Journal of Chromatography A, 1218 (2011) 3337–3344

Contents lists available at ScienceDirect

Journal of Chromatography A

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

Untargeted metabolite discovery in kinetic data from multi-dose intervention studies

Sonja Peters[a,b,*], Hans-Gerd Janssen[a,b], Gabriel Vivó-Truyols[b]

[a] Unilever Research and Development, Advanced Measurement and Data Modelling, P.O. Box 114, 3130 AC Vlaardingen, The Netherlands
[b] University of Amsterdam, Analytical-Chemistry Group, van't Hoff Institute for Molecular Sciences, Postbus 94157, 1090 GD Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Available online 18 November 2010

Keywords:
Metabolic profiling
Biomarker discovery
Trend analysis
Dose–response profiles
Gut fermentation
Polyphenols

ABSTRACT

A new strategy for biomarker discovery is presented that is based on multi-dose kinetic metabolomics data. Gas chromatography–mass spectrometry (GC–MS) data sets recorded in the full scan mode are scanned for compounds showing a meaningful trend following the different doses and sampling time points. From a biological point of view, a meaningful trend denotes a compound that responds similarly at all doses and follows a smooth trend along the time points. This type of information can be used to distinguish relevant metabolites from those compounds not following the expected trends. The method is based on analysing the time and dosage trends of each compound via principal component analysis. As only local information is analysed at a time (meaning no correlation with other metabolites is taken into account), the proposed model flags relevant metabolites even if their trend is different from that of any other compound. The new method is therefore an attractive way to reduce the long list of detected compounds in a metabolomics sample set to include only those having the expected smooth time profile that is common for all doses. The new strategy is tested on a sample set obtained from a gut fermentation study of a polyphenol-rich diet. For this study, the initial list of over 25,000 potentially interesting features was reduced to less than 250, thus significantly reducing the expensive and time-consuming manual examination.

1. Introduction

Metabolic profiling and metabolomics are rapidly gaining importance in pharmaceutical and nutritional intervention studies. Metabolomics is the comprehensive study of the metabolome, i.e. it involves the comprehensive identification and quantification of all metabolites present in biological systems such as plants, animals or humans. When gas chromatography (GC) is used as the analytical method, the metabolic fingerprint includes small molecules only. Yet, metabolic profiles easily contain thousands of compounds. However, for a given research question, not all of these may be of interest. Thus, the major difficulty in metabolomics usually is the extraction of relevant information from the immensely complex sample profiles recorded.

The study of the metabolome is normally performed in two steps. In a first step, comprehensive metabolic profiles are recorded at two occasions, i.e. before and after an intervention. The goal here is to scan which compounds change due to an intervention. In a second step, a kinetic study can be performed targeting specific metabolites of interest. Kinetic studies are sometimes carried out in conjunction with multiple doses in order to establish the relationship between concentration profiles and dosage levels. While such targeted analyses allow monitoring target compounds very sensitive and selective, all other possible metabolic changes are missed. With the development of new and better analytical instrumentation, the two historically separate studies, metabolic profiling (i.e. untargeted analysis) and targeted kinetic or dose–response studies, can be combined in a single analysis. Realising this, a new strategy for biomarker discovery utilising sample sets intended only for target-compound analysis becomes apparent. Clearly, this calls for new data-analytical methods that are able to use the pre-knowledge on the expected trends and therefore extract only those compounds following these trends from the complex data sets.

A widely used method for untargeted data analysis is Principal Component Analysis (PCA). In PCA, the original variables are projected onto so-called latent variables or principal components that point in the direction of the highest variance present in the data. PCA is generally not able to take into account the structure of complex experimental designs. For example, when multiple sources of variation exist in data possessing a high number of uncorrelated variables, the principal components will point to an average direction, complicating the interpretation of the PCA model.

* Corresponding author at: Unilever Research and Development, Advanced Measurement and Data Modelling, P.O. Box 114, 3130 AC Vlaardingen, The Netherlands. Tel.: +31 10 4606397; fax: +31 10 4605310.
E-mail address: sonja.peters@unilever.com (S. Peters).

0021-9673/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.chroma.2010.11.023

Incorporation of pre-knowledge on the data structure into the model can be done by combining ANOVA (Analysis of Variance) with simultaneous component analysis (SCA) [1,2] or PCA [3]. More recently, this methodology of ANOVA and SCA, also named ASCA, has been further enhanced by combining it with Parallel Factor Analysis (PARAFAC) [4]. PARAFAC as a stand-alone method [5] or (*n*-way) Partial Least Squares (PLS) [6] are also commonly applied to analyse such higher dimensional structured data sets.

While the above mentioned multivariate methods are very useful for many types of applications, they have one disadvantage in our case: multivariate methods are based on the analysis of correlations *between* variables (metabolites). Such correlations are of less interest in kinetic dose–response studies. The pre-knowledge available here is that potentially interesting metabolites would (i) have a reasonable time profile and (ii) show the same kinetic trend over all doses. Furthermore, the time profiles may differ per metabolite, meaning that any data-analysis method should be applied locally, i.e. on one potential metabolite at a time.

In earlier work we have utilised the idea of looking for compounds meeting expected trends as a new method for metabolite discovery in time-series GC–MS data [7]. In that work, for each detected compound at a time, the time profile was investigated to see whether it followed a pre-defined smooth kinetic trend for the majority of volunteers in the study. This concept of using pre-knowledge on the time profiles of metabolites was also applied in the current study, which includes multiple doses and time points. Interesting metabolites now are those having a smooth time trend that is common for all doses. The tool to do so is by fitting a PCA model locally and monitoring the magnitude of the variance explained by the first loading and its smoothness. From both features, a short list of potentially interesting metabolites is obtained. This (reduced) list should be evaluated manually based on the research question. The procedure is tested on kinetic dose–response profiles obtained from samples of a gut fermentation study used to explore the influence of a polyphenol-rich diet on the human metabolic profiles.

2. Experimental

2.1. Sample background and preparation

The samples used in this study originated from a gut microbial fermentation experiment that investigated microbial-induced metabolic changes after polyphenol digestion. In short, a polyphenol mixture was subjected to a simulated digestion by gut microbes in an in-vitro Simulator of the Human Intestinal Ecosystem (SHIME) [8] at three different levels. Samples were then collected at 15 time points (1, 2, 3, 4, 5, 8, 9, 10, 11, 12, 13, 14, 18, 21, 24 h) for each of the three polyphenol dosage levels (“low, middle and high”). The fermentations were carried out at the University of Ghent. Experimental details on the gut fermentations can be found elsewhere [9].

All samples were prepared for GC–MS analysis as described in Ref. [10]. In short, after a first centrifugation step, the acidified samples were extracted three times with ethyl acetate. The combined organic layers were evaporated to dryness and derivatised using N,O-*bis*[trimethylsilyl]trifluoroacetamide (BSTFA, Sigma–Aldrich, Zwijndrecht, The Netherlands). The gas chromatographic analysis included a 1:10 hot-split injection (1 μL injection volume, 280 °C injector temperature) and a temperature-programmed separation from 45 °C to 350 °C at 8 °C min^{-1}. The column was a VF-17 ms (30 m × 0.25 mm, df = 0.1 μm) from Varian (Varian, Middelburg, The Netherlands). The gas chromatograph used was an Agilent 7890A with a 5975 quadrupole-MS analyser (both from Agilent, Amstelveen, The Netherlands). Full-scan mass spectra were recorded in the mass window from 50 to 600 Da in the electron-impact (EI) mode at 70 eV. The MS source and the GC–MS interface were kept at 200 and 320 °C, respectively.

Table 1
Parameter values used in MetAlign for baseline correction and alignment.

Parameter	Settings
Retention time begin/end (scans)	1–1500
Maximum amplitude	5,500,000
Peak slope factor	0.5
Peak threshold factor	1
Peak threshold	1000
Average peak width at half height	5
Scale on marker peak	211 Da/348
Initial peak search criteria	0 – 4 1500 – 8

The internal standard was *trans*-cinnamic acid-d_6 (Sigma–Aldrich). This internal standard was used to correct for analytical variation. Additionally, a reference standard mixture containing phenolic acids spiked to a comparable matrix (i.e. QC sample) was systematically analysed throughout the whole series of samples in order to identify other sources of variance in the analysis series. Phenolic acids were selected as they were known to be the major metabolites formed in this study. The relative standard deviation of these selected compounds spiked at a concentration of 4 μg/mL was typically around 7% for peak heights, after normalisation by the internal standard.

2.2. Data pre-processing

In total, 45 GC–MS chromatograms were obtained and subjected to MetAlign [11] for data pre-processing. MetAlign performs background and noise removal by pruning the original data in order to contain only those retention time/mass traces that are likely to originate from a chromatographic peak. The retained retention time/mass pairs belonging to different samples are then aligned by using either “rough alignment” or an “iterative alignment” procedure. For our data set, both options were tested and the rough alignment resulted in a satisfactory alignment of all samples in a timely manner. A detailed description of MetAlign is out of the scope of this article and the authors refer to the MetAlign manual for more details [11]. Due to the peak-picking process, one compound can be described by several rows (e.g. one retention time/several masses), depending on the parameter settings in the software as well as the peak’s chromatographic or mass spectrometric properties (e.g. intensity or fragmentation pattern). Peak picking and alignment are defined by parameters given by the user that need to be properly selected. A guideline on parameter selection has been published by Peters et al. [12]. The software finally performs a normalisation of the samples on an internal standard, automatically located by its retention time and mass. The parameters selected for peak picking and alignment are given in Table 1. A schematic of the obtained peak table for our sample set is shown in Fig. 1. This schematic is the input matrix for our method presented in this work.

The obtained aligned peak lists were imported into Matlab 2008 (The Mathworks, Natick, MA, USA) for further processing. In a first step, the data was reduced to only include retention time/mass pairs between 9 and 22 min and having a mass-to-charge ratio of 70 Da and higher. These windows were known from previous studies to contain most relevant biological information.

2.3. Analysis of kinetic dose–response profiles

Fig. 2 shows a schematic of the re-arrangement of the data set before principal component analysis. Consider first the

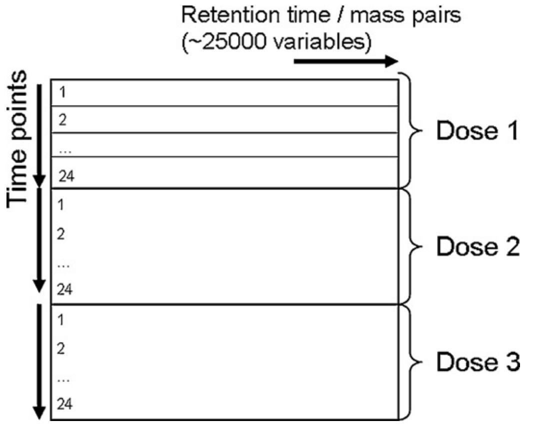

Fig. 1. Schematic of the sample set used in this study. Samples were taken at 15 time points between 1 h and 24 h for three doses resulting in 45 chromatograms as rows in the data matrix and around 25,000 columns corresponding to the retention time/mass pairs "detected" by MetAlign.

chromatographic signal at a single retention time/mass pair for all 45 chromatograms (corresponding to all 15 sampling time points at 3 doses). This data subset can be rearranged to a 3 × 15 matrix; the rows correspond to the doses and the columns to the 15 sampling time points. This matrix is submitted to PCA, as explained below. Before PCA, the data is pre-processed by standard-normal variate (SNV) transformation [13]. In SNV scaling, the data is first centred row-wise and then scaled (row-wise) to unit standard deviation. Centring is commonly performed to remove offsets, while scaling will give all variables within a row, i.e. dose, the same variance thus they will get an equal chance to participate in the PCA model. Column-wise mean-centring was not applied as this would disrupt the structure of the data too much. One should note that one of the core means in the method presented here is to check the smoothness of the loadings along the different sampling time points (see below). As each column corresponds to one sampling time point, column-wise mean-centring would shift the relative position of each sampling point, thus modifying the value of the smoothness. An in-depth discussion on data pre-processing is out of the scope of this article and more details can be found in literature [14].

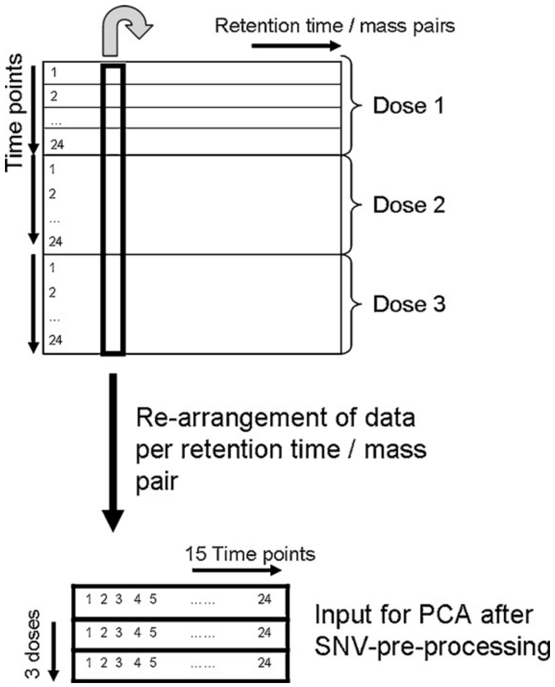
Retention time / mass pairs
Time points
1
2
...
24
Dose 1
Dose 2
Dose 3
Re-arrangement of data per retention time / mass pair
15 Time points
1 2 3 4 5 24
3 doses
Input for PCA after SNV-pre-processing

Fig. 2. Schematic overview on the method presented here. See text for further explanation.

After pre-processing, PCA is performed with the subset of the data. As the columns of the matrix correspond to sampling time points, the first loading of the PCA model corresponds to a description of the time profile that describes the maximum variance of the data, i.e. the common trend explaining all three doses the best. Two features of this common trend are monitored: (i) the variance explained by this common trend (relative to the total variance of the matrix) and (ii) its smoothness. A compound is considered relevant when both features are high. First, the relative variance explained by the first loading informs about how common this trend is. In other words, a low relative variance would mean that the amount of information retained by this first loading is low and this compound does not show a common pattern among the different doses. Secondly, the loadings should be smooth. An interesting peak or metabolite (experiencing a change due to the intervention) should not show too many maxima and valleys during the 24 h following the intervention. If it does, it is likely that the variation experienced by this peak is due to a random variation and hence the compound is not of interest.

Autocorrelation constitutes one way to measure the smoothness of the loadings: the smoother the signal, the closer the autocorrelation is to 1. If only random variation is present in the data, autocorrelation should be zero or near-zero. Negative autocorrelation would be obtained for data series showing high frequency noise. In our case, we are interested in selecting those retention time/mass pairs whose loadings of the first principal component show a smooth trend, meaning they will have an autocorrelation value close to 1. Various autocorrelation functions have been described in literature. In this study, the function by Vivó-Truyols et al. [15] (Eq. (1)) was selected:

$$\rho = 1 - \left\{ 0.5 \times \frac{\sum (y_i - y_{i-1})^2}{\sum (y_i)^2} \cdot \frac{n}{n-1} \right\} \tag{1}$$

where y_i corresponds to the value of the first PCA loading at the ith sampling time point, and n to the number of sampling time points. The values of y should be previously centred by the mean of y calculated over all time points. Autocorrelation functions assume that the sampling time points are equally spaced. This condition is not fulfilled in this work. However, as we are not interested in absolute values of ρ but in relative ones (comparing to other cases measured at exactly the same sampling points), the condition of having equally-spaced sampling points could be ignored. Note that ρ is partially described by the sum of squares of each time-point y_i and its previous time point y_{i-1}. This type of equation is also called 'lag 1'-autocorrelation [16] as only the intensity of the previous time point is being taken into account. Lag 1 was selected as higher lags did not improve the results (data not shown).

3. Results and discussion

The key feature of our new approach for metabolite discovery is the inclusion of pre-knowledge on combined concentration profiles and dosage levels. Only biomarkers that have similar concentration profiles for all doses are of interest and thus, data analysis on the concentration profiles of individual dosages is not the most logical choice. A data analysis method is presented below that takes time profiles and dosage information simultaneously into account.

3.1. Local PCA

As shown in the schematic in Fig. 2, PCA is performed on the data matrix of one retention time/mass pair at a time. For each

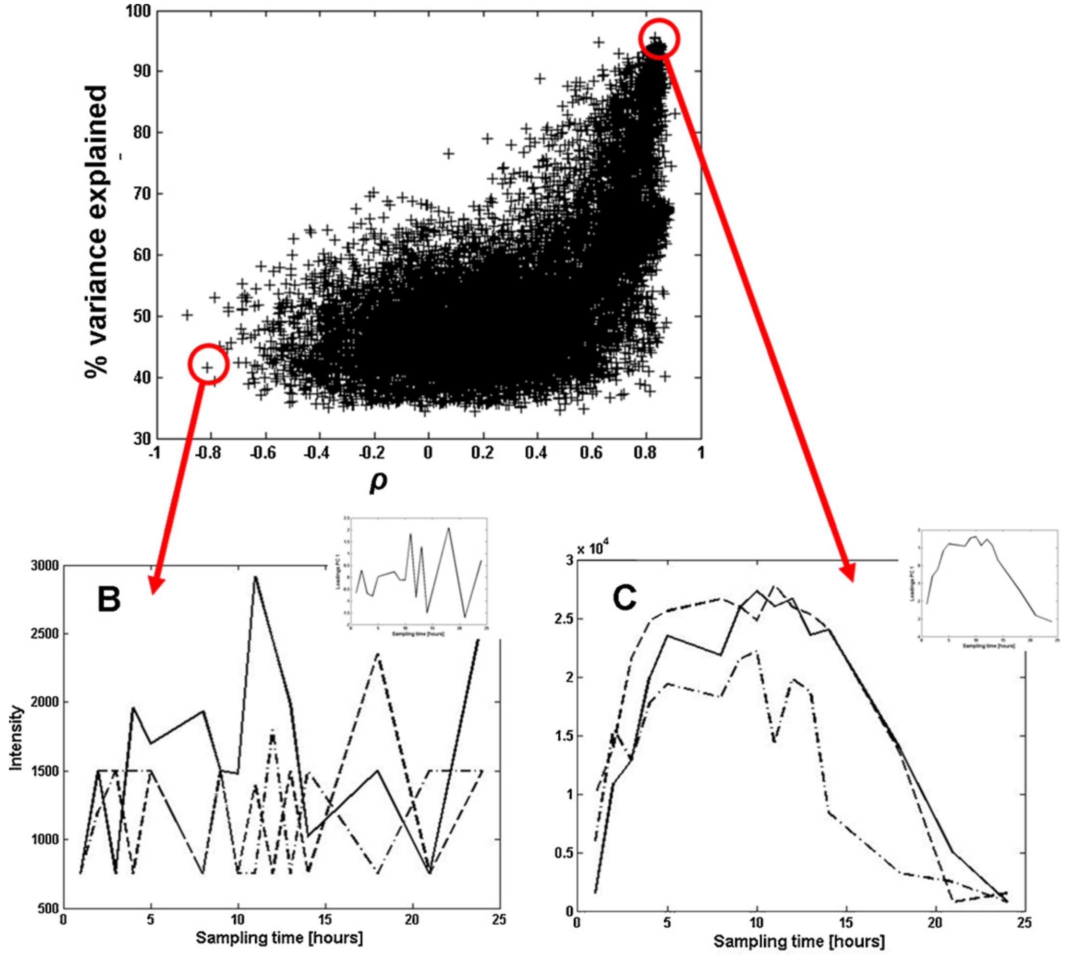

Fig. 3. Values of ρ and % of variance explained obtained for each retention time/mass pair after applying the local PCA method described in Section 2.3. Each cross in this figure corresponds to one retention time/mass pair. The time trends of two retention time/mass pairs (encircled) are depicted in 3B and 3C. The dash-dotted line corresponds to dose 1, the dashed line corresponds to dose 2, and the solid line to dose 3. Inserted in the figures are the corresponding loading plots.

pair, a ρ-value as well as a number for the percentage of explained variance is obtained. These two can be plotted against each other as shown in Fig. 3. Here, each point in the figure corresponds to one retention time/mass pair. Several interesting observations can be made in this plot. Firstly, there are a lot of retention time/mass pairs which are associated with a negative or 'close to zero' autocorrelation value. Due to the low number of sampling points, random behaviour yields ρ-values that are not exactly zero. Additionally, high-frequency noise (yielding negative ρ-values) may be introduced in the data due to the pre-processing software. The negative ρ-values are not of interest in this case, which becomes clear when investigating the loadings of a retention time/mass pair associated with a negative or zero value of autocorrelation (Fig. 3B). The loadings shown are not smooth and when the three doses are overlaid, only noise is detected for this retention time/mass pair.

The most relevant retention time/mass pairs are those that result in a PCA model which first loading has a high autocorrelation (high ρ-value) and which first principal component describes most of the variance. High autocorrelation in the first loading corresponds to a smooth evolution of this compound over time and a high value of explained variance means that the shape of this evolution is common to all doses. Retention time/mass pairs with high ρ and high variance explained are located at the top right corner of Fig. 3. An example is shown in Fig. 3C. The overlaid dosage curves clearly show a common trend for all doses (which makes the explained variance with the first PC high) and the corresponding loadings (insert in Fig. 3C) are smooth (making the autocorrelation value high). The choice on how many retention time/mass pairs are to be (manually) checked is somewhat arbitrary and needs to be tested for each sample set. Table 2 gives an overview on the number of markers obtained with various, reasonable thresholds. As can be seen when decreasing the cut-off point for the ρ-value from 0.85 to 0.7 at a threshold of 90% explained variance, the number of detected pairs increases from 26 to 250. When the threshold for relative explained variance was lowered to 70%, many more retention time/mass pairs were detected, especially in combination with a ρ-cut-off of 0.7. These thresholds yield more than 1000 pairs being flagged as "potentially of interest". The fourth column gives an estimation of how many potential metabolites are represented by the number of retention time/mass pairs that have to be verified. These now need to be manually reviewed for the likelihood that they indeed represent possible new metabolites of interest.

Table 2
Combinations of ρ-values and percentage variance explained and their respective number of retention time/mass pairs found to be of interest. An estimation of the number of metabolites corresponding to these pairs is also given.

ρ	% var. expl.	Features	Compounds
0.85	90	26	7
0.8	90	222	18
0.7	90	250	23
0.85	70	79	24
0.7	70	1182	100

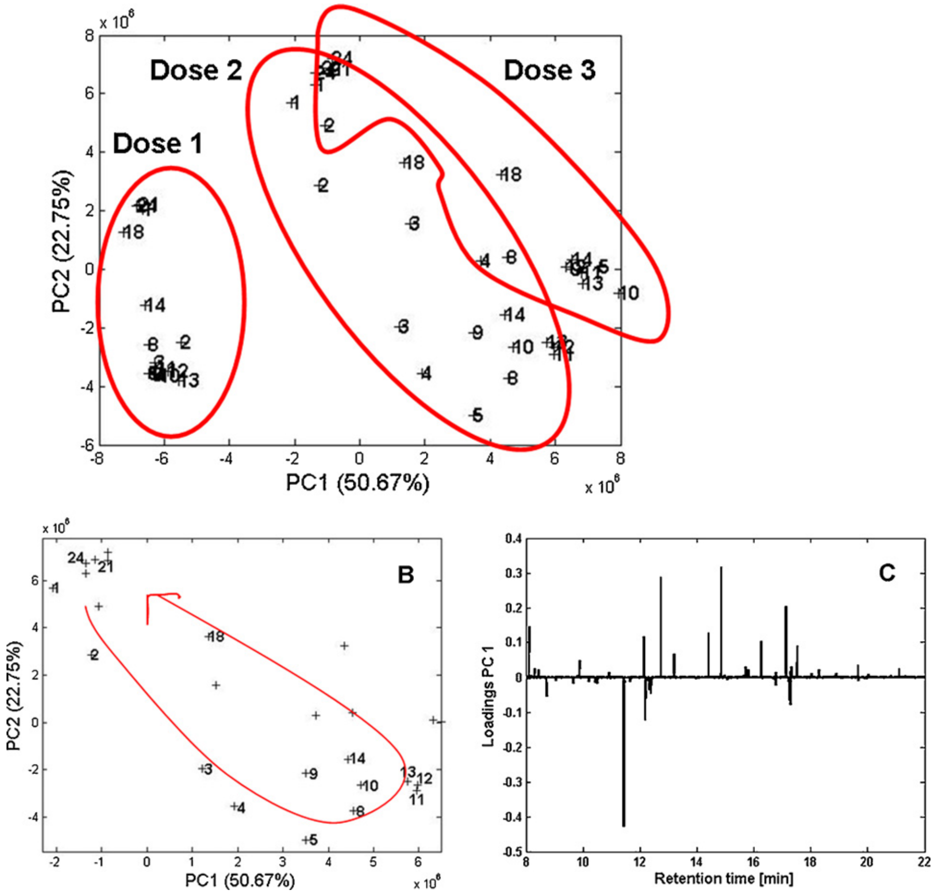

Fig. 4. PCA score plot obtained when applying PCA to the whole data set. Numbers correspond to the sampling time points and encircled regions correspond to doses. Captured variances are 50.67% (PC 1) and 22.75% (PC 2). The score plot for dose 2 is depicted in detail in 4B. The arrow depicts the direction of increasing sampling time. 4C depicts the loadings plot for the first PC.

Additionally, the thresholds should be established in such a way that a reasonable (low) number of false positives are obtained. In other words, the number of time/mass pairs that are wrongly flagged as potential biomarkers (when in fact they are not) should be low. One way to measure the type I error is performing random permutation tests [17]. In this case, we have permuted the metabolite labels so that in each PCA model the dosages of different metabolites are included. For each permutation, one ρ-value and one value for the percentage explained variance is obtained.

For every set of thresholds as presented in Table 2, a false discovery rate, fdr, can now be calculated using Eq. (2):

$$fdr = \frac{f(perm)/N(perm)}{f(real)/N(real)} \tag{2}$$

where $f(perm)$ and $f(real)$ are the number of pairs found to be potentially of interest when the data is permuted (perm) and non-permuted (real). $N(perm)$ equals the number of permutations and $N(real)$ equals the number of retention time/mass pairs in the original data set. For 2 million permutations and using a threshold of 0.85 as the minimum ρ-value and including only compounds which PCA model can explain more than 90% of variance with their first principal component, 26 retention time/mass pairs or about seven compounds could be identified as potentially interesting metabolites (see Table 2). When the class labels are permuted, only one retention time/mass pair is found above these two thresholds, resulting in a false discovery rate of 9.4e^{-4}. For the other thresholds presented in Table 2, the false discovery rates are 0.0024, 0.0030, 0.0342 and 0.0506, respectively. These values can also be used to (additionally) verify which threshold combination will result in the most meaningful and reliable markers. As can be seen only the last combination of thresholds ($\rho = 0.7$ and % explained variance greater than 70%) results in a false discovery rate higher than 5%. Also, using 90% rather than 70% explained variance as threshold results in a roughly 10-fold lower fdr, indicating a higher confidence in the markers obtained using those thresholds.

3.2. Global PCA

In order to compare the results from the previous section with standard techniques, the data set as presented in Fig. 1 was also subjected as a whole to principal component analysis. In this case, the data was mean-centred column-wise (i.e. per retention time/mass pair) before PCA analysis. Fig. 4 shows the score plot of the first two principal components obtained, with the first PC describing 50.67% of the variance and the second PC 22.75%. It can be seen that the variance captured by the first PC mainly corresponds to dose effects (see circled groups in the figure) while the second PC mainly describes the effect of sampling time. However, for both effects, no clear separations can be obtained. Fig. 4B is a zoom into Fig. 4 for the second dose with the arrow pointing along the direction of the sampling time points. For this dose, the time trend is visible with the first and the last sampling time points being clustered together. This suggests that metabolites that have similar starting and end concentrations, e.g. metabolites that are first formed and then disappear at the time-scale of the experiment, have the strongest impact on the PCA model. However, for the other doses, no such clear time trend can be detected, meaning that the variance captured by the first PC is not only based on sampling time points.

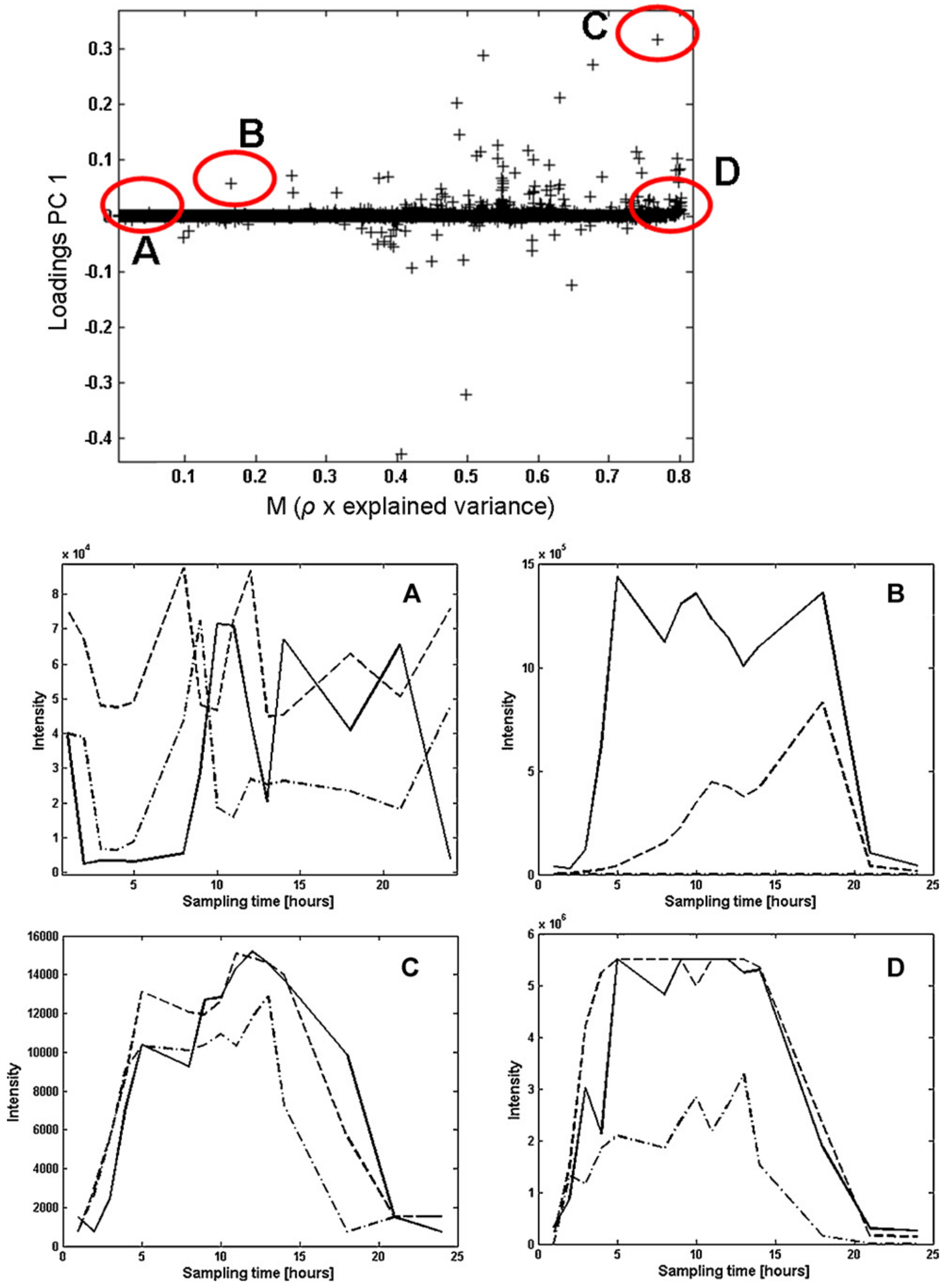

Fig. 5. Loadings of the first PC obtained by global PCA vs. the single value *M* obtained for local PCA. *M* is obtained by multiplying *ρ* by the % of variance explained. Each point in the plot corresponds to one retention time/mass pair. Circles (A–D) correspond to particular retention time/mass pairs whose kinetic dose–response profiles are depicted in detail below. In 5A–5D, the dash-dotted lines correspond to dose 1, dashed lines correspond to dose 2 and the solid lines to dose 3.

The loadings plot of PC 1 is depicted in Fig. 4C. In this figure, all retention time/mass pairs that have an absolute value of the loading value above the noise level are interesting possible metabolites for further (manual) assessment. That is because their impact on the PCA model is the largest. However, in contrast to the local PCA model, in global PCA we are fitting a PCA model to the data set as a whole. The model therefore tries to find latent variations in sampling time and dose that can be explained by a combination of metabolites. In other words, the model finds a “common pattern” that is able to explain the variability in the data at all doses and at all sampling times. The obtained “interesting metabolites”, i.e. those which are flagged by a large absolute value of the loading, are therefore difficult to interpret biologically as the variability in the data is due to different sources (time and dose).

3.3. Comparison of the two methods

Fig. 5 shows the results of both PCA methods plotted against each other. For simpler visualisation, the autocorrelation value and the relative variance explained by the local PCA model are multiplied to a single value *M*. Therefore, *M* constitutes a single number collecting both the smoothness of the first loading and its importance. In Fig. 5, the loading of the first PC from the global PCA method (as seen in Fig. 4C) is plotted against *M*. Each point in the plot corresponds to one retention time/mass pair. Following the discussions in Sections 3.1 and 3.2, both methods highlight interesting retention time/mass pairs when a high *M* (*X*-axis) or a high absolute value of the loading (*Y*-axis) is obtained. However, as will be discussed below, different subsets of retention time/mass pairs are being flagged as "interesting" by the two methods. In fact, four groups of extremes can be found in Fig. 5. The kinetic dose–response curves of an example for each group are depicted in Figs. 5A–D: retention time/mass pairs that have a low *M*-value and a loading close to zero (5A), pairs that have a high loading (positive or negative) and a low *M*-value (5B), pairs that have a high loading (positive or negative) and a high *M*-value (5C) and those that have a high *M*-value but a loading close to zero (5D). Retention time/mass pairs in group A are of least interest as they will not be flagged as "interesting" by any of the methods. They only describe noise as can be clearly seen in Fig. 5A. Pairs in group B will only be flagged with the global PCA model as their loading is rather high while their *M*-value is low. That means that they are important for the overall PCA model, but do not exhibit a smooth, common trend for all doses. This becomes clear when overlaying the kinetic dose–response curves of one retention time/mass pair in that group (5B). Clearly, no common trend can be observed and when the loadings for this pair are investigated no smooth time trend can be seen (plot not shown). Fig. 5C gives an example of a kinetic dose–response curve for a pair present in group C (high *M*-value, high loadings). The pairs in this group will be flagged as "of interest" with respect to both PCA models. It can be seen that a (rather) smooth trend is obtained for all doses, which explains why it is picked up by the local PCA method. When the global PCA model is considered, it can only be said that this compound shows a high correlation with other compounds that are important for the PCA model. The opposite can be said for pairs found in group D. These pairs do not contribute much to the global PCA model, i.e. they would not be picked up with the global PCA method; however, a smooth, common trend can be seen for all doses (see Fig. 5D).

3.4. Biological interpretation

Any data-analytical method is best validated biologically, meaning that the results obtained must be interpretable and answering the biological question of the study. Table 3 gives an overview on the possibly interesting metabolites found by both methods. For applying PCA locally, all metabolites were selected which are associated with a ρ-value of greater 0.8 and 90% explained variance. For the global PCA method, the threshold selected was an absolute value of the loading above 0.03. Note that this number is derived visually from the loadings plot (Fig. 4C). While there is generally a good agreement between the two methods some metabolites are only found when PCA is applied locally, while others are only flagged by the overall PCA method. Knowing the background of the present (polyphenol) study, phenolic acids were expected as major metabolites, which was confirmed by the mass spectra of the compounds in the reduced peak lists. Fig. 6 shows the kinetic profiles of two of them, vanillic acid (6A) and 4-hydroxyphenylacetic acid (6B). Vanillic acid (ρ = 0.84, % variance explained = 92.37%) is flagged by the local PCA method only. Clearly, all three curves have a common trend and a smooth loading (insert in Fig. 6A). 4-Hydroxyphenylacetic acid (ρ = 0.78, % variance explained = 65.74%) is flagged by the global PCA method only. This is easy to understand as our requirement was a common profile for all three doses and, as can be seen in Fig. 6B, this is not the case for this metabolite. For two of the doses, no clear response profile is obtained, while for the third, a typical appearance–disappearance curve can be observed. Thus, the percentage of variance explained is rather low. The loadings obtained with the local PCA method for this metabolite (insert

Table 3
List of metabolites flagged by the local PCA and global PCA methods. Local PCA was performed using ρ >0.8 and >90% of explained variance as thresholds. With global PCA, a threshold of 0.03 for the absolute value of the loadings was used. A cross in the corresponding column (local PCA or global PCA) means that the corresponding metabolite was flagged by the corresponding PCA method.

Retention time	Compound ID	Local PCA	Global PCA
8.12	Phenylacetic acid	X	X
8.31	Catechol	X	
8.73	Unknown		X
9.88	Phenylpropionic acid	X	X
10.2	Unknown	X	
11.09	Unknown	X	
11.44	Unknown		X
11.65	Pyrogallol	X	
12.16	3-Hydroxybenzoic acid	X	X
12.75	3-Hydroxyphenylacetic acid	X	X
13.01	4-Hydroxybenzoic acid	X	
13.2	4-Hydroxyphenylacetic acid		X
14.43	3-Hydroxyphenylpropionic acid		X
14.88	4-Hydroxyphenylpropionic acid	X	X
14.95	Vanillic acid	X	
15.16	Unknown	X	
15.7	3,4-Dihydroxybenzoic acid	X	X
15.8	3,4-Dihydroxyphenylacetic acid	X	X
16.24	Unknown		X
16.76	Syringic acid	X	
17.12	2-Hydroxyphenylvaleric acid		X
17.25	*p*-Coumaric acid	X	X
17.29	3-*o*-Methylgallic acid	X	X
17.55	Gallic acid	X	X
19.67	Unknown		X

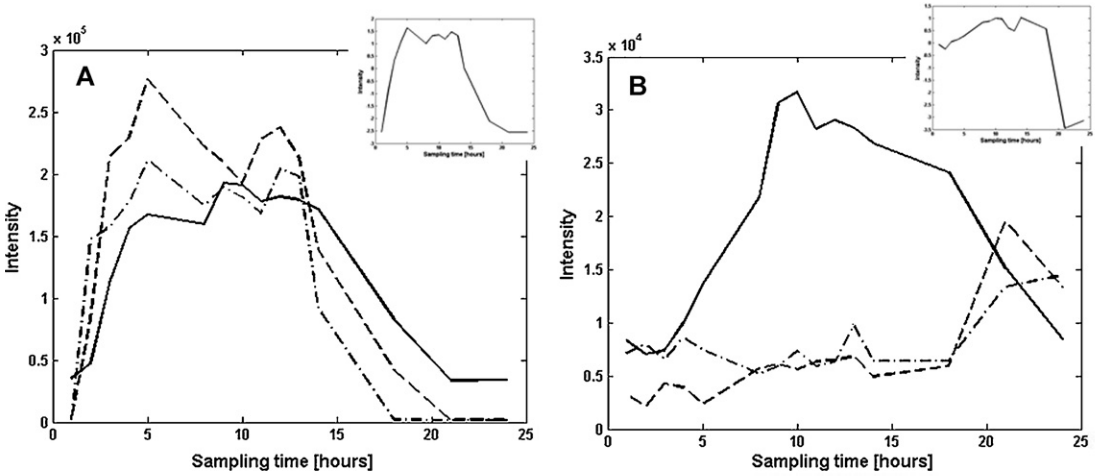

Fig. 6. Kinetic dose–response profiles of two metabolites, vanillic acid (part A) and 4-hydroxyphenylacetic acid (part B). The dash-dotted lines correspond to dose 1, dashed lines correspond to dose 2 and the solid lines to dose 3.

in Fig. 6B) are difficult to interpret and seem to be dominated by the response profile of the third dose. Generally, the global PCA method is flagging compounds which correlations are described well by the first principal component. The understanding of why certain compounds have high loadings in the global PCA method is not easy as the global PCA model is based on mixed effects (time and dose). With the local PCA method, however, the interpretation of the results is very straight-forward and the only factor to consider is how many possible metabolites may be included in the (manual) analysis of the results, i.e. which thresholds to select.

4. Conclusions

Scanning data sets for compounds that meet expected trends is an interesting route for data analysis in metabolomics. In the current kinetic multi-dose study, an interesting metabolite was defined as having a smooth trend over sampling time that is common for all doses. When principal component analysis is performed locally, i.e. on one compound at a time, the information obtained can be used to extract compounds defined to be of interest while dismissing all others. Therefore, a quick analysis of the kinetic curves of all compounds present in the sample set can be carried out.

Using the newly developed method a list of thousands of possible metabolites could be reduced to just 18 compounds that meet the set of pre-knowledge criteria (i.e. concentration and dosage profiles). In order to validate that possible markers are not obtained due to chance, permutation tests were performed. When applying PCA to the sample set as a whole as commonly performed, some of these markers were missed, while others were selected that do not follow a specific pattern with dose and/or time. With our method, easily interpretable results are obtained and laborious (manual) data analysis is greatly reduced.

Acknowledgment

Huub Hoefsloot is gratefully acknowledged for his advice and assistance in the validation of the statistical methods developed.

References

[1] A.K. Smilde, J.J. Jansen, H.C.J. Hoefsloot, R.-J.A.N. Lamers, J. van der Greef, M.E. Timmerman, Bioinformatics 21 (2005) 3043.
[2] J.J. Jansen, H.C.J. Hoefsloot, J. van der Greef, M.E. Timmerman, J.A. Westerhuis, A.K. Smilde, J. Chemometr. 19 (2005) 469.
[3] P. de B. Harrington, N.E. Vieira, J. Espinoza, J.K. Nien, R. Romero, A.L. Yergey, Anal. Chim. Acta 544 (2005) 118.
[4] J.J. Jansen, R. Bro, H.C.J. Hoefsloot, F.W.J. van den Berg, J.A. Westerhuis, A.K. Smilde, J. Chemometr. 22 (2008) 114.
[5] R. Bro, Chemometr. Intell. Lab. 38 (1997) 149.
[6] H. Antti, T.M.D. Ebbels, H.C. Keun, M.E. Bollard, O. Beckonert, J.C. Lindon, J.K. Nicholson, E. Holmes, Chemometr. Intell. Lab. Syst. 73 (2004) 139.
[7] S. Peters, H.-G. Janssen, G. Vivó-Truyols, Anal. Chim. Acta 663 (2010) 98.
[8] K. Molly, M. Vande Woestyne, W. Verstraete, Appl. Microbiol. Biotechnol. 39 (1993) 254.
[9] X. Tzounis, J. Vulevic, G.G.C. Kuhnle, T. George, J. Leonczak, G.R. Gibson, C. Kwik-Uribe, J.P.E. Spencer, Br. J. Nutr. 99 (2008) 782.
[10] C.H. Grün, F.A. van Dorsten, D.M. Jacobs, M. Le Belleguic, E.J.J. van Velzen, M.O. Bingham, H.-G. Janssen, J.P.M. van Duynhoven, J. Chromatogr. B 871 (2008) 212.
[11] A. Lommen, http://www.metalign.wur.nl/UK (last accessed July 2010).
[12] S. Peters, E.J.J. van Velzen, H.-G. Janssen, Anal. Bioanal. Chem. 394 (2009) 1273.
[13] R.J. Barnes, M.S. Dhanoa, S.J. Lister, Appl. Spectrosc. 43 (1989) 772.
[14] R. Bro, A.K. Smilde, J. Chemometr. 17 (2003) 16.
[15] G. Vivó-Truyols, P.J. Schoenmakers, Anal. Chem. 78 (2006) 4598.
[16] D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. De Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of chemometrics and Qualimetris: Part A, Elsevier, Amsterdam, 2003, p. 594.
[17] C.M. Rubingh, S. Bijlsma, E.P.P.A. Derks, I. Bobeldijk, E.R. Verheij, S. Kochhar, A.K. Smilde, Metabolomics 2 (2006) 53.