MSPtool: A Versatile Tool for Mass Spectrometry Data Preprocessing

F. Gullo, G. Ponti, A. Tagarelli
DEIS
University of Calabria
Via P. Bucci, 41c
87036 Rende (CS) — Italy
{fgullo,gponti,tagarelli}@deis.unical.it

G. Tradigo, P. Veltri
Department of Experimental and Clinical Medicine
University Magna Græcia of Catanzaro
Viale Europa
88100 Località Germaneto (CZ) — Italy
gtradigo@si.deis.unical.it, veltri@unicz.it

Abstract

Preprocessing mass spectrometry (MS) data has been recognized as a crucial preliminary phase in order to perform data management and knowledge discovery tasks on mass spectra. The huge dimensionality and heterogeneity of MS data make mandatory the use of tools that are able to guide the user in the MS preprocessing task. However, most MS preprocessing tools are typically designed to perform only some preprocessing steps and are strictly coupled with MS data analysis modules.

In this paper, we present Mass Spectra Preprocessing tool (MSPtool)*, a user-friendly versatile tool for preprocessing MS data. MSPtool provides the user with a wide set of MS preprocessing steps by means of an easy-to-use graphical interface. Also, this tool has been embedded in a time-series-based framework for MS data clustering.*

1. Introduction

Mass Spectrometry (MS) is a powerful analytical technique that can be applied to tissue or serum samples in order to extract interesting information from biological samples [2]. A mass spectrometer is able to produce and separate ions of different masses from a sample, so that the output spectral data consists of a vector of counts, where each count is the number of ions hitting the spectrometer detector during a small, fixed interval of time. A mass spectrum is typically represented as a plot of ion abundance (*intensity*) versus the mass-to-charge ratio (*m/z*). By analyzing mass spectra it is possible to identify macromolecules contained in the original compounds by associating (portions of) proteins to their peak expressions in a spectrum.

MS data preprocessing has been recognized as a mandatory phase in mass spectra data analysis. The need for preprocessing mass spectra arises since *(i)* data obtained from a mass spectrometer have very large dimensionality and *(ii)* MS data are naturally corrupted by various noisy factors. Several research studies have been proposed on the development of preprocessing steps for MS data (e.g., [9, 10]), and in some cases they have focused on specific steps, such as baseline subtraction [4, 1, 3], peak identification [12, 13], and peak alignment [8, 11, 1].

Also, there has been recently a growing interest for developing MS data preprocessing systems that are able to fulfill the following main requirements: filtering data and highlighting relevant spectra portions w.r.t. non-relevant ones (e.g., noise), and allowing the user to perform the various preprocessing stages iteratively and interactively.

In this paper, we present *Mass Spectra Preprocessing tool (MSPtool)*, a graphical tool for preprocessing mass spectrometry data. The main features of our tool can be summarized as follows:

- *Wide set of supported preprocessing operations.* MSPtool is designed to cover most MS data preprocessing steps that have been recognized as the most relevant in the literature.
- *Efficiency.* MSPtool guarantees high performance in executing MS preprocessing experiments, by adopting fast algorithms for each step. This allows for efficiently dealing with high dimensional data.
- *Support for user interaction.* Our tool makes the user able to monitor and control the whole preprocessing task. In particular, the user can choose

which preprocessing steps have to be performed and their execution order, and she/he can properly set the parameters involved into each step.

- *Ease-to-use.* MSPtool provides a user-friendly graphical interface and a simple wizard which guides the user in each preprocessing step.
- *Web-based access.* MSPtool makes use of the Java Web Start technology (JWS), which allows for launching the tool directly from the Web. A beta version of MSPtool is available at the following web address: http://polifemo.deis.unical.it/~ gtradigo/jnlp/msptool/

MSPtool has been integrated into a general framework for knowledge discovery in MS data, and specifically for MS data classification. Within this view, we have tested our MSPtool as a MS preprocessing unit into a time-series-based framework for clustering MS data [7]. Experimental results obtained by the clustering framework have shown this capability of MSPtool in performing effective preprocessing on mass spectra.

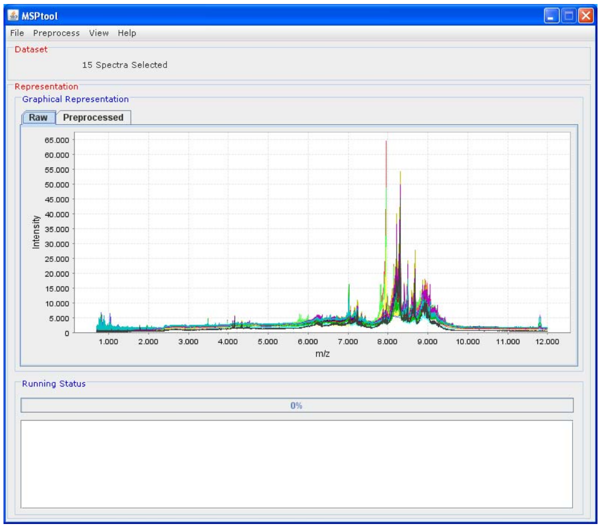
The rest of the paper is organized as follows. The next section presents the prominent steps involved into a general MS data preprocessing task. Section 3 describes our MSPtool. Section 4 discusses the use of MSPtool as a preprocessing module integrated into a framework for clustering MS data. Finally, Section 5 concludes the paper.

2. MS data preprocessing

A raw spectrum outputted from a mass spectrometer is substantially a combination of three components: the true signal, a baseline signal, and noise [9]; in particular, the true signal contains biological information, whereas the base intensity level (baseline) varies from point to point across the m/z axis, so that intensity values that are under the baseline represent ground noise and should be hence filtered out. Separating and reconstructing such individual components from a raw spectrum is a hard task, since their analytical forms are not known. Thus, spectra usually need to be subject to one or more preprocessing operations, in order to make them amenable to further analysis phases.

Since the variety of spectrometry platforms, experimental conditions and clinical studies, there exists a number of preprocessing operations (see, e.g., [9, 10]). While there has not been shared agreement about a general-purpose preprocessing scheme, a reasonable list of preprocessing steps on mass spectra can be given as follows:

- *calibration*, which is used to map the observed time of flight into the inferred mass-to-charge ratio;



15 Spectra Selected

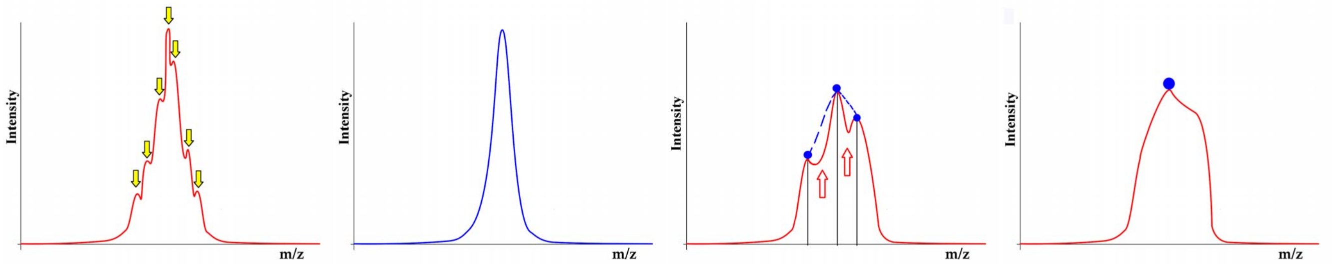
Figure 1. A sample MSPtool screenshot

- *filtering or denoising*, which aims to reduce random noise generated by electronic or chemical causes;
- *baseline correction*, which is in order to recognize and filter out the baseline signal of mass spectra;
- *normalization*, which makes peak intensities understandable over a uniform range;
- *peak detection*, which is in charge of locating specific proteins or peptides on the identified locations on the m/z axis and typically involves an assessment of the spectra local maxima and their signal-to-noise ratio (S/N);
- *peak quantification*, which represents each detected peak by means of a concise information (e.g., peaks heights or areas);
- *peak matching/alignment*, which aims to recognize which peaks in different samples correspond to the same biological molecule.

MSPtool supports most of the MS preprocessing steps reported in the above list. A description of the capabilities of our tool is given in the next section.

3. MSPtool overview

MSPtool is a Java$^{\text{TM}}$ developed tool that implements various operations of MS data preprocessing (Figure 1). This tool offers its features visually in order to assist the user in performing an MS preprocessing



(a) (b) (c) (d)

Figure 2. Peak smoothing: (a) example M-peaks and (b) the corresponding ideal peak; (c) three local M-peaks and (d) the resulting profile after smoothing

task, i.e., observing the raw spectra, selecting an appropriate sequence of preprocessing steps, and choosing the parameter setting for each of the selected preprocessing steps. Figure 3 shows a screenshot of the last step of the preprocessing wizard, which reports a summary of the preprocessing setting; in this step, it is also possible to change the order of the selected preprocessing operations.

MSPtool is able to deal with various formats storing the raw spectrum/spectra to be preprocessed, including plain-text files, comma separated values files (CSV), XML data.

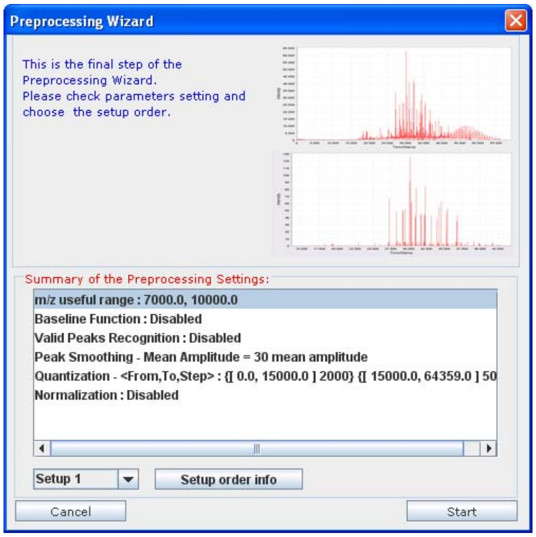
In the following, the various preprocessing steps for MS data involved into the MSPtool wizard are described.

Range cut. This step provides a cut of the *m/z* range of the spectra, in order to filter out those portions of the spectra that do not contain relevant biological information.

Peak smoothing. Peak smoothing falls into the category of peak detection/quantification steps. This step deals with smoothing peak profiles in the spectra and is accomplished to reconstruct the theoretical Gaussian profile of the peaks.

An ideal peak profile is comprised of two parts: a monotonic ascending side and a monotonic descending side. We call *M-peak* a peak in a spectrum having its intensity higher than both the previous and the next point, i.e. a local maximum in the spectrum (Figure 2 (a)–(b)).

The peak smoothing algorithm has a parameter w_p (*peak amplitude*). w_p is a function of the mass spectrometer resolution, and can be initially set to the average width of peaks in the spectrum. Nevertheless, the user can change the value of w_p according to the data features. Basically, the algorithm works as follows: first, it detects all the M-peaks in the spectrum; each M-peak (except the last one) is compared with the next M-peak. If the distance between these two M-peaks is lower than $w_p/2$ then either a descending phase or an ascending phase can occur, and the spectrum is modified such that the resulting peak has the expected pseudo-Gaussian shape for both the ascending and the descending sides (Figure 2 (c)–(d)).



This is the final step of the
Preprocessing Wizard.
Please check parameters setting and
choose the setup order.

Summary of the Preprocessing Settings:

m/z useful range : 7000.0, 10000.0
Baseline Function : Disabled
Valid Peaks Recognition : Disabled
Peak Smoothing - Mean Amplitude = 30 mean amplitude
Quantization - <From,To,Step> : {[0.0, 15000.0] 2000} {[15000.0, 64359.0] 50
Normalization : Disabled

Figure 3. Preprocessing Wizard - Summary of the preprocessing settings

Valid peaks recognition. Valid peaks recognition is a further step of peak detection/quantification. This step aims to recognize as valid peaks the local maxima into a mass spectrum that satisfy specific requirements. In particular, the algorithm for valid peaks recognition implemented into MSPtool takes into account the signal-to-noise ratio (S/N) and works as follows: for each spectrum, the S/N at each local maximum of the spectrum is firstly computed by taking the ratio of the intensity at the maximum to the local noise estimate; then, only the local maxima having S/N greater than a user-defined threshold (*multiplicative factor*) are recognized as valid peaks. The non-valid peaks in a spectrum are discarded from the further analysis.

Baseline correction. This step aims to identify the baseline signal in the spectra and filter out all spectra intensity values below the baseline. In this module, the user can choose the function suitable for approximating the baseline (i.e., the baseline function) and setting the parameters for each function. In particular, in MSPtool the baseline functions available are four: *linear function*, *logarithmic function*, *exponential function* and *piecewise linear function*. The first three functions approximate the baseline as a linear, logarithmic and exponential function, respectively, whereas the definition of the piecewise linear function is as follows. The m/z range of each spectrum is divided into a user-defined number of equally-sized windows. The final piecewise linear function is composed by a number of linear functions, each of them properly defined according to the associated window. In particular, for each window, the corresponding linear function is computed by solving a line fitting problem to the local minima in the window.

Quantization. This step performs a quantization of the spectra, that is a discretization of the original intensity values according to specific quantization levels. A non-uniform quantization model is used in the MPStool in such a way that two or more ranges in the intensity axis are identified and subject to different fine-grained quantization.

Normalization. Spectra normalization changes spectra shapes by transforming original intensity values into new ones proportionally calculated according to a certain fixed range.

MSPtool implements several normalization techniques, including z-normalization and min-max normalization. The former subtracts the mean over all the spectra intensities from each intensity value and then divides this difference by the standard deviation over all the spectra intensities; the latter scales the intensity values such that, for each m/z and over all the spectra, the smallest intensity value becomes zero and the largest intensity becomes one.

4. Integrating MSPtool into a framework for MS data analysis

MS preprocessing is a starting point for performing profitable further analysis on MS data, which typically is in order to classify spectra and discover useful knowledge that can aid clinicians in early detecting disease-related biological states. Therefore, it is important to validate a MS preprocessing module by assessing its adaptability in an MS-based knowledge discovery system.

For this purpose, MSPtool has been integrated into a time-series-based framework for clustering MS data, whose conceptual architecture is depicted in Figure 4.

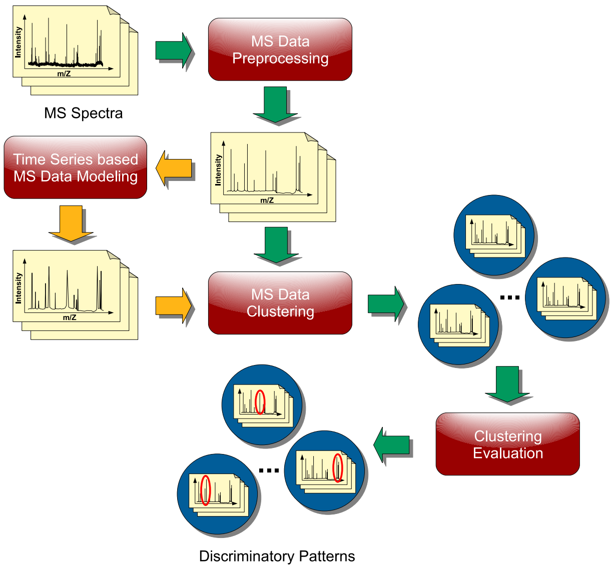
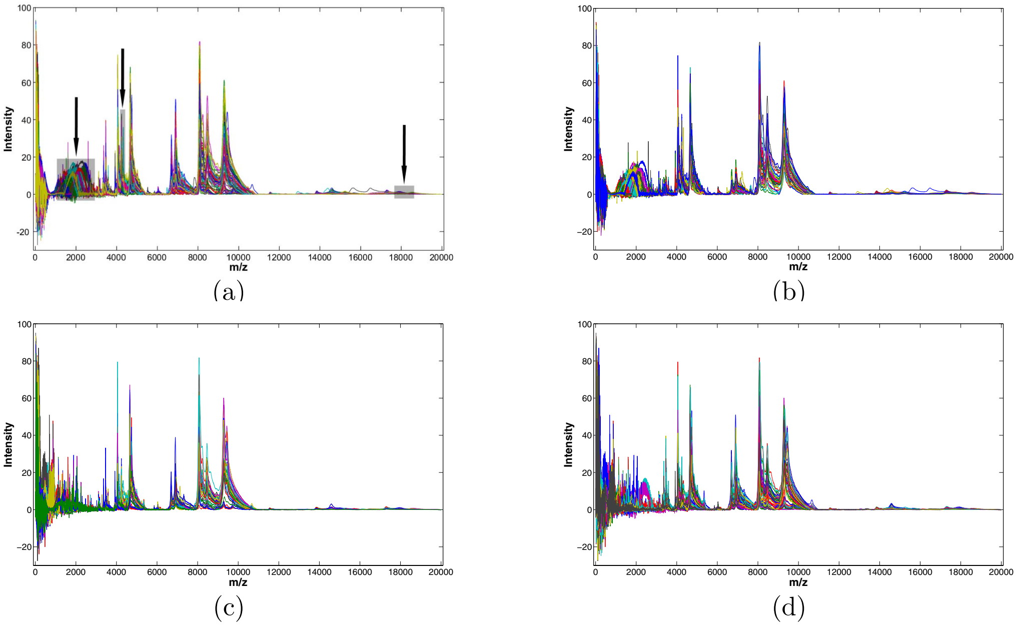


Figure 4. Conceptual architecture of the time-series-based framework for clustering MS data

In the following, we show the preprocessing results provided by MSPtool in preparing MS data with respect to a clustering analysis performed by the time-series-based framework. We illustrate the results obtained on a real dataset, `ProstateCancer`, a SELDI-TOF dataset available from the NCI's Center for Cancer Research.[1] `ProstateCancer` contains low resolution spectra used in a study on prostate cancer [6]. It is composed by 322 spectra, each of which contains 15,154 (m/z, intensity) couples. Data are assigned to 4 classes: cancer and PSA level > 10 ng/ml (43 spectra), cancer and PSA level within [4..10] ng/ml (26

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



(a)

(b)

(c)

(d)

Figure 5. Clusters vs. natural classes from `ProstateCancer`: (a) cluster and (b) class of cancer with PSA>10 ng/ml; (c) cluster and (d) class of no evidence of disease

spectra), benign and PSA level > 4 *ng/ml* (190 spectra), no evidence of disease and PSA level < 1 *ng/ml* (63 spectra).

On this dataset, we identified the following sequences of operations as different *preprocessing setups*: *(S1)* range cut, peak smoothing, valid peaks recognition, normalization; *(S2)* range cut, normalization; *(S3)* range cut, peak smoothing, valid peaks recognition; *(S4)* range cut; *(S5)* range cut, normalization. It should be noted that the baseline correction step is missing, since the spectra in the dataset have been already provided with the baseline subtracted.

We measured the accuracy reached by the time-series-based clustering framework on `ProstateCancer` using such preprocessing setups, in terms of two popular validity measures, namely *F-measure* [5] and *Entropy*. Experimental results have shown that MS data preprocessing allowed for increasing the accuracy in clustering spectra from 15% to 27% with respect to the case in which no preprocessing operation is performed.

From a qualitative evaluation viewpoint, we compared the definite cancer and no-evidence-of-disease clusters to their respective groupings in the reference classification (Figure 5)—for the sake of comparison, original spectra have been displayed for both clusters and classes. In `ProstateCancer`, as discovered by authors of the study in [6], there is a number of main discriminatory patterns, mostly distributed in the early m/z values. Figures 5(a) and (b) plot spectra belonging to the cluster and the class of definite cancer, respectively. At a first glance, the two plots look quite similar, suggesting that most cancer spectra have been correctly recognized. Also, we have highlighted on Figure 5(a) some of the most evident trends that distinguish the cancer conditions from the healthy ones. Analogously, we can observe similar graphs for the healthy cluster/class (Figures 5(c)/(d)). The interested reader can find details about the framework, the experimental evaluation methodology and further experimental results in [7].

5. Conclusion

In this paper we presented MSPtool, a user-friendly, graphical tool for preprocessing mass spectrometry data. Major features of MSPtool are the support for a wide set of preprocessing steps, the capability of performing MS preprocessing by allowing the user to fully control the whole process, and a Web-based access.

Moreover, MSPtool has been integrated into a framework for clustering MS data. The experimental results obtained by this framework have shown that MSPtool is able to profitably support the user in performing preprocessing of MS data.

Acknowledgements. Authors would like to thank Ettore Bruzzese and Simone Cafiero for their support.

References

[1] A.C. Sauve and T.P. Speed. Normalization, Baseline Correction and Alignment of High-Throughput Mass Spectrometry Data. In *Proc. Genomic Signal Processing and Statistics Conference*, 2004.

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

[3] A.F. Ruckstuhl, M.P. Jacobson, R.W. Field, and J.A. Dodd. Baseline subtraction using robust local regression estimation. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 68:179–193, 1999.

[4] B. Williams, S. Cornett, B.M. Dawant, A. Crecelius, B. Bodenheimer, and R. Caprioli. An algorithm for baseline correction of MALDI mass spectra. In *Proc. ACM Southeast Regional Conf.*, pages 137–142, 2005.

[5] C.J. van Rijsbergen. *Information Retrieval.* Butterworths, 1979.

[6] E.F. Petricoin 3rd, D.K. Ornstein, C.P. Paweletz, A. Ardekani, P.S. Hackett, B.A. Hitt, A. Velassco, C. Trucco, L. Wiegand, K. Wood, C.B. Simone, P.J. Levine, W.M. Linehan, M.R. Emmert-Buck, S.M. Steinberg, E.C. Kohn, and L.A. Liotta. Serum Proteomic Patterns for Detection of Prostate Cancer. *Journal of the National Cancer Institute*, 94(20):1576–1578, 2002.

[7] F. Gullo, G. Ponti, A. Tagarelli, G. Tradigo, P. Veltri. A Time Series Based Approach for Classifying Mass Spectrometry Data. In *Proc. Computer-Based Medical Systems (CBMS)*, pages 412–420, 2007.

[8] J.W.H. Wong, G. Cagney, and H.M. Cartwright. SpecAlign - processing and alignment of mass spectra datasets. *Bioinformatics*, 21(9):2088–2090, 2005.

[9] K.R. Coombes, K.A. Baggerly, and J.S. Morris. *Pre-Processing Mass Spectrometry Data.* Fundamentals of Data Mining in Genomics and Proteomics. Kluwer, Boston, 2007.

[10] M. Wagner, D. Naik, and A. Pothen. Protocols for Disease Classification from Mass Spectrometry Data. *Proteomics*, 3(9):1692–1698, 2003.

[11] N.O. Jeffries. Algorithms for alignment of mass spectrometry proteomic data. *Bioinformatics*, 21(14):3066–3073, 2005.

[12] W.E. Wallace, A.J. Kearsley, and C.M. Guttman. An Automated Peak Identification/Calibration Procedure for High-Dimensional Protein Measures From Mass Spectrometers. *Analytical Chemistry*, 76(9):2446–2452, 2004.

[13] Y. Yasui, D. McLerran, B.L. Adam, M. Winget, M. Thornquist, and Z. Feng. An Operator-Independent Approach to Mass Spectral Peak Identification and Integration. *Journal of Biomedicine and Biotechnology*, 4:242–248, 2003.