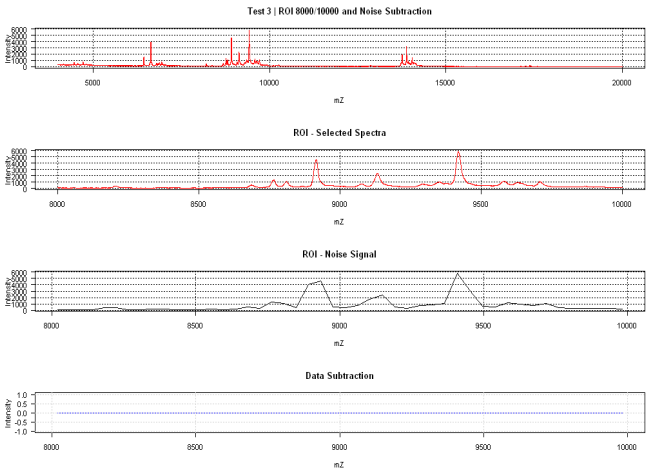

Figure 2: SpADS single spectrum viewer. From top to bottom: given spectra, selected ROI on the given spectra, selected Roi on given on the noise signal, preprocessed spectrum. In this example the same spectrum was used as noise resulting in a 0 function.

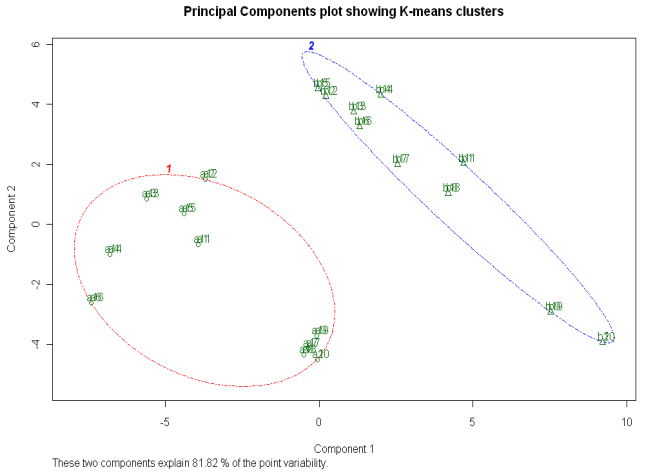

Figure 3: Clustering of the two specimen after a pre-processing with a binning window of 1000 mZ values. In this pre-processed datasets were not selected any ROI, threshold and noise signal.

The same test was performed on a region of the spectra of interest, namely the region in which, visually, one has the most information and which is located between 8000 and 10000 mZ. Even in this case, it is able to make the clustering with a perfect separation of the two sets of data, Figure 5. After application on test set, SpADS and cluster analysis were performed for the "NAPPA case". Noise subtraction and peak alignment were performed to overcome master mix effect that is, usually, to spread out protein spectra on the chart.

To overcome this trouble different attempts were performed. Mainly, SpADS/k means were tested for all the known spectra of p53, CDK2 and so on, with and without noise subtraction, respectively Figures 6 and 7. As showed in Figure 7 more distinct cluster results after the subtraction of master mix appear. These results are available even in related Tables 1 and 2.

In Table 2, with a color code, sample cluster assignment is showed. Assignment in Table 2 is given based on calculated frequentist cluster probability assignment. This was done in a very straightforward way for red, pink and blue clusters assigned to PTPN11-SH2, Src-SH2 and p53 proteins respectively while green cluster was assigned for exclusion since it was not possible to discriminate this latter one.

After all the so far discussed attempts a run with 80 spectra (CDK2, PTPN11-SH2, p53, Src-SH2 and a to f) was performed and k means result are shown in Figures 8 and 9, moreover a cluster probability was

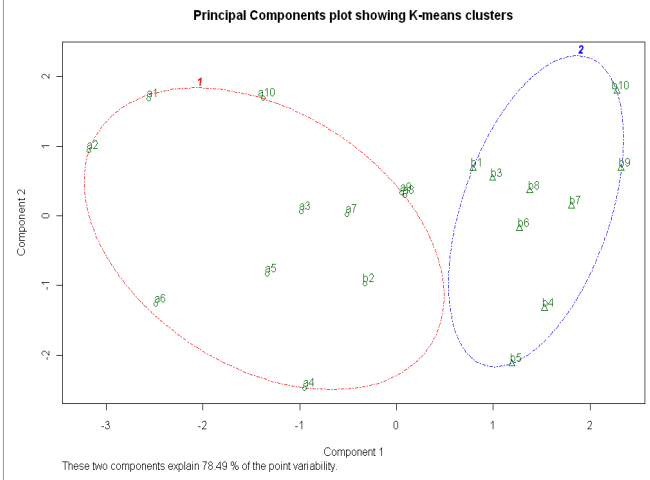

Figure 4: Clustering of the two specimen after a pre-processing with a binning window of 2000 mZ values. In this pre-processed datasets were not selected any ROI, threshold and noise signal.

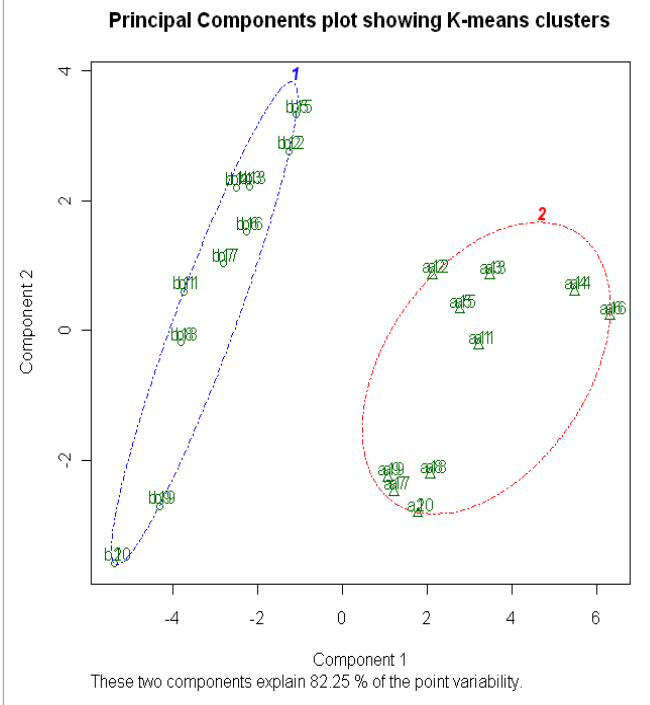

Figure 5: Clustering of the two specimen after a pre-processing with a binning window of 1000 mZ values. In this pre-processed datasets were selected ROI 8000-10000.

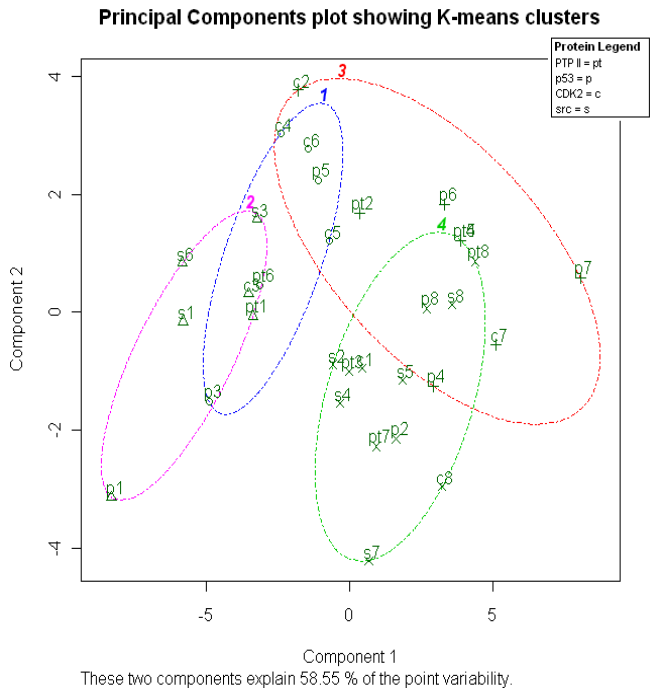

Figure 6: Clustering solution for a specimen of 32 protein samples of raw data. Only binning preprocessing function was performed before cluster analysis run. On the bottom of the red cluster (cluster 3) Master Mix spectra ("mm") and Noise spectra ("noise") are shown in the same point.

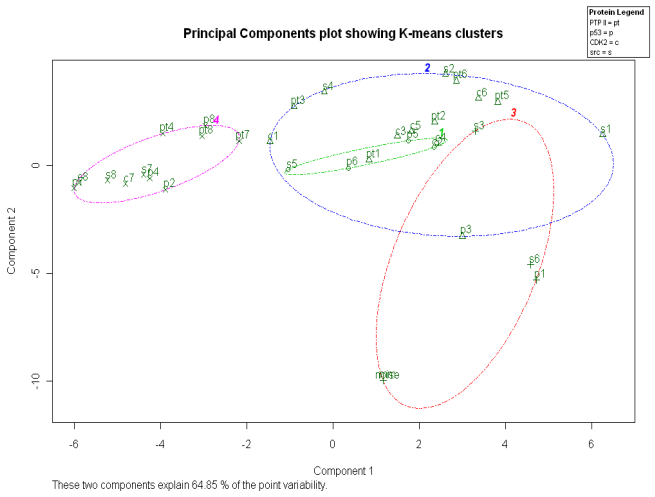

Figure 7: Clustering solution for a specimen of 32 protein samples in which Master Mix spectrum was subtracted with preprocessing function of noise reduction for every single protein spectra. Specimen is composed of 4 proteins and 8 samples for each protein. Proteins are: CDK2, PTP II, p53 and SRC. Green cluster is for X labeled samples, Red cluster for + labeled samples, Pink cluster for Δ labeled sample and Blue cluster for O labeled samples.

then computed and it is given in Tables 3 and 4. Even for this latter case eight spectra per sample were used. The result of this acquisition is of a specimen of 80 spectra. Cluster analysis shows us a much messed chart (Figure 8) and, thus, cluster assignment was done only by a manual computation of the clustering probability as shown in Table 4 [18].

Conclusions

In this manuscript the first version of SpADS is introduced and discussed. SpADS is an R script that provides preprocessing functions

Assigned Cluster Color	Green	Red	Pink	Blue
Assigned Symbol	**X**	**+**	**Δ**	**O**
P53 (p)	2	3	1	2
PTP (pt)	3	3	1	1
SRC (s)	5	0	3	0
CDK2 (c)	2	2	1	3
Total Number of Samples	12	8	6	6

Table 1: Number of protein samples assigned to clusters. Cluster analysis is based on a specimen of 32 protein samples, 8 samples for each protein type in which noise reduction were performed previously. This results are based on SPADS output showed in Figure 7.

	Cluster Assignment Probability (%)			
Assigned Cluster Color	**Green**	**Red**	**Pink**	**Blue**
Assigned Symbol	**X**	**+**	**Δ**	**O**
P53 (p)	**16,67**	37,5	16,67	33,3
PTP (pt)	25	**37,5**	16,67	16,67
SRC (s)	41,6	0	**50 ± 28**	0
CDK2 (c)	16,67	25	16,67	**50**

Table 2: Cluster probability assignment for each known protein sample. Statistics are based on the SPADS results given in Figure 7. Probability is calculated over the total number of samples.

Assigned Cluster Color	Pink	Blue	Red	Green
Assigned Symbol	**X**	**+**	**Δ**	**O**
P53 (p)	1	1	3	3
PTP (pt)	0	3	4	1
SRC (s)	3	2	0	3
CDK2 (c)	0	3	1	4
A	1	3	2	2
B	2	3	2	1
C	2	4	2	0
D	0	2	4	2
E	1	1	2	4
F	2	5	0	1
Total Number of Samples	12	26	20	22

Table 3: Number of protein samples assigned to clusters. Cluster analysis is based on a specimen of 80 protein samples. Specimen is composed of 4 proteins, as showed in previous results. Protein called A, B, C, D, E and F have an unknown distribution on the MS/SNAP sample during acquisition. This results are based on SPADS/k means output showed in Figure 8.

	Cluster Assignment Probability (%)			
Assigned Cluster Color	**Pink**	**Blue**	**Red**	**Green**
Assigned Symbol	**X**	**+**	**Δ**	**O**
P53 (p)	8.3	3.84	15	15
PTP (pt)	0	11.5	**20**	4.5
SRC (s)	**25**	7.6	0	15
CDK2 (c)	0	11.5	5	**18.18**
A	8.3	**11.5**	10	9.1
B	**16.67**	11.5	10	4.5
C	**16.67**	15.38	10	0
D	0	7.6	**20**	9.1
E	8.3	3.84	10	**18.18**
F	16.67	**19.23**	0	4.5

Table 4: Cluster probability assignment for each known protein sample on a specimen of 80 samples. Statistics are based on the SPADS/k means results given in Figure 8. Probability is calculated over the total number of samples.

for non-conventional MS acquisitions and it aims to produce more compact datasets while preprocessing functions are applied to those data. Its outcome could be used as input for other software packages that implement data mining algorithms, and in this manuscript, we tested its application with an R implementation of the k means clustering algorithm. Moreover, its application to non conventional MS acquisition can overcome background signal expression, trouble faced in particular when a NAPPA/SNAP expression system is used. To test this latter application cluster analysis based on different protein specimen was performed. In particular three different datasets were used, the first composed of two different proteins spectra, acquired with a classical MS method, while the second and third specimens where composed of spectra acquired by a modified MS label-free technique based on NAPPA SNAP. In particular, the second dataset was composed of known protein in a known displacement over the array, while the third dataset was composed of known protein displaced only partially in known positions. This latter test was performed in order to evaluate the discriminate power of SpADS coupled with k means implementation.

As showed, for the first test case, after SpADS application, k means was still able to recognize the two spectra families in a very straightforward way, as showed in Figures 3-5.

For the NAPPA cases, instead, a "semi automatic" cluster analysis has been performed to distinguish proteins acquired by MS/SNAP samples. Evaluation of the results needs human intervention. In some cases it is really difficult to distinguish cluster of same proteins, such in case of Figure 6. To solve this latter problem, authors of this manuscript performed a noise subtraction preprocessing given as input the master mix spectra as noise spectra. This latter usually can provide datasets that after a cluster analysis admit user to distinguish protein spectra on the base of a cluster probability that appear more discriminant than without noise subtraction, Figure 7. In our hands SpADS coupled with k means cluster analysis provide the following recognition: B and C are supposed to be Src-SH2, D is supposed to be PTPN11-SH2 and E is supposed to be CDK2. For what concern p53 it is only possible to recognize it and recognition of this latter was done only after exclusion

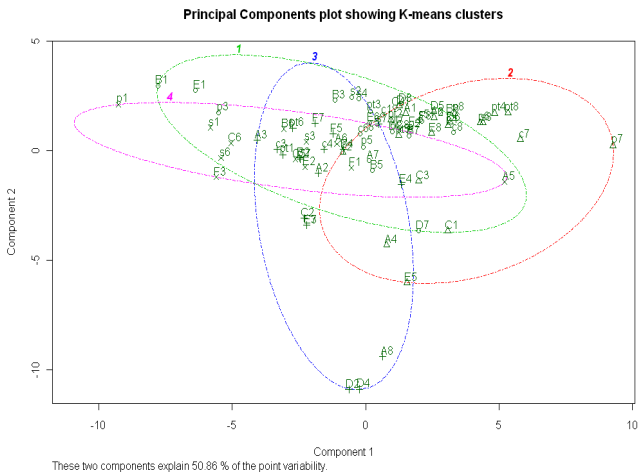

Figure 8: Clustering solution for a specimen of 80 protein samples in which Master Mix spectrum was subtracted with pre-processing function of noise reduction for every single protein spectra. Specimen is composed of 4 proteins. Proteins are: CDK2, PTP II, p53 and SRC. A, B, C, D, E and F are the same protein but with an unknown displace to the user.

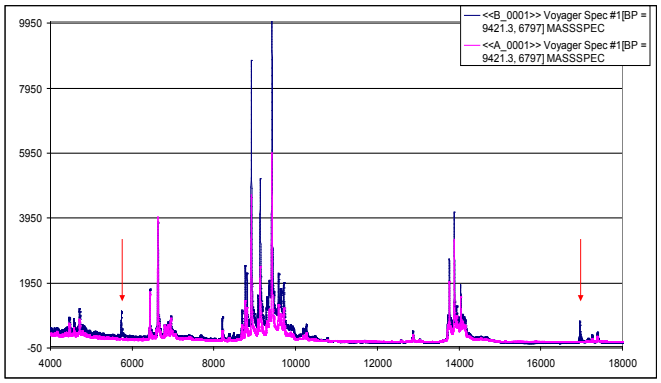

Figure 9: Mass Spectra of two unknown sample, a (pink) and b (blue), differing only in two peaks, at 5740 and 16958 Da/C and later identified as a mixture of small proteins. The red arrows point the only differences between the two spectra.

of the other protein spectra, Figure 8. This means that recognition without human intervention and classical MS analysis is not possible even using preprocessing and k means algorithms.

Summarizing, while the results obtained in the former case (Figures 3-5) appears comforting, there is no straightforward certainty in the latter case. However, it may be concluded that this uncertainty cannot be due to a bad software implementation neither to a bad samples acquisition. Indeed, as is well known, the applications of mining algorithms is strongly problem dependent and a possible way out to overcome this troubles could be to use SpADS coupled with other implementations of clustering algorithms or classification or with more direct background subtraction from the signal. Finally, MS acquisition was done using standard MS acquisition methods, while author think that ad hoc hardware development, and expression protocols, have to be implemented has was done at software side [5]. The results obtained by bioinformatics, however, are encouraging even with a low number of spectra because they partially fix the trouble and show the way to go in future development of this emerging technique. Moreover, authors consider that the elements of the specimen were chosen for human intervention, there is therefore the need to implement heuristics to allow determining a priori which data are suitable for the purposes of processing.

Acknowledgments

This work was supported by a PhD fellowship to Luca Belmonte and by grants to FEN (Fondazione Elba Nicolini) and to Professor Claudio Nicolini of the University of Genova by the FIRB Italnanonet (RBPR05JH2P) from MIUR (Ministero dell'Istruzione, Universita` e Ricerca; Italian Ministry for Research and University). SpADS script can be downloaded free of charge from http://www.ibf.unige.it/SpADS.

References

1. Bantscheff M, Schirle M, Sweetman G, Rick J, Kuster B (2007) Quantitative mass spectrometry in proteomics: a critical review. Anal Bioanal Chem 389: 1017-1031.
2. Spera R, LaBaer J, Nicolini C (2011) Mass Spectrometry Detection of Nucleic Acid Programmable Protein Array. Journal of Mass Spectrometry 46: 960-965.
3. Nicolini C, Bragazzi N, Pechkova E (2012) Nanoproteomics enabling personalized nanomedicine. Adv Drug Deliv Rev 64: 1522-1531.
4. Nicolini C, Labaer J (2010) Functional Proteomics and Nanotechnology-based Microarrays, Pan Stanford Series on Nanobiotechnology 2: 1-308.
5. Spera R, Festa F, Belmonte L, Chong S, Pechkova E, et al. (2013) Mass Spectrometry and Florescence Analysis of SNAP-NAPPA Arrays Expressed Using E. Coli Cell Free Expression System.

6. Christin C, Bischoff R, Horvatovich P (2011) Data processing pipelines for comprehensive profiling of proteomics samples by label-free LC-MS for biomarker discovery. Talanta 83: 1209-1224.

7. Polpitiya AD, Qian WJ, Jaitly N, Petyuk VA, Adkins JN, et al. (2008) DAnTE: a statistical tool for quantitative analysis of omics data. Bioinformatics 24: 1556-1558.

8. Strohalm M, Hassman M, Kosata B, Kodícek M (2008) mMass data miner: an open source alternative for mass spectrometric data analysis. Rapid Commun Mass Spectrom 22: 905-908.

9. Martens L, Vandekerckhove J, Gevaert K (2005) DB Tool kit: processing protein databases for peptide-centric proteomics. Bioinformatics 21: 3584-3585.

10. Drahos L, Vékey K (2001) Mass Kinetics: a theoretical model of mass spectra incorporating physical processes, reaction kinetics and mathematical descriptions. J Mass Spectrom 36: 237-263.

11. Taylor JA, Walsh KA, Johnson RS (1996) Sherpa: a Macintosh-based expert system for the interpretation of electrospray ionization LC/MS and MS/MS data from protein digests. Rapid Commun Mass Spectrom 10: 679-687.

12. Pappin DJ, Hojrup P, Bleasby AJ (1993) Rapid identification of proteins by peptide-mass fingerprinting. Curr Biol 3: 327-332.

13. Peaks

14. Gopalakrishnan V, William E, Ranganathan S, Bowser R, Cudkowic ME, et al. (2004) Proteomic data mining challenges in identification of disease specific biomarkers from variable resolution mass spectra.

15. Forgy EW (1965) Cluster analysis of multivariate data: efficiency vs interpretability of classifications. Biometrics 21: 768-769.

16. Peeples, Matthew A (2011) R Script for K-Means Cluster Analysis.

17. Hartigan JA, Wong MA (1979) A K-means clustering algorithm. Applied Statistics 28: 100-108.

18. Lloyd SP (1982) Least squares quantization in PCM. Technical Note, Bell Laboratories. IEEE Transactions on Information Theory 28: 128-137.