Anal Bioanal Chem (2013) 405:5063–5074
DOI 10.1007/s00216-013-6884-3

RESEARCH PAPER

Integrating multiple analytical platforms and chemometrics for comprehensive metabolic profiling: application to meat spoilage detection

Yun Xu · Elon Correa · Royston Goodacre

Received: 1 February 2013 / Accepted: 28 February 2013 / Published online: 20 March 2013

Abstract Untargeted metabolic profiling has become a common approach to attempt to understand biological systems. However, due to the large chemical diversity in the metabolites it is generally necessary to employ multiple analytical platforms so as to encompass a wide range of metabolites. Thus it is beneficial to find chemometrics approaches which can effectively integrate data generated from multiple platforms and ideally combine the strength of each platform and overcome their inherent weaknesses; most pertinent is with respect to limited chemistries. We have reported a few studies using untargeted metabolic profiling techniques to monitor the natural spoilage process in pork and also to detect specific metabolites associated with contaminations with the pathogen *Salmonella typhimurium*. One method used was to analyse the volatile organic compounds (VoCs) generated throughout the spoilage process while the other was to analyse the soluble small molecule metabolites (SMM) extracted from the microbial community, as well as from the surface of the spoiled/contaminated meat. In this study, we exploit multi-block principal component analysis (MB-PCA) and multi-block partial least squares (MB-PLS) to combine the VoCs and SMM data together and compare the results obtained by analysing each data set individually. We show that by combining the two data sets and applying appropriate chemometrics, a model with much better prediction and importantly with improved interpretability was obtained. The MB-PCA model was able to combine the strength of both platforms together and generated a model with high consistency with the biological expectations, despite its unsupervised nature. MB-PLS models also achieved the best over-all performance in modelling the spoilage progression and discriminating the naturally spoiled samples and the pathogen contaminated samples. Correlation analysis and Bayesian network analysis were also performed to elucidate which metabolites were correlated strongly in the two data sets and such information could add additional information in understanding the meat spoilage process.

Keywords Multi-block principal component analysis · Multi-block partial least squares · Data fusion · Correlation analysis · Bayesian network · Pork spoilage · *Salmonella typhimurium*

Introduction

Untargeted metabolic profiling is a methodology which aims to profile a broad spectrum of metabolites in a given biological system of interest. It provides biochemically rich information for various purposes such as biomarker discovery, understanding the mechanism(s) of physiological processes or assigning unknown gene function [1–3]. However, the main challenge in untargeted metabolic profiling is that it is desirable to measure as much of the metabolome as possible, yet there is a large chemical diversity in these small molecules (metabolites) so that it is very difficult, if not impossible, to analyse the full metabolome using a single analytical platform

Published in the topical collection *Metabolomics and Metabolite Profiling* with guest editors Rainer Schuhmacher, Rudolf Krska, Roy Goodacre and Wolfram Weckwerth.

Electronic supplementary material The online version of this article (doi:10.1007/s00216-013-6884-3) contains supplementary material, which is available to authorized users.

Y. Xu (✉) · E. Correa · R. Goodacre
School of Chemistry, Manchester Institute of Biotechnology,
University of Manchester, 131 Princess Street,
Manchester M1 7DN, UK
e-mail: yun.xu-2@manchester.ac.uk

R. Goodacre
Manchester Centre for Integrative Systems Biology,
Manchester Institute of Biotechnology,
University of Manchester, 131 Princess Street,
Manchester M1 7DN, UK

due to the limitations of each analytical approach. Fingerprinting techniques such as Fourier transformed infrared (FT-IR), Raman, nuclear magnetic resonance spectroscopy and direct infusion mass spectrometry (DIMS) can provide a fingerprint profile of a broad range of metabolites, although it is often difficult to obtain an unambiguous identification of each of the metabolites assayed. Hyphened analytical platforms such as gas chromatography–mass spectroscopy (GC-MS) or liquid chromatography–mass spectroscopy (LC-MS) are better choices when it comes to metabolites identification. However, both GC-MS and LC-MS are only capable of analysing a limited range of chemicals using one particular experimental setting. For example, conventional GC-MS analysis can only analyse volatile to semi-volatile metabolites, thus to enable GC-MS to analyse non-volatile metabolites, a suitable derivatization process is required. Unfortunately, it is very difficult to develop a GC-MS method which can analyse both volatile/semi-volatile metabolites and those small molecules that need to be derivatized *simultaneously*; therefore, it is usual that two separate experiments are undertaken to cover both types of chemical diversity. Thus, to find a model which is able to integrate the data generated by different analytical platforms and/or experiments together is highly valuable in untargeted metabolomics.

In our previous work, we have reported two metabolic profiling studies aimed at meat spoilage detection. The aim of these studies was to test various metabolic profiling techniques for their ability to monitor food spoilage using pork as the meat substrate, as well as to detect deliberate contamination with a bacteria food pathogen that causes salmonellosis. *Salmonella typhimurium* was used as the model microorganism in these studies. In one article, we successfully demonstrated that it was possible to model the spoilage process of pork by analysing the volatile organic compounds (VoCs) collected from the pork samples incubated under room temperature over a period of 72 h [4]. It was also possible to discriminate the naturally spoiled pork samples from the ones contaminated by *S. typhimurium* after 24 h incubation. In a separate report, we showed that similar discriminatory ability could also be obtained by analysing the small molecule metabolites (SMM) extracted from the biomass and meat substrate harvested from the surface of the pork samples [5]. Both experiments used GC-MS as the metabolomics analytical platform. The difference is that the VoCs samples were collected using a polydimethylsilicone (PDMS) patch as described in [6] and analysed after thermal desorption by GC-MS directly. By contrast, the small molecule metabolites were extracted from the biomass harvested from the surface of the pork samples using sterilised swabs and followed by a freeze–thaw cycle extraction as described in [7]. The extracted metabolites were then derivatized in a two-step process where the first involved methoximation and the second stage silylation with subsequent analysis by GC-MS. The results from the data analysis on both data sets independently showed a common trend which was that the metabolic profiles of spoiled pork samples had changed significantly as spoilage progressed, whilst at each time point when the samples were taken there were also differences in the metabolic profiles between the naturally spoiled samples and those that had been contaminated by *S. typhimurium*.

In the present study, we attempted to integrate the two data sets, which contain information from two different classes of metabolites, together by employing a multi-block PCA (MB-PCA) model for unsupervised exploratory analysis. We also employed multi-block partial least squares regression (MB-PLS-R) and multi-block partial least squares for discriminant analysis (MB-PLS-DA) models for predicting the spoilage progression and the type of the samples respectively. With a broader range of the metabolites taken into account, the results showed that the MB-PCA model had significantly improved interpretability compared to the classical PCA models applied to each data set individually. The MB-PLS model had also showed greatly improved prediction accuracy compared to the classical PLS-R and PLS-DA models. In addition, correlation analysis was also performed to find the most correlated metabolites between the two of the data sets and such correlation can be visualised by Bayesian networks in an intuitive way and provided a list of potentially interesting metabolites which may have connections between each other and worth for further studies.

Experiments

Cultures and chemicals

S. typhimurium strain 4/74 was kindly provided by Professor Tim Brocklehurst (The Institute of Food Research, Norwich, UK). The strain was sub-cultured on Lab M LAB028 blood agar plates (Lab M Ltd., Lancashire, UK). A single colony was inoculated into nutrient broth (250 mL) and incubated at 37 °C for 16 h, which resulted in a culture of $\sim 5\times 10^{7}\,\mathrm{cfu\,mL^{-1}}$. An aliquot (5 mL) of the culture was harvested by centrifugation at 4,810×*g* for 10 min. The supernatant was removed and the pellet was re-suspended in 50 mL of sterile saline solution (0.9 % NaCl, *w*/*v*) and centrifuged again at 4,810×*g* for 10 min. This process was repeated two further times. The pellet was then re-suspended in 50 mL of sterile saline solution and used for the artificial contamination of the pork.

Sample collection

A total of 24 boneless pork chops (weight 200–300 g) were purchased from a local supermarket. Each pork chop was then cut laterally into two pieces (so called butterflying) to provide two near-sterilised surfaces for the study. For each pair of the matched two pieces of the meat from the same pork chop, one was used as control and to which 1 mL of sterilised saline solution was added; the other piece was used as an artificial contamination surface and to which 1 mL of saline suspension of *S. typhimurium* (*vide supra*) was added and spread across the surface by using a sterile plastic loop. Each piece of the pork was then placed in a large glass Petri dish lined with sterilised filter paper (Whatman grade 40 cat. no. 90-7501-06) to which 2 mL of sterilised saline solution was added to act as a moisture source and prevent the surface of the meat from drying out. The Petri dishes were then sealed and placed into an incubator for incubation. At 0, 24, 48 and 72 h after the contamination, 6 pork chops; i.e., 12 pieces were taken out of the incubator for the sampling.

VoC samples collection

A piece of PDMS patch which was cut from silicone elastomer sheet (cat no. 751-624-16; Goodfellow Cambridge Ltd.) with a dimension of 20×15×50 mm and pre-conditioned as described in [4] was used for the VoCs sampling. Passive sampling of the headspace was achieved where one PDMS patch was stuck onto the underside of the front cover of the Petri dish, due to its natural adhesiveness, and the Petri dish was then resealed and placed back in the incubator for 1 h. The patch was removed after the 1 h incubation and placed into a thermal desorption tube and analysed by GC-MS. All the samples were analysed within <24 h after the collection.

SMM extraction

For each piece of pork sample that had had its VoCs samples collected by the PDMS patch, the biomass on the surface was subsequently harvested by using two sterilised swabs. The biomass was then transferred directly into 1 mL of ice-cold methanol stored on dry ice (−48 °C). The suspension was extracted with three freeze–thaw cycles (frozen in liquid nitrogen and allowed to thaw on dry ice). The suspensions were then centrifuged at 16,060×*g*, at −9 °C for 5 min. The supernatants were immediately lyophilised, derivatised as described below and subjected to GC-MS analysis.

GC-MS analysis

More detailed information about the GC-MS settings can be found in [4, 5]; for brevity only a brief description is given here:

The PDMS patches were analysed directly through a Markes International Unity 1 thermal desorption unit which was connected to a Varian CP 3800 gas chromatograph coupled to a 2200 quadrupole ion trap mass analyzer. An Agilent HP-5 (60 m×0.25 mm×0.25 μm) was used as the analytical column. The mass range used was from 40 to 400 DA with a scan rate of 1.03 scans/s.

The SMM extracted from the biomass on the surface of the pork samples was subsequently derivatized as the following. An aliquot of 1,000 μL of each metabolite extract was spiked with 100 μL of internal standard solution (0.19 mgmL^{-1} succinic d_4 acid, 0.27 mL^{-1} malonic d_2 acid and 0.22 mgmL^{-1} glycine d_5 in HPLC grade water) and then lyophilised in a HETO VR vacuum centrifuge attached to a HETO CT/DW cooling trap (Thermo Life Sciences, Basingstoke, UK). An aliquot (50 μL) of 20 mgmL^{-1} *O*-methylhydroxylamine solution in pyridine was added and heated at 60 °C for 45 min followed by adding an aliquot (50 μL) of MSTFA (*N*-acetyl-*N*-(trimethylsilyl)-trifluoroacetamide) and then heating at 60 °C for 45 min. The derivatized samples were subsequently analysed by employing an Agilent 6890 GC coupled to a LECO Pegasus III Time of Flight (TOF) mass spectrometer. A Supelco DB-50 (30 m×0.25 mm×0.25 μm) was used as the analytical column. The mass range used was 40–600 DA. Comparing to the quadrupole mass spectrometer which has a rather slow scanning rate, the TOF spectrometer can acquire up to 500 mass spectra per second although that would generate a massive amount of data and be very difficult to analyse. In this study, the acquisition rate was set to ten mass spectra per second.

CGC-MS deconvolution

The data generated by both GC-MS instruments were exported to netCDF files and imported to MATLAB 2009a (MathWorks, MA, USA). The raw GC-MS data were then deconvolved by using a hierarchical multivariate curve resolution (H-MCR) procedure as described in [4]. After the H-MCR deconvolution, the VoCs data had generated a total number of 63 unique peaks while the small molecule metabolite extraction data had a total number of 200 unique peaks.

The data were then ready for analysis and we have followed MSI guidelines in reporting data processing [8].

Multi-block principal component analysis and multi-block partial least squares

MB-PCA is the extension of the commonly used PCA model which aims to combine *multiple* data matrices of different origins together and gain a “consensus” view of all the data matrices incorporated into the model. There are several MB-PCA algorithms reported in the literature and each algorithm has its own properties, a comprehensive review can be found in [9]. In this study, an algorithm named CPCA-W, which was proposed by Westerhuis et al. [10], has been used. Regardless of the algorithm, a MB-PCA model usually consists of 3 main components: (1) a super scores matrix, $\boldsymbol{T_t}$ (2) c pairs of blocks scores $\boldsymbol{T_b}$ and loadings $\boldsymbol{P_b}$ matrices and (3) a block weights vector W_t, where c is the number of data sets (i.e. blocks) incorporated into the MB-PCA model. The $\boldsymbol{T_t}$ matrix represents the common trend across all the data matrices incorporated into the model; each pair of $\boldsymbol{T_b}$ and $\boldsymbol{P_b}$ represents the unique pattern of each block under the "consensual" view revealed in $\boldsymbol{T_t}$ and the block weights vector W_t represents the relative contribution of each block to the common trend showed in $\boldsymbol{T_t}$.

The advantage of MB-PCA modelling is that by including multiple data matrices, which hopefully all contain a common trend of interest, the interpretability can be improved; i.e., a better agreement between the mathematical factors (i.e., principal components) and the underlying biological factors can be expected. Another main advantage of MB-PCA is that by integrating multiple data matrices into a unified model, the loadings matrices $\boldsymbol{P_b}$ are *directly* comparable to each other and thus it is easy to identify the highly correlated variables between the blocks and this could be a very useful feature to find the metabolites found by different analytical platforms which are closely related, e.g. in the same pathway.

Similar to MB-PCA, as an extension of classical PCA, MB-PLS [11–13] is the extension of PLS modelling which had been widely used in both regression and classification applications. There were several algorithms to construct a MB-PLS model and a detailed discussion can be found in [12, 13]. In this study, we employed an algorithm which deflate the response block (e.g. spoilage time) Y with the super scores to build the MB-PLS models [13]. Two MB-PLS models were used, one regression model (MB-PLS-R) was used to predict the spoilage time while another classification model (MB-PLS-DA) was used to predict the sample type, i.e. whether the sample was naturally spoiled or contaminated with *S. typhimurium*. For comparison, classic PLS-R and PLS-DA [14] were also applied to each data set for spoilage time and sample type prediction respectively. All the MB-PLS and classic PLS models were built and validated by using a double cross-validation procedure as described in [15]. In this procedure, one sample was left out as the independent testing sample and the remaining samples were used as the training set. The MB-PLS and classic PLS models were built on the training set, the number of PLS components were chosen by using another leave-one-out cross-validation performed on the training set only. The sample been left out was then predicted by the model. This procedure was repeated until every sample had been left out once as the independent testing sample. The performances of MB-PLS-R and PLS-R were measured by calculating the cross-validated correlation coefficient Q^2 [15] and root-mean-squares error of cross-validation (RMSECV) based on the testing samples while the performances of the MB-PLS-DA and PLS-DA models was measured by calculating the correct classification rate (CCR) of the testing samples which is the percentage of the samples been correctly classified.

In multi-block modelling, an appropriate block weighting step is generally required in addition to other commonly used data pre-processing methods such as mean centre, scaling etc. This is because different data blocks may have very different scales and/or different number of variables. Without an appropriate block weighting, the multi-block model could be dominated by one particular block which had significantly higher variance than the other blocks and lose its advantages of being able of utilise multiple blocks. The method of choice for block weighting depends on each particular application. One may chose to apply an appropriate scaling factor to each block so that all the blocks would have equal variance after the weighting; or one can put more weights on some of the more "useful" blocks, if such prior information is available. In this study, there were significant differences both in the number of variables and in the scale of the data between the VoCs and the SMM extraction data sets, although it was unknown that which block would provide more useful information about the spoilage progression and/or pathogen contamination. Thus, we employed an equal-variance block weighting method. This was done by first autoscaling each data block so that every variable has a mean of 0 and a standard deviation of 1. Then, a block scaling factor which is the inverse of the square root of the number of variables of that block was applied to each block.

Correlation analysis and Bayesian network

In addition to the MB-PCA modelling, correlation analysis is also useful as it can be employed to find out which metabolites in each block were most correlated to each other as these mostly correlated metabolites may have great biological interest; e.g., they may arise from the same metabolic pathway—although we are of course aware that correlation does not necessarily equate to causality [16] and care is needed in interpretation. Thus, we performed a pair-wise correlation analysis between the two data sets using Pearson correlation coefficient and generated a correlation coefficient heat map for a global view of the connections of all the metabolites. Then, we selected the metabolites pairs having an absolute value of the correlation coefficient greater than 0.8 for probabilistic Bayesian network (BN) analysis.

A BN is a statistical method that learns the probabilistic relationships between measured variables, such as metabolites, and represents those relationships via a directed acyclic graph (DAG). The DAG is constructed from a numerical data matrix (samples are rows and metabolite (variables) are columns) and each node in the graph represents a variable from the data. An edge linking two nodes indicates a statistical correlation between them. To find the DAG that best represents the correlations from the data matrix, the algorithm tests several DAGs, or structures, which represent different relationships between the nodes and selects the DAG that maximizes a preselected scoring metric. The winning DAG is then used to represent graphically the structure of the most likely relationships among the variables of that data matrix. The strength of those relationships can be estimated by the conditional probabilities based on the data and the associations encoded by the structure of the DAG. In the present work, however, we do not use such information as the objective of the model is only the visualization of the associations between metabolites collected on two different platforms and from two different metabolite classes (VoCs vs. SMM).

Learning the structure of a BN is an NP-hard problem [17], and this is one of the most challenging tasks in dealing with BNs. Many algorithms developed to this end use a heuristic search procedure and a scoring metric to find the network that best fits the data. The task of the scoring metric is to evaluate the goodness-of-fit of a candidate DAG to the data. The task of the search procedure is to modify the DAG structures iteratively until a suitable structure is found. To search for the network structure (DAG), we used a greedy search algorithm (GSA) which, at each stage of the process, makes a locally optimal choice. First, the GSA starts with an empty network, containing only nodes but no edges. Second, the algorithm tests all possible new network structures that differ from the current network structure by the addition of a single edge. The new network structure with the highest score, according to the preselected scoring metric, becomes the current network. The process is repeated until no new edge addition improves the score of the network. To evaluate the score of the candidate networks we used the Bayesian Dirichlet method described in [18]. As the search for the best BN model is an NP-hard problem, the number of metabolites (variables) included in the model needs to be limited to a small manageable number. From previous empirical models and from the literature [19], we observed that networks with more than approx. 30 nodes become very messy and difficult to interpret. Therefore, we selected only pairs of metabolites having an absolute correlation coefficient (≥0.8) to be included in the BN analysis. This is a subjective threshold but worked well for this particular application as only 20 metabolites met this criterion. However, we make no claim that this is an optimum variable selection criterion. The scripts for the BN analysis were developed in-house using MATLAB v. 2009a (The MathWorks Inc., Natick, MA, USA) and are available from the authors on request.

Contrary to intuition, the direction of the arrow on a BN does not necessarily imply a cause–effect relationship between variables; indeed, a BN model is not a “causal network”. A causal network is a graphical model with an explicit requirement that the relationships (arrow directions) be causal [20]. Causal networks are out of the scope of this paper and are a subject for future work. Therefore, we have intentionally omitted arrowheads from the final BN graph for two reasons: (1) to avoid this possible misleading cause–effect interpretation which is not the objective of this work; and (2) because bidirectional relationships, which are likely between metabolites, cannot be represented by a DAG. Therefore, the interpretation of the network (Fig. 5) is straightforward. Edges linking nodes represent association between them which was detected during the model building process based on Bayesian Dirichlet method presented in [18]. After the network structure is built, we then computed the Pearson correlation coefficient between linked nodes and show the corresponding values beside each edge. Blue edges represent positive correlation coefficients and red edges represent negative ones.

The BN analysis provides complementary information to the MB-PCA results. However, there is no direct association or dependency between both methods. The objective of the BN is to represent association between metabolites in a graphical model which should simplify visual interpretation of the results (correlation between metabolites). One of the main differences between a heat map and a BN is the way in which associations between metabolites are computed. On the heat map associations are computed based on Pearson correlation coefficient whereas the on the BN they are based on the score (goodness-of-fit) of the network according to the Bayesian Dirichlet method [18]. As both methods are looking for highly correlated metabolites, it is expected that they will detect similar associations (correlated metabolites) and their results will overlap as it is the case for this dataset. For the sake of comparison, after building the network topology, we also displayed the Pearson correlation coefficients on the BN. One of the advantages of the BN over the heat map is that the associations on the graphical model are much easier to visualise and interpret.

Results and discussion

The PCA scores plots of both VoCs and SMM data which were analysed separately are shown in Fig. 1 while the super

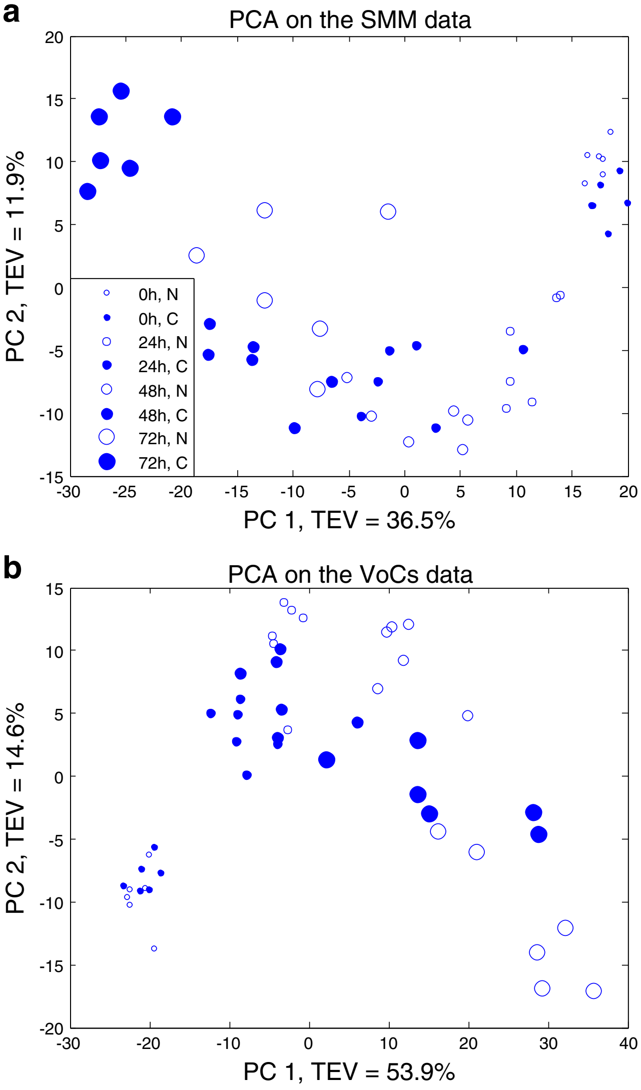
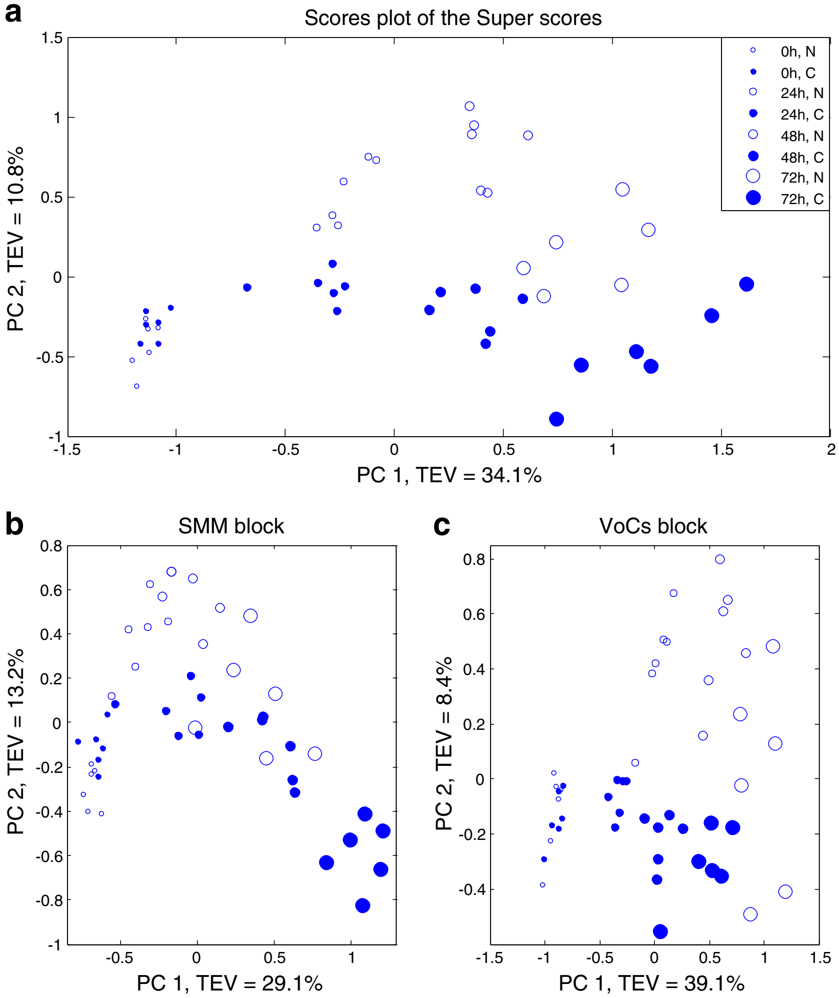


Fig. 1 PCA scores plots from **a** extracted small metabolites after chemical derivatisation and **b** volatile organic compounds sampled using thermal desorption. *N* naturally spoiled sample; *C* samples contaminated with *S. typhimurium*

scores and block scores of the MB-PCA from the combined VOCs *plus* SMM data are given in Fig. 2. Comparing the results of the two methods it is very easy to see that although the pattern showed by the PCA scores plots are very similar to that from MB-PCA, where in both analysis types there is a clear trajectory corresponding to the progression of spoilage, the separation between the two types of samples (natural spoilage samples *versus* the ones contaminated by *S. typhimurium*) are more readily observed in the MB-PCA (Fig. 2a); with the exception of time 0 h, which is perhaps expected as no microbial growth will have occurred yet. In the PCA models (Fig. 1) the first two PCs appear to have contributions from both factors of interest (i.e., the progression of spoilage and the type of sample), and this makes the interpretation of the loadings plot rather tricky. By contrast, the super scores plot of the MB-PCA model (Fig. 2a) showed significantly improved interpretability with the first PC modelling one biological factor of interest (the progression of spoilage) while the second PC models the other factor; namely the two different sample types. Thus in the MB-PCA model, there was a better consistency between the mathematical factors (i.e. PCs) and the biological factors. The block scores of the VoCs and SMM block also depict a similar trend. However, there is an intriguing difference between the two blocks: the block scores of the VoCs block showed better separation between the four time points compared to that of the SMM block, whilst rather surprisingly in the last time point the separation between the two types of samples (natural spoilage vs. *S. typhimurium* contaminated) appeared to be worse than the previous two time points; contrary to this the SMM block showed less separation between the different time points in general, albeit the separation between the two types of sample seemingly became better over time. The percentage of total explained variance (TEV) of each PC in each block also showed the same trend; i.e., the TEV% of PC 1 of the VoC block is significantly larger than that in the SMM block while the opposite is true for PC 2.

It is also interesting to compare the block loadings plots of the two blocks. Thanks to the better consistency between the projected mathematical MB-PCA factors and the biological factors, the interpretation of the block loadings becomes much easier. In the VoCs block loadings plot (Fig. 3b), the most interesting variables are located at the top-right corner of the figure which were the VoCs which increased significantly over time which resulted in significant and positive loadings of the first PC; these VoCs were also more abundant in the natural spoiled samples than in the *S. typhimurium* contaminated samples. It is interesting to see that there were very few variables in the bottom left corner. This indicates that natural spoiled samples were generally more VoC rich than the samples contaminated by *S. typhimurium* which could be the results of relatively larger bio-diversity in the natural spoiled samples. Only one third of the variables had negative PC 1 loadings and even fewer of those also had significant PC 2 loadings (either positive or negative). This indicates that the decreasing-over-time VoCs are generally not likely to be the ones of interest and were probably environment-related VoCs. When the VoCs generated from the spoilage process increased over time these VoCs were less and less absorbed by the silicone patch and thus resulted in a decreasing trend. By contrast, the block loadings of the SMM block (Fig. 3a) is very different. It appeared that variables were more polarised across PC 2 than PC 1 together and there were also a few variables had significant PC 2 loadings but very low PC 1 loadings. This may indicate that there were a considerable number of metabolites which showed distinguished difference between the two types of samples but did not show a drastic

Fig. 2 CPCA-W scores plot. **a** Super scores of combined data along with the individual block scores for **b** the small metabolites block and **c** the VOC block. *N* Naturally spoiled sample; *C* Samples contaminated with *S. typhimurium*



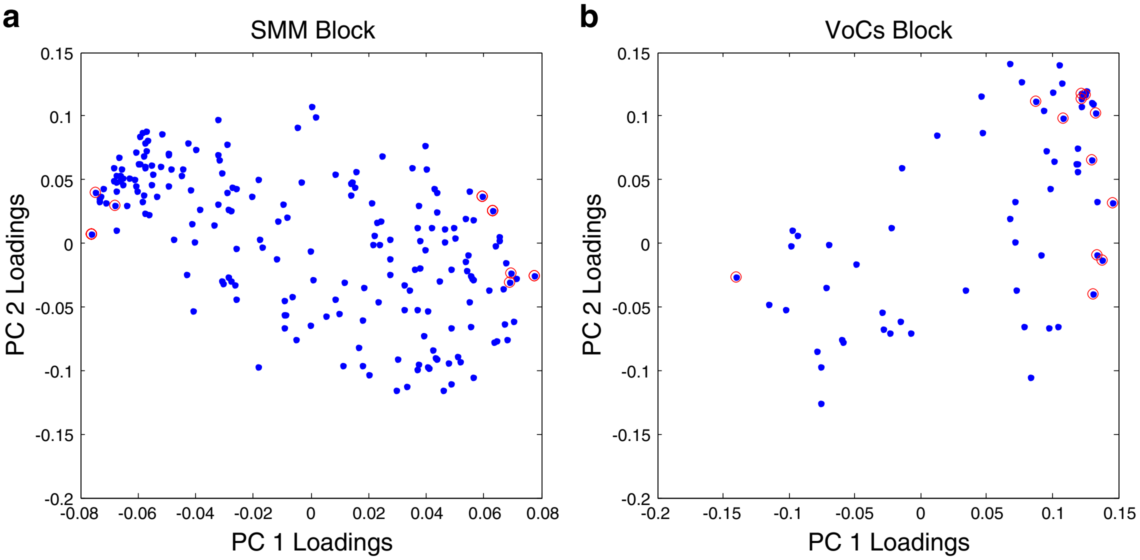


Fig. 3 MB-PCA block loadings plot for **a** the small metabolites block and **b** the VoC block. The metabolites that are circled are the ones selected for probabilistic Bayesian network analysis

changing-over-time trend like most of the VoCs features did. In addition, PC 2 had approximately balanced positive and negative loadings which indicate that unlike the VoCs block, in which the VoCs were generally more abundant in the natural spoiled samples, the SMM block contained metabolites which were more abundant in either type of samples. The best way to verify this is to examine the box-whisker plot of the metabolites which had the significant loadings. The annotated loadings plots and the box-whisker plot of a few selected metabolites with significant loadings are provided in the Electronic Supplementary Material (Figures S1–S11).

The subtle difference between the two blocks revealed by MB-PCA could at least partially be attributed to the sampling techniques we used for the study. The reason the VoCs block did not show a better separation between the two types of samples on the last time could be the results of excessive amount VoCs had been generated over a prolonged spoilage process. Since there were no unique VoC which would only be found in one type of sample, it was the relative difference in the abundance of the discriminate the samples. When all VoCs were excessively available in the headspace, the relative difference between the different VoCs decreased and resulted in less difference between the two types of samples. The less separation between different time points in the SMM block was also at least partially caused by the sampling. At 48 h after the spoilage, the biomass on the surface was too abundant to be extracted sufficiently by using two swabs. Thus, the extraction efficiency decreased at later time points and this could be the reason why in SMM block there were less separation between the late time points (24 h and afterwards). However, since the two types of samples had potentially very different microorganism communities on the surface and they were extracted in an exhaustive manner (unlike the partial sampling method used in the headspace VoCs sampling), the SMM block had shown clear separation between the two types of samples at each time point. Therefore, the results revealed by MB-PCA had shown that the two metabolic profiling methods were in fact complementary to each other and the MB-PCA model were able to utilise the information provided by both methods and gave a "global" model which showed almost perfect separation between different time points and also the two types of sample at each individual time point as showed in Fig. 2a.

It is worth noting that there is a caveat in MB-PCA models. As demonstrated by Westerhuis et al. [13], the information between the blocks could get mixed after deflation and thus the interpretations of the loadings of late PCs could be tricky. In MB-PLS models, this could be overcome by deflating the response block ***Y*** instead while it is not possible for MB-PCA. This is not a serious problem for this study as the factor of interest had been satisfactorily modelled by the first two PCs. Although for more difficult problems which cannot be sufficiently modelled by the first a few PCs of MB-PCA, a supervised MB-PLS model might be a better choice.

Table 1 The results of supervised prediction

	Spoilage time prediction		Sample type prediction
	Q^2	RMSECV	Cross-validated CCR (%)
MB-PLS	0.9533	5.5596	87.5
PLS on the SMM data	0.8097	11.7062	62.5
PLS on the VoCs data	0.9587	5.4511	58.3

The results of MB-PLS and PLS modelling were given Table 1. The PLS-R model applied to the VoC data set showed an impressive predictive performance with Q^2=0.9587 and RMSECV=5.4511. The PLS-R model applied to the SMM data set showed a relatively worse performance with Q^2=0.8097, RMSECV=11.7062. As discussed above, this could at least be partially attributed to the limitation of the sampling procedure. The PLS-DA model applied to the SMM data set had showed a slightly

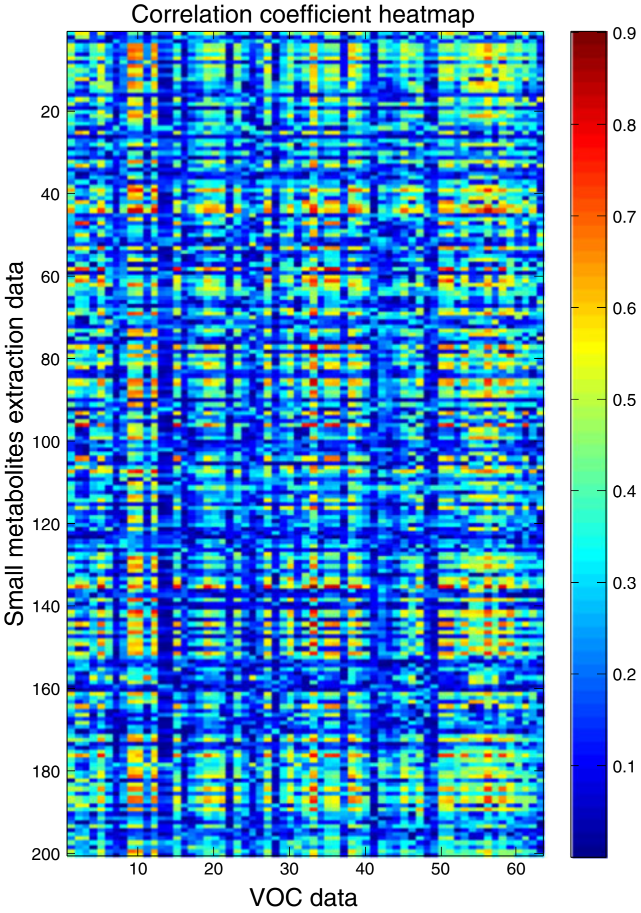


Fig. 4 Heat map of the correlation analysis. Pearson correlation analysis was performed to generate this plot. The *colour bar* shows the absolute values of the correlation coefficients

better performance (CCR=62.5 %) than that applied to the VoCs data set (CCR=58.3 %). Given the microbial consequences of the spoilage, one could expect the spoilage progression would be the most dominating factor in this study which was particularly true for the VoCs data set. Therefore, the PLS-R model applied to the VoCs data set had showed a better predictive performance. However, the type of sample classification had been approved to be a more difficult problem which is probably mainly caused by the fact that spoilage progression had been the most dominating factor, both PLS-DA models had only showed rather moderate predictive accuracies. On contrast, the MB-PLS models had showed the best over-all performance in which the MB-PLS-R model for spoilage progression modelling had a Q^2= 0.9533 and a RMSECV=5.5596 which was comparable to that of the PLS-R model applied to the VoCs data while the MB-PLS-DA model had showed a much better performance than both PLS-DA models with the prediction accuracy of 87.5 %. Most of the misclassified samples (5 out 6) were the ones at time 0 h which is rather expected since at the start the spoilage process had little effect to the sample. This has clearly demonstrated that for modelling complex systems, employing a multi-block approach which integrates multiple related data sets together could improve the predictive performance of the supervised model significantly compare to the ones which analyse each data set separately.

It is worth noting that even the PLS-DA models, which were applied to a single data set had generated a rather disappointing prediction accuracy for sample types, showed good separation between the two types of samples had been achieved by employing a more sophisticated model which were able to explicitly incorporate the experiment design into the model. In [5], PARAFAC 2 was applied to the SMM data set while in [4], the data was re-arranged and a MB-PCA model was fitted. By using these models, the two interacting factors (i.e. spoilage progression and the type of sample) were separated and satisfactory separations between the two types of samples were observed. Nevertheless, this study had showed that incorporating multiple data sets together is another way to increase the power of models and could be particularly useful for those applications when it is difficult to find a model which can fit the experiment design well.

The next stage was to explore additional methods to map both sets of data. The heat map of the correlation analysis is provided in Fig. 4. A noticeable feature in the heat map is that there were many 1-to-several correlations between the two data sets which indicates a very complicated network of these metabolites discovered by the two analytical platforms. There were 22 pairs of metabolites that obtained greater than 0.8 absolute values of correlation coefficients which involved 20 unique metabolites in which 8 were from the SMM block and 12 from VoCs block. We have putatively identified (using MSI terminology [21]) a few of these metabolites through mass spectra matching with the NIST02

Table 2 Tentative identification of the top correlated metabolites

Variable id.	Identification	Match factor	Rev. match factor	CAS#
SMM 44	Alanine	850	851	338-69-2
SMM 58	1,4-Butanediamine	769	821	110-60-1
SMM 85	Unknown	–	–	–
SMM 86	Ribitol	775	815	488-81-3
SMM 96	Tyramine	769	820	51-67-2
SMM 135	galactose	812	816	59-23-4
SMM 144	unknown	–	–	–
SMM 149	Linoelaidic acid	818	872	80969-37-5
VoC 2	Propanoic acid, 2-methyl-, pentyl ester	766	855	2050-01-3
VoC 5	4,5-Dimethyl-1-hexene	804	824	16106-59-5
VoC 9	Unknown	–	–	–
VoC 10	5-Methyl-pyrimidine	814	885	2036-41-1
VoC 12	Dimethyl disulfide	787	887	624-92-0
VoC 15	3-Dodecene	898	912	7239-23-8
VoC 33	Unknown	–	–	–
VoC 35	1-Butanol, 3-methyl-acetate	846	847	123-92-2
VoC 36	Unknown	–	–	–
VoC 51	Unknown	–	–	–
VoC 56	Dimethyl trisulfide	922	929	3658-80-8
VoC 58	2-Octanone	815	861	111-13-7

mass library. A putative identification was made if either the matching factor or the reverse match factor was greater than 800 and the results are given in Table 2. It is rather surprising to see that there were more VoCs than SMMs in the top correlated metabolite pairs since the SMM data have a total of 200 variables while the VoCs block only have 63. Also considering there were 200×63 (12,600) correlation coefficients in total, only 0.17 % of those had obtained a correlation coefficient greater than 0.8, this has seemingly contradicted with the very similar results of PCA obtained from each of the data set. The loadings plots of the MB-PCA provided a possible reason for this which is that the majority of the VoCs were increasing rapidly over time while there were many SMMs which did not show such a rapidly increasing trend (at least partially due to the limitation of the sampling technique used as discussed above). Such disparity may be the reason why there were not many metabolites between the two blocks that obtained high correlation coefficient. This can be verified by finding where these top correlated metabolites were located in the block loadings plots as shown in Fig. 3, which are the ones marked by circles in the figure. It is easy to see that all of these metabolites have significant PC 1 loadings in their corresponding block.

The top 20 most correlated metabolites were then analysed using Bayesian Networks and the results are shown in Fig. 5. The structure of the BN revealed some clear patterns from the data. First, it indicated that the metabolites selected from the VoCs data are more highly correlated amongst themselves (VoCs–VoCs intra-correlations) than the metabolites from the SMM data (SMMs–SMMs intra-correlations). This could be one of the reasons why there are more VoCs than SMMs in the top 20 most correlated metabolites. Other interesting features related to the VoCs and revealed by the network are: (1) the clustering of the VoCs into two main subgroups, subgroup 1 composed by nodes (9, 10, 12 and 56) and subgroup 2 composed by the other eight VoCs; and (2) the fact that subgroup 1 is connected to subgroup 2 by a unique metabolite represented by node 56 (dimethyl trisulfide; Table 2). In a future study, these metabolites could be further identified (i.e.

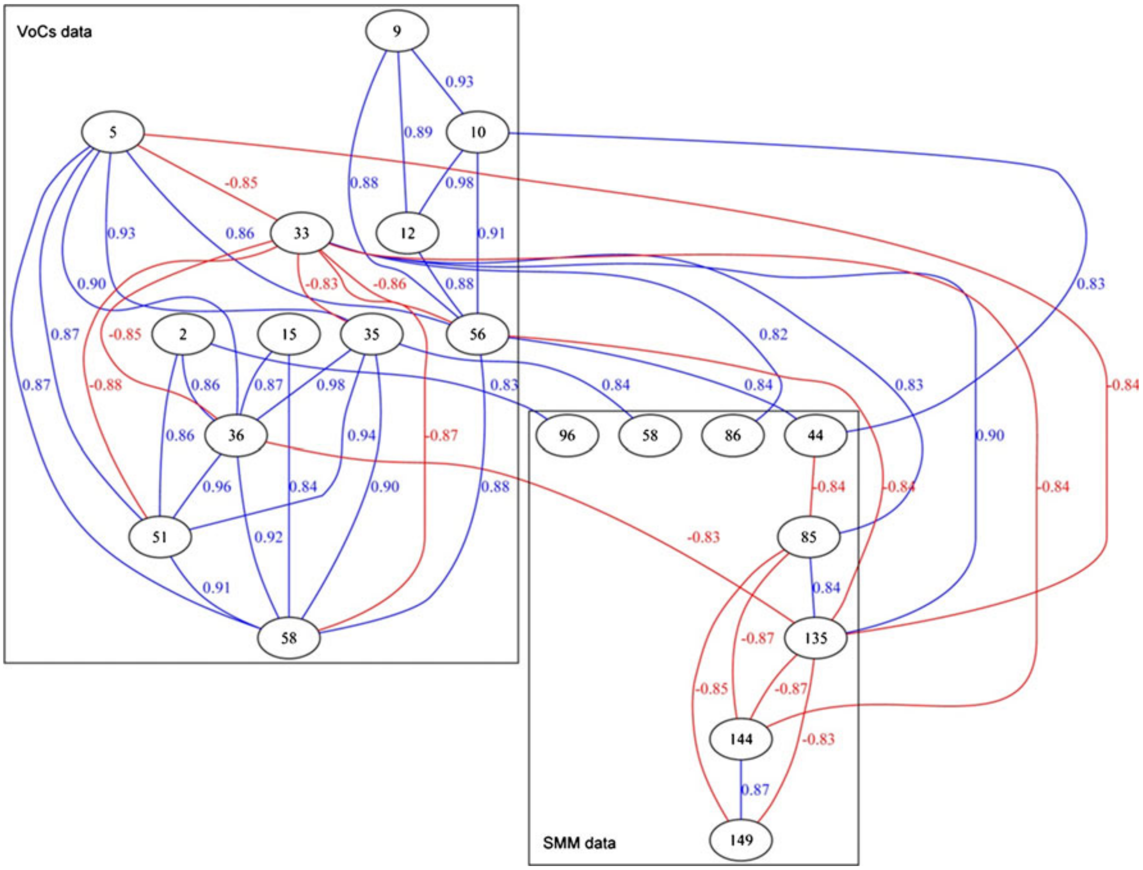


Fig. 5 Bayesian networks of the top correlated metabolites. The nodes of the network are clustered according to the data sets. The 12 nodes representing metabolites selected from the VoCs data are shown in the box labelled as "VoCs data" and the 8 nodes representing metabolites selected from the SMM data are shown in the box labelled as "SMM data"; the numbers within the nodes refer to the metabolite features and the putative identities are provided in Table 2. Positive correlations are represented by *blue edges* and negative ones by *red edges*. The number beside each edge represents the Pearson's correlation coefficient between the nodes (metabolites) linked by that edge. For each of the correlation coefficients shown on the network the *p* value is <0.001

definitive identification against a matched standard [21]) and the biological relevance of these correlations verified; note that one of the immediate neighbours of node 56 (trimethyl sulphide) is dimethyl disulphide (node 12) with a correlation coefficient of +0.88 so some inferences should be possible.

Concerning the main objective of this work, data integration, the network clearly detected the most relevant extra-correlations between VoCs and SMMs. For instance, the metabolite represented by node 33 (unknown identity) from the VoCs data is linked to four SMM metabolites, namely, 85 and, 144 (both unidentified) and 135 and 144; which are putatively identified as ribitol, a pentose alcohol, and galactose, a C4 epimer of glucose and so both are of sugar origins. Similarly, the metabolite represented by node 135 from the SMM data is correlated to four VoC metabolites (36, 56, 33 and 5), and three out of these four correlations are negative, except for VoC 33. The high number of extra-correlations observed for VoC 33 and SMM 135 are corroborated by the heat map depiction of the correlation analysis (see Fig. 4). Through mass spectra matching, SMM 135 could be tentatively identified as galactose while the identity of VoC 33 unfortunately remains unknown. If these metabolites could all be definitively identified, such information could provide important clues for further biological interpretation of the results which is the next step of this study. Finally, the BN analysis also revealed a higher number of extra-metabolite (between different platforms) correlations between VoCs–SMMs, than intra-metabolite correlations between SMMs–SMMs. This may be because, as the focus of this study is data integration, the 20 metabolites were selected solely based on extra-metabolite correlations between VoCs–SMMs. Nevertheless, the agreement between the results from the correlation analysis and the BN approach combined with its informative graphical model (Fig. 5) demonstrates that the BN analysis is a useful tool for the analysis of integrated data that have been collected by different analytical platforms.

Conclusions

In this study, we have demonstrated the advantages of using various chemometrics techniques to integrate the data generated by different analytical platforms together over the ones which applied to one data set only. The MB-PCA has established its ability to combine the strength of each of the data block together and generated a model with much better consistency with the biological expectations while still retaining its unsupervised nature. This improved consistency also made the interpretation of the model much easier. In addition, the block scores/loadings have revealed intriguing differences between the two analytical platforms which were very useful for improving the experiment design for further studies. The supervised MB-PLS models were also applied to the data sets and satisfactory results were obtained, especially the MB-PLS-DA had showed much better prediction accuracy than the PLS-DA models which were applied to a single data set

The BN model, which in this work was also applied as an unsupervised method, detected and displayed the most relevant correlations amongst the two different data blocks analysed. The main use of the BN in this study is to provide an intuitive view of the connections between the variables. The correlation analysis along with BN has indicated a list of potentially interesting metabolites from the volatile and liquid components of the food. These are most pertinent to natural spoilage and contamination with a specific food pathogen and we believe that with more metabolite identifications it will be possible to covert such findings into improved biological knowledge of the microbial action on meat. This will be the next stage of this study.

Acknowledgments We thank Dr. William Cheung, Dr. Catherine Winder and Dr. Warick Dunn for providing their support with generating the data. We are also grateful to Professor Tim Brocklehurst for provision of *S. typhimurium* strain 4/74. RG and YX thank the EU Framework 6 programme Biotracer for funding. E.C. and R.G. also acknowledge the EU Commonsense (www.fp7projectcommonsense.eu/) project (Grant 261809) financed by the European Commission under the seventh Framework Programme for Research and Technological Development.

Reference

1. Ellis DI, Goodacre R (2012) Metabolomics-assisted synthetic biology. Curr Opin Biotech 23:22–28
2. Dunn WB, Broadhurst D, Atherton HJ, Goodacre R, Griffin JL (2011) Systems level studies of mammalian metabolomes: the roles of mass spectrometry and nuclear magnetic resonance spectroscopy. Chem Soc Rev 40:387–426
3. Goodacre R, Vaidyanathan S, Dunn WB, Harrigan GG, Kell DB (2004) Metabolomics by numbers—acquiring and understanding global metabolite data. Trends Biotechnol 22:245–252
4. Xu Y, Cheung W, Winder CL, Goodacre R (2010) VOC-based metabolic profiling for food spoilage detection with the application to detecting *Salmonella typhimurium* contaminated pork. Anal Bioanal Chem 397:2439–2449
5. Xu Y, Cheung W, Winder CL, Dunn WB, Goodacre R (2011) Metabolic profiling of meat: assessment of pork hygiene and contamination with *Salmonella typhimurium*. Analyst 136:508–514
6. Riazanskaia S, Blackburn G, Harker M, Taylor D, Thomas CLP (2008) The analytical utility of thermally desorbed polydimethylsilicone membranes for in-vivo sampling of volatile organic compounds in and on human skin. Analyst 133:1020–1027
7. Winder CL, Dunn WB, Schuler S, Broadhurst D, Jarvis RM, Stephens GM, Goodacre R (2008) Global metabolic profiling of *Escherichia coli* cultures: an evaluation of methods for quenching and extraction of intracellular metabolites. Anal Chem 80:2939–2948
8. Goodacre R, Broadhurst D, Smilde A, Kristal BS, Baker JD, Beger R, Bessant C, Connor S, Capuani G, Craig A, Ebbels T, Kell DB, Manetti C, Newton J, Paternostro G, Somorjai R, Sjöström M,

Trygg J, Wulfert F (2007) Proposed minimum reporting standards for data analysis in metabolomics. Metabolomics 3:231–241

9. Smilde AK, Westerhuis JA, de Jong S (2003) A framework for sequential multiblock component methods. J Chemometr 17:323–337
10. Westerhuis JA, Kourti T, Macgregor JF (1998) Analysis of multiblock and hierarchical PCA and PLS models. J Chemometr 12:301–321
11. Kowalski BR, Wangen LE (1989) A multiblock partial least squares algorithm for investigating complex chemical systems, J Chemometr 3:3–20
12. Westerhuis JA, Coenegracht PMJ (1997) Multivariate modelling of the pharmaceutical two-step process of wet granulation and tableting with multiblock partial least squares. J Chemometr 11:379–392
13. Westerhuis JA, Smilde AK (2001) Deflation in multiblock PLS. J Chemometr 15:485–493
14. Wold S, Sjöström M, Eriksson L (2001) PLS-regression: a basic tool of chemometrics. Chemometr Intell Lab 58:109–130
15. Westerhuis JA, Hoefsloot HCJ, Smit S, Vis SD, Smilde AK, van Velzen EJJ, van Duijnhoven JPM, van Dorsten FA (2008) Assessment of PLSDA cross validation. Metabolomics 4:81–89
16. Camacho D, de la Fuente A, Mendes P (2005) The origin of correlations in metabolomics data. Metabolomics 1:53–63
17. Bouckaert RR (1994) In Lopez de Mantaras, R., D. Poole, D. (ed) Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann, San Francisco, CA
18. Cooper GF, Herskovits EA (1992) Bayesian method for the induction of probabilistic networks from data. Mach Learn 7:299–347
19. de Campos CP, Zeng Z, Ji Q (2009) Structure learning of Bayesian networks using constraints. ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning, p. 113–120, Montreal, Quebec, Canada. ISBN: 978-1-60558-516-1
20. Pearl J (2000) Causality: models, reasoning and inference. Cambridge University Press, Cambridge
21. Sumner LW, Amberg A, Barrett D, Beger R, Beale MH, Daykin C, Fan TW-M, Fiehn O, Goodacre R, Griffin JL, Hardy N, Higashi R, Kopka J, Lindon JC, Lane AN, Marriott P, Nicholls AW, Reily MD, Viant M (2007) Proposed minimum reporting standards for chemical analysis. Metabolomics 3:211–221