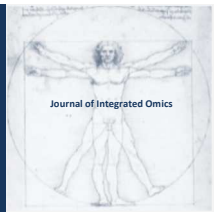

JOURNAL OF INTEGRATED OMICS
A METHODOLOGICAL JOURNAL
HTTP://WWW.JIOMICS.COM

ORIGINAL ARTICLE | DOI: 10.5584/jiomics.v6i1.196

Implementing effective machine learning-based workflows for the analysis of mass spectrometry data

Hugo López-Fernández*[,a], Miguel Reboiro-Jato[a], José A. Pérez Rodríguez[b], Florentino Fdez-Riverola[a], Daniel Glez-Peña[a]

[a]Escuela Superior de Ingeniería Informática, University of Vigo, Edificio Politécnico, Campus Universitario As Lagoas s/n, 32004 Ourense, Spain. [b]CFR: Centro de Formación e Recursos de Ourense, Rúa Universidade s/n, 32005 Ourense, Spain.

Received: 9 May 2016 **Accepted:** 8 June 2016 **Available Online**: 19 June 2016

ABSTRACT

Mass spectrometry using matrix assisted laser desorption ionization coupled to time of flight analyzers (MALDI-TOF MS) has become very popular during the last decade due to its high speed, sensitivity and robustness for accurately detecting proteins and peptides. This allows quickly analyzing large sets of samples in one single batch and doing high-throughput proteomics. In this scenario, bioinformatics methods and computational tools play a key role in MALDI-TOF MS data analysis, as they are able to correctly handle the large amount of raw data generated with the goal of discovering new knowledge and extracting useful conclusions.

A typical MALDI-TOF MS data analysis workflow consists of three main stages: data acquisition, preprocessing and analysis. Although the most popular use of this technology is to identify proteins through their peptides, analyses that make use of artificial intelligence (AI), machine learning (ML), and statistical methods are of particular interest to conduct biomarker discovery, automatic diagnosis, and knowledge discovery.

In this introductory work, the potential of these techniques is explored and novel solutions based on the application of AI, ML, and statistical methods are reviewed. In addition, an integrated software platform that supports full MALDI-TOF MS data analysis is presented with the goal of facilitating the work of proteomics researchers without advanced bioinformatics skills.

Keywords: Computational mass spectrometry, Machine learning, MALDI-TOF MS data, intelligent analysis workflow.

1. Introduction

In the last years, high-throughput proteomic data analysis using matrix assisted laser desorption ionization coupled to time of flight analyzers based mass spectrometry (MALDI-TOF MS) has been an active research area due to its high speed, sensitivity and robustness for detecting proteins and peptides. Within this technique, large sets of samples are analyzed quickly in one single batch. In this context, bioinformatics methods and computational tools play a key role in MALDI-TOF MS data analysis, since they can handle the vast amount of raw data generated, supporting the application of complex analysis with the goal of finally extracting new knowledge and useful conclusions [1].

A common MALDI-TOF MS data analysis workflow is characterized by three main stages: (*i*) data acquisition, (*ii*) preprocessing, and (*iii*) analysis. This standardized workflow starts with the acquisition and management of raw data that must be preprocessed to obtain clean peak lists, suitable for being used as input of the analysis stage [2]. Despite its apparent simplicity, each of these three main stages is composed by smaller steps, and different solutions and approaches have been proposed to address them in the last years [3]. Regarding the analysis stage, the most popular use of MALDI-TOF MS is to identify proteins through their peptides, a process known as *peptide-mass fingerprinting* (PMF). For this application scenario, the mass spectrum must be preprocessed for obtaining a list of peptide experimental masses, which can be searched against a database to identify target proteins. Nevertheless, analyses that make use of artificial intelligence (AI), machine learning (ML), and

***Corresponding author:** Dr. Hugo López-Fernández, Escuela Superior de Ingeniería Informática, University of Vigo, Edificio Politécnico, Campus Universitario As Lagoas s/n, 32004 Ourense, Spain. Phone number: +34 988387027. Email Address: hlfernandez@uvigo.es

statistical methods can also be executed in order to perform biomarker discovery, automatic diagnosis and knowledge discovery [1,4,5], taking peak lists as input. AI and ML methods have demonstrated their usefulness when applied to many different biomedical, biological and *omics* problems.

It is also important to note that when working with MALDI-TOF MS data, low quality spectra may be occasionally generated. For example, spectra showing a low number of m/z values in comparison with other spectra, or containing many unique m/z values not present in their sibling replicates. These noisy spectra may easily lead to many different types of errors or most severe incorrect conclusions. To prevent such a scenario, a quality control (QC) step, which may be performed between the preprocessing and the analysis tasks, should be considered.

Based on our previous experience in the field [6–8], the present work reviews the most important aspects for correctly implementing such machine learning-based workflows for the analysis of MALDI-TOF MS data. The core workflow analyzed, shown in Figure 1, is also implemented by Mass-Up [9], an open-source software platform freely available to the scientific community.

2. Preprocessing

Preprocessing of MALDI-TOF MS data is a decisive stage that transforms raw data into a suitable input for further analysis. In this context, inadequate or incorrect preprocessing methods can result in a biased dataset, hindering the process of reaching meaningful biological conclusions [10]. In such a situation, preprocessing becomes critical since raw data contains signals coming from the real peptides/proteins, as well as signals derived from several forms of noise (e.g. chemical, electronic factors, etc.). The specific goals of this phase are (*i*) to remove noisy peaks without discarding any of the true peaks and (*ii*) to determine both m/z and intensity values with the best accuracy [11]. Since there is no standard MS data preprocessing pipeline, some authors proposed different guidelines to establish a design/data analysis protocol [12,13]. After reviewing these guidelines, we proposed the following core preprocessing steps: (*i*) baseline correction, (*ii*) smoothing, (*iii*) peak detection and (*iv*) peak alignment. While the first two steps aims to remove noise, peak detection is a feature extraction process able to select true (i.e. peptide/protein-related) peaks from a given spectrum. Finally, peak alignment (also referred as peak matching) consists on determining which peaks correspond to the same peptide/protein in different samples. As a result of this phase, all the aligned peaks have the same mass values in all spectra and therefore, they are comparable and suitable for further machine learning analyses.

Additionally, our proposed workflow also incorporates a complementary filtering step that is closely related with the matching process. This step allows the creation of a

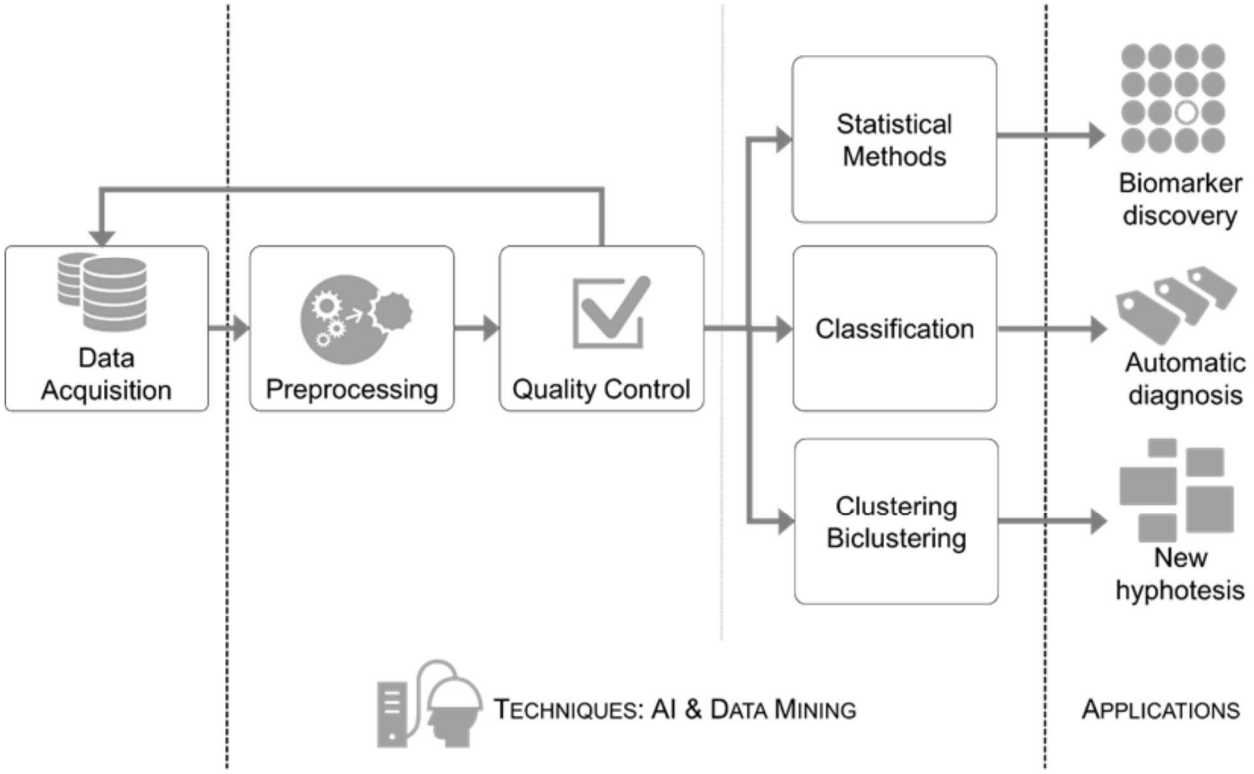

Figure 1. Machine learning-based workflow for the analysis of MALDI-TOF MS data.

consensus spectrum for a sample, which summarizes its replicates in one single spectrum. In our approach, the Percentage of Presence (POP) parameter allows the user to set the number of replicates where an m/z value must be present in order to be considered a valid consensus m/z value.

3. Machine learning-based analyses

Each preprocessed spectrum (or peak list) contains a finite number of peaks. A biomarker analysis can be done by using some adapted statistical methods that led to identify which of those peaks are associated with factors of interest [14].

Automatic diagnosis given a set of previously classified samples, is a supervised ML problem [15]. For example, given an unlabeled serum sample from an individual, which can come from one or more replicates (i.e. spectra), the purpose of classification could be to assign it to a specific diagnostic group (e.g. healthy or diseased). In this case, a classification model is built from a set of labeled samples using the intensity or the presence/absence of the different peaks (m/z) as input features [1]. It is important to note that when intensity values are used, the dataset must be normalized in order to make intensities comparable. Data used to build this model is called *training* data. The model is then used to predict the class of those unlabeled samples.

Common types of ML supervised techniques are, among others: (*i*) *Bayesian classifiers*, such as *Naïve Bayes*, which are based on Bayes theorem; (*ii*) *rule-based learners*, which are based on the creation of human-readable rules that could explain why certain samples belong to a class; (*iii*) *decision trees*, which are based on tree-like structures that organize the knowledge to discriminate between samples and predict their class; (*iv*) *random forests*, which use several decision trees to predict the class of each sample; (*v*) *support vector machines* (SVMs), such as *Sequential Minimal Optimization* (SMO), which are based in the concept of linear separability between classes; and (*vi*) *artificial neural networks* (ANNs), which simulate brain's operation in order to build the model and predict the class of each sample [1]. In algorithms such as rule-based learners or decision trees, it is also possible to consider some specific peaks as biomarkers used to separate the target classes. Despite the fact that these algorithms take peak lists as input, they can still contain noisy, irrelevant or redundant peaks, which can reduce the accuracy of the underlying classifiers. To mitigate these symptoms, *feature selection* can be applied prior to the use of classification algorithms, generating a cleaner dataset on which apply them. Feature selection methods can also be used to discover potential biomarkers.

In contrast with supervised machine learning, in unsupervised classification (or *clustering*), samples do not have associated class labels and they consist in grouping together samples with similar peak profiles. The main clustering approaches are: (*i*) *partition clustering* (e.g. *K-means* algorithm), (*ii*) *hierarchical clustering*, and (*iii*) *mixture models* [15]. These techniques are characterized by the fact that they perform a one-dimensional clustering using samples' attributes. A specific sub-type of clustering, called *biclustering* (or *co-clustering),* is able to perform a two dimensional clustering, that is, clusters are modeled with both samples and samples' attributes. These unsupervised techniques lead to the creation of new hypotheses (e.g. proposed groups) that must be further explored and evaluated.

4. Results

The straightforward workflow proposed in this work is implemented by Mass-Up [9], our all-in-one open software development for MALDI-TOF MS knowledge discovery fully covering the whole data analysis workflow. Mass-Up is an AIBench-based application [16] that allows researchers to easily manage and visualize raw data or peak lists, preprocess data, and execute different types of analyses such as (*i*) biomarker discovery, (*ii*) clustering, (*iii*) biclustering, (*iv*) three-dimensional PCA visualization and (*v*) classification of large sets of spectra data. This section briefly outlines the most relevant aspects of each analysis stage, from preprocessing to advanced machine learning-based analysis.

As commented before, preprocessing of raw data is a critical stage needed to generate a suitable input for further analysis in form of clean peak lists. Since inadequate or incorrect preprocessing methods can hinder the achievement of meaningful biological conclusions [8], Mass-Up includes state-of-the-art algorithms supporting the main preprocessing steps: (*i*) baseline correction, (*ii*) smoothing, (*iii*) peak detection and (*iv*) peak alignment. Mass-Up provides *Top Hat*, *SNIP*, *Convex Hull*, and *Median* algorithms for baseline correction from the MALDIquant package [17]. Regarding smoothing, the *moving average window* and *Savitzky-Golay* methods, both from the MALDIquant library, are offered. Additionally, Mass-Up supports two m/z selection alternatives: the *CWT-based* method implemented in MassSpecWavelet and a *SNR-based* method provided by MALDIquant. Concerning peak matching algorithms, Mass-Up includes a sequential procedure based on a sliding window (*Forward*, an in-house development) and a clustering based approach from MALDIquant.

When analyzing MALDI-TOF MS data, low quality spectra can be occasionally obtained (e.g. spectra showing a significant lower or higher number of m/z values in comparison with other spectra). These kind of spectra could lead to the achievement of incorrect conclusions or even hinder them. In order to prevent this possibility, a quality control (QC) step was included between the preprocessing and the analysis tasks. This QC procedure has two targets: *replicates*, a low-level analysis on the replicates of each sample; and *samples*, a high-level analysis with extra information about the intra-sample m/z matching process.

An important aim of MALDI-TOF MS analyses is

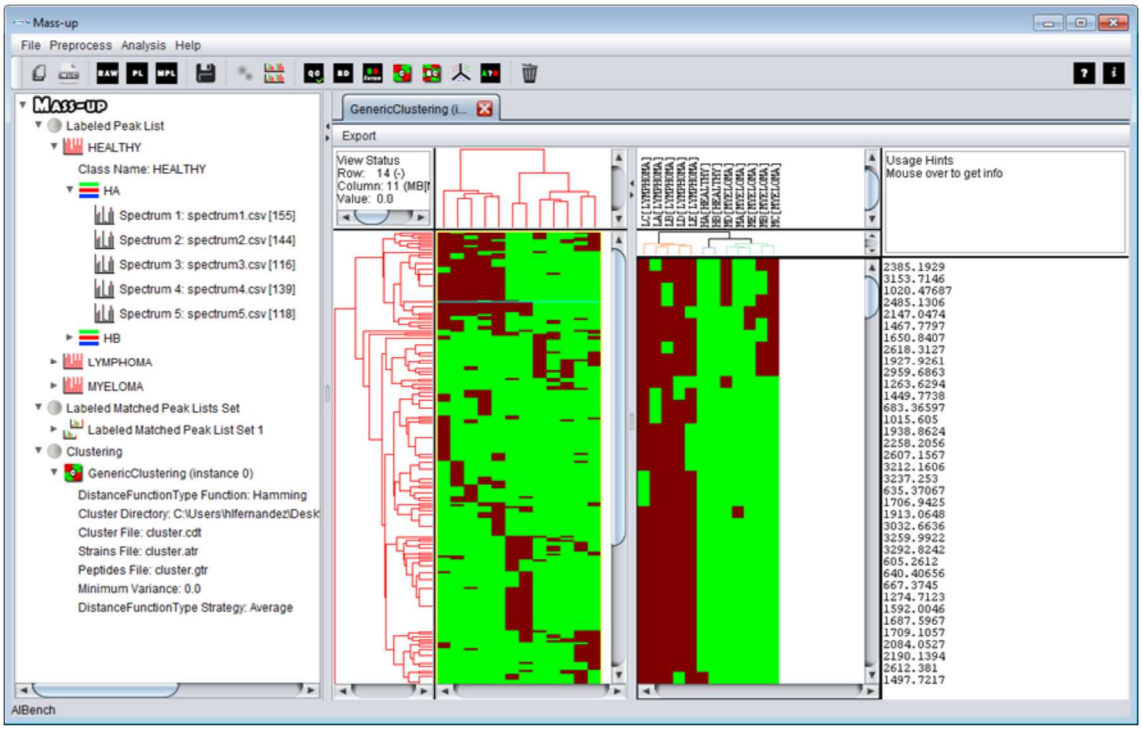

Figure 2. Mass-Up application showing a clustering analysis. The example dataset (available at http://sing.ei.uvigo.es/mass-up/downloads/datasets/Cancer-Dataset.zip) is composed of sera from 5 patients with lymphoma, sera from 5 patients with myeloma, and sera from 2 healthy donors. As the dendrogram illustrates, the three conditions are correctly separated since all the samples of each condition are grouped together.

biomarker discovery, that is, the identification of peptides or proteins of a sample able to differentiate specific conditions such as diseases or infections [18]. Following the recommendations given by McDonald [19], Mass-Up includes four different tests of independence (i.e. Fisher's exact test, Yates' chi-square test, Randomization test, Chi-square test) that allow users to identify those peaks that can be potential biomarkers to differentiate the conditions. The test applied in each analysis depends on both the number of samples and the number of peaks present in the dataset.

PCA is a mathematical procedure that can be applied to reduce the dimensionality of a set of samples containing eventually correlated variables (i.e. m/z values), by creating a set of values of linearly uncorrelated variables called principal components (PC). These PC can be used to represent the samples in a 3-dimensional space. By simply assigning different colors to samples' conditions, users can visually identify if there is a separation between conditions and, therefore, they are distinguishable.

Cluster analysis allows finding groups of samples with similar spectral profiles in the dataset. As an unsupervised technique, it allows discovering hidden or previously unknown subgroups of unlabeled samples. When applied to labeled data, it allows researchers to check if the different conditions previously identified in the dataset are separable by means of this technique (see Figure 2). Mass-Up includes an in-house development of an agglomerative, bottom-up hierarchical clustering algorithm.

In previous studies we have proposed a novel workflow for the application of biclustering to MALDI data [20], a simultaneous clustering on both rows and columns. Mass-Up supports this workflow allowing researchers to apply different biclustering algorithms such as Bimax and BiBit and inspect results in an intuitive biclustering viewer.

Finally, sample classification is the ability to predict the label of a sample given a training set of labeled samples, therefore, the capacity of producing a diagnosis machine [21]. Mass-Up provides an interface adapted from the Weka software allowing users to configure a specific classifier and evaluate its performance using different validation schemes. Through this operation, users can determine which classifier performs best for the dataset under study. As a result, users can: (*i*) analyze the performance of the classifier using different statistical measurements (e.g. accuracy, kappa, precision, recall, etc.) and (*ii*) make ROC analyses per condition.

Mass-Up is freely available at http://sing.ei.uvigo.es/mass-up/, where users can find installers for Windows and Linux/MacOS systems along with detailed tutorials, manuals and sample datasets.

5. Concluding Remarks

In this work, we have explored machine learning-based workflows for the analysis of MALDI-TOF mass spectrometry data. The proposed approach enhances typical

MALDI-TOF MS data analysis workflows by adding a quality control step after preprocessing and, by supporting the application of different ML approaches.

With Mass-Up, a multiplatform open-source tool implementing such workflow is provided to the scientific community. Its usefulness is demonstrated by the increasing number of studies that use our solution [22–24] and by the fact that it has been included in public mass spectrometry software repositories and projects, such as MASSyPup(64), the Mass Spectrometry Live Linux, a Puppy Linux based Live distribution that groups several tools focused on the analysis of MS data. A strength of Mass-Up is that it comes within a friendly graphical user interface designed to allow proteomics researchers analyze MALDI-TOF MS data without the need to be bioinformatics experts.

Acknowledgements

H. López-Fernández was supported by a pre-doctoral fellowship from Xunta de Galicia. SING group thanks CITI (*Centro de Investigación, Transferencia e Innovación*) from University of Vigo for hosting its IT infrastructure.

References

1] A.L. Swan, A. Mobasheri, D. Allaway, S. Liddell, J. Bacardit, OMICS J. Integr. Biol. 17 (2013) 595–610. DOI: 10.1089/omi.2013.0017

2] J.S. Morris, K.A. Baggerly, H.B. Gutstein, K.R. Coombes, in: A.J. Rai (Ed.), Urin. Proteome, Humana Press, Totowa, NJ, 2010, pp. 143–166.

3] Y. Perez-Riverol, R. Wang, H. Hermjakob, M. Müller, V. Vesada, J.A. Vizcaíno, Biochim. Biophys. Acta BBA - Proteins Proteomics (n.d.). DOI: 10.1016/j.bbapap.2013.02.032

4] R.A. McDonald, P. Skipp, J. Bennell, C. Potts, L. Thomas, C.D. O'Connor, Expert Syst Appl 36 (2009) 5333–5340. DOI: 10.1016/j.eswa.2008.06.133

5] R. Tibshirani, T. Hastie, B. Narasimhan, S. Soltys, G. Shi, A. Koong, Q.-T. Le, Bioinformatics 20 (2004) 3034–3044. DOI: 10.1093/bioinformatics/bth357

6] H. López-Fernández, M. Reboiro-Jato, D. Glez-Peña, J.R. Méndez Reboredo, H.M. Santos, R.J. Carreira, J.L. Capelo-Martínez, F. Fdez-Riverola, J. Integr. Bioinforma. 8 (2011) 171. DOI: 10.2390/biecoll-jib-2011-171

7] M. Galesio, H. López-Fdez, M. Reboiro-Jato, S. Gómez-Meire, D. Glez-Peña, F. Fdez-Riverola, C. Lodeiro, M.E. Diniz, J.L. Capelo, Steroids 78 (2013) 1226–1232. DOI: 10.1016/j.steroids.2013.08.014

8] M. Reboiro-Jato, D. Glez-Peña, J.R. Méndez-Reboredo, H.M. Santos, R.J. Carreira, J.L. Capelo, F. Fdez-Riverola, Building Proteomics Applications with the Aibench Application Framework, 2011.

9] H. López-Fernández, H.M. Santos, J.L. Capelo, F. Fdez-Riverola, D. Glez-Peña, M. Reboiro-Jato, BMC Bioinformatics 16 (2015). DOI: 10.1186/s12859-015-0752-4

10] K.R. Coombes, K.A. Baggerly, J.S. Morris, (2007) 79–99.

11] H. Shin, M.K. Markey, J. Biomed. Inform. 39 (2006) 227–248. DOI: 10.1016/j.jbi.2005.04.002

12] R. Armañanzas, Y. Saeys, I. Inza, M. García-Torres, C. Bielza, Y. van de Peer, P. Larrañaga, IEEEACM Trans. Comput. Biol. Bioinforma. IEEE ACM 8 (2011) 760–774. DOI: 10.1109/TCBB.2010.18

13] A. Barla, G. Jurman, S. Riccadonna, S. Merler, M. Chierici, C. Furlanello, Brief. Bioinform. 9 (2008) 119–128. DOI: 10.1093/bib/bbn008

14] P. Roy, C. Truntzer, D. Maucort-Boulch, T. Jouve, N. Molinari, Brief. Bioinform. 12 (2011) 176–186. DOI: 10.1093/bib/bbq019

15] P. Larrañaga, B. Calvo, R. Santana, C. Bielza, J. Galdiano, I. Inza, J.A. Lozano, R. Armañanzas, G. Santafé, A. Pérez, V. Robles, Brief. Bioinform. 7 (2006) 86–112.

16] D. Glez-Peña, M. Reboiro-Jato, P. Maia, M. Rocha, F. Díaz, F. Fdez-Riverola, Comput. Methods Programs Biomed. 98 (2010) 191–203. DOI: 10.1016/j.cmpb.2009.12.003

17] S. Gibb, K. Strimmer, Bioinformatics 28 (2012) 2270–2271. DOI: 10.1093/bioinformatics/bts447

18] E.P. Diamandis, Mol. Cell. Proteomics MCP 3 (2004) 367–378. DOI: 10.1074/mcp.R400007-MCP200

19] John H. McDonald, Handbook of Biological Statistics, 2nd ed., Sparky House Publishing, Baltimore, Maryland, 2009.

20] H. López-Fernández, M. Reboiro-Jato, S.C. Madeira, R. López-Cortés, J.D. Nunes-Miranda, H.M. Santos, F. Fdez-Riverola, D. Glez-Peña, in: M.S. Mohamad, L. Nanni, M.P. Rocha, F. Fdez-Riverola (Eds.), 7th Int. Conf. Pract. Appl. Comput. Biol. Bioinforma., Springer International Publishing, 2013, pp. 145–153.

21] M. Hilario, A. Kalousis, C. Pellegrini, M. Müller, Mass Spectrom. Rev. 25 (2006) 409–449. DOI: 10.1002/mas.20072

22] C. Fernández-Costa, M. Reboiro-Jato, F. Fdez-Riverola, C. Ruiz-Romero, F.J. Blanco, J.-L. Capelo-Martínez, Talanta 125 (2014) 189–195. DOI: 10.1016/j.talanta.2014.02.050

23] J.E. Araújo, T. Santos, S. Jorge, T.M. Pereira, M. Reboiro-Jato, R. Pavón, R. Magriço, F. Teixeira-Costa, A. Ramos, H.M. Santos, Anal Methods 7 (2015) 7467–7473. DOI: 10.1039/C5AY00620A

24] R. López-Cortés, J. Formigo, M. Reboiro-Jato, F. Fdez-Riverola, F.J. Blanco, C. Lodeiro, E. Oliveira, J.L. Capelo, H.M. Santos, Talanta (n.d.). DOI: 10.1016/j.talanta.2015.06.043