WEIGHTED NMF FOR HIGH-RESOLUTION MASS SPECTROMETRY ANALYSIS

Rémi Dubroca[(1)], *Christophe Junot*[(2)], *Antoine Souloumiac*[(1)]

[(1)]CEA, LIST,
Laboratoire d'Outils pour l'Analyse de Données
[(2)]CEA, iBiTec-S,
Service de Pharmacologie et d'Immunoanalyse, Laboratoire d'Etude du Métabolisme des Médicaments
F-91191 Gif-sur-Yvette, France.
{remi.dubroca, christophe.junot, antoine.souloumiac}@cea.fr

ABSTRACT

Metabolomics studies are designed to identify, measure, and interpret complex profiles of small organic molecules in different biological samples. Recent high-resolution mass spectrometers, thanks to accurate mass measurements, enable a more reliable compounds identification. The mass spectra matrices considered are nonnegative quantities exhibiting intensity-dependent noise. Among the different Nonnegative Matrix Factorization (NMF) algorithms, the weighted one seems able to handle, with a suitable weighting, this kind of noise. In this paper we compare the performance of NMF and Weigthed NMF (WNMF) in the extraction of different compounds in mass spectrometry data. Numerical experiments on simulated and real data sets illustrate the good behavior and the usefulness of this new method for high-resolution mass spectrometry.

Index Terms— mass spectrometry, nonnegative matrix factorization, multiplicative algorithms, weighted low-rank approximation

1. INTRODUCTION

Nonnegative matrix factorization (NMF) is a useful method as soon as we consider measured nonnegative quantities issued from additive combination of different features [1, 2]. NMF has been used in a large scope of applications as face representation [1] or music analysis [3]. To evaluate the accuracy of the factorization, numerous cost functions exist and the most applied are squared Euclidean distance or Kullback Leibler divergence [1]. Other cost functions rely on generalized divergence such as α or β divergences [4, 5]. The choice of a relevant cost function is made through the analysis of the noise in the data or the data itself.

In this paper, we consider the analysis of mass spectra data. In recent Orbitrap mass spectrometer [6, 7], an intensity-dependent noise appears in the data [8, 9] but due to the complexity of the measure and the novelty of this technology, no noise model has been rigorously proposed. Itakura Saito divergence has been studied as a good choice for factorization of music signals in a multiplicative gamma noise case [3]. In the case of spectrometry data, the lack of information about the mixing model of the compounds, leads to a simple model modified by weighting dealing with uncertainty in the mixing or noise model as the Weighted Nonnegative Matrix Factorization (WNMF). Through the use of weighting matrix, these approachs deal with missing values in distance metrics [10] or collaborative prediction [11] and psychoacoustical masking model [12].

NMF has been recently applied to other technology of mass spectrometer with different mixing model [13]. In Matrix-Assisted Laser Desorption/Ionisation Time-Of-Flight (MALDI-TOF) mass spectrometry [14], sparse coding is highlighted as a simpler ion peak detector in the extracted compounds features. Including uncertainty information coming from replicated measurements in a Least Squares Nonnegative Matrix Factorization (LS-NMF) [15] improves the biological grouping of genes in DNA microarray.

In this paper, we investigate the use of NMF for analyzing the data acquired by the combination of high-performance liquid chromatography (HPLC) and high-resolution mass spectrometry (MS). HPLC-MS is widely used for metabolomics, the study of the collection of small molecules occuring in biological media.

The traditional HPLC-MS analyses go through the use of methods in software like XCMS [16] or MZmine [17]. The first stage in these analyses is the feature detection or peak detection in the retention time / mass-to-charge ratio plane (two dimensions), and feature matching or peak grouping based on compound ions. The interest of the NMF approach is to replace this first stage and thus its capability to be included in a classical workflow. Furthermore, this approach begins with the grouping of features without additional informations, hence the peak detection can be done in the extracted compounds in retention time dimension and mass-to-charge ratio dimension separately.

This paper is organized as follows: Section 2 recalls the NMF and weighted NMF algorithms. In Section 3, we describe the HPLC-MS data and we derive an adapted weighting matrix. The two considered algorithms are then compared in the experimental section 4 and we conclude on the usefulness of our approach.

2. WEIGHTED NMF

We consider a matrix $\mathbf{X} \in \mathbb{R}_+^{T\times M}$ i.e. with nonnegative elements, $\forall t, m \in [1,\dots,T]\times[1,\dots,M], \mathrm{X}_{tm} \geqslant 0$ or $\mathbf{X} \geqslant 0$. The NMF problem is to find the matrices $\mathbf{A} \in \mathbb{R}_+^{T\times K}$ and $\mathbf{S} \in \mathbb{R}_+^{K\times M}$ with nonnegative elements such that:

$$\mathbf{X} \approx \mathbf{AS} = \sum_{k=1}^{K} \mathbf{A}_{.k}\mathbf{S}_{k.}, \tag{1}$$

where K is the number of sources or the rank of the factorization. K is assumed known or overestimated. NMF can be seen as similar decomposition like Singular Value Decomposition (SVD) or Independent Component Analysis (ICA) [18] for non-negative data [19] with less independence constraints. The conventional way to find $\mathbf{A}$ and $\mathbf{S}$ is to minimize a cost function C dealing with a measure of fit D between $\mathbf{X}$ and $\mathbf{AS}$:

$$\begin{aligned} \min_{\mathbf{A}\geqslant 0,\mathbf{S}\geqslant 0} C(\mathbf{A},\mathbf{S}) &= \min_{\mathbf{A}\geqslant 0,\mathbf{S}\geqslant 0} D(\mathbf{X}|\mathbf{AS}), \\ D(\mathbf{X}|\mathbf{AS}) &= \sum_{t=1}^{T}\sum_{m=1}^{M} d(\mathrm{X}_{tm}|(\mathrm{AS})_{tm}), \end{aligned} \tag{2}$$

D is a separable function and d is a scalar function. C is minimized in an iterative alternating way. At each iteration, C is minimized with respect to S with A fixed, then C is minimized with respect to A with S fixed. The measure of fit between $\mathbf{X}$ and $\mathbf{AS}$ can be a distance or a divergence.

A measure of fit with Euclidian distance yields:

$$\begin{aligned} D_{EUC}(\mathbf{X}|\mathbf{AS}) &= \tfrac{1}{2}\|\mathbf{X}-\mathbf{AS}\|^2 \\ &= \tfrac{1}{2}\sum_{t=1}^{T}\sum_{m=1}^{M}(\mathrm{X}_{tm}-(\mathrm{AS})_{tm})^2, \end{aligned} \tag{3}$$

A first approach to minimize (3) has been introduced by [1] with the multiplicative update algorithm:

NMF EUC
Input: $\mathbf{X} \in \mathbb{R}_+^{T\times M}$, K
Output: $\mathbf{A} \in \mathbb{R}_+^{T\times K}$ and $\mathbf{S} \in \mathbb{R}_+^{K\times M}$

1. Initial values: $\mathbf{A}_0$, $\mathbf{S}_0$.
2. Repeat:
   - $\mathbf{S}_{k+1} = \mathbf{S}_k \odot \frac{\mathbf{A}_k^T\mathbf{X}}{\mathbf{A}_k^T(\mathbf{A}_k\mathbf{S}_k)}$
   - $\mathbf{A}_{k+1} = \mathbf{A}_k \odot \frac{\mathbf{X}\mathbf{S}_{k+1}^T}{(\mathbf{A}_k\mathbf{S}_{k+1})\mathbf{S}_{k+1}^T}$
3. Solutions : $\mathbf{A}$ and $\mathbf{S}$.

where $\odot$ and — are respectively element-wise multiplication and division. The updates of $\mathbf{S}$ and $\mathbf{A}$ are then done with a multiplication by a nonnegative factor, hence preserving the nonnegativity.

To deal with uncertainty in the model and/or the data $\mathbf{X}$, the cost function D_{EUC} can be weighted, giving:

$$D_{WEUC}(\mathbf{X}|\mathbf{AS}) = \frac{1}{2}\sum_{t=1}^{T}\sum_{m=1}^{W}\mathrm{W}_{tm}(\mathrm{X}_{tm}-(\mathrm{AS})_{tm})^2, \tag{4}$$

where $\mathbf{W} \in \mathbb{R}_+^{T\times M}$ is a nonnegative weighting matrix. Multiplicative updates are proposed in [10]:

WNMF EUC
Input: $\mathbf{X} \in \mathbb{R}_+^{T\times M}$, $\mathbf{W} \in \mathbb{R}_+^{T\times M}$, $0 \leqslant \mathbf{W} \leqslant 1$ and K
Output: $\mathbf{A} \in \mathbb{R}_+^{T\times K}$ and $\mathbf{S} \in \mathbb{R}_+^{K\times M}$

1. Initial values: $\mathbf{A}_0$, $\mathbf{S}_0$.
2. Repeat:
   - $\mathbf{S}_{k+1} = \mathbf{S}_k \odot \frac{\mathbf{A}_k^T(\mathbf{W}\odot\mathbf{X})}{\mathbf{A}_k^T(\mathbf{W}\odot\mathbf{A}_k\mathbf{S}_k)}$
   - $\mathbf{A}_{k+1} = \mathbf{A}_k \odot \frac{(\mathbf{W}\odot\mathbf{X})\mathbf{S}_{k+1}^T}{(\mathbf{W}\odot\mathbf{A}_k\mathbf{S}_{k+1})\mathbf{S}_{k+1}^T}$
3. Solutions : $\mathbf{A}$ and $\mathbf{S}$.

3. HIGH RESOLUTION HPLC-MS

In this paper, we will study data acquired by the combination of HPLC and high-resolution LTQ-Orbitrap instrument [6, 7].

The liquid chromatography realises a separation of the different compounds in a biological or chemical sample giving a time dimension in the acquired data. Chromatographically resolved compounds are ionized in the source of the mass spectrometer, and then discriminated in the analyzer according to their mass-to-charge ratio (m/z). This gives the m/z dimension of the data. The abundance (positive value) of the trapped ions are measured in a detector.

In the model (1), the matrix $\mathbf{X} \in \mathbb{R}_+^{T \times M}$ stacks the T different m/z spectra $\mathbf{X}_{t.}$ row vectors of length M, measured over time. The NMF performs the separation of the K compounds, each defined by an elution profile $\mathbf{A}_{.k}$ and a mass spectrum $\mathbf{S}_{k.}$.

The noise in the abundance measurement has an absolute part and an intensity-dependent one [8, 9]. We consider the model:

$$\mathrm{X}_{tm} = \mathrm{Y}_{tm} + \mathrm{B}_{tm}, \tag{5}$$

with B_{tm}, X_{tm}, Y_{tm}, noise, noise corrupted abundance and true abundance. We observed on real data that the noise variance can be approximately written as:

$$\mathrm{Var}(\mathrm{B}_{tm}) \approx \sigma_A^2 + \sigma_M^2 \mathrm{Y}_{tm}^2, \tag{6}$$

where σ_A and σ_M are empirically tuned. Consequently, it is natural to normalize this signal dependent noise variance via the following weighting matrix $\mathbf{W}$:

$$\mathbf{W} = \frac{1}{\sqrt{1 + (\alpha \mathbf{X}_S)^2}}, \tag{7}$$

where $\sqrt{.}$ and $.^2$ are applied element-wise and $\alpha = \sigma_M / \sigma_A$. $\mathbf{X}_S$ is the temporally smoothed (filtering of the columns) matrix $\mathbf{X}$ and α deals with the predominance of the intensity-dependent (multiplicative) noise over the absolute one (additive). This weighting matrix is non negative and takes its values between 0 and 1.

4. EXPERIMENTAL RESULTS

The HPLC-MS data sets (simulated and real) treated here are mixtures of commercial chemical compounds [20]. We compare performance of Euclidian distance based algorithms, with or without weighting, denoted thereafter NMF and WNMF. The two algorithms are initialized with the same starting points:

- $\mathbf{A}_0$ is initialized with the elution profiles of the compounds determined by the previous analysis [20]. In a real world scenario, these elution profiles can be read for the most abundant compounds from the most intensive peaks in $\mathbf{X}$.
- $\mathbf{S}_0$ is randomly initialized.

We set the number of iterations for the two algorithms to 200. Our experiments show that the best compromise is obtained with the α parameter in (7) set to 10^{-5}.

4.1. Simulated Data set

In [20], the full mass spectra extraction of the chemical compounds is done with flow-injection analysis (FIA), that is, direct injection of each compound separately. This method limits the ion suppression phenomenon occuring during the ionization of different mixed compounds and allows the full mass spectra extraction of all the compounds.

Combining the full mass spectra from FIA-MS and the previously extracted elution profile from HPLC-MS, we generate realistic compound analysis presented in Fig. 1. We present the four different columns of the matrix $\mathbf{A}$ (elution profile), rows of the matrix $\mathbf{S}$ (mass spectrum) and the contribution of each source (compound) to the matrix $\mathbf{X}$. We display retention time between 11 and 13 minutes and mass-to-charge ratio between 80 and 680 m/z.

In the contribution view, we see more easily the isotopic clusters (chromatographic peaks close in m/z) in the retention time / mass-to-charge ratio plane. Even with intensities really lower than some main lone peaks, these groups are more visible because of the thickness of the m/z peaks.

The mixing is done according to the following model:

$$\mathbf{X} = [\mathbf{AS} \odot \mathbf{N}_M + \mathbf{N}_A]_0, \tag{8}$$

where the non linear operator $[v]_0$ used to avoid negative values is defined as $[v]_0 = \max\{v, 0\}$. $\mathbf{N}_M$ and $\mathbf{N}_A$ are randomly drawn matrices following $\mathcal{N}(1, \sigma_M^2)$ and $\mathcal{N}(0, \sigma_A^2)$ normal distributions respectively. They are accounting respectively for intensity-dependent and absolute part of noise.

Fig. 2 presents the Mean Squared Error (MSE) between the extracted and simulated mass spectra for differents σ_M with $\sigma_A = 1000$. σ_M varies between 1×10^{-4} and 1×10^{-1}. 100 Monte-Carlo realizations have been done for each value of σ_M. The mean of the MSE is taken over the four mixed sources. WNMF produces better results than NMF even in the near pure additive noise case ($\sigma_M = 1 \times 10^{-4}$).

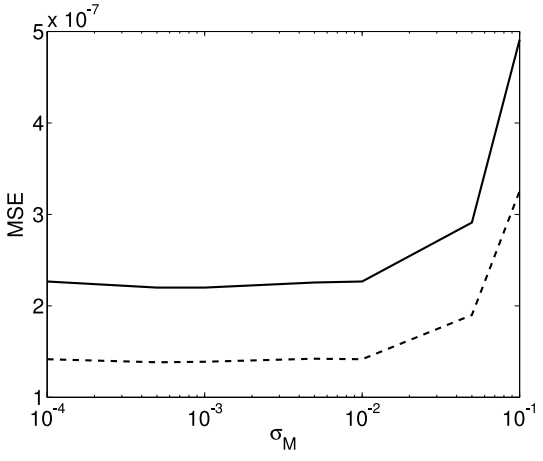

Fig. 2. Average MSE of the four extracted mass spectra over 100 Monte-Carlo realizations: NMF (solid lines) and WNMF (dashed lines)

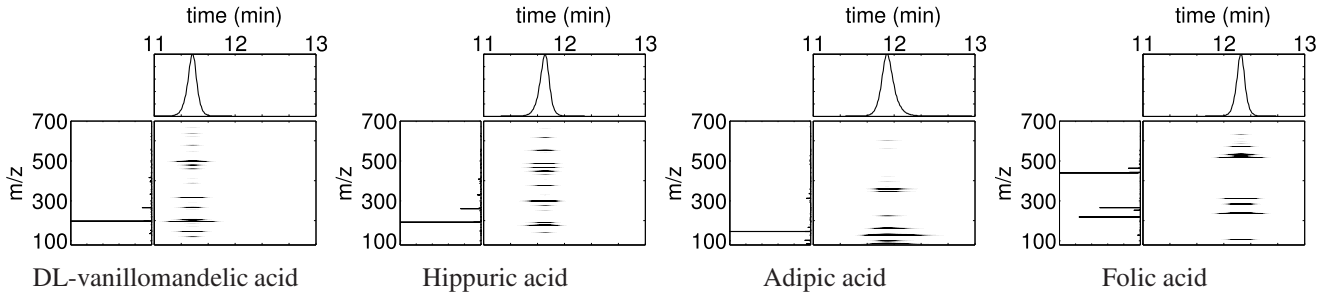

Fig. 1. Simulated HPLC-MS spectra: maximum amplitude are 1 for the mass spectra and 5×10^6 for elution profiles

4.2. Real Data set

We study the same time span for the real data set. Some of the theoretically expected ions were not detected, due to low ionization recoveries for some compounds, and perhaps also to ion suppression phenomenon, the mass-to-charge ratio spanned the 60-450 m/z.

Fig. 3 shows the analysed spectrum. This set consists of five compounds, so we set the number of sources K to 5. The four first columns of $\mathbf{A}_0$ are initialized with the elution profiles of the known compounds, the last column and $\mathbf{S}_0$ are randomly inititialized.

The separated elution profiles are drawn in Fig. 4 for NMF and WNMF. The fourth known compound is not present in the result given by the NMF, even if one of the column of $\mathbf{A}_0$ is initialized with its elusion profile. The NMF algorithm spreads the third known compound on two components, overfitting its elusion profile. The overfitting of high intensity peaks happens at the expense of the accuracy of the lower intensity peaks. The recognition of a compound from its ionized forms is easier when we have more ions in the mass spectrum. The WNMF exhibits a better behavior as it gives all the expected compounds in the component.

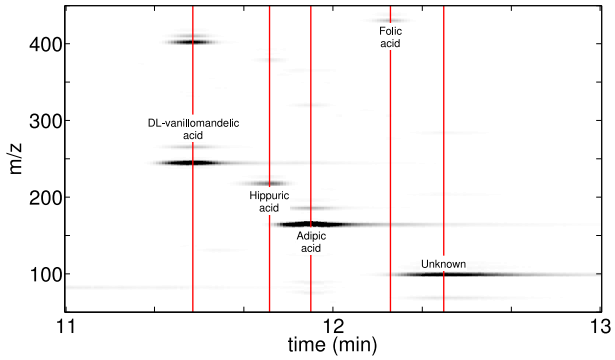

Fig. 3. HPLC-MS spectrum.

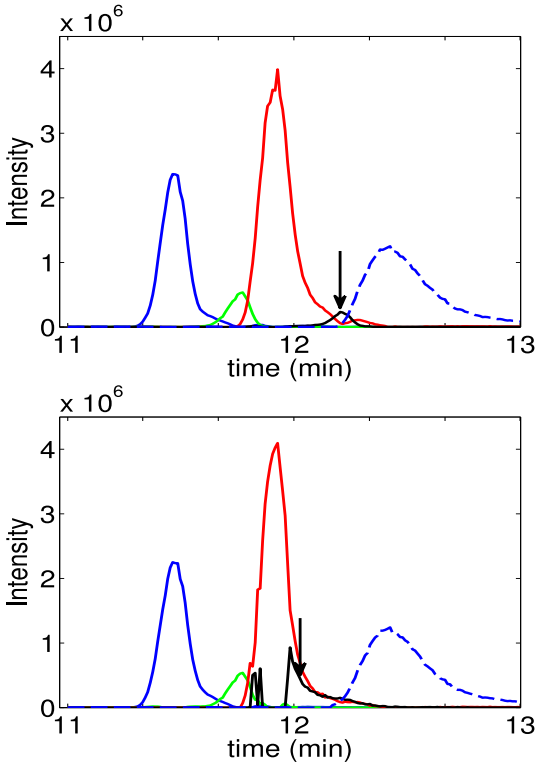

Fig. 4. Separated elusion profiles, real data set: recognized compounds (solid lines) and unknown compound (dashed line), from top to bottom, WNMF and NMF.

5. CONCLUSION

We have presented the problem of high resolution HPLC-MS analysis using nonnegative matrix factorization. The weighted nonnegative matrix factorization was adapted to the intensity-dependent noisy spectrum data. Experiments on simulated and real data sets show a promising behavior of the WNMF over the classical NMF: the non overfitting decomposition allows to discover more compounds and more accurately. We expect to get even better results with algorithms using non multiplicatives updates as the Alternating Nonnegative Least Squares update.

6. REFERENCES

[1] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, Oct. 1999.

[2] A. Cichocki, R. Zdunek, and S. Amari, "Csiszars divergences for non-negative matrix factorization: Family of new algorithms," in *Proc. ICA 2006*, Mar. 2006, pp. 32–39.

[3] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, Mar. 2009.

[4] A. Basu, I. R. Harris, N. L. Hjort, and M. C. Jones, "Robust and efficient estimation by minimising a density power divergence," *Biometrika*, vol. 85, no. 3, pp. 549–559, Sep. 1998.

[5] M. Minami and S. Eguchi, "Robust blind source separation by Beta divergence," *Neural Computation*, vol. 14, no. 8, pp. 1859–1886, Aug. 2002.

[6] A. Makarov, "Electrostatic axially harmonic orbital trapping: A high-performance technique of mass analysis," *Analytical Chemistry*, vol. 72, no. 6, pp. 1156–1162, Mar. 2000.

[7] A. Makarov, E. Denisov, A. Kholomeev, W. Balschun, O. Lange, K. Strupat, and S. Horning, "Performance evaluation of a hybrid linear ion trap/Orbitrap mass spectrometer," *Analytical Chemistry*, vol. 78, no. 7, pp. 2113–2120, Apr. 2006.

[8] C. Hundertmark, R. Fischer, T. Reinl, S. May, F. Klawonn, and L. Jnsch, "MS-specific noise model reveals the potential of iTRAQ in quantitative proteomics," *Bioinformatics*, vol. 25, no. 8, pp. 1004–1011, Apr. 2009.

[9] F. P. Breitwieser, A. Müller, L. Dayon, T. Köcher, A. Hainard, P. Pichler, U. Schmidt-Erfurth, G. Superti-Furga, J.-C. Sanchez, K. Mechtler, K. L. Bennett, and J. Colinge, "General statistical modeling of data from protein relative expression isobaric tags," *Journal of Proteome Research*, vol. 10, no. 6, pp. 2758–2766, Jun. 2011.

[10] Y. Mao and L. K. Saul, "Modeling distances in large-scale networks by matrix factorization," in *Proc. IMC 2004*, Oct. 2004, pp. 278–287.

[11] Y.-D. Kim and S. Choi, "Weighted nonnegative matrix factorization," in *Proc. ICASSP 2009*, Apr. 2009, pp. 1541–1544.

[12] J. Nikunen and T. Virtanen, "Noise-to-mask ratio minimization by weighted non-negative matrix factorization," in *Proc. ICASSP 2010*, Mar. 2010, pp. 25 –28.

[13] P.W. Siy, R.A. Moffitt, R.M. Parry, Y. Chen, Y. Liu, M.C. Sullards, A.H. Merrill, and M.D. Wang, "Matrix factorization techniques for analysis of imaging mass spectrometry data," in *Proc. BIBE 2008*, Oct. 2008, pp. 1–6.

[14] T. Alexandrov, K. Steinhorst, O. Keszoecze, and S. Schifflerangvill, "Peak detection in mass spectrometry data using sparse coding," in *Proc. COMPSTAT 2010*, Aug. 2010, p. 373.

[15] G. Wang, A. Kossenkov, and M. Ochs, "LS-NMF: A modified non-negative matrix factorization algorithm utilizing uncertainty estimates," *BMC Bioinformatics*, vol. 7:175, Mar. 2006.

[16] H. P. Benton, D. M. Wong, S. A. Trauger, and G. Siuzdak, "XCMS2: Processing tandem mass spectrometry data for metabolite identification and structural characterization," *Analytical Chemistry*, vol. 80, no. 16, pp. 6382–6389, Aug. 2008.

[17] T. Pluskal, S. Castillo, A. Villar-Briones, and M. Oresic, "MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data," *BMC Bioinformatics*, vol. 11:395, Jul. 2010.

[18] P. Comon, "Independent component analysis, a new concept?," *Signal Processing*, vol. 36, no. 3, pp. 287–314, Apr. 1994.

[19] M.D. Plumbley, "Conditions for nonnegative independent component analysis," *IEEE Signal Processing Letters*, vol. 9, no. 6, pp. 177–180, Jun. 2002.

[20] Y. Xu, J.-F. Heilier, G. Madalinski, E. Genin, E. Ezan, J.-C. Tabet, and C. Junot, "Evaluation of accurate mass and relative isotopic abundance measurements in the LTQ-Orbitrap mass spectrometer for further metabolomics database building," *Analytical Chemistry*, vol. 82, no. 13, pp. 5490–5501, Jun. 2010.