Journal of Pharmaceutical and Biomedical Analysis 51 (2010) 373–381

Contents lists available at ScienceDirect

Journal of Pharmaceutical and Biomedical Analysis

journal homepage: www.elsevier.com/locate/jpba

Methodological aspects for metabolome visualization and characterization A metabolomic evaluation of the 24 h evolution of human urine after cocoa powder consumption

R. Llorach-Asunción [a], O. Jauregui [b], M. Urpi-Sarda [a], C. Andres-Lacueva [a,*]

[a] Nutrition and Food Science Department, XaRTA-INSA, INGENIO-CONSOLIDER program Fun-C-Food CSD2007-063. Pharmacy Faculty, University of Barcelona, Barcelona, Spain
[b] Scientific and Technical Services, University of Barcelona, Barcelona, Spain

ARTICLE INFO

Article history:
Received 11 February 2009
Received in revised form 16 June 2009
Accepted 18 June 2009
Available online 25 June 2009

Keywords:
Metabolomics
Nutrition
Metabolome visualization
Clustering
Metabolite cluster
Food metabolome

ABSTRACT

The LC–MS based metabolomics studies are characterized by the capacity to produce a large and complex dataset being mandatory to use the appropriate tools to recover and to interpret as maximum information as possible. In this context, a combined partial least square discriminat analysis (PLS-DA) and two-way hierarchical clustering (two-way HCA) using Bonferroni correction as filter is proposed to improve analysis in human urinary metabolome modifications in a nutritional intervention context. After overnight fasting, 10 subjects consumed cocoa powder with milk. Urine samples were collected before the ingestion product and at 0–6, 6–12, 12–24 h after test-meal consumption and analysed by LC-Q-ToF. The PLS-DA analysis showed a clear pattern related to the differences between before consumption period and the other three periods revealing relevant mass features in this separation, however, a weaker association between mass features and the three periods after cocoa consumption was observed. On the other hand, two-way HCA showed a separation of four urine time periods and point out the mass features associated with the corresponding urine times. The correlation matrix revealed complex relations between the mass features that could be used for metabolite identifications and to infer the possible metabolite origin. The reported results prove that combining visualization strategies would be an excellent way to produce new bioinformatic applications that help the scientific community to unravel the complex relations between the consumption of phytochemicals and their expected effects on health.

1. Introduction

The purpose of metabolomics is to assess metabolic changes in a comprehensive and global manner in order to explain biological functions or provide detailed biochemical responses of cellular systems [1]. Metabolomics experiments are based on a step-by-step protocol including data acquisition, data extraction, multivariate analysis (MVA), identification of relevant biomarkers and finally, but of equal importance, the biological interpretation of such biomarkers. This workflow has been reviewed [2–4] and a representation as a metroline of this pipeline has now been proposed as an open-access site by the NUGO organization (http://www.nugo.org/metabolomics). A major attempt to identify, develop and disseminate best practice in all aspects of metabolomics has been made by the Metabolomics Standards Initiative (http://msi-workgroups.sourceforge.net).

Besides the other technological and methodological aspects included in the metabolomics workflow [5] there is a growing interest in improving the tools for the visualization and interpretation of metabolome [6]. Basically, a good visualization of data provides useful information for understanding how experimental conditions could modify metabolome. This has been partially resolved by the use of visualization tools, already used by other omics [7], which are often the same as those proposed for multivariate statistical analysis [8]. The most used are hierarchical clustering (HCA), principal component analysis (PCA), partial least squares discriminant analysis (PLS-DA) and their related variations such as orthogonal projection of latent structures (OPLS) and orthogonal signal correction PLS-DA (OSC-PLSDA [9,10]. This filter was first introduced by Wold et al. [11] and afterwards several groups have published [12,13] various variants of the OSC algorithm with the idea of remove structured or systematic variation in responses from matrix X that is not correlated or orthogonal to Y. This filter has been successfully applied in metabolomics studies [14,15]. In addition, other chemometric techniques such as random forest [16] or support vector machines (SVM) have been used in metabolomics studies [17].

* Corresponding author. Tel.: +34 934034840; fax: +34 934035931.
E-mail address: candres@ub.edu (C. Andres-Lacueva).

0731-7085/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.jpba.2009.06.033

Usually PCA, PLS-DA, OSC-PLSDA, OPLS allows to visualize the data as a scatter plots termed scores plot (samples) and loadings plot (markers). In this context, Wiklund et al. [6] has proposed a new scatter plots to improve the visualization and interpretation termed S-plot and SUS-plot. Furthermore, other analysis and visualization tools such as Cytoscape, MetaNetter, VANTED or MeltDB are also available [18–21]. Likewise, the analysis and visualization of LC–MS metabolomics datasets can be achieved using commercial and/or freely available software [5].

An important factor regarding the data analysis is the high number of mass features correlations that exist within the datasets. The meaning and the origin of such correlations have already been revised [7,22]. The localization of these correlations could assist metabolome analysis for at least three reasons [9]. First, the correlated variables are directly of interest due to could suggest a possible biochemical connection or biomarkers of a particular experimental condition. Second, detecting particulars group behaviors could point out the variables that are due to effects that may not be of interest, perhaps gender-related changes, therefore they can be excluded from further analyses which may in turn simplify and improve sample clustering [9]. Third, these correlations could also imply analytical effects such as in-source fragmentations in LC–MS data that usually results as production of fragments or adducts (MS signal redundancy [23]), and their identification could be very useful in the task of metabolite identification.

Metabolomics has been successful in several fields such as toxicology, medicine and pharmacology [24] and it has now become important in the field of nutrition [25–27]. Among external and internal factors that could affect metabolome, diet has been shown as a very important external factor because it produces notable changes in urine composition [28]. These modifications could be related to the presence of exogenous metabolites coming from food components such as phytochemicals (food metabolome) or from microbiota metabolism [25]. Metabolomics studies of food metabolome may lead to the discovery of new phytochemical metabolites and new markers of phytochemical intake [29].

Cocoa and cocoa derived foods have been considered as important source of phytochemicals such as phenolic compounds mainly flavan-3-ols [30] and alkaloids mainly theobromine [31]. The health benefits of cocoa products have been demonstrated in human studies. Recently Cooper et al. [32] have been review the last decade of the studies relating the consumption of cocoa and health. In this context, cocoa consumption has been related with the improvement of the antioxidant status, reduction of inflammation and correlate and with the reduction of heart disease risk.

The aim of this work was to explore the capacity to improve metabolome visualization and interpretation using a strategy based on a combined multivariate analysis with PLS-DA and two-way HCA. In addition, the ability to cluster mass features to characterize metabolites has been tested. For this purpose a real dataset, provided by a nutritional intervention study focused on studying urinary metabolome modifications during the 24 h period after cocoa powder consumption, was analyzed.

2. Materials and methods

2.1. Solvent and standards

The following chemicals were obtained commercially: theobromine and hippuric acid were purchased from Sigma–Aldrich (St. Louis, MO), water for chromatographic separation was purified with a Milli-Q Gradient A10 system (Millipore, Schwalbach, Germany), and acetonitrile of HPLC grade came from Scharlab (Barcelona, Spain).

2.2. Cocoa samples

The soluble cocoa powder used in the study contained 57% of carbohydrates (sucrose, 46%; starch, 1%; complex carbohydrates, 10%), 16% of fiber, 5.4% of fat, 14.1% of protein, 3.97% of moisture, 1.3% of theobromine, 0.13% of caffeine and 2% of ash. The phenolic composition (mean ± SD) of the cocoa powder was determined according to the methodology of Andres-Lacueva et al. [33] and Roura et al. [34]: 23.1% of monomers with 0.71 ± 0.09 mg/g of (−)-epicatechin and 0.21 ± 0.01 mg/g of (+)-catechin, 13.4 % of dimers, among which 0.64 ± 0.06 mg/g of procyanidin B2, 63.6% of 3–8 mers [35,36] and flavonols including 33.87 μg/g isoquercitrin, 5.74 μg/g quercetin, 4.33 μg/g quercetin-3-glucuronide and 36.32 μg/g quercetin-3-arabinoside. The total polyphenolic content was 11.51 ± 0.95 mg catechin/g cocoa powder.

2.3. Subjects and study design

Ten healthy volunteers (5 women and 5 men) between 18 and 50 years old with a corporal mass index of 21.6 ± 2.1 were recruited. After overnight fasting, they were provided with 40 g of cocoa powder with 250 ml of milk. Urine samples were obtained before consumption (0 h) and during the 0–6 (6 h), 6–12 (12 h) and 12–24 h (24 h) periods after test meal consumption. The volunteers remained in the clinical ward for over 6 h to avoid the possibility of transgressing the proscribed diet in the first study period. For the remaining 18 h, all the volunteers followed a standardized polyphenol-free diet (as they had done the day before the study). None reported any history of heart disease, homeostatic disorders or other medical issues, nor received any medication or vitamin supplements. All gave written informed consent before their inclusion in the trial, and the Institutional Review Board of the Hospital Clinic of Barcelona (Spain) approved the study protocol. Participants were instructed to abstain from vitamin supplements, drugs, alcoholic beverages and any polyphenol-rich foods for at least 48 h before and during the test day. A list of allowed and forbidden foods and two menus were given to all participants to help them to follow the polyphenol-free diet strictly the day before the study. The urine samples were stored at −80 °C until analysis.

2.4. Sample preparation

The urine samples were thawed before analysis and centrifuged for 5 min at 12,000 × *g*. A 50 μL aliquot of the supernatant was diluted with 50 μL of Milli-Q water and vortex mixed; the resulting solution was transferred to a 96-well plate autosampler for HPLC-QToF analysis.

2.5. HPLC-QToF analysis

The chromatography was performed on an Agilent 1200 RRLC (Agilent) using an RP-18 Luna 5 μm, 50 × 2.0 mm (Phenomenex, Torrance, CA). The mobile phase consisted of (A) 0.1% HCOOH and (B) acetonitrile 0.1% HCOOH. The flow rate was 600 μL/min and the injection volume was 15 μL. A linear gradient with the following proportions (v/v) of phase B (*t*, %B) was used: (0, 1), (4, 20), (6, 95), (7.5, 95), (8, 1), (15, 1). The HPLC system was coupled to a hybrid quadrupole time-of-flight QSTAR Elite (Applied Biosystems/MDS SCIEX). The MS acquisition was performed in positive ionization and full scan (70–700 amu) modes. Spray parameters were IS +4000, DP 80; FP 380; DP2 10; IRD 6; IRW 5, TEM 400 °C with N_2 as curtain (CUR = 50) and nebulizer (NEB = 60). The QTOF was calibrated with reserpine (1 pmol/μL) using the ions at *m/z* 195.1651 and 609.2812.

2.6. Data processing

LC–MS data was analyzed using the MarkerView™ 1.2 software (Applied Biosystems, MDS SCIEX, Toronto, Ontario, Canada), which performed feature extraction by peak finding for each sample and alignment using mass and retention time windows for the peaks. Peak detection was performed using a minimum peak width of 1 ppm, a noise threshold of 5 and a subtraction multiple factor of 1.5. Alignment used 10 ppm mass tolerance and 0.1 min retention time tolerance. In addition, it was required that variables be present in at least five samples (half the number of samples of each class). With these parameters a dataset containing 3000 mass features was obtained including all detected mass such as isotopes and adducts.

2.7. ANOVA data filter

A filtering procedure using ANOVA was used to remove possible mass features shared by all of the urine samples that could produce a common pattern of non-informative signals. In this context, one-way ANOVA with a Bonferroni correction was used to compare the time effect on metabolome modifications. A probability level of $p < 0.05$ was considered statistically significant. Statistical analysis was performed using the R environment [37].

2.8. Multivariate analysis (MVA)

2.8.1. Partial least squares discriminant analysis (PLS-DA) and orthogonal signal correction-PLS-DA (OSC-PLSDA)

Partial least squares discriminant analysis (PLS-DA) is a supervised analysis that finds directions in a multivariate space for maximum separation of observations (urine samples) belonging to different classes (time of urine collection).

In order to test the possible improvement of data analysis, data were pre-processed using orthogonal signal correction filter (OSC) [11] with time as a correction factor. The OSC filter remove structured or systematic variation in responses from matrix X (dataset) that is not correlated or orthogonal to Y (time) .The number of OSC-component was selected as proposed by Wold et al. [11]. The application of this filter was by a corresponding tool provided by SIMCA-P 11.5 software.

Prior to analysis and in both methods, the dataset was mean centered and Pareto scaled (each variable was weighted according to $1/\sqrt{SD}$). This scaling effectively increases the importance of low concentration metabolites in the resultant models, but not to an extent where the noise significantly contributes to the model [38]. The data was visualized by scatter plots termed scores or loadings plots, where each point on the score plot represented an individual urine sample and each point on the loadings plot represented a mass feature. Moreover, directions in the score plot correspond to directions in the loadings plot [10].

The quality of the models was evaluated by the goodness-of-fit parameter (R^2X), the proportion of the variance of the response variable that is explained by the model (R^2Y) and the predictive ability parameter (Q^2), which was calculated by a seven-round internal cross-validation of the data using a default option of the SIMCA-P+ 11.5 software. The values of $Q^2 < 0$ suggests a model with no predictive ability, and $0 < Q^2 < 1$ suggests some predictive character, with the reliability increasing as Q^2 approaches 1 [39]. These parameters are largely used in metabolomics studies.

Validation and the degree of over fitting for PLS-DA models, were determined by randomly permuting ($n = 999$) [40,41] the order of Y (urine collection time) and fitting separate models to the permuted Y's extracting the same number of components as the original model. The correlation coefficient between the original Y and the permuted Y is plotted against the cumulative R^2 and Q^2 and a regression line is calculated. The y-axis intercept (R^2 and Q^2 when the correlation coefficient is zero) is an indication of over fit; generally the R^2- and Q^2-intercept limits for a valid model should be less than 0.4 and 0.05, respectively, as previously suggested by van der Voet [42].

2.8.2. Two-way hierarchical clustering analysis (two-way HCA)

The dataset composed of the significant ions obtained as described in the Section 2.7 was submitted for two-way HCA analysis using PermutMatrix version 1.9.3 [43]. PermutMatrix is a freely available program (http://www.lirmm.fr/~caraux/PermutMatrix/EN/index.html) enabling evaluation of the overall major HCA methods. Two-way clustering means that the mass features (rows) and urine samples (columns) are clustered simultaneously to obtain groups of samples and mass features that behave as similar as possible [44]. This characteristic produces results that are easily visualized and interpreted. Two-way HCA were carried out using Pearson correlation, and aggregation of the observations was performed with Ward's method [45]. A heat map of intensities and a matrix of mass feature correlations were obtained to visualize and characterize urinary metabolome.

2.9. Ion characterization

To relate the molecular fragments, adducts as well as isotopes normally constituting a chromatogram, to their quasi-molecular metabolites, a modification of multivariate mass spectra reconstruction based on the HCA of ions using a Pearson coefficient was used [14,46]. This approach is based on two principles: first, since fragmentation and adduct production of ions occurs after chromatographic separation, all fragments and adducts will appear at the same retention time to their quasi-molecular ion. Secondly, the high correlation between molecular fragments, adducts and isotopes of a given quasi-molecular ion will be consistently recovered throughout sets of various metabolic profiles. In this context a dataset with 364 significant ions was subjected to HCA. Mass differences between ions from each cluster were analyzed in reference to mass losses or adduct masses often observed in mass spectrometry [47], as well as to characteristic neutral losses of important types of phase II conjugates [48]. The LC–MS behavior of metabolites was also compared with that proposed by the MassBank Database (www.massbank.jp) and Human Metabolome Database (HMDB; www.hmdb.ca).

2.10. Metabolite identification

Markers (Table 1) were further identified on the basis of their exact mass, which was compared to those registered in the Human Metabolome Database (HMDB; www.hmdb.ca) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) (http://www.genome.jp/) (ΔmDa $\leq$ 3 mDa) using an in-house *R* script for *R* environment [37], and to those expected metabolites of cocoa as described in the literature [49–52]. In addition, the source fragmentation behavior obtained from the characterization of ions (Table 1) was used as complementary information to identify the metabolites as proposed by Plumb et al. [53].

3. Results and discussion

Urine samples from ten volunteers were collected before (0 h), and 6 (6 h), 12 (12 h) and 24 h (24 h) after 40 g of cocoa powder with 250 mL of milk intake. This protocol yielded a total of 40 urine samples. An LC–MS fingerprint of these samples was obtained. After peak alignment and markers detection a dataset with 3000 mass features, containing their *m/z*, retention time and intensity of each mass features, was submitted for multivariate analysis in

Table 1
Assignation of identified mass features.

Metabolite cluster	Mass feature cluster[a]	Retention time	Detected mass	Metabolite putative identification	Assignation	Theoretical mass
a	MC1	0.63	166.0718 149.0459 167.0738 150.0547	N-methylguanine	$[M+H]^+$ $[M+H\ NH_3]^+$ ($^{13}C\ [M+H]^+$) ($^{13}C\ [M+H\text{-}NH_3]^+$)	166.0723
b	MC2	4.18	226.0711 151.0392 227.0756 152.0418	Vanilloylglycine	$[M+H]^+$ $[M+H\text{-glycine}]^+$ ($^{13}C\ [M+H]^+$) ($^{13}C\ [M+H\text{-glycine}]^+$)	226.0709
c	MC3	5.33	385.1131 209.0818 149.0603 367.1070	Dihydroxyphenyl valerolactone glucuronide	$[M+H]^+$ $[M+H\text{-GlcA}^b]^+$ $[M+H\text{-GlcA-}C_2H_3O_2]^+$ $[M+H\text{-}H_2O]^+$	385.1129
d	MC3	2.88	170.0449 95.0114 171.0472 124.0376 152.0338	Furoylglycine	$[M+H]^+$ $[M+H\text{-glycine}]^+$ ($^{13}C\ [M+H]^+$) $[M+H\text{-}(CO_2H_2)]^+$ $[M+H\text{-}H_2O]^+$	170.0447
e	MC4	2.18	167.0575 168.0579 333.1036 124.0491	7-Methylxanthine	$[M+H]^+$ ($^{13}C\ [M+H]^+$) $[2M+H]^+$ $[M+H\text{-CHNO}]^+$	167.0563
		2.62	167.0586 168.0608	3-Methylxanthine	$[M+H]^+$ ($^{13}C\ [M+H]^+$)	167.0563
f	MC4	3.63	181.0720 182.0729 138.0660	Theobromine	$[M+H]^+$ ($^{13}C\ [M+H]^+$) $[M+H\text{-CHNO}]^+$	181.0719
g	MC5	4.97	206.0469 178.0512 207.0501	Xanthurenic acid	$[M+H]^+$ $[M+H\text{-CO}]^+$ ($^{13}C\ [M+H]^+$)	206.0453

[a] As noted in Fig. 2.
[b] GlcA: glucuronic acid.

order to visualize and characterize urinary metabolome modifications.

3.1. Partial least squares discriminant analysis (PLS-DA) and orthogonal signal correction-PLS-DA (OSC-PLSDA)

A PLS-DA model with two latent variables was obtained with an R^2X, R^2Y and Q^2 of 0.25, 0.59 and 0.52, indicating low quality of the model (Fig. 1A). This result was related with the tight grouping of classes 12 h and 24 h reflecting subtle metabolome differences between both classes. Hence, the robustness of the classes was evaluated using the SIMCA-P tool for this purpose, finding that the classes 0 h and 6 h were robust groups whilst the classes 12 h and 24 h showed very weak robustness. It should be noted that no 3D scores plot was carried out because the model was constructed with only two latent variables which represent the axis of scatter plots (scores and loadings plots)

In an attempt to improve this analysis an OSC filter was carried out before the PLS-DA [14,15]. Data filtration was carried out by removing two orthogonal components representing 17.88 of the total variation in the original model. The PLS-DA analysis of the filtered data resulted in a three-latent variables model characterized by R^2X, R^2Y and Q^2 values of 0.31, 0.94 and 0.76, indicating that this model was able to distinguish between the four classes. This data showed an improvement of the quality parameters comparing with PLS-DA without OSC filtering. In addition a permutation test ($n = 999$) was carried out to validate the model. This test showed an R^2 intercept of 0.35 and a Q^2 intercept of −0.37 (Supplementary Fig. 1), validating the model. In this context, Fig. 1B depicts a time-related mean trajectory of samples, with a clear separation of four sample classes. Additionally, both score plots also showed that urinary metabolome 24 h after ingestion of cocoa did not reach the baseline (0 h), indicating an incomplete recovery 24 h after cocoa intake.

With regards to the visualization of the markers patterns, a loadings plot related to the score plot was investigated. It should be noted that the distance of a mass feature from the origin in this plot represents the influence of that ion on the PLS components and therefore on the classification in both the PLS-DA and OSC-PLS-DA scores plots. The scores plot of the PLS-DA model showed differences between before consumption period (0 h) and the other three periods, however, not association between mass features and three periods after cocoa consumption was observed (Fig. 1C). Only some markers could be associated with the samples corresponding with the 6 h period (Fig. 1C). With regards to the 3D loadings plot of the OSC-PLSDA model, it showed a slight association between mass features and samples in the corresponding class (Fig. 1D). The 3D loadings plot is a possible solution; however, due to the large number of variables (3000) it is not a feasible option for handling. In fact, this kind of tool is scarcely used even if the scores plot is presented in 3D [54], and only a few works showed the 3D loadings plot [55].

3.2. Two-way hierarchical clustering analysis (two-way HCA)

Pursuing the aim to improve the visualization and interpretation of metabolome, the differences in metabolite signatures related to the 24 h evolution of urinary metabolome after consumption of cocoa were analyzed and visualized by a two-way HCA based on the Pearson correlation coefficient and Ward's method as a method of aggregation. Unsupervised clustering using all metabolites did not result in a good separation between samples (data not shown). Therefore, an ANOVA test ($p < 0.05$) with a Bonferroni correction was used as a filter. This statistical filter allows the detection of

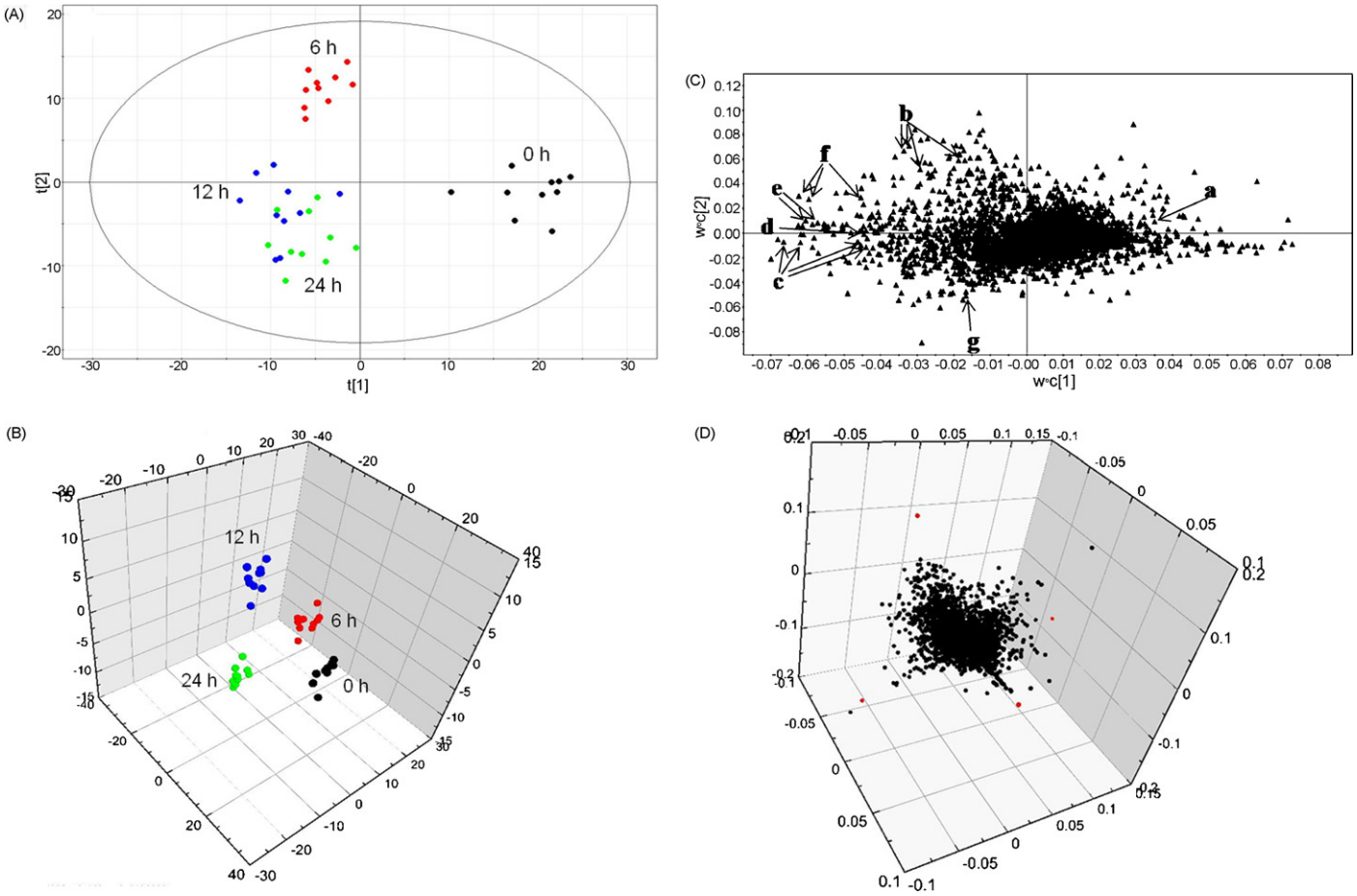

Fig. 1. 2D PLS-DA score plot (A) and loadings plot (C) and 3D OSC-PLSDA scores plot (B) and loadings plot (D) of LC–MS deriving from the urine samples collected at 0 h (black sphere), 6 h (red sphere), 12 h (blue sphere) and 24 h (green sphere) after the consumption of 40 g of cocoa powder. In addition, the all quasi-molecular ions and daughter ions of metabolite cluster b, c and f (Table 1) have been noted in the PLS-DA loadings plots.

mass features related to the possible differences between the tested classes, providing a well-known index to assess variable importance for further analysis before subjecting them to further visualization tools. After statistical filtering the 12.14% (364 mass features) were kept as significant markers and 87.96% of the data was removed from the dataset. These values were similar to those obtained by Boccard et al. [56] where, using an ANOVA filtering ($p < 0.01$), about 90% of the total original information (LC-ToF/MS data) was removed before being submitted for PCA analysis in order to obtain the principal components for a subsequent HCA. In this context, Denkert et al. [57] used the Welch *t*-test as a statistical filter for subsequent clustering, in the context of improving the visualization of the differences in metabolite signatures between borderline tumors and ovarian carcinomas. The HCA using the ANOVA filter data ($p < 0.05$) has the ability to classify the samples into two main groups corresponding with 0 h (before cocoa consumption) and the other periods (6, 12 and 24 h) (Fig. 2). Subsequently, this cluster was divided into 2 cluster levels, separating the time 6 h from the 12 and 24 h (Fig. 2). The final subdivisions succeeded in separating the 12 and 24 h classes being agreed with previous OSC-PLSDA analysis.

Investigation of the clustering behavior of the mass features showed that two main clusters were obtained after their hierarchical clustering, one termed MC1 and another that presented several divisions, providing four clusters termed MC2, MC3, MC4 and MC5 (Fig. 2). These patterns suggested that the cluster MC1 would reflect the signatures that were meaningful at the baseline of the study (endogenous metabolites). As depicted in Fig. 2, the clusters MC2, MC3 and MC5 were related to the samples at 6, 12 and 24 h, respectively. Likewise, the cluster MC4 would group those mass features present all of the time after cocoa consumption. Overall, the clusters MC2-5 represent the mass features that would reflect the evolution of metabolome during the 24 h after cocoa powder consumption, including endogenous metabolites response of the human metabolism, and the exogenous metabolites, mainly phytochemicals provided by cocoa powder which has recently been proposed as a part of food metabolome [27].

3.3. Metabolite characterization and identification

The LC–MS dataset is formed by a mixture of ions where, with the quasi-molecular ion, it is possible to find different fragment ions coming from the ionization source [53] as well as adducts formed in the source [23]. Several in-source fragmentations such as a loss of water or loss of carboxylic moiety have been described. In addition, some in-source losses that are often similar in MS/MS experiments, such as loss of glucuronic acid (−176 amu) or even loss of glycine (−75 amu) have been described [53,58]. Also, some adducts with Na, K and NH_3 could be present in the MS dataset. Following a published strategy [46] a clustering of ions based on the correlation between the mass features in the dataset was used. As shown in Fig. 3 a high correlation is detected in the clusters. A detailed analysis revealed that quasi-molecular ions as well as the in-source fragments and adduct ions are grouped in a small cluster (Fig. 3). This result shows that as a general rule the "*metabolites*" are formed by a number of ions produced by the MS technique (Table 1). This data concurs with that proposed recently by Mocco

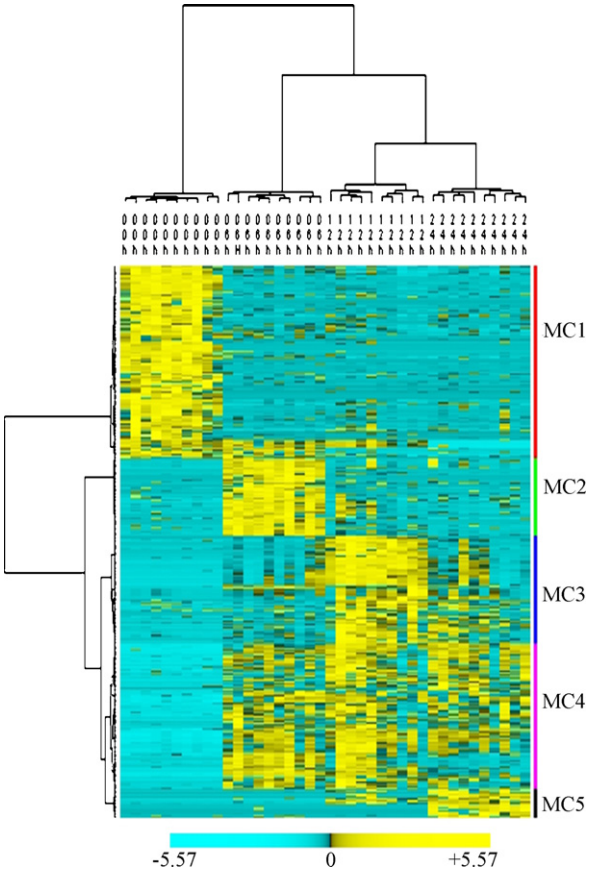

Fig. 2. Two-way hierarchical clustering analysis (processed with PermutMatrix according to the Pearson distance and Ward's aggregation method). Heat map representation of the clustered data matrix in which each colored cell represents the intensity of appropriate mass feature in one sample, according to the color scale at the bottom of the figure. Rows: mass features (364, $p < 0.05$ with a Bonferroni correction). Columns: urine samples before cocoa consumption (0 h) and 6 h, 12 and 24 h following cocoa consumption (40 samples).

et al. [59]. The authors showed that after analysis of the correlations in an LC–MS dataset containing 3.374 mass signals the highest positive correlations were found for mass signals belonging to the same metabolite, such as adducts and fragments. In addition, Chen et al. [14] proposing a correlation with Pearson coefficient allow to select quasi-molecular ion in the first step of biomarker identification.

The non-quasi-molecular ions could reach the same or even more statistical relevance than the quasi-molecular ions; therefore, the correct assignation of these fragments or isotopes is important to avoid mistaken metabolite identification [23]. In this context, the ions from the Table 1 have been place in the loading plot in order to show their relevance in the PLS-DA model. As depicted in Fig. 1C daughter ions presents similar relevance than the quasi-molecular ions.

The metabolite cluster **a** profile was assigned as described in Table 1. A loss of 17 uma is observed from the ion at *m/z* 166 yielding a daughter ion at *m/z* 149. According with Levsen et al. [48] this loss could reflects a loss of NH_3. The database research proposed that quasi-molecular ion (*m/z* 166) could be identified either as 7, 3 or 1-methylguanine. Concerning these metabolites, information available in HMDB confirms that the ion at *m/z* 149 is the main product in MS/MS experiment (HMDB metabocard: HMDB03282 and HMDB00897). MassBank database proposed the ion at *m/z* 149 as main fragment in MS/MS experiment with collision energy at 10 V (MassBank accession number: KO003407 and KO003412). Moreover, Catinella et al. [60] showed that MS/MS spectra of both 7 and 3-methylguanine reveal a fragmentation pattern where the ion at *m/z* 149 was an important fragment ion. Therefore the metabolite cluster was putatively identified as N-methylguanine. Methylguanines are a methylated purine bases. These metabolites haven been detected in normal urine however changes in their levels has been related with pathological conditions such as leukemia, tumors and immunodeficiencies [61].

The metabolite cluster **b** showed a quasi-molecular ion at *m/z* 226 that showed a loss of 75 uma yielding a daughter ion at *m/z* 151. This behavior has been previously associated to a loss of glycine moiety [48]. In this context, a well-known urine metabolite, hippuric acid (N-benzoylglycine) was used as standard to test the MS protocol in order to verify the ability to produce this loss of glycine. As shown in Fig. 4 the mass spectrum of hippuric acid was mainly characterized by cleavage of the amide bond with loss of glycine (−75 amu). With lower abundance there was also a loss of water (18 Da), a ^{13}C isotope and Na adduct. The similarity between the hippuric acid MS behavior and the metabolite clusters suggested that this metabolite corresponded to glycine conjugate. After database interrogation, the in-house phytochemical database proposed that probably the ion at *m/z* 226.0711 would be the vanilloylglycine (226.0709). Taking into account the LC–MS pattern and the database suggestion this metabolite clusters was putatively identified as vanilloylglycine. This compound has been related to the metabolism of vanillin, vanillic acid and ferulic acid [62], all of them present in cocoa powder. Moreover, this behavior of glycine conjugates has been observed in other studies and was used as useful information for the identification of cinnamoylglycine and hippuric acid [63].

The metabolite cluster **c** showed a quasi-molecular ion at *m/z* 385 with a four daughter ions at *m/z* 209, *m/z* 367 and *m/z* 149. This LC–MS pattern could be associated with a glucuronide conjugate. In fact, the loss of 176 amu from the ion at *m/z* 385 yields the ion at *m/z* 209. This loss has been demonstrated that this could be achieved by both in-source fragmentation and MS/MS experiments [58] and has been proposed as a useful tool for the identification of glucuronide derivatives [53]. This metabolite cluster was well associated with the samples at 12 h (MC3 cluster, Fig. 2) that imply an important lag time after cocoa consumption. Moreover, it was not detected in samples at 0 h. This behavior could suggest a possible origin from the microbiota metabolism. In fact, several metabolites derived from ring-fission of catechin are formed in the large intestine by the colonic microbiota metabolism [52,64,65]. According with this LC–MS behavior and with data of our in house phytochemicals database, the quasi-molecular ion at *m/z* 385 was putatively identified as glucuronide derivative of dihydroxy-phenylvalerolactone.

Cluster **d** presents a profile with a ^{13}C isomer and three daughter ions (Table 1). The LC–MS pattern was similar to metabolite cluster b and hippuric acid, with an ion (*m/z* 95) provided by the loss of 75 uma from the quasi-molecular ion (*m/z* 170) suggesting also a possible glycine derivative. After database interrogation a glycine derivative termed furoylglycine was proposed (HMDB metabocard: HMDB00439). The HMDB also provide the MS/MS information of thousands of metabolites. In this context, the MS/MS information shows that this metabolite has a very similar behavior to the metabolite cluster d. In fact the MS/MS recorded in the HMDB shows the ions at *m/z* 152, *m/z* 124 and *m/z* 95 as the main daughter ions of the ion at *m/z* 170. Therefore, this metabolite cluster has been related to the metabolite furoylglycine. The precursors of this metabolite are dietary furan derivatives which are part of the aromatic compounds of cocoa powder [66].

Concerning the metabolite cluster **e** an in-source loss of 43 amu was observed for the ion at *m/z* 167, with retention time of 2.18 min, giving rise to the product ion at *m/z* 124. This pattern is similar to that proposed by HMDB for the compound 7-methylxanthine (HMDB metabocard HMDB01991). In accordance with its exact mass and its LC–MS behavior, the ion at 167 was tentatively iden-

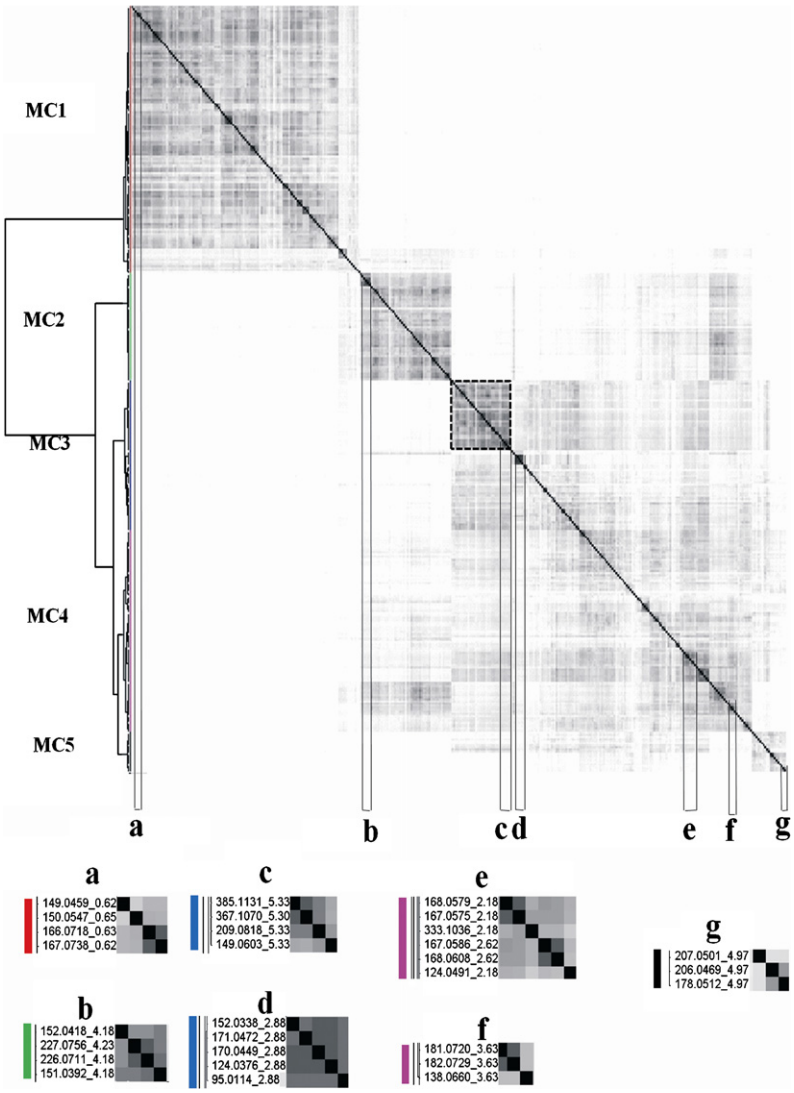

Fig. 3. Metabolite correlation matrix. Correlations between metabolites are shown in grayscale: the darker the colour gray, the higher the percentage of similarity between metabolite excretion patterns. Black lines indicate a zoom to show the small cluster related to the black square. Precise assignation of zoomed cluster in Table 1. Dotted line indicates a high correlated subcluster into MC3.

tified as 7-methylxantine. In this context, loss of 43 amu was not detected for the ion at *m/z* 167 with retention time of 2.62 (Table 1, metabolite cluster e) suggesting that this metabolite was probably the 3-methylxanthine which showed an LC–MS behavior where the ion at *m/z* 124 has low intensity (HMDB metabocard HMDB01886). In addition, this shift in the retention time also suggests these identifications [31,67]. These compounds derived from the metabolism of alkaloids such as caffeine or theobromine [31]. In this context, metabolite cluster **f** also presents a loss of 43 amu giving an ion at *m/z* 138 from the ion at *m/z* 181. This MS pattern had been previously proposed for theobromine [60,68], the main alkaloid in cocoa powder [31].

The metabolite cluster **g** presents a quasi-molecular ion with similar fragmentation behavior to that proposed by Plumb et al. [53] for xanthurenic acid. The authors showed that the daughter ion at *m/z* 178 results from the loss of CO from the acid group. Also, other daughter ions at *m/z* 188 and *m/z* 160 were not detected in our data. In addition, the MS/MS behavior proposed by both MassBank and HMDB was similar to that proposed in Table 1. Therefore, this metabolite cluster has been associated with xanthurenic acid. This compound is a metabolite of the tryptophan degradation pathway (KEGG database, tryptophan pathway ko00380).

3.4. Investigation of metabolite origin

A further example of the capability of this approach is to infer the possible metabolic source of mass features. As described elsewhere, high correlation does not imply a direct biological link, and proximity in the cluster does not necessarily reflect an underlying pathway structure [7,22,69]. Nevertheless, the study of correlations combined with the two-way clustering could be very useful in discovering the possible source (endogenous, exogenous or microbial metabolites) of corresponding metabolites. As depicted in Fig. 3, a high level of correlation is observed between the mass features grouped in MC1 and any correlations were detected with the mass features of other clusters. In this way, whereas the components of MC2 have a high level of correlation, very little correlation has been detected with clusters MC3 to 5. This result and their trend highly related with the samples at 6 h, suggests that, even if it is provided by the diet it is evident that a particular metabolism is responsible for their production. A subcluster into the MC3 cluster showed a high correlation (Fig. 3). This subcluster is formed by different metabolite clusters and one of them is the metabolite cluster **c** that it has been identified as a glucuronide conjugate of dihydrophenylvalerolactone a colonic microbiota polyphenol metabolite (Table 1).

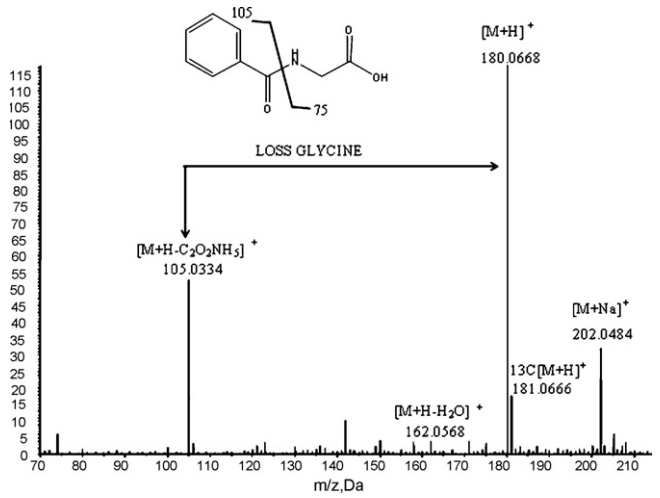

Fig. 4. LC–MS-QTOF spectra in positive ionization mode of commercial standard hippuric acid.

Inspection of this subcluster showed that other ions could be related with other microbial metabolites (data not shown) suggesting that the whole cluster has a microbial origin. In fact, the microbial metabolome is an important part of the urinary metabolome [25]. Finally, a good level of correlation has been detected between the components of MC3, MC4 and MC5 (Fig. 3). The published data in this field is usually related to genetic modifications in plants [70] or microorganisms [44] and, as far as we know, little information about the application of this approach to deduce the possible origin of metabolite signatures in nutrition intervention studies is available, even if the identification of food metabolome seems to be very important for further correlations with their expected biological activities and to produce a new markers of consumption [27].

4. Conclusions

Nowadays Metabolomics Standards Initiative develops important efforts to improve technical and methodological aspects (http://msi-workgroups.sourceforge.net). The ability to visualize and to interpret metabolomic data is without doubt a significant 'bottleneck' in metabolomics studies. In this context, the loadings plot is an efficient tool showing the real contribution of all the mass features in the model construction and therefore providing valuable information about the relevance of this variable; however, it does not provide information about the possible relationships of mass features. On the other hand, two-way cluster provides valuable detailed information about the relationship between mass features and their corresponding classes; however, the information about the relevance of each mass features in the construction of clusters or subcluster divisions is sometimes not evident. Since these approaches provide complementary data, we propose an analysis, visualization and interpretation methodology based on the combination of both tools during metabolomics studies. The strength of this methodology was proven by its ability to pointing out relevant mass features related to time depend on urinary modifications after cocoa consumption.

The correlation matrix has proved to be an interesting tool providing information about the correlations in the dataset and to infer the possible origin of metabolites. An important conclusion is that one metabolite is not characterized by an unique mass feature rather by several mass features. Some detected correlations were assigned as in-source fragmentation or adduct formation and were very useful in the putative identification of metabolite clusters (Table 1). We agree that this procedure is not sufficient for final identification; however, it allows us to obtain an initial identification that could be considered strong enough for those metabolites that have been largely characterized by different techniques over recent years (i.e. hippuric acid). Fortunately, databases such as MassBank, Metlin (http://metlin.scripps.edu) and HMDB have tried hard to compile experimental information related to the NMR, GS–MS, LC–MS, MS/MS of thousands of metabolites.

The mass features correlations could also explain biological relations probably not related to a concrete pathway but rather to the metabolic source. In this context, two-way clustering is an interesting tool for discovering the possible source of metabolites. According to experiment design and the results obtained, a number of mass features were well related to the time before cocoa consumption, suggesting that more of them could be considered as endogenous metabolites. On the other hand, the other clusters would group both the endogenous metabolites reflecting the consumption of phytochemicals (markers of effect) and exogenous metabolites (markers of intake) provided by the diet. Confirmation of this hypothesis is nowadays hampered by the scarcity of information about phytochemical metabolites; mainly those coming from the microbiota metabolism. The reported results prove that combining visualization strategies would be an excellent way to produce new bioinformatic applications that help the scientific community to unravel the complex relations between the consumption of phytochemicals and their expected effects on health.

Acknowledgments

This research was supported by Spanish national grants, CICYT's (AGL: 2004-08378-C02-01/02, and 2006-14228-C03-02) and Grupo Consolider-Ingenio 2010 Fun-C-Food (CSD2007-063). R LL (F.I.S. CD06/00161) and MU-S (FPI-fellow) thank to the Spanish Ministry of Science and Innovation its support.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jpba.2009.06.033.

References

[1] O. Fiehn, B. Kristal, B. van Ommen, L.W. Sumner, S.A. Sansone, C. Taylor, N. Hardy, R. Kaddurah-Daouk, Omics 10 (2006) 158–163.
[2] M. Brown, W.B. Dunn, D.I. Ellis, R. Goodacre, J. Handl, J.D. Knowles, S. O'Hagan, I. Spasić, D.B. Kell, Metabolomics 1 (2005) 39–51.
[3] K. Dettmer, P.A. Aronov, B.D. Hammock, Mass. Spectrom. Rev. 26 (2007) 51–78.
[4] X. Lu, X. Zhao, C. Bai, C. Zhao, G. Lu, G. Xu, J. Chromatogr. B Analyt. Technol. Biomed. Life Sci. 866 (2008) 64–76.
[5] M. Katajamaa, M. Oresic, J. Chromatogr. A 1158 (2007) 318–328.
[6] S. Wiklund, E. Johansson, L. Sjostrom, E.J. Mellerowicz, U. Edlund, J.P. Shockcor, J. Gottfries, T. Moritz, J. Trygg, Anal. Chem. (2007).
[7] D. Camacho, A. de la Fuente, P. Mendes, Metabolomics 1 (2005) 53–63.
[8] P. Mendes, Brief Bioinf. 3 (2002) 134–145.
[9] G. Ivosev, L. Burton, R. Bonner, Anal. Chem. 80 (2008) 4933–4944.
[10] J. Trygg, E. Holmes, T. Lundstedt, J. Proteome Res. 6 (2007) 469–479.
[11] S. Wold, H. Antti, F. Lindgren, J. Öhman, Chemometr. Intell. Lab. 44 (1998) 175–185.
[12] T. Fearn, Chemometr. Intell. Lab. 50 (2000) 47–52.
[13] J.A. Westerhuis, S. de Jong, A.K. Smilde, Chemometr. Intell. Lab. 56 (2001) 13–25.
[14] J. Chen, X. Zhao, J. Fritsche, P. Yin, P. Schmitt-Kopplin, W. Wang, X. Lu, H.U. Haring, E.D. Schleicher, R. Lehmann, G. Xu, Anal. Chem. 80 (2008) 1280–1289.
[15] C. Ducruix, D. Vailhen, E. Werner, J.B. Fievet, J. Bourguignon, J.C. Tabet, E. Ezan, C. Junot, Chemometr. Intell. Lab. 91 (2008) 67–77.
[16] M. Beckmann, D.P. Enot, D.P. Overy, J. Draper, J. Agric. Food. Chem. 55 (2007) 3444–3451.
[17] S. Mahadevan, S.L. Shah, T.J. Marrie, C.M. Slupsky, Anal. Chem. 80 (2008) 7562–7570.
[18] Arita, Brief Funct. Genomic Proteomic 3 (2004) 84–93.
[19] F. Jourdan, R. Breitling, M.P. Barrett, D. Gilbert, Bioinformatics 24 (2008) 143–145.
[20] B.H. Junker, C. Klukas, F. Schreiber, BMC Bioinf. 7 (2006) 109.
[21] H. Neuweger, S.P. Albaum, M. Dondrup, M. Persicke, T. Watt, K. Niehaus, J. Stoye, A. Goesmann, Bioinformatics 24 (2008) 2726–2732.

[22] R. Steuer, Brief Bioinf. 7 (2006) 151–158.
[23] E. Werner, J.F. Heilier, C. Ducruix, E. Ezan, C. Junot, J.C. Tabet, J. Chromatogr. B Analyt. Technol. Biomed. Life Sci. 871 (2008) 143–163.
[24] J.C. Lindon, E. Holmes, J.K. Nicholson, Febs. J. 274 (2007) 1140–1151.
[25] M.J. Gibney, M. Walsh, L. Brennan, H.M. Roche, B. German, B. van Ommen, Am. J. Clin. Nutr. 82 (2005) 497–503.
[26] S. Rezzi, Z. Ramadan, L.B. Fay, S. Kochhar, J. Proteome Res. 6 (2007) 513–525.
[27] D.S. Wishart, Trends Food Sci. Technol. 19 (2008) 482–493.
[28] M.C. Walsh, L. Brennan, E. Pujos-Guillot, J.L. Sebedio, A. Scalbert, A. Fagan, D.G. Higgins, M.J. Gibney, Am. J. Clin. Nutr. 86 (2007) 1687–1693.
[29] A. Fardet, R. Llorach, A. Orsoni, J.F. Martin, E. Pujos-Guillot, C. Lapierre, A. Scalbert, J. Nutr. 138 (2008) 1282–1287.
[30] C. Andres-Lacueva, M. Monagas, N. Khan, M. Izquierdo-Pulido, M. Urpi-Sarda, J. Permanyer, R.M. Lamuela-Raventos, J. Agric. Food. Chem. 56 (2008) 3111–3117.
[31] N. Rodopoulos, L. Hojvall, A. Norman, Scand. J. Clin. Lab. Invest. 56 (1996) 373–383.
[32] K.A. Cooper, J.L. Donovan, A.L. Waterhouse, G. Williamson, Br. J. Nutr. 99 (2008) 1–11.
[33] C. Andres-Lacueva, R.M. Lamuela-Raventos, O. Jauregui, I. Casals, M. Izquierdo-Pulido, J. Permanyer, LC*GC Eur. (2000) 902–905.
[34] E. Roura, C. Andres-Lacueva, R. Estruch, R.M. Lamuela-Raventos, Clin. Chem. 52 (2006) 749–752.
[35] E. Roura, M.P. Almajano, M.L. Bilbao, C. Andres-Lacueva, R. Estruch, R.M. Lamuela-Raventos, Free Radic. Res. 41 (2007) 943–949.
[36] U.S. Department of Agriculture. USDA Database for the Proanthocyanidin Content of Selected Foods. http://www.nal.usda.gov/fnic/foodcomp, 2004.
[37] R. Ihaka, R. Gentleman, J. Comput. Graph Stat. 5 (1996) 299–314.
[38] G.L. Jones, E. Sang, C. Goddard, R.J. Mortishire-Smith, B.C. Sweatman, J.N. Haselden, K. Davies, A.A. Grace, K. Clarke, J.L. Griffin, J. Biol. Chem. 280 (2005) 7530–7539.
[39] A.M. Weljie, R. Dowlatabadi, B.J. Miller, H.J. Vogel, F.R. Jirik, J. Proteome. Res. 6 (2007) 3456–3464, 6.
[40] B.J. Blaise, J. Giacomotto, B. Elena, M.E. Dumas, P. Toulhoat, L. Segalatgal, L. Emsley, PNAS 104 (2007) 19808–19812.
[41] M. Piotto, F.M. Moussallieh, B. Dillmann, A. Imperiale, A. Neuville, C. Brigand, J.P. Bellocq, K. Elbayed, I.J. Namer, Metabolomics (2008), doi:10.1007/s11306-008-0151-1 (Epub).
[42] H. van der Voet, Chemometr. Intell. Lab. 25 (1994) 313–323.
[43] G. Caraux, S. Pinloche, Bioinformatics 21 (2005) 1280–1281.
[44] J. Hageman, R. van den Berg, J. Westerhuis, M. van der Werf, A. Smilde, Metabolomics 4 (2008) 141–149.
[45] B. Meunier, E. Dumas, I. Piec, D. Bechet, M. Hebraud, J.F. Hocquette, J. Proteome. Res. 6 (2007) 358–366.
[46] Y. Tikunov, A. Lommen, C.H. de Vos, H.A. Verhoeven, R.J. Bino, R.D. Hall, A.G. Bovy, Plant Physiol. 139 (2005) 1125–1137.
[47] Metabolomics-Fiehn-Lab, http://fiehnlab.ucdavis.edu/staff/kind/Metabolomics/MS-Adduct-Calculator.
[48] K. Levsen, H.M. Schiebel, B. Behnke, R. Dotzer, W. Dreher, M. Elend, H. Thiele, J. Chromatogr. A 1067 (2005) 55–72.
[49] M. Urpi-Sarda, M. Monagas, N. Khan, R.M. Lamuela-Raventos, C. Santos-Buelga, E. Sacanella, M. Castell, J. Permanyer, C. Andres-Lacueva, Anal Bioanal. Chem. 394 (2009) 1545–1556.
[50] A. Fardet, R. Llorach, J.F. Martin, C. Besson, B. Lyan, E. Pujos-Guillot, A. Scalbert, J. Proteome. Res. 7 (2008) 2388–2398.
[51] L.Y. Rios, M.P. Gonthier, C. Remesy, I. Mila, C. Lapierre, S.A. Lazarus, G. Williamson, A. Scalbert, Am. J. Clin. Nutr. 77 (2003) 912–918.
[52] S. Sang, M.J. Lee, I. Yang, B. Buckley, C.S. Yang, Rapid Commun. Mass Spectrom. 22 (2008) 1567–1578.
[53] R.S. Plumb, K.A. Johnson, P. Rainville, B.W. Smith, I.D. Wilson, J.M. Castro-Perez, J.K. Nicholson, Rapid Commun. Mass Spectrom. 20 (2006) 1989–1994.
[54] L.K. Schnackenberg, J. Sun, P. Espandiari, R.D. Holland, J. Hanig, R.D. Beger, BMC Bioinf. 8 (Suppl. 7) (2007) S3.
[55] A. Prieto, O. Zuloaga, A. Usobiaga, L. Bartolomé, L.A. Fernández, N. Etxebarria, E. Ciprain, A. Alonso, Marine Pollution Bull. 56 (2008) 2094–2099.
[56] J. Boccard, E. Grata, A. Thiocone, J.Y. Gauvrit, P. Lanteri, P.A. Carrupt, J.L. Wolfender, S. Rudaz, Chemometrics Intell. Lab. Syst. 86 (2007) 189–197.
[57] C. Denkert, J. Budczies, T. Kind, W. Weichert, P. Tablack, J. Sehouli, S. Niesporek, D. Konsgen, M. Dietel, O. Fiehn, Cancer Res. 66 (2006) 10795–10804.
[58] Z. Yan, G.W. Caldwell, W.J. Jones, J.A. Masucci, Rapid Commun. Mass. Spectrom. 17 (2003) 1433–1442.
[59] S. Moco, J. Forshed, R. De Vos, R. Bino, J. Vervoort, Metabolomics 4 (2008) 202–215.
[60] S. Catinella, L. Rovatti, M. Hamdan, B. Porcelli, B. Frosi, E. Marinello, Rapid Commun. Mass Spectrom. 11 (1997) 869–874.
[61] M.R. Chao, C.J. Wang, H.H. Yang, L.W. Chang, C.W. Hu, Rapid Commun. Mass. Spectrom. 19 (2005) 2427–2432.
[62] Z. Zhao, M.H. Moghadasian, Food Chem. 109 (2008) 691–702.
[63] Y. Zhen, K.W. Krausz, C. Chen, J.R. Idle, F.J. Gonzalez, Mol. Endocrinol. 21 (2007) 2136–2151.
[64] X. Meng, S. Sang, N. Zhu, H. Lu, S. Sheng, M.J. Lee, C.T. Ho, C.S. Yang, Chem. Res. Toxicol. 15 (2002) 1042–1050.
[65] C. Li, M.J. Lee, S. Sheng, X. Meng, S. Prabhu, B. Winnik, B. Huang, J.Y. Chung, S. Yan, C.T. Ho, C.S. Yang, Chem. Res. Toxicol. 13 (2000) 177–184.
[66] J.S. Bonvehí, Eur. Food Res. Technol. 221 (2005) 19–29.
[67] M.C. Di Pietro, D. Vannoni, R. Leoncini, G. Liso, R. Guerranti, E. Marinello, J. Chromatogr. B Biomed. Sci. Appl. 751 (2001) 87–92.
[68] M. Thevis, G. Opfermann, O. Krug, W. Schänzer, Rapid Commun. Mass Spectrom. 18 (2004) 1553–1560.
[69] G. Gipson, K. Tatsuoka, B. Sokhansanj, R. Ball, S. Connor, Metabolomics 4 (2008) 94–103.
[70] U. Roessner, L. Willmitzer, A.R. Fernie, Plant Physiol. 127 (2001) 749–764.