analytical chemistry

Letter

pubs.acs.org/ac

New Approach for Chemometric Analysis of Mass Spectrometry Data

Frutos C. Marhuenda-Egea,*[,†] Rubén D. Gonsálvez-Álvarez,[‡] Belén Lledó-Bosch,[‡] Jorge Ten,[§] and Rafael Bernabeu[§]

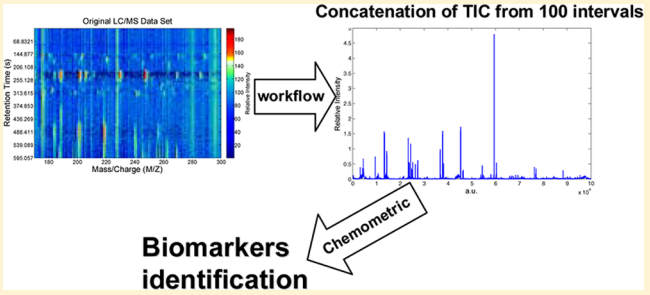
[†]Department of Agrochemistry and Biochemistry, University of Alicante, Apartado 99, E-03080 Alicante, Spain

[‡]Department of Molecular Biology, Instituto Bernabeu, Alicante, Spain

[§]Department of Reproductive Medicine, Instituto Bernabeu, Alicante, Spain

Supporting Information

ABSTRACT: The search of metabolites which are present in biological samples and the comparison between different samples allow the construction of certain biochemical patterns. The mass spectrometry (MS) methodology applied to the analysis of biological samples makes it possible for the identification of many metabolites. Each obtained signal (m/z) is characteristic of a particular metabolite. However, the mass data (m/z) interpretation is difficult because of the large amount of information that they contain. In this work, we present a relatively simple tool that allows us to deal with the whole of the mass information from the chemometric analysis. The statistical analysis is a key stage in order to identify the metabolites involved in a particular biochemical pattern. We transformed the mass data matrix in a vector. By having the data as a vector, it was possible to keep all the information and also avoid the signals overlapping, which is the major problem when the total ion chromatogram (TIC) is obtained. In the approach proposed here, the mass data (m/z) matrix was split in 100 different TIC in order to avoid the signal overlapping. The 100 chromatograms were concatenated in a vector. This vector, which can be plotted as a continuous (2D pseudospectrum), greatly simplifies for one to understand the subsequent dimensional multivariate analysis. To validate the method, 19 samples from two human embryos culture medium were analyzed by high-pressure liquid chromatography−mass spectrometry (HPLC−MS). Our methodology would be applied to the obtained raw data. Later on, a multivariate analysis was conducted using a robust principal components analysis interval (robPCA) and interval partial least squares algorithm (iPLS). The results obtained allow one to differentiate the two sample populations undoubtedly, although their composition was similar.



Currently, the study of metabolites present in biological samples appears to be the key to the interpretation of certain biochemical cycles. Metabolomics is the branch of the biochemical sciences that studies the low molecular weight molecules. Sophisticated analytical techniques such as mass spectrometry (MS) coupled to a chromatographic technique (gas chromatography (GC), high-pressure liquid chromatography (HPLC), ...) and/or nuclear magnetic resonance (NMR) are required.[1] Mass spectrometry is a technique broadly used in the study of biological samples. By choosing between different ionization methods (i.e., electrospray ionization (ESI), matrix-assisted laser desorption ionization (MALDI)...) and using different mass analyzers (time-of-flight (TOF), ion trap.....), it is possible to obtain valuable information in order to identify the sample under study. By coupling MS and HPLC (HPLC−MS), each signal obtained on the chromatogram (which has a characteristic retention time) has linked a molecular mass (m/z). In addition, HPLC−MS allows the quantification, and it is a reproducible technique. For the identification of metabolites, it is necessary to use a database on the basics of accurate mass and tandem mass spectrometry (MS/MS) data of model compound for structural confirmation.[2−5]

The raw data obtained from these analyses are very complex due to the large amount of information that they contained. An adequate selection of the mass analyzer enables the selection of certain molecular mass range, which greatly simplifies the amount information. However, in a biological sample there are a lot of low molecular weight metabolites which (m/z) lie within a similar range. In order to understand this biochemical information of mass-spectrometry data, several software tools have been developed (e.g., metaXCMS,[6] XCMS,[7,8] Mzmine,[9,10] and MathDAMP[11]). This type of programs identifies the variations in the relative intensities of the signals (features) obtained in the analysis of different groups of samples. These signals can therefore serve as biomarkers in a given process (e.g., sick/healthy[12]). These software tools worked with the total mass spectrum. For the beginners in metabolomic research, their internal operations could be difficult to understand.

Received: November 12, 2012
Accepted: February 8, 2013
Published: February 8, 2013

   dx.doi.org/10.1021/ac303255h | *Anal. Chem.* 2013, 85, 3053−3058

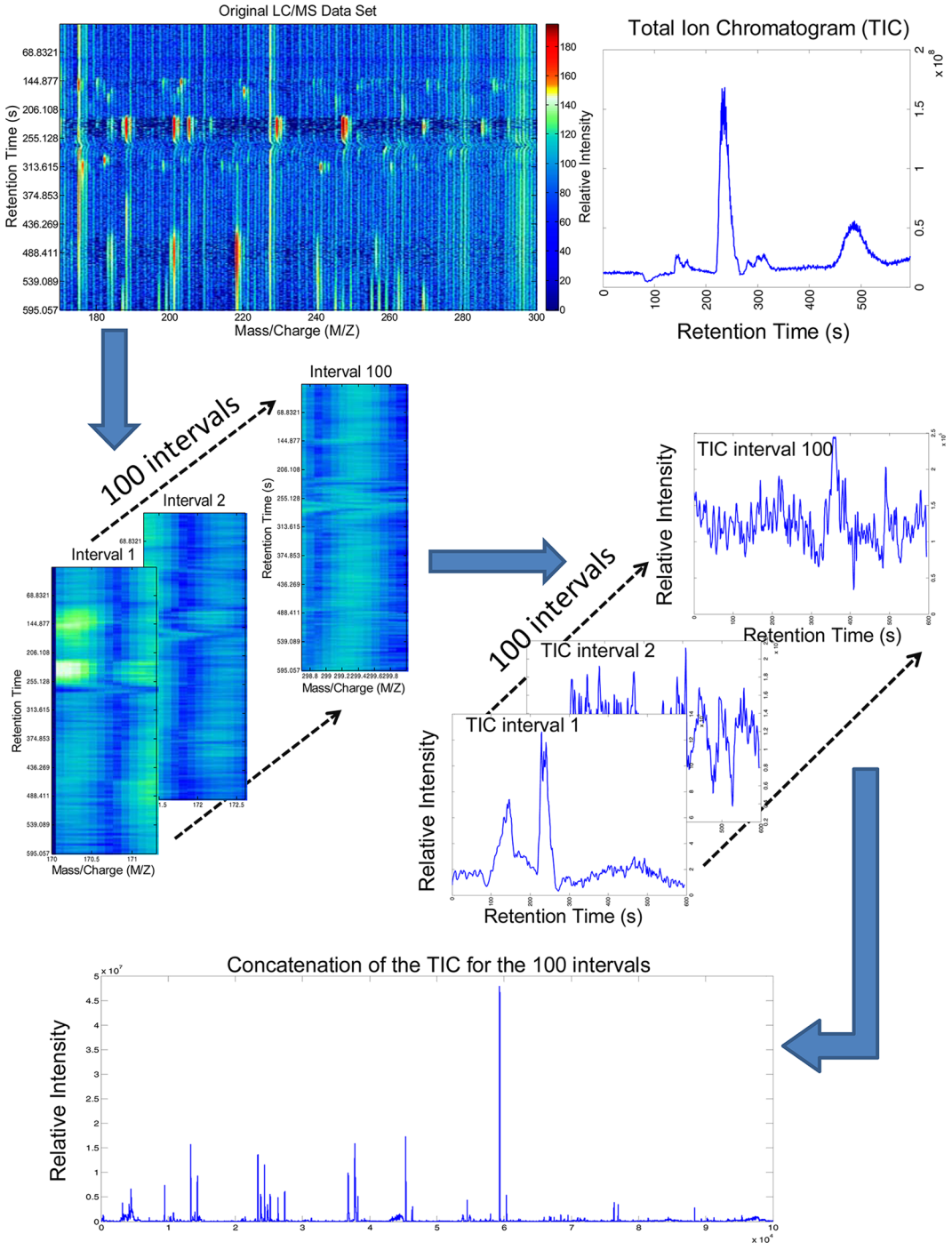


Figure 1. Raw mass data transformation to 2D pseudospectra. Raw data was visualized as a LC−MS data set and a total ion chromatogram (TIC). Raw data was divided in 100 intervals in the mass/charge (m/z) dimension. After that, the TIC for each interval was concatenated in a unique vector (2D pseudospectra).

Other studies proposed to look for unregulated metabolites present in certain biological patterns, as phenotypes, but the comparison of these patterns is very complicated due to the large number of relevant elements (i.e., metabolites) in the proposed study. A software tool such as MetaXCMS[6] proposes a reduction of total data in order to work only with the relevant data for the study. The mass-spectrometry data are not metabolites but features. The question to identify a feature as a metabolite is another problem, such as it is described above.

A novel approach for the analysis of mass-spectrometry features without data reduction is presented in this work. Our aim was to include in the statistical analysis all the chemical information contained in the mass data because a priori the relevant information was very difficult to select.

In this way we would not lose any molecular mass (selected range), although a very small signal gain. The mass-spectrometry data (matrix ($m \times n$)) are transformed in a pseudo-2D spectrum (vector ($1n$)). There are a lot of chemometric tools for working with data in the format of 2D spectra data. To work with these pseudo-2D spectra simplified preprocessing (i.e., alignment) and, hence, multivariate analysis. Moreover the relative signal intensity obtained was directly

dx.doi.org/10.1021/ac303255h | *Anal. Chem.* 2013, 85, 3053−3058

proportional to the concentration, so that the metabolites could be quantified.

WORKFLOW

The mass spectrometer raw data are saved as <folder_name>.D. These folders contain several files (data file saved as <analysis.yep> file). It was necessary to convert the raw data (<analysis.yep> file) in to mzXML format, using CompassXport (Bruker Daltonics free software, http://www.bdal.com/service-support/software-support-downloads.html). Once the data are in mzXML format, it is possible to make an initial visual data inspection by using software such as Msight (http://web.expasy.org/MSight/), Openchrom (http://www.openchrom.net), or MZmine[13] (Mzmine also has the possibility to perform statistical analysis). These softwares are available online for free.

After this first data inspection, the mzXML data file was opened with MatLab using the MZXMLREAD function. This function reads the XML document into a MATLAB structure:

```
mzXML_name = mzxmlread('name_datafile.mzXML')
```

The MZXML2PEAKS function extracts the list of peaks from each scan into a cell array (*peaks*) and their respective retention time into a column vector (*time*), in order to work easier with the mass data:

```
[peaks_name,time_name]
  = (mzxml2peaks(mzXML_name)
```

To visualize the data set, MATLAB can create a common grid in the mass/charge dimension. By choosing the appropriate parameters for the MSPPRESAMPLE function, we can ensure that the resolution of the spectra was not lost:

```
[MZ, Y] = msppresample(peaks_name, 5000)
```

Regarding the ion intensities matrix, *Y*, it is possible to create a colored heat map. The MSHEATMAP function automatically adjusts the color bar utilized to show the statistically significant peaks with hot colors and the noisy peaks with cold colors. When we work with heat maps, it is common to display the ion intensities logarithm, which enhances the dynamic range of the color map (mass range from 170 to 300 amu) (Figure 1):

```
fh1 = msheatmap(MZ, time log(Y), 'resolution', .1,
    'range', [170 300]
```

As a next optional step to improve the image, we can apply a Gaussian filter in the chromatographic direction to smooth the whole data set:

```
Gpulse = exp(−.1*(−10:10).^ 2)
        /sum(exp(−.1*(−10:10).^ 2))
YF = convn(Y, Gpulse, 'same')
```

In the Matlab document "Visualizing and Preprocessing Hyphenated Mass Spectrometry Data Sets for Metabolite and Protein/Peptide Profiling" (www.mathworks.es), it is possible to obtain a more detailed information and more data treatments.

In this smoothed ion intensities matrix, *YF*, the *x* axis as *m*/*z*, the *y* axis as retention time in the HPLC, and the *z* axis as intensity of the signal (Figure 1). The contour maps analysis is very difficult and complex. The *YF* matrix dimension was resolution × scan number. The resolution was fixed in 5000 and the scan number was 1400. The *YF* matrix was divided in 100 intervals in the resolution dimension (*m*/*z* axis) that corresponded to a mass working range from 170 to 300 amu.

For each interval, the total ion chromatogram (TIC) was obtained, adding the 50 row batches (we have 5000 rows in the *YF* matrix (5000 × 1400)). After this, the 100 TIC were added one after other one. Finally, a linear data representation similar to spectroscopy data (vector (1 × *m*)), 2D pseudospectrum was performed (Figure 1). In this way, overlapping among features present in the total ion chromatogram were avoided. The algorithm performed had the instructions next:

```
mzXML_name = mzxmlread('name.mzXML')

[peaks_name, time_name]
  = mzxml2peaks(mzXML_name)

[MZ, Y] = msppresample(peaks_5000, 'Range',
  [170, 300])

Gpulse = exp( − .1*( − 10:10).^2)
  /sum(exp( − 1*( − 10:10).^2))
YF = convn(Y, Gpulse, 'same')

Intervals(1,:) = sum(YF(1:50, Intervals 1:1400))
Intervals(2,:) = sum(YF(51:100, Intervals 1:1400))
Intervals(3,:) = sum(YF(101:150, Intervals 1:1400))

Intervals(99,:) = sum(YF(4901:4951, Intervals 1:1400))
Intervals(100,:) = sum(YF(4951:5000, Intervals 1:1400))

Sample_name = cat(2, Intervals(1,:), Intervals(2,:),
  Intervals(3,:), ..., Intervals(99,:), Intervals(100,:))
```

The next step, in our approach was a peak alignment between the sample data, using an algorithm for this task (icoshift). When the peaks were aligned, robust principal components analysis with intervals (robPCA) and partial least-squares with intervals (iPLS) were used in order to classify the samples (iToolboox). The selection of the multivariate tool depends on previous design of the experience or on the particular research approach (e.g., classification tools or regression tools).

EXPERIMENTAL SECTION

Here we describe the experimental application of our methodology. We have considered samples of two different human embryo culture media: CM (Cook Medical) and CCM (Vitrolife, Goteborg, Sweden) media.

The following groups of samples were compared: (1) three days culture medium (9 samples) and (2) five days culture medium (10 samples). The culture media described above were obtained from the clinic. The samples are the remaining culture media after the embryos were transferred into the uterus. The in vitro fertility (IVF) protocol was described in the Supporting Information. The samples for analysis were prepared as follows: 25 μL of culture medium (CM or CCM), 10 μL of citrulline 1 mM as internal standard, and 15 μL of H_2O miliQ were mixed in a tube and microinjected into the HPLC−MS system. The

dx.doi.org/10.1021/ac303255h | Anal. Chem. 2013, 85, 3053−3058

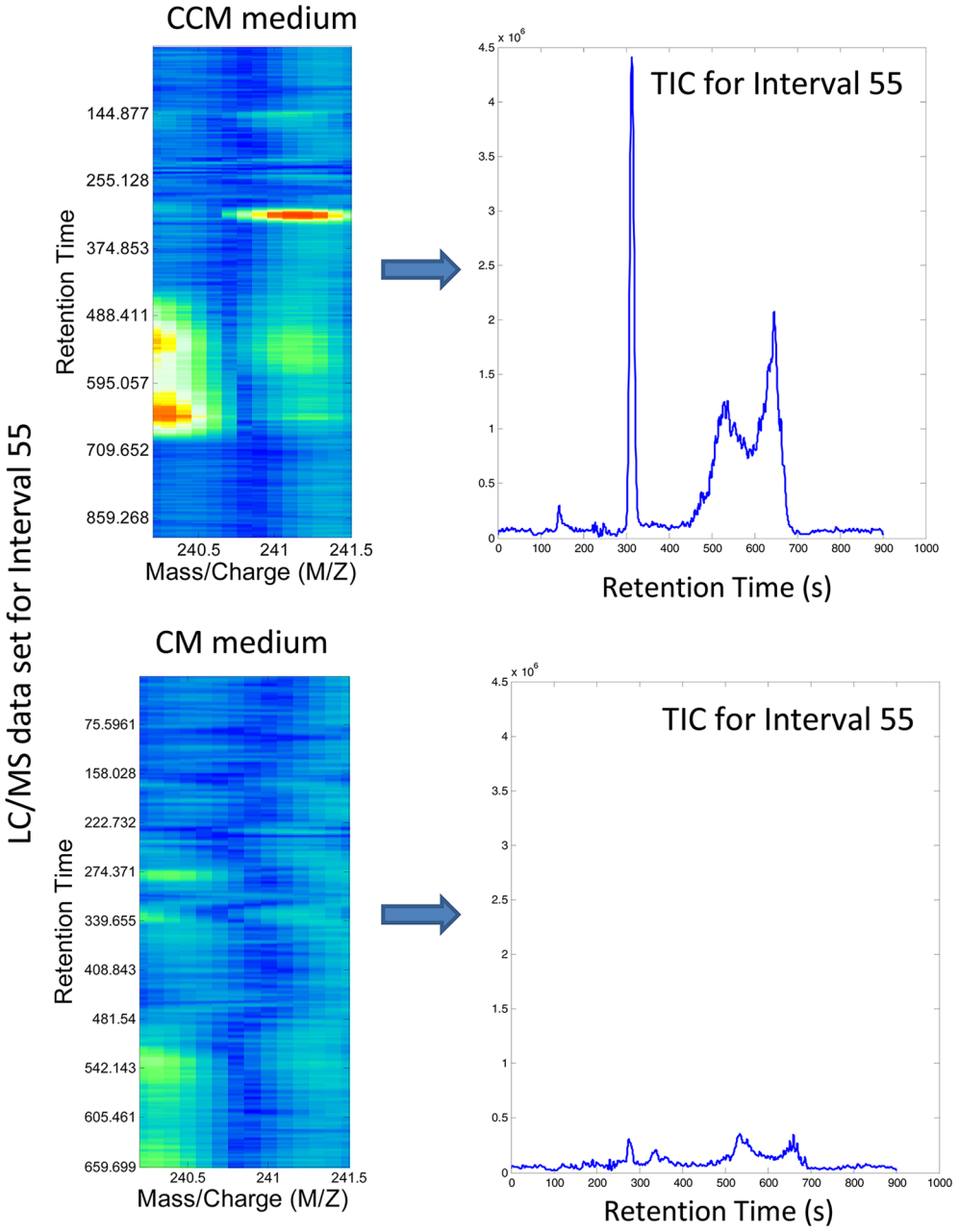


Figure 2. LC–MS data set and a total ion chromatogram (TIC) for interval 55 from CCM (figure top) and CM (figure bottom).

liquid chromatography coupled to tandem mass spectrometry with an electrospray ionization (LCESI–MS) analysis was performed as described in (Supporting Information).[14,15]

MATLAB version 6.5 from MathWorks was used for the calculations, and the iToolbox (iPLS) and icoshift (peak alignment) are available from http://www.models.kvl.dk. Multivariate data analysis by robust principal component analysis (robPCA) was carried out using the LIBRA toolbox (available at www.wis.kuleuven.ac.be/stat/robust.html).

RESULTS AND DISCUSSION

After processing the raw data as described in Workflow section, we obtained a 2D-pseudospectrum composed by the concatenation of the 100 TIC, which is displayed in Figure 1. Having all the 2D-pseudospectra data from the samples, a matrix ($m \times n$, where m was the sample and n was the 2D-pseudospectrum data) was built. In the next step, the 2D-pseudospectra were aligned (icoshift algorithm). With the aligned data, we started the multivariate statistical analysis.

In Matlab there are a lot of chemometric tools for the analysis of spectroscopic data (i.e., principal components analysis (PCA), regression analysis, independent components (IC), genetic algorithms, etc.) The aligned data also can be exported (or copy and paste) to other pieces software such as Excel, R, SPSS, The Unscrambler, etc. The aligned data was only a matrix ($m \times n$, where m was the sample and n was the alienated 2D-pseudospectra).

PCA is a data visualization method that it is useful to observe groupings within multivariate data. The data is represented in n dimensional space, where n is the number of variables, and it is reduced into a few principal components, which are descriptive dimensions that describe the maximum variation within the data.[16] Principal components (PC) can be displayed as a "scores" plot. The score plot is useful to observe any groupings

dx.doi.org/10.1021/ac303255h | *Anal. Chem.* 2013, 85, 3053–3058

in the data set. PCA models are constructed using all the samples under study. The coefficients by which the original variables must be multiplied to obtain the PC are called "loadings". The numerical value of a loading for a given variable on a PC shows how much the variable has in common with that component. The differences between the two human embryo culture media were found by multivariate analysis using PCA (Figure S-1 in the Supporting Information). The top three principal components (PCs) captured 98.8% of the variability. The analysis of PCA results (Figure S-1 in the Supporting Information) shows the dependence of the human embryo culture media samples on the PC loadings. The loadings plots, when viewed as line plots, resemble a 2D-pseudospectra data and they can be interpreted as such (Figure S-1 in the Supporting Information). The loadings of PC1 were determined for different signals (Figure S-1 in the Supporting Information).

The iPLS models[17,18] for 2D-pseudospectra data displayed good values of RMSECV (0.0672) by human embryo culture media, only with the interval 55 between 54 001 to 55 000 points (Figure S-2 in the Supporting Information). The peaks selected by the multivariante data analysis is a feature in the original raw data (the peak with m/z 241.1). The 100 intervals used in the transformation used have a length of 100 000 points with these particular data (see the Workflow section). With our data, the m/z resolution was fixed in 5000 points in the range of 170 to 300 m/z; therefore, each 50 points we had a increase around of 1.3 m/z. If we selected the range between X to X + 1000 (interval length) in the multivariate data analysis, the features were in the m/z range of Y to Y + 1.3. These intervals can be showed as the original mass spectra with the total ion chromatogram for the interval (Figure 2). In other words, the peaks selected in the different multivariate data analysis were real features in the raw data. It can be considered as the more important question in our approach.

CONCLUSIONS

The transformation of the raw data presented here simplified greatly the chemometric analysis of the metabolomic data from HPLC−MS. A 2D pseudospectrum was performed with several TIC, specifically 100. The 2D pseudospectra from the samples can be used in different statistical software, in order to solve metabolomic questions, such as finding metabolites associated with a particular phenotype. Our metabolomic workflow offers a very useful and easy method to identify features with biological relevance prior to work for chemical identification. An important question in metabolomic research should be to make efforts in order to standardize the methodology. Our metabolomic workflow moves in this direction.

ASSOCIATED CONTENT

Supporting Information

Additional information as noted in text. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: frutos@ua.es.

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This paper is dedicated to José González, Ph.D. (†1967−2012), Professor of Physical Chemistry, Alicante University. This work has been supported by grants from Instituto Bernabeu (Grant INSTITUTOBERNABEU1-08I) and the University of Alicante (Grant UAUSTI09-08) Project.We thank Dr. J. L. Todolí and M.Phil. B. Gomis for the manuscript revision and Dr. P. Candela for the technical support.

REFERENCES

(1) Viant, M. R.; Sommer, U. *Metabolomics*. Published online March 9, **2012**; http://link.springer.com/content/pdf/10.1007%2Fs11306-012-0412-x.

(2) Warwick B. Dunn, W. B.; Erban, A.; Weber, R. J. M.; Creek, D. J.; Brown, M.; Breitling, R.; Hankemeier. T.; Goodacre, R.; Neumann, S.; Kopka, J.; Viant, M.R.. *Metabolomics*. Published online May 26, **2012**; http://link.springer.com/content/pdf/10.1007%2Fs11306-012-0434-4.

(3) Horai, H.; Arita, M.; Kanaya, S.; Nihei, Y.; Ikeda, T.; Suwa, K.; Ojima, Y.; Tanaka, K.; Tanaka, S.; Aoshima, K.; Oda, Y.; Kakazu, Y.; Kusano, M.; Tohge, T.; Matsuda, F.; Sawada, Y.; Hirai, M. Y.; Nakanishi, H.; Ikeda, K.; Akimoto, N.; Maoka, T.; Takahashi, H.; Ara, T.; Sakurai, N.; Suzuki, H.; Shibata, D.; Neumann, S.; Iida, T.; Tanaka, K.; Funatsu, K.; Matsuura, F.; Soga, T.; Taguchi, R.; Saito, K.; Nishioka, T. *J. Mass Spectrom.* **2010**, *45*, 703−714.

(4) Wishart, D. S.; Tzur, D.; Knox, C.; Eisner, R.; Guo, A. C.; Young, N.; Cheng, D.; Jewell, K.; Arndt, D.; Sawhney, S.; Fung, C.; Nikolai, L.; Lewis, M.; Coutouly, M. A.; Forsythe, I.; Tang, P.; Shrivastava, S.; Jeroncic, K.; Stothard, P.; Amegbey, G.; Block, D.; Hau, D. D.; Wagner, J.; Miniaci, J.; Clements, M.; Gebremedhin, M.; Guo, N.; Zhang, Y.; Duggan, G. E.; Macinnis, G. D.; Weljie, A. M.; Dowlatabadi, R.; Bamforth, F.; Clive, D.; Greiner, R.; Li, L.; Marrie, T.; Sykes, B. D.; Vogel, H. J.; Querengesser, L. *Nucleic Acids Res.* **2007**, *35*, D521−526.

(5) Smith, C. A.; Maille, G. O.; Want, E. J.; Qin, C.; Trauger, S. A.; Brandon, T. R.; Custodio, D. E.; Abagyan, R.; Siuzdak, G. METLIN: A Metabolite Mass Spectral Database. *Proceedings of the 9th International Congress of Therapeutic Drug Monitoring and Clinical Toxicology*, Louisville, KY, 2005; pp 747−751.

(6) Tautenhahn, R.; Patti, G. J.; Kalisiak, E.; Miyamoto, T.; Schmidt, M.; Lo, F. Y.; McBee, J.; Baliga, N. S.; Siuzdak, G. *Anal. Chem.* **2011**, *83*, 969−700.

(7) Smith, C.; Want, E.; O'Maille, G.; Abagyan, R.; Siuzdak, G. *Anal. Chem.* **2006**, *78*, 779−787.

(8) Tautenhahn, R.; Böttcher, C.; Neumann, S. *BMC Bioinf.* **2008**, *9*, 504.

(9) Katajamaa, M.; Oresic, M. *J. Chromatogr., A* **2007**, *1158*, 318−328.

(10) Pluskal, T.; Castillo, S.; Villar-Briones, A.; Oresic, M. *BMC Bioinf.* **2010**, *11*, 395.

(11) Baran, R.; Kochi, H.; Saito, N.; Suematsu, M.; Soga, T.; Nishioka, T.; Robert, M.; Tomita, M. *BMC Bioinf.* **2006**, *7*, 530.

(12) Lindon, J. C., Nicholson, J. K., Holmes, E., Eds. *The Handbook of Metabonomics and Metabolomics*; Elsevier B.V.: Amsterdam, The Netherlands, 2007.

(13) Pluskal, T.; Castillo, S.; Villar-Briones, A.; Orešič, M. *BMC Bioinf.* **2010**, *11*, 395.

(14) Thiele, B.; Füllner, K.; Stein, N.; Oldiges, M.; Kuhn, A. J.; Hofmann, D. *Anal. Bioanal. Chem.* **2008**, *391*, 2663−2672.

(15) Marhuenda-Egea, F. C.; Gonsálvez-Álvarez, R.; Martínez-Sabater, E.; Lledó, B.; Ten, J.; Bernabeu, R. *Metabolomics* **2011**, *7*, 247−256.

(16) Esbensen, K. H. *Multivarite Data Analysis −in Practice, An Introduction to Multivariate Data Analysis and Experimental Design*, 5th ed.; CAMO Software AS: Oslo, Norway, 2006.

(17) Martínez-Sabater, E.; Bustamante, M. A.; Marhuenda-Egea, F. C.; El-Khattabi, M.; Moral, R.; Lorenzo, E.; Paredes, C.; Gálvez, L. N.; Jordá, J. D. *J. Agric. Food Chem.* **2009**, *57*, 9613−9623.
(18) Winning, H.; Viereck, N.; Norgaard, L.; Larsen, J.; Engelsen, S. B. *Food Hydrocolloid.* **2007**, *21*, 256−266.