Anal. Chem. **2005,** *77,* 6729−6736

Fusion of Mass Spectrometry-Based Metabolomics Data

Age K. Smilde,* Mariët J. van der Werf, Sabina Bijlsma, Bianca J. C. van der Werff-van der Vat, and Renger H. Jellema

TNO Quality of Life, P. O. Box 360, 3700 AJ Zeist, The Netherlands

A general method is presented for combining mass spectrometry-based metabolomics data. Such data are becoming more and more abundant, and proper tools for fusing these types of data sets are needed. Fusion of metabolomics data leads to a comprehensive view on the metabolome of an organism or biological system. The ideas presented draw upon established techniques in data analysis. Hence, they are also widely applicable to other types of X-omics data provided there is a proper pretreatment of the data. These issues are discussed using a real-life metabolomics data set from a microbial fermentation process.

Recent years have seen a revival of system theory thinking in biology. Although system theory is already a relatively old discipline, there is an increasing awareness that this way of thinking is very fruitful for solving fundamental biological problems.[1] This revival is also stimulated by advanced analytical chemical methods for measuring key biological compounds, such as DNA, mRNA (transcriptomics), proteins (proteomics), and metabolites (metabolomics). Systems biology is by now a new field in biology,[2,3] and studies emerge integrating X-omics measurements in different fields of (medical) biology.[4,5]

One of the X-omics technologies that is increasingly used in systems biology is metabolomics. There are different terms for measuring metabolites in a system, such as metabolic profiling, metabolic fingerprinting, metabonomics, and metabolomics. A nice overview of these terms including their definitions is given by Fiehn.[6] In metabolomics, the goal is to measure as many metabolites as possible in a biological system. This gives a holistic view of the metabolism of the system studied and creates opportunities for studying metabolic networks.[7,8] Moreover, metabolites form a crucial part of the link between genotype and phenotype.[6,9]

To measure metabolites in a biological system, samples have to be taken and analyzed. There are different instrumental methods available for analyzing such samples, e.g., NMR and mass spectrometry. For metabolomics, mass spectrometry (MS) in combination with a separation method, e.g., gas chromatography (GC/MS) or liquid chromatography (LC/MS), is a very powerful and often used instrumental method.[6,10]

In GC/MS and LC/MS, different classes of compounds are measured although some groups of compounds are measured with both GC and LC (i.e., sugar nucleotides, aromatic amino acids, sugar phosphates, and certain nucleosides such as adenosine). Some components can be differentiated better on one platform than the other, e.g., sugar phosphates. This group of compounds results in one large peak in LC analysis whereas in GC analysis the sugar phosphates are more or less separated from each other in retention time. Due to these complementarities, it is worthwhile to consider merging GC/MS and LC/MS measurements performed on the same samples to obtain a more complete overview of the metabolites in the samples. As an example, the two analytical methods allow the detection of 93% of commercially available metabolites of the in silico metabolome of *Bacillus subtilis* (unpublished results) and are expected to have similar coverage when analyzing the metabolome of *Escherichia coli*.

Data fusion is the part of data analysis studying fusion of data sets of different origins. Fusion of data can be done in different ways and on different levels.[11] This also has repercussions for fusion of metabolomics data. Fusion of different MS-based metabolomics methods is, however, not straightforward. Especially the megavariate nature of the data (i.e., a very high variables-to-sample ratio) deserves attention.

An approach of fusing MS-based metabolomics data is illustrated by an example from microbial metabolomics. We are interested in selection and ranking of targets that are limiting bioproduct formation. To this end, we aim to identify metabolites correlating with bioproduct formation.[12] We recently performed a study in which our aim was to identify bottlenecks in phenylalanine production in *E. coli* NST 74 and to use this information to construct improved production strains. The NST 74 strain was cultivated under different environmental conditions, and from

* To whom correspondence should be addressed. E-mail: smilde@voeding.tno.nl. Phone: +31 30 694 4527. Fax: +31 30 694 4894.

(1) Capra, F. *The web of life*; Anchor Books: New York, 1996.
(2) Kitano, H. *Science* **2002**, *295*, 1662−4.
(3) Ideker, T.; Galitski, T.; Hood, L. *Annu, Rev. Genomnics Hum. Genet.* **2001**, *2*, 372.
(4) Ideker, T.; Thorsson, V.; Ranish, J. A.; Christmas, R.; Buhler, J.; Eng, J. K.; Bumgarner, R.; Goodlett, D. R.; Aebersold, R.; Hood, L. *Science* **2001**, *292*, 929−34.
(5) Clish, C. B.; Davidov, E.; Oresic, M.; Plasterer, T. N.; Lavine, G.; Londo, T.; Meys, M.; Snell, P.; Stochaj, W.; Adourian, A.; Zhang, W.; Morel, N.; Neumann, E.; Verheij, E.; Vogels, J. T.; Havekes, L. M.; Regnier, F.; van der Greef, J.; Naylor, S. *Omics* **2004**, *1*, 3−13.
(6) Fiehn, O. *Plant Mol. Biol.* **2002**, *48*, 155−71.
(7) Oltvai, Z. N.; Barabasi, A. L. *Science* **2002**, *298*, 763−4.
(8) Steuer, R.; Kurths, J.; Fiehn, O.; Weckwerth, *Bioinformatics* **2003**, *19* (8), 1019−26.
(9) Goodacre, R.; Kell, D. B. In *Metabolic profiling: its role in biomarker discovery and gene function analysis*; Harrigan, G. G., Goodacre, R., Eds.; Kluwer Academic Publishers: Boston, 2003.
(10) van der Werf, M. J.; Jellema, R. H.; Hankemeier, T. *J. Ind. Microbiol. Biotechnol.* **2005**, *32*, 234−52.
(11) Roussel, S.; Bellon-Maurel, V.; Roger, J. M.; Grenier, P. *Chemom. Intell. Lab. Syst.* **2003**, *65*, 209−19.
(12) van der Werf, M. J. *Trends Biotechnol.* **2005**, *23*, 11−6.

10.1021/ac051080y CCC: $30.25
Published on Web 09/17/2005

these fermentations of in total 28 samples, the metabolomes were analyzed using a GC/MS and LC/MS method in the "holistic" mode: screening for as much metabolites as possible. Subsequently, the aim was to identify metabolites that correlate in a robust way with the phenylalanine yield, as these might be indicative for those parts of the metabolism that limit phenylalanine production in this strain. Moreover, it is the goal to rank these leads based upon the strength of the correlation, as it is too time-consuming and expensive to follow up or validate all potential leads.[12,13] This example shows the potential of combining data fusion techniques with subsequent multivariate analysis tools.

EXPERIMENTAL SECTION

Background of the Data Set. *E. coli* NST 74, the phenylalanine overproducing strain, and *E. coli* W3110, the wild-type strain, were obtained from the ATCC (ATCC31884, Manassas, VA). Cultures were grown batch at 30 °C in a Bioflow II (New Brunswick Scientific) bioreactor containing 2 L of MMT12 medium[14] with 30 g/L glucose as carbon source. The culture was inoculated with 5% (v/v) of a preculture grown for 16 h on the same medium (200 rpm; 30 °C). A constant pH (pH 6.5) was maintained by the automatic titration with 4 M KOH and 1 M HCl. The oxygen tension was maintained at 30% by automatic increase of the stirring speed in the fermentor. Samples were taken from the bioreactor after 16, 24, 40, and 48 h and immediately quenched at −45 °C in methanol as previously described by de Koning and van Dam.[15] Variations in this standard fermentation protocol were introduced by changing one of the default conditions, resulting in a screening experiment (see Table 1). This design was chosen in a first exploratory step. Experiment 1 is the standard fermentation, and the alterations in the default conditions are listed in Table 1 together with the numbering of the experiments. Experiment 8 did not give satisfactory results and was therefore omitted.

The intracellular metabolites were extracted from the samples by chloroform extraction, as described by de Koning and van Dam.[15] At appropriate time points, different quality and internal standards were added. After extraction, the sample was divided in two portions. The LC/MS sample was deproteinized. Subsequently, both the GC/MS and LC/MS samples were lyophilized.

LC/MS. Lyophilized metabolome samples were dissolved in 0.1 mL of methanol/water. Samples (25 μL) were separated on a reversed-phase column (Chrompack Inertsil 5 μm ODS-3 100 × 3 mm, Middelburg, The Netherlands) using a 40-min linear gradient from 5 mM hexylamine (pH 6.3) to 100% of 90% methanol/10 mM ammonium acetate (pH 8.5) at a flow rate of 0.4 mL/min.

To obtain the yield, culture supernatants (5 μL) were separated on a HILIC TSK Amide 80 column (100 × 2 mm, 5 μm) using a linear gradient from 95% MeCN/5% 10 mM NH4Ac (pH 5.5) to 5% MeCN/95% 10 mM NH4Ac (pH 5.5) at a flow rate of 0.15 mL/min and a column temperature of 30 °C.

Compounds were detected by electrospray ionization (ESI; negative ion mode) using a Finnigan LTQ linear ion trap mass spectrometer. During data acquisition, the mass spectrometer probe voltage was maintained at 3−4 kV and the heated capillary was kept at 250 °C. For detection and quantification an MS2 method (m/z 169 → 147) was used.

Table 1. Experimental Design

no.	fermentn times (h)	carbon source	phosphate concn[a]	oxygen (%)	pH	strain
1.1	16	glucose	1	30	6.5	NST 74
1.2	24					
1.3	40					
1.4	48					
2.2	32	glucose	1	30	6.5	NST 74
3.3	40	glucose	1	2	6.5	NST 74
4.1	16	glucose	1	unknown	6.5	NST 74
4.2	24					
4.3	40					
4.5	64					
5.1	16	glucose	3	30	6.5	NST 74
5.2	24					
5.3	40					
5.4	48					
6.1	16	glucose	1/3	30	6.5	NST 74
6.2	24					
6.3	40					
7.2	24	succinate	1	30	6.5	NST 74
7.3	40					
7.4	48					
8	---	---	---	---	---	---
9.1	16	glucose	1	30	6.5	W3110
9.2	24					
9.3	40					
9.4	48					
10.1	16	glucose	1	30	7	NST 74
10.2	24					
10.3	40					
10.4	48					

[a] Phosphate relative to the reference concentration in experiment 1 (13.2 mM).

GC/MS. Lyophilized metabolome samples were derivatized using a solution of ethoxyamine hydrochloride in pyridine as the oximation reagent followed by silylation with *N*-trimethyl-*N*-trimethylsilylacetamide principally as described by ref 16. The 1-μL aliquots of the derivatized samples were injected in splitless mode on a HP5-MS 30m × 0.25 mm × 0.25 μm capillary column using a temperature gradient from 70 to 325 °C at a rate of 15 °C/min. Gas chromatographic analyses were performed using a 6890 gas chromatograph (Agilent Technologies, Palo Alto, CA). A Pegasus III time-of-flight mass spectrometer (Leco Corp., St. Joseph, MI) was used as detector. Detection was performed using MS detection in electron impact (EI) mode (70 eV).

Preprocessing. The data from the GC/MS and LC/MS analyses were delivered as .cdf files. These GC data were subsequently preprocessed using software packages Impress version GC/MS[17] and WinLin version V2.4[18] (TNO Quality of Life, Location Zeist, The Netherlands). The LC data were preprocessed using software packages Impress V1.2, WinLin V2.4 and an in-house-developed peak-picking algorithm. Subsequently, the output from Winlin (in the form of mass.retentiontime and intensity information) were used as the input for subsequent data analysis. The LC/MS data

(13) Stephanopoulos, G.; Alper, H.; Moxley, J. *Nat. Biotechnol.* **2004**, *22*, 1261−7.
(14) Tribe, D. E. Novel microorganism and methods. U.S. Patent 4,681,852], 1983.
(15) de Koning, W.; van Dam, K. *Anal. Biochem.* **1992**, *204*, 118−23.
(16) Fiehn, O.; Kopka, J.; Trethewey, R. N.; Willmitzer, L. *Anal. Chem.* **2000**, *72*, 3573−80.
(17) van der Greef, J.; Vogels, J. T. W. E.; Wulfert, F.; Tas, A. C. Method and system for identifying and quantifying chemical components of a mixture. U.S. Patent 2004267459, 2004.
(18) Vogels, J. T. W. E.; Tas, A. C.; Venekamp, J.; van der Greef, J. *J. Chemom.* **1996**, *10*, 425−38.

set consisted of 56 measurements (28 duplicates) and 8525 variables, and the GC/MS data set consisted of 56 measurements (28 duplicates) and 24 095 variables. The duplicate measurements were combined after correction for internal standard. In this way, a reduced data set was obtained for the LC/MS set of size 28 × 8525 and for the GC/MS set of size 28 × 24095.

Both data sets contained many zeros. For the LC/MS data set, 80% of the data contained a zero, and for the GC/MS set, 60% of the data contained a zero. To reduce the number of zeros present, the following procedure was applied, which will be referred to as the "80% rule". Every sample can be assigned to a certain experiment (experiments 1−10, with the exception of experiment 8, which is not present). For both data sets mentioned, a variable is kept if the variable has a nonzero value for at least 80% of all samples for at least one experiment. After the 80% rule, the LC/MS data set is of size 28 × 2532 and the GC/MS set is of size 28 × 12553, which indicates a large reduction of variables.

After this 80% rule, there are still many zeros left in the data set. For the LC/MS data set, 50% of the data contained a zero, and for GC/MS, 35% of the data contained a zero. These zeros are artificial cutoffs from the software, and these are replaced by the minimum value of that particular data set nonequal to zero divided by two. This is done to avoid a large gap between the lowest real value in the data set and the artificial zero.

Data analysis was performed in the Matlab environment (version 6.5.1 Release 13, The Mathworks, 2003) using home-written routines and the PLS Toolbox (version 3.0.2, Eigenvector Research, 2003).

THEORY

When considering fusion of MS-based metabolomics data, the first decision to be made is the level of fusion. Second, preprocessing of the individual MS data sets is important, since it can dramatically influence the end result. Third, the choice of the subsequent multivariate data analysis method is important. Fourth, an operational definition of "robust correlation" is needed and a way to select variables. The four issues are dealt with in separate subsections.

The following notation is used. Scalars are written as lowercase italic characters (e.g., x); vectors as bold lowercase characters (e.g., $\mathbf{x}$); matrices as bold uppercase characters (e.g., $\mathbf{X}$); and three-way arrays as bold underlined characters (e.g., $\underline{\mathbf{X}}$). The characters i, j, and k are used as indices which run from 1 to I, J, and K, respectively.

Levels of Data Fusion. On the lowest level, data fusion comes down to simply concatenating the matrices of measurements in such a way that the samples are the shared mode. Hence, all variables measured on the samples are simply put next to each other. There are indications in the literature that this is not the optimal way of fusion.[11] Moreover, in the present case, this would result in a data matrix of size 28 samples and 15 085 variables with a very unfavorable samples-to-variables ratio.

It is expected that many of the variables are not of primary importance for the yield. Therefore, a variable screening method is used based on ideas on preliminary variable selection in PLS. This method ranks the individual variables according to their covariance with the response (yield) and using a cutoff point.[19] The data matrices are subsequently concatenated for obtaining the model. This way of processing the data is sometimes called midlevel fusion.

It is also possible to make two separate models, one for the GC/MS data and one for the LC/MS data. The predicted values can then be combined, e.g., by averaging. In classification models, combining is often done with voting schemes. This is high-level data fusion. However, such combining of results has two disadvantages in the current case: (i) it does not give transparent models, i.e., interpretation of model results is difficult, and (ii) correlations between measurements in both blocks are not taken into account. Moreover, essentially the same types of compounds (metabolites) are measured with both methods, and there is no a priori reason not to put them on an equal footing. Thus, high-level fusion is not considered here.

Preprocessing of MS Data prior to Fusion. In the ideal case, each component measured within a sample would end up in the data matrix as one number representing the concentration of that component within that specific sample. This would be possible if overlapping peaks can be separated from each other in a mathematical manner, which is called deconvolution or peak resolving.[20] A number of commercial packages are available, but after testing these on a benchmark data set, none of them were found to work satisfactory (unpublished results). Hence, deconvolution was not applied and the preprocessed data were used (see Experimental Section).

Prior to modeling, the data were mean-centered columnwise in order to remove offsets.[21] To put all measured intensities on an equal footing, range scaling was applied, which means that the measured intensity was divided by the range of those intensities over all samples. This type of scaling also has the advantage of removing instrumental response factors from the data, generating relative concentrations for each variable (see Appendix). Hence, the result is a (centered) matrix of samples times relative concentrations. A property of range scaling is that all levels of variation of the metabolites are treated equally. This may or may not be desired. The advantage of range scaling (putting all variables on an equal footing) was deemed to be important enough to use this type of scaling.

Multiblock Methods. Several multivariate analysis tools are available for simultaneous analysis of multiple sets of data. These methods are called multiset or multiblock methods or models. A class of multiblock models are so-called latent variable multiblock models. Such models work by extracting latent variables from each block and consider these latent variables as information carriers. In a theoretical comparison of different sequential multiblock component methods (in which each component is calculated sequentially), consensus-PCA (CPCA) came out as the best alternative.[22] For regression models, multiblock-PLS (MB-PLS) is a viable alternative.[23] Both methods have proven their usefulness in various applications.[24,25]

Both CPCA and MB-PLS work on properly concatenated and matricized three-way arrays. For simplicity, assume the availability of a GC/MS data set ($\underline{\mathbf{X}}_1$), an LC/MS data set ($\underline{\mathbf{X}}_2$), and a response vector yield ($\mathbf{y}$). The GC/MS data set $\underline{\mathbf{X}}_1$ has size $I \times J_1 \times K_1$,

(19) Höskuldsson, A. *Chemom. Intell. Lab. Syst.* **2001**, *55*, 23−38.

(20) Tauler, R.; Kowalski, B. R.; Flemming, S. *Anal. Chem.* **1993**, *65*, 2040−7.
(21) Bro, R.; Smilde, A. K. *J. Chemom.* **2003**, *17*, 16−33.
(22) Smilde, A. K.; Westerhuis, J. A.; De Jong, S. *J. Chemom.* **2003**, *17*, 323−37.

where I is the number of samples, J_1 is the number of mass channels, and K_1 is the number of GC scans. When preprocessed properly (e.g., removing retention time shifts), such a three-way array has a trilinear structure.[26,27] The same holds for the LC/MS data set $\underline{\mathbf{X}}_2$ with size $I \times J_2 \times K_2$, where J_2 is the number of mass channels and K_2 is the number of LC scans. Note that J_1 does not equal J_2 since a different type of MS is used hyphenating the GC or LC. The yield vector $\mathbf{y}$ has size $I \times 1$, assuming a single biological response. The common mode of all the arrays is the I-mode, since, all measurements are performed on the same samples.[28] This observation has consequences for the subsequent models.

In principle, three-way analysis methods such as PARAFAC and Tucker decompositions could be used to analyze $\underline{\mathbf{X}}_1$ and $\underline{\mathbf{X}}_2$.[27,29−31] There are also three-way multiblock methods available[32,33] that have shown their usefulness in chemical engineering.[34] However, the characteristics of metabolomics data are different from chemical or chemical engineering data, the most striking differing feature being the contribution of very many chemical components to the signal in metabolomics data. Hence, it is not immediately clear how standard three-way methods perform on these data. Since the focus of the current paper is on fusion methodology, the three-way arrays are matricized to two-way arrays to allow for the use of two-way methods.

The matricization operation results in data sets X_1 ($I \times J_1K_1$) for the GC/MS data and in X_2 ($I \times J_2K_2$) for the LC/MS data sets. In both cases, new variables are formed by combining the retention times and m/z values. For simplicity, these combined variables are indexed by $j_1 = 1, ..., J_1$ for the GC/MS data, and by $j_2 = 1, ..., J_2$ for the LC/MS data. For analyzing the consensus between the GC/MS and LC/MS data sets, a CPCA model can be built:

$$\mathbf{X}_1 = \mathbf{T}\mathbf{P}_1' + \mathbf{E}_1$$
$$\mathbf{X}_2 = \mathbf{T}\mathbf{P}_2' + \mathbf{E}_2 \qquad (1)$$

where $\mathbf{T}$ ($I \times R$) represents the scores on the R common (consensus) components, $\mathbf{P}_1$ ($J_1 \times R$) and $\mathbf{P}_2$ ($J_2 \times R$) are the loadings of the blocks $\mathbf{X}_1$ and $\mathbf{X}_2$ on these R consensus components, and $\mathbf{E}_1$, $\mathbf{E}_2$ contain the residuals. Equation 1 can be written in a more compact way by using concatenation:

$$\mathbf{X} = [\mathbf{X}_1\ \mathbf{X}_2] = \mathbf{T}[\mathbf{P}_1'\ \mathbf{P}_2'] + [\mathbf{E}_1\ \mathbf{E}_2] \qquad (2)$$

and the least-squares solution for $\mathbf{T}$, $\mathbf{P}_1$, and $\mathbf{P}_2$ can be calculated by performing a principal component analysis on $\mathbf{X}$. For details on the algorithms of CPCA and its properties, see elsewhere.[22]

Multiblock-PLS can also be written in terms of the concatenated matrix $\mathbf{X}$:

$$\mathbf{X} = [\mathbf{X}_1\ \mathbf{X}_2] = \mathbf{T}\mathbf{P}' + \mathbf{E}$$
$$\mathbf{y} = \mathbf{T}\mathbf{q} + \mathbf{f}$$
$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} \qquad (3)$$

where $\mathbf{T}$ are the scores (different from the CPCA scores, but for convenience using the same symbol), $\mathbf{P}$ the X-block loadings, $\mathbf{q}$ the y-block loadings, $\mathbf{b}$ the vector of regression coefficients, and $\mathbf{E}$, $\mathbf{f}$ residuals in the X and y-block, respectively. Hence, an ordinary PLS1 is performed using $\mathbf{X}$ and $\mathbf{y}$; after properly rearranging the loadings and scores, results per block can be obtained. For details regarding multiblock-PLS and its algorithms, see elsewhere.[23]

The number of variables is much larger for the GC/MS data than for the LC/MS data. This does not reflect, however, the fact that GC/MS measures more metabolites. This difference is due to the properties of the used GC/MS in which an ionization technique (EI) is used that produces more ions per metabolite (10−40) than the ESI used for the LC/MS measurements, which results generally in 3−10 ions per metabolite. Since multiblock methods are sensitive to such differences, block scaling was applied, meaning that both blocks of data (GC/MS and LC/MS) were given the same sum of squares prior to modeling. This can be done with a simple overall scaling constant for each block of measurements.

Robust Correlation and Variable Selection. To measure the correlation between the X-variables and the y-variable, the multiple correlation coefficient from the regression of y on X could be used (R^2_{fit}). This is, however, not a good idea since this coefficient is far too optimistic due to overfitting. This will be shown in the results. A better measure is the cross-validated multiple correlation coefficient (R^2_{cv}), which can be calculated easily from the cross-validation calculations. This will also be shown in the results.

Variable selection is a difficult problem in situations with a high variable-to-sample ratio. Classical variable selection methods in regression analysis, e.g., forward selection and stepwise regression,[35] fail due to the high number of variables. In the PLS literature, several alternatives have been suggested, e.g., using the regression coefficients (the b values, see eq 3) or the VIP statistic.[36,37] These latter two methods will be discussed in the Results section.

RESULTS AND DISCUSSION

Consensus-PCA. A high correlation between a column in X_1 and X_2 can mean two things: either the same metabolite is measured on the different instruments or two metabolites are measured (one with GC/MS and the other one with LC/MS) that are highly correlated due to their association in the biology by an enzymatic reaction with an equilibrium constant close to 1.

(23) Westerhuis, J. A.; Kourti, T.; MacGregor, J. F. *J. Chemom.* **1998**, *12*, 301−21.
(24) Skagerberg, B.; MacGregor, J. F.; Kiparissides, C. *Chemom. Intell. Lab. Syst.* **1992**, *14*, 341−56.
(25) Lopes, J. A.; Menezes, J. C.; Westerhuis, J. A.; Smilde, A. K. *Biotechnol. Bioeng.* **2002**, *80*, 419−27.
(26) Sanchez, E.; Kowalski, B. R. *J. Chemom.* **1988**, *2*, 265−80.
(27) Smilde, A. K.; Bro, R.; Geladi, P. *Multi-way analysis. Applications in the chemical sciences*; John Wiley & Sons: Chichester, 2004.
(28) Tauler, R.; Smilde, A. K.; Kowalski, B. R. *J. Chemom.* **1995**, *9*, 31−58.
(29) Harshman, R. A. *UCLA Work. Pap. Phonetics* **1970**, *16*, 1−84.
(30) Carroll, J. D.; Chang, J. *Psychometrika* **1970**, *35*, 283−319.
(31) Tucker, L. R. *Psychometrika* **1966**, *31*, 279−311.
(32) Bro, R. *J. Chemom.* **1996**, *10*, 47−61.
(33) Smilde, A. K.; Westerhuis, J. A.; Boqué, R. *J. Chemom.* **2000**, *14*, 301−31.
(34) Boque, R.; Smilde, A. K. *AIChE J.* **1999**, *45*, 1504−20.
(35) Draper, N. R.; Smith, H. *Applied regression analysis*, 3rd ed.; John Wiley & Sons: New York, 1998.
(36) Wold, S.; Sjöström, M.; Eriksson, L. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109−30.
(37) Perez-Enciso, M.; Tenenhaus, M. *Hum. Genet.* **2003**, *112*, 581−92.

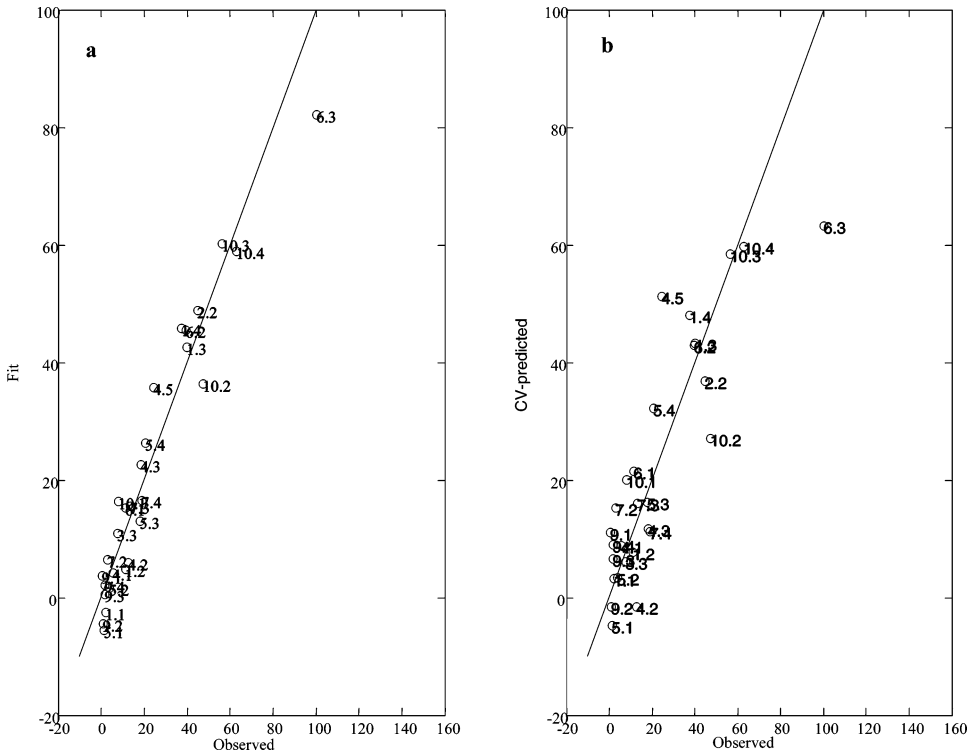

Figure 1. Parity plots for the low-level MB-PLS model: (a) fit vs observed and (b) CV predicted vs observed. The used sample labels correspond to the experiment numbers listed in Table 1.

Only a detailed look at the mass spectrometry (identification using databases) and of the biological pathways present in the organism studied can distinguish between both cases. This also has repercussions for the interpretation of the CPCA and MB-PLS methods.

To get an impression of the amount of overlap between the GC/MS and LC/MS measurements, techniques are available to calculate correlations between matrices.[38] Directly applying such techniques would, however, lead to spurious results because of the high dimensionality of the data sets.[22] Likewise, applying directly canonical correlation analysis between X_1 and X_2 would lead to trivial results due to the fact that both X_1 and X_2 have more columns than rows and are of full rank (28). Hence, their column spaces overlap. To avoid trivial results, first a dimension reduction using PCA is performed to reduce the data sets to their essential variability. This also filters out some noise in both data sets thereby focusing more on essential correlations. After that, criteria for correlating the reduced matrices can be used. Of these, the RV coefficient measures the degree of overlap between matrices allowing for rotation and isotropic scaling.[38,39] Stated otherwise, the RV coefficient measures to what degree the LC/MS and GC/MS data set give the same view on the samples. Table 2 summarizes the results.

After three components, the RV coefficient flattens off to ~60%. This means that there is overlap between the GC/MS and LC/MS data, but there is also a substantial nonoverlapping part. This is confirmed by the amounts of explained variation of the consensus components in both GC/MS and LC/MS. These amounts are also around 40−60% for 4−6 principal components. Hence, both types of data have overlap but also clearly unique parts.

(38) Ramsay, J. O.; Ten Berge, J. M. F.; Styan, G. P. H. *Psychometrika* **1984**, *49*, 403−23.
(39) Robert, P.; Escoufier, ? *Appl. Stat.* **1976**, *25*, 257−65.

Table 2. Summary of (Dis)similarities between the GC/MS and LC/MS Data

PC[a]	GC/MS[b]	LC/MS[c]	RV[d]
1	16.8	12.3	0.17
2	30.4	23.1	0.46
3	42.4	30.5	0.60
4	48.8	37.1	0.59
5	55.9	42.8	0.59
6	60.2	49.4	0.60

[a] The principal component number. [b] The amount of explained variation of the consensus component in the GC/MS data. [c] The amount of explained variation of the consensus component in the LC/MS data. [d] The RV coefficient: a matrix correlation coefficient (for more explanation, see the text).

Multiblock-PLS. The multiblock calculations start with using the full data set, i.e., all variables (15 085; 12 553 for GC/MS and 2532 for LC/MS) for all samples (28). The parity plots (fit versus observed values) are shown in Figure 1. The fit (Figure 1a) is judged with the R^2_{fit} value (0.93). Cross-validation is used to select the number of latent variables and to obtain the cross-validated multiple correlation coefficient. Cross-validation was performed using the venetian blind method with four splits. In the first step, six samples were left out and the remaining samples were used to build a PLS model. The model was used to predict the class assigment of the "left out" samples. Next, six other samples were left out. This was repeated until all samples were left out once. Cross-validation indicated a MB-PLS model with two components, and the cross-validated predictions are shown in Figure 1b, which resulted in a R^2_{CV} of 0.75. This is low-level fusion, without any variable selection. The corresponding values for the PLS analyses of the separate GC/MS and LC/MS data were $R^2_{fit} = 0.91$; $R^2_{CV} = 0.61$ (3 LVs) and $R^2_{fit} = 0.99$; $R^2_{CV} = 0.78$ (4 LVs), respectively.

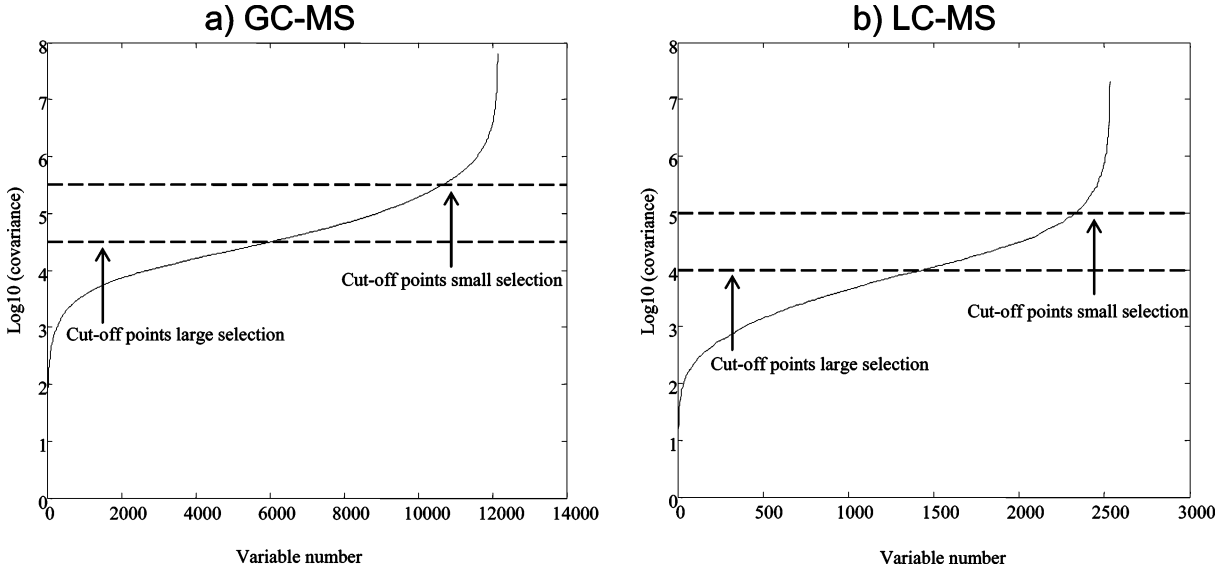

Figure 2. Covariance plots for GC/MS (a) and LC/MS (b).

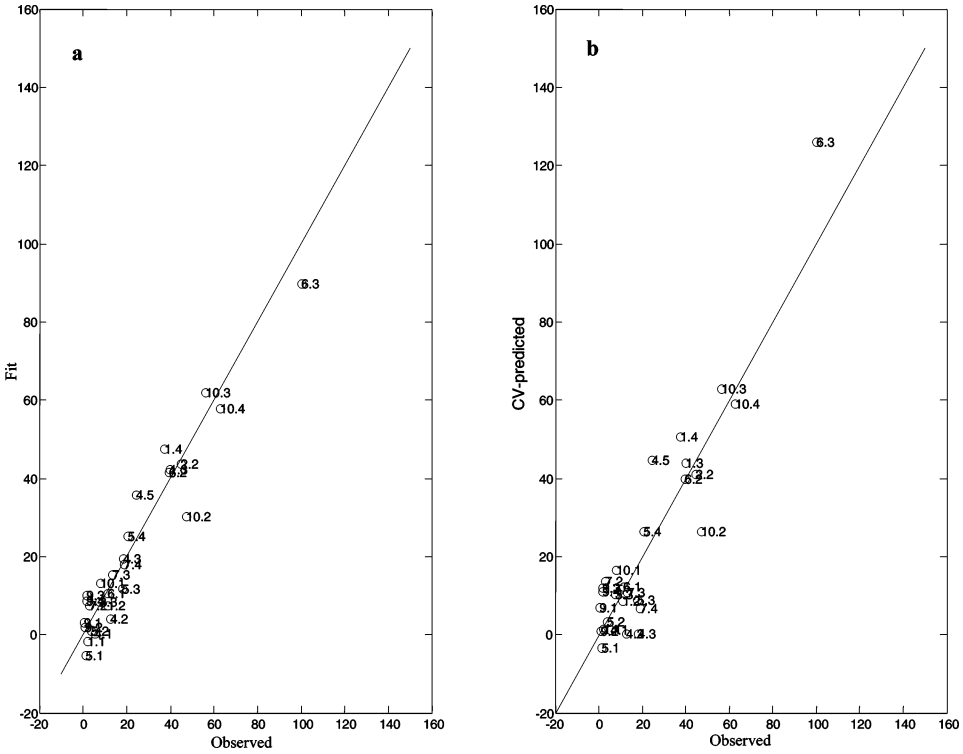

Figure 3. Parity plots for the medium-level MB-PLS model: (a) fit vs observed and (b) CV-predicted vs observed. The used sample labels correspond to the experiment numbers listed in Table 1.

Hence, in terms of performance, MB-PLS gives results comparable to the single LC/MS analysis.

The covariance plots for GC/MS and LC/MS are shown in Figure 2a and b on a logarithmic scale. Two cutoff points are chosen visually, which results in a "small" and a "large" selection. The large selection reduces the number of GC/MS variables to 6103 and the number of LC/MS to 1089. The resulting R^2_{fit} and R^2_{CV} are 0.93 and 0.82, respectively, again using two MB-PLS components based on cross-validation. Cutting down on the number of variables even further (small selection: 1491 GC/MS and 195 LC/MS variables) gives an R^2_{fit} and R^2_{CV} of 0.93 and 0.86, respectively, again using two MB-PLS components based on cross-validation. The parity plots are shown in Figure 3. This is medium-level data fusion, and it clearly pays to do a variable importance screening prior to PLS model building.

The final MB-PLS model has a high R^2_{CV} value (0.86). Moreover, since this is a cross-validated value, it is not prone to overfitting and can be considered as a robust correlation. The model simultaneously considers the LC/MS and GC/MS measurements, but its multiblock structure allows for a separate interpretation of the different measurement blocks. The next step would be to investigate the model parameters to obtain indications for the importance of metabolites in terms of their relationship with yield. The VIP statistic could be used for this purpose, but

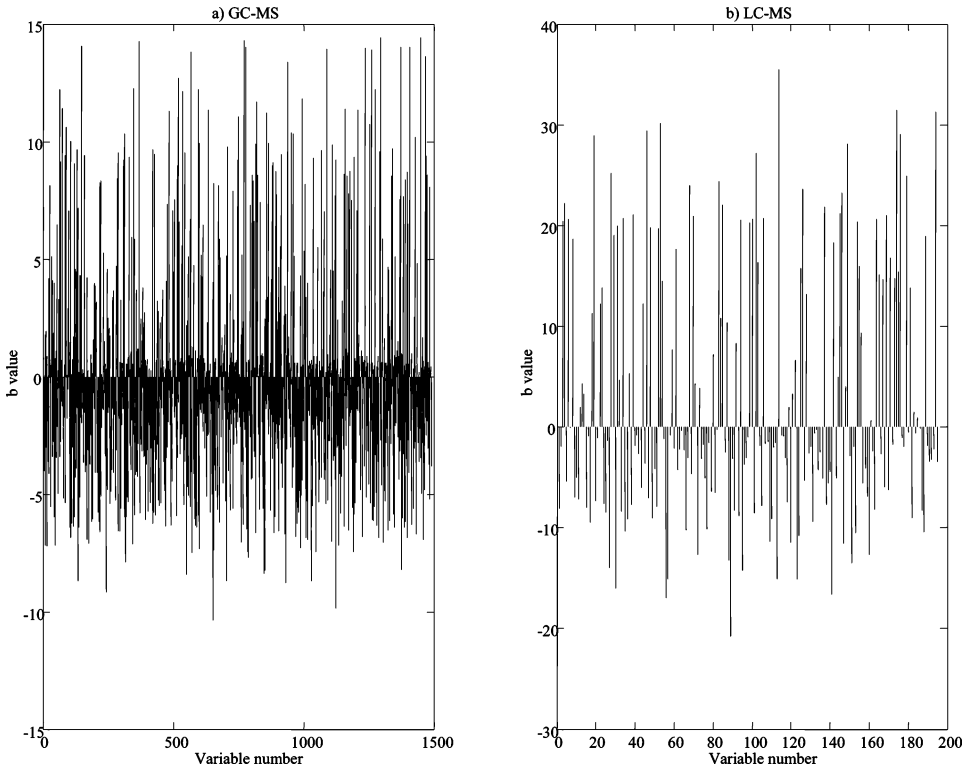

Figure 4. Calculated *b* values for the GC/MS (a) and LC/MS (b).

there are conflicting equations in the literature.[36,37] We found that the VIP statistic (as implemented in ref 37) and the absolute size of the *b* values gave comparable results in terms of the variable importance. Hence, we decided to concentrate on the *b* values.

The *b* values for the GC/MS and LC/MS platforms are shown in Figure 4 for the final model. Note that the sizes of the *b* values can only be compared *within* a block and not *between* blocks. Ranking of the variables for their importance with respect to the yield can now be done using the largest absolute values of the *b* values per block.

CONCLUSIONS

With the availability of mass spectrometry as a standard analysis tool in metabolomics, the question rises how to combine disparate sets of data obtained from different types of spectrometry. A framework is provided for such a fusion where special attention is paid to the preprocessing of the data. Such a preprocessing is crucial for a successful fusion of mass spectrometry-based metabolomics data.

GC/MS and LC/MS partly measure different metabolites as is shown by matrix correlations. Hence, it is worthwhile fusing these data sets in order to obtain a comprehensive view on the metabolome of the microorganism under study and of biological systems in general. If analyzed in a correct way, the fused data correlate highly and robustly with the yield of the fermentation reaction (with an R^2_{CV} of 0.86). This opens the way for finding important metabolites and, finally, aids in identifying targets for strain improvement.

The proposed framework uses well-established methods of multivariate data analysis, which are widely available. Hence, it can be used relatively straightforwardly in all types of metabolomics studies.

APPENDIX

Suppose there are I samples ($i = 1, ..., I$) and J ($j = 1, ..., J$) metabolites. The concentration of metabolite j in sample i is c_{ij} and $x_{LC,ij}$, $x_{GC,ij}$ are the responses of the LC/MS and GC/MS instruments on these concentrations, respectively. Clearly, c_{ij} is equal for the GC/MS and LC/MS. The relationships between the concentrations and responses are given by

$$x_{LC,ij} = a_{LC,j}c_{ij} + b_{LC,j}$$
$$x_{GC,ij} = a_{GC,j}c_{ij} + b_{GC,j} \quad (4)$$

where $a_{LC,j}$ and $b_{LC,j}$ are the response factors for LC/MS (depending on metabolite j), and similarly, $a_{GC,j}$ and $b_{GC,j}$ are the response factors for the GC/MS.

Mean-centering is performed columnwise, hence, across the index i. Since both $b_{LC,j}$ and $b_{GC,j}$ are constant across this index, centering removes these as offsets.[21] Likewise, range scaling removes $a_{LC,j}$ and $a_{GC,j}$ as will be shown below.

Range scaling is done within the j mode. Suppose (without loss of generality) that c_{2j} is the highest concentration and c_{3j} the lowest of metabolite j over all the samples. Then the range for variable j in the LC/MS data is

$$a_{LC,j}c_{2j} - a_{LC,j}c_{3j} = a_{LC,j}(c_{2j} - c_{3j}) \quad (5)$$

and eq 4 after range scaling (dropping the constant $b_{LC,j}$ for convenience) becomes

$$\frac{a_{LC,j}c_{ij}}{a_{LC,j}(c_{2j} - c_{3j})} = \frac{c_{ij}}{(c_{2j} - c_{3j})} \quad (6)$$

and it is clear that the response factor $a_{LC,j}$ drops out of the equation. Similar calculations can be shown for the GC response factors. Centering and range scaling removes the response factors and expresses the instrumental readouts as concentrations relative to their range over the experimental conditions.

ACKNOWLEDGMENT

We thank Roelie Bijl, Karin Rochat, and Karin Overkamp for generating the biological samples; Leon Coulier, Bas Muilwijk, Leo van Stee, Richard Bas, and Thomas Hankemeier for analyzing the metabolomes; and Ivana Bobeldijk-Pastorova for useful suggestions.

Received for review June 20, 2005. Accepted August 2, 2005.

AC051080Y