High-Throughput Data Analysis for Detecting and Identifying Differences between Samples in GC/MS-Based Metabolomic Analyses

ARTICLE *in* ANALYTICAL CHEMISTRY · OCTOBER 2005

Impact Factor: 5.64 · DOI: 10.1021/ac050601e · Source: PubMed

10 AUTHORS, INCLUDING:

Pär Jonsson
Umeå University
30 PUBLICATIONS **1,754** CITATIONS
SEE PROFILE

Annika Johansson
Umeå University
13 PUBLICATIONS **983** CITATIONS
SEE PROFILE

Johan Trygg
Umeå University
125 PUBLICATIONS **7,325** CITATIONS
SEE PROFILE

Henrik Antti
Umeå University
89 PUBLICATIONS **5,895** CITATIONS
SEE PROFILE

High-Throughput Data Analysis for Detecting and Identifying Differences between Samples in GC/MS-Based Metabolomic Analyses

Pär Jonsson,[†] Annika I. Johansson,[‡] Jonas Gullberg,[‡] Johan Trygg,[†] Jiye A,[§] Bjørn Grung,[||] Stefan Marklund,[§] Michael Sjöström,[†] Henrik Antti,[†] and Thomas Moritz*[,‡]

Research Group for Chemometrics, Organic Chemistry, Department of Chemistry, Umeå University, SE-901 87 Umeå, Sweden. Umeå Plant Science Center, Department of Forest Genetics and Plant Physiology, Swedish University of Agricultural Sciences, SE-901 87 Umeå, Sweden, Department of Medical Biosciences, Clinical Chemistry, Umeå University Hospital, Umeå University, SE-901 85, Umeå, Sweden, and Department of Chemistry, University of Bergen, N-5007, Norway

In metabolomics, the objective is to identify differences in metabolite profiles between samples. A widely used tool in metabolomics investigations is gas chromatography−mass spectrometry (GC/MS). More than 400 compounds can be detected in a single analysis, if overlapping GC/MS peaks are deconvoluted. However, the deconvolution process is time-consuming and difficult to automate, and additional processing is needed in order to compare samples. Therefore, there is a need to improve and automate the data processing strategy for data generated in GC/MS-based metabolomics; if not, the processing step will be a major bottleneck for high-throughput analyses. Here we describe a new semiautomated strategy using a hierarchical multivariate curve resolution approach that processes all samples simultaneously. The presented strategy generates (after appropriate treatment, e.g., multivariate analysis) tables of all the detected metabolites that differ in relative concentrations between samples. The processing of 70 samples took similar time to that of the GC/TOFMS analyses of the samples. The strategy has been validated using two different sets of samples: a complex mixture of standard compounds and *Arabidopsis* samples.

In the postgenomics era, the complementary use of life science technologies that enable transcriptomic, proteomic, and metabolomic developments to be analyzed in detail in the same biological systems has revolutionized biological investigations. Instead of relating biological phenomena to a small number of variables, it is now possible to investigate biological systems with a global analytical approach. This is often called systems biology, and the purpose is to study biological materials as integrated systems of genetic, protein, metabolic, cellular, and pathway events that are in constant flux and interdependent.[1−3] Besides solving the difficulties involved in integrating the different types of data sets, it must be possible to perform the individual parts of such investigations rapidly and to generate results that are reliable and easy to interpret. In metabolomics, the ultimate goal is to identify and quantify every metabolite in a biological system.[4−6] Although this is not yet technologically feasible, relevant metabolic profiles of different samples can today be compared and contrasted. For instance, in various toxicological and clinical investigations, NMR spectroscopy together with pattern recognition and chemometric tools has been shown to be very useful[7,8] However, mass spectrometry-based metabolomics has received increasing interest since it combines the sensitivity of MS analysis with the possibility to identify metabolites in the usually complex samples that are analyzed. Gas chromatography−mass spectrometry (GC/MS) has been shown in a large number of publications to be very useful for purposes such as identifying metabolic changes in transgenic plants,[9,10] while LC−ESI-MS has, for example, been used in the classification of yeast mutants by footprinting[11] and for analyzing responses to nutritional stresses in *Arabidopsis*.[12] Capillary electrophoresis coupled to MS has also been shown to be a potentially important metabolomic approach.[13] The overall purpose of metabolomic analysis is to identify metabolites that explain the metabolic differences between samples, e.g., between genotypes or between healthy and nonhealthy tissues. Comparisons of samples normally involve the use of appropriate multivariate

* Corresponding author. E-mail: thomas.moritz@genfys.slu.se.
[†] Department of Chemistry, Umeå University.
[‡] Swedish University of Agricultural Sciences.
[§] Department of Medical Biosciences, Umeå University.
[||] University of Bergen.

(1) Kitano, H. *Science* **2002**, *295*, 1662−1664.
(2) Nicholson, J. K.; Holmes, E.; Lindon, J. C.; Wilson, I. D. *Nat. Biotechnol.* **2004**, *22*, 1268−1274.
(3) van der Greef, J.; Stroobant, P.; van der Heijden, R. *Curr. Opin. Chem. Biol.* **2004**, *8*, 559−565.
(4) Weckwerth, W. *Annu. Rev. Plant Biol.* **2003**, *54*, 669−689.
(5) Goodacre, R.; Vaidyanathan, S.; Dunn, W. B.; Harrigan, G. G.; Kell, D. B. *Trends Biotechnol.* **2004**, *22*, 245−252.
(6) Fiehn, O. *Plant Mol. Biol.* **2002**, *48*, 155−171.
(7) Brindle, J. T.; Antti, H.; Holmes, E.; Tranter, G.; Nicholson, J. K.; Bethell, H. W. L.; Clarke, S.; Schofield, P. M.; McKilligin, E.; Mosedale, D. E.; Grainger, D. J. *Nat. Med.* **2002**, *8*, 1439−1444.
(8) Nicholson, J. K.; Wilson, I. D. *Nat. Rev. Drug Discovery* **2003**, *2*, 668−676.
(9) Roessner-Tunali, U.; Hegemann, B.; Lytovchenko, A.; Carrari, F.; Bruedigam, C.; Granot, D.; Fernie, A. R. *Plant Physiol.* **2003**, *133*, 84−99.
(10) Weckwerth, W.; Loureiro, M. E.; Wenzel, K.; Fiehn, O. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 7809−7814.
(11) Allen, J.; Davey, H. M.; Broadhurst, D.; Heald, J. K.; Rowland, J. J.; Oliver, S. G.; Kell, D. B. *Nat. Biotechnol.* **2003**, *21*, 692−696.
(12) Hirai, M. Y.; Yano, M.; Goodenowe, D. B.; Kanaya, S.; Kimura, T.; Awazuhara, M.; Arita, M.; Fujiwara, T.; Saito, K. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 10205−10210.

10.1021/ac050601e CCC: $30.25
Published on Web 00/00/0000

statistical tools such as principal component analysis (PCA), hierarchical cluster analysis,[14] discriminant analysis,[11] partial least-squares projections to latent structures (PLS) analysis,[15] independent component analysis,[16] or correlative network analysis.[17] Although many of the analytical techniques used in metabolomics analysis, e.g., GC/MS, can be high-throughput techniques, there are a number of constraints regarding processing of the data files and subsequently in the detection of differences between samples. One of the main problems associated with the data analysis is that, prior to the multivariate analysis, relative levels (responses) of the components (metabolites/MS peaks) must be calculated for all samples. If not, direct injection MS methodology is used (see Allen et al.[11]), LC/MS or GC/MS analysis involves peak detection algorithms[18,19] or preferentially mathematical curve resolution procedures,[20−24] which is often named deconvolution. Deconvolution of peaks obtained in analyses of complex samples is extremely difficult to automate completely. This is mainly due to problems associated with separating completely overlapping components, peak alignment and peak matching between samples, and other issues such as determining the number of components in the chromatograms that need to be resolved and detecting minor components that are close to their detection limits. Recently, a method for rapid comparison of multiple GC/MS samples was presented that can identify parts of the GC/MS chromatogram that differ between samples.[15] Hence, this technique can minimize the amount of time spent on curve resolution and allow more time to be spent on identifying metabolites explaining the differences between the samples. Although this method is fast compared to the traditional methodologies used for comparing samples, there is an urgent need to improve and automate all the steps involved in analyzing the data generated in metabolomic GC/MS or LC/MS analyses (see, for instance, van der Greef et al.[3]). Otherwise, the data processing will continue to be a bottleneck in metabolomic analysis.

In this study, we describe a semiautomated strategy for analyzing GC/MS metabolomic data where the outcome from the data analysis of nonprocessed GC/MS files is the mass spectra explaining the differences between groups of samples. The mass spectra can then be subjected to an automated library search, and thereby, an identification of metabolites that differ between sample groups can be obtained on-line from the GC/MS analysis. The strategy is exemplified and validated with two sets of samples: a complex mixture of standard compounds and *Arabidopsis* samples.

EXPERIMENTAL SECTION

Chemicals. Most of the standard compounds and the oximation reagent methoxyamine hydrochloride were obtained from Sigma (St. Louis, MO). *N*-Methyl-*N*-trimethylsilyltrifluoroacetamide plus 1% TMCS was obtained from Pierce (Rockford, IL), chloroform (Analytical Reagent) from Riedel-de Haën (Seelze, Germany), methanol from J.T. Baker (Deventer, The Netherlands), heptane (high-purity solvent) from Burdick & Jackson (Muskegon, MI), and pyridine from Regis (Morton Grove, IL). The 11 stable isotope reference compounds, ([$^2H_4$]-succinic acid, [$^{13}C_5$,^{15}N]-glutamic acid, [$^2H_7$]-cholesterol, [1,2,3-$^{13}C_3$]-myristic acid, [$^{13}C_5$]-proline, and [$^{13}C_4$]-disodium α-ketoglutarate were purchased from Cambridge Isotope Laboratories (Andover, MA), [$^{13}C_6$]-glucose, [$^{13}C_{12}$]-sucrose,[$^{13}C_4$]-hexadecanoic acid, and [$^2H_4$]-1,4-butane-diamine·2HCl, were from Campro (Veenendaal, The Netherlands), and 2-hydroxy-[$^2H_6$]-benzoic acid was from Icon (Summit, NJ).

Sampling, Extraction, and Derivatization of *Arabidopsis* Samples. Wild-type *Arabidopsis thaliana* (Ler), ga1-3, gai-t6 ga1-3, rga-24 ga1-3, and rga-24 gai-t6 ga1-3 mutant lines were grown in soil at 22 °C under long day conditions (16-h photoperiod), with a photon density of 130 μE m^{-2} s^{-1}. After three weeks, the rosette leaves were harvested, frozen immediately in liquid nitrogen, and stored at −80 °C until analysis. Each sample was extracted and derivatized according to the method described by Gullberg et al.[25]

GC/MS Analysis. One microliter of each derivatized sample was injected splitless by an Agilent 7683 autosampler (Agilent, Atlanta, GA) into an Agilent 6890 gas chromatograph equipped with a 10 m × 0.18 mm i.d. fused-silica capillary column with a chemically bonded 0.18-μm DB 5-MS stationary phase (J&W Scientific, Folsom, CA). The injector temperature was 270 °C, the septum purge flow rate was 20 mL min^{-1}, and the purge was turned on after 60 s. The gas flow rate through the column was 1 mL min^{-1}; the column temperature was held at 70 °C for 2 min, then increased by 40 °C min^{-1} to 320 °C, and held there for 2 min. The column effluent was introduced into the ion source of a Pegasus III time-of-flight mass spectrometer, GC/TOFMS (Leco Corp., St. Joseph, MI). The transfer line and the ion source temperatures were 250 and 200 °C, respectively. Ions were generated by a 70-eV electron beam at an ionization current of 2.0 mA, and 30 spectra s^{-1} were recorded in the mass range 50−800 m/z with unit resolution. The acceleration voltage was turned on after a solvent delay of 170 s. The detector voltage was 1660 V.

Analysis of GC/MS Data. Nonprocessed MS files from GC/TOFMS analysis were exported in CSV or NetCDF format to MATLAB software 6.5 (Mathworks, Natick, MA), where all data-pretreatment procedures, such as baseline correction and chromatogram alignment, time-window setting, and MCR were performed using custom scripts. Multivariate analysis was performed with SIMCA-P+ 10.5 software (Umetrics AB, Umeå, Sweden). All calculations were performed with a normal PC (Pentium IV, 2.80-GHz CPU, 1.0-MB RAM).

(13) Soga, T.; Ohashi, Y.; Ueno, Y.; Naraoka, H.; Tomita, M.; Nishioka, T. *J. Proteome Res.* **2003**, *2*, 488−494.
(14) Sumner, L. W.; Mendes, P.; Dixon, R. A. *Phytochemistry* **2003**, *62*, 817−836.
(15) Jonsson, P.; Gullberg, J.; Nordstrom, A.; Kusano, M.; Kowalczyk, M.; Sjostrom, M.; Moritz, T. *Anal. Chem.* **2004**, *76*, 1738−1745.
(16) Scholz, M.; Gatzek, S.; Sterling, A.; Fiehn, O.; Selbig, J. *Bioinformatics* **2004**, *20*, 2447−2454.
(17) Steuer, R.; Kurths, J.; Fiehn, O.; Weckwerth, W. *Biochem. Soc. Trans.* **2003**, *31*, 1476−1478.
(18) Andreev, V. P.; Rejtar, T.; Chen, H. S.; Moskovets, E. V.; Ivanov, A. R.; Karger, B. L. *Anal. Chem.* **2003**, *75*, 6314−6326.
(19) Windig, W.; Phalp, J. M.; Payne, A. W. *Anal. Chem.* **1996**, *68*, 3602−3606.
(20) Halket, J. M.; Przyborowska, A.; Stein, S. E.; Mallard, W. G.; Down, S.; Chalmers, R. A. *Rapid Commun. Mass Spectrom.* **1999**, *13*, 279−284.
(21) Shao, X. G.; Wang, G. Q.; Wang, S. F.; Su, Q. D. *Anal. Chem.* **2004**, *76*, 5143−5148.
(22) Sinha, A. E.; Fraga, C. G.; Prazen, B. J.; Synovec, R. E. *J, Chromatogr., A* **2004**, *1027*, 269−277.
(23) Idborg-Bjorkman, H.; Edlund, P. O.; Kvalheim, O. M.; Schuppe-Koistinen, I.; Jacobsson, S. P. *Anal. Chem.* **2003**, *75*, 4784−4792.
(24) Manne, R.; Grande, B. V. *Chemom. Intell. Lab. Syst.* **2000**, *50*, 35−46.
(25) Gullberg, J.; Jonsson, P.; Nordstrom, A.; Sjostrom, M.; Moritz, T. *Anal. Biochem.* **2004**, *331*, 283−295.

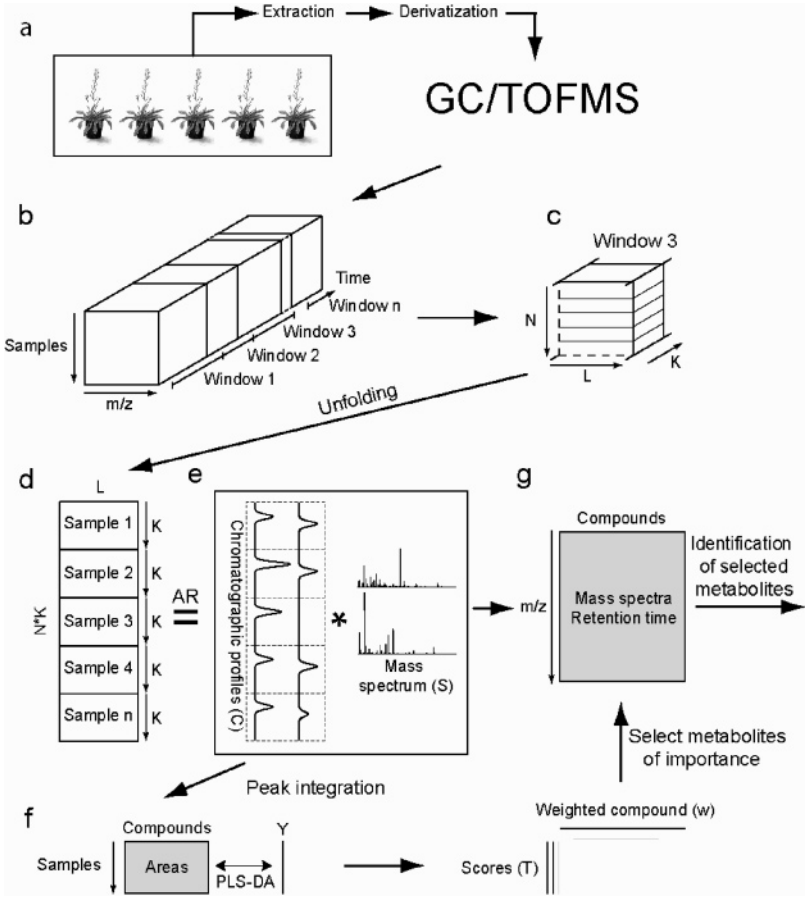

Figure 1. Summary of the metabolomic analysis concept used in the present investigation. (a) Extraction and GC/MS analysis. (b−e) The hierarchical MCR method for automatically resolving GC/TOFMS data. (f) Multivariate analysis, e.g., PCA and PLS-DA. (g) Export of the mass spectra that explain the differences between samples, according to multivariate analysis, for mass spectra library searches. See also Supporting Information for MCR method.

ChromaTOF (2.12) software (Leco Corp., St. Joseph, MI) was used for comparisons with the MCR method. Automatic peak detection and mass spectrum deconvolution with the ChromaTOF software were performed using a peak width set to 1.5 s. To obtain accurate peak areas for the deconvoluted components, unique quantification masses for each component were specified and the samples were reprocessed. The obtained peak areas for quantification masses were used for comparison with the concentration values obtained with the MCR method and for calculation of peak areas for internal standards.

RESULTS

Methodology and Hierarchical Multivariate Curve Resolution. The analyses were performed using a GC/TOFMS instrument (Figure 1a). The main advantage of GC/TOFMS systems is that they enable spectra to be accumulated rapidly, thereby increasing the speed of GC/MS analyses,[26] making them ideal for the analysis of complex mixtures, such as metabolomic samples.[10] In this case, short analytical cycles of ~15 min/sample were used, resulting in high throughputs with close to 90 analyses/24 h. In addition, the mass spectra are homogeneous over the peak profile, facilitating deconvolution and data interpretation.

Multivariate curve resolution[27] (MCR) was used to automatically extract information from all analyzed samples simultaneously, providing mass spectra, retention times, and relative amounts of all compounds in each sample. The method, including the whole metabolomics strategy, is summarized in Figure 1. The presented approach involves smoothing, background reduction, alignment, time-window setting, MCR, multivariate modeling, and export of mass spectra to library search software. All steps, including the import of raw data into the processing software are computer automated except for the time-window setting, which is done manually. The MCR method has been used earlier to resolve pure profiles in an industrial process,[28] where multiple runs were analyzed simultaneously, and to resolve multiple samples of second-order data.[27] In these examples, the data were globally resolved using the MCR method. This was possible due to the fact that the total rank was quite low (often below 10). In complex GC/MS data with several hundreds of components, as used in the present investigation, it is not possible to use a global resolving approach due to the complexity in the data. Instead, a hierarchical approach is applied, where each time window is resolved separately.

In the present investigation, we calculated the peak areas for internal standard using the ChromaTOF software. [$^{13}C_5$]-Proline-

(26) Veriotti, T.; Sacks, R. *Anal. Chem.* **2001**, *73*, 4395−4402.

(27) Tauler, R. *Chemom. Intell. Lab. Syst.* **1995**, *30*, 133−146.

(28) Tauler, R.; Kowalski, B.; Fleming, S. *Anal. Chem.* **1993**, *65*, 2040−2047.

TMS was used for correction of recovery and was imported into Matlab for correction of differences in recovery and sample weights, respectively. An alternative to correct for recovery differences prior the MCR is to make the correction after the MCR. The smoothing, background reduction, chromatogram alignment, and time-window setting were done as described by Jonsson et al.[15] To reduce the noise, all m/z channels were filtered using moving averages (each of seven time points). Background reduction was performed by subtracting the minimum value of each m/z channel from all other values in the same m/z channel. These operations make the start and end points of peaks or peak clusters easier to detect, which is important when dividing the chromatograms into time windows. Smoothing and background reduction were done for each sample individually. The next step in the preprocessing was to align the chromatograms, which is also important for the time-window setting. Several methods have been developed for correcting retention drifts (e.g., Duran et al.[29,30]). We have used the method published by Malmquist and Danielsson[31] due to its simplicity. This method aligns chromatograms by finding the maximal covariance between the chromatograms. After alignment of chromatograms, the data were divided into time windows (Figure 1b). The edges between the windows were set at globally low intensity points; i.e., time points where the intensity was low for all samples, and consequently, the size of the windows was not uniform, varying between 1.53 and 13.07 s in the investigations described below. The time-window setting step was done manually by visually inspecting plots of all of the total ion currents or base peak chromatograms superimposed on each other. This procedure gives all samples the same edges. These globally low intensity points were also used to remove interfering systematic background, by linear interpolation between the globally low intensity points. The fitted line was then subtracted from the data for all samples individually. For a more thorough description of each step, see Jonsson et al.[15]

The next step in the data processing was to resolve the data acquired in each time window into mass spectral and chromatographic information for each compound (peak). The data in each time window can be seen as a data cube of size $N \times K \times L$, where N is the number of samples, K is the number of time points (scans), and L is the number of m/z channels (Figure 1c). The cubes were then unfolded to form a data matrix $\mathbf{X}$ of size $(N \times K) \times L$ (Figure 1d). This data matrix was then resolved (deconvoluted) using alternating regression (AR). AR is an iterative method that alternates between two operations until convergence.[32]

$$\mathbf{C} = \mathbf{X} \times \mathbf{S} \times (\mathbf{S}^{\mathrm{T}} \times \mathbf{S})^{-1}$$

$$\mathbf{S} = \mathbf{X}^{\mathrm{T}} \times \mathbf{C} \times (\mathbf{C}^{\mathrm{T}} \times \mathbf{C})^{-1}$$

where $\mathbf{C}$ is the chromatographic profile and $\mathbf{S}$ is the mass spectral profile. The superscripts T and -1 denote matrix transposition and matrix inversion, respectively.

We have chosen to use the algorithm PURE by R. Tauler and A. de Juan (available at http://www.ub.es/gesq/mcr/ndownload.htm) for detecting the purest mass channels in the data and used the ion chromatograms of these mass channels as starting estimates of $\mathbf{C}$. PURE uses the SIMPLISMA algorithm by Windig and Guilement[33] and results in more rapid convergence than using only random numbers as start estimates. $\mathbf{C}$ will be of size $(N \times K)$ times the number of resolved components ($\mathbf{R}$; i.e., the number of components with a common mass spectrum) and $\mathbf{S}$ of size $L \times R$. Negative values in $\mathbf{C}$ or $\mathbf{S}$ are not allowed, and such values will be replaced by zeros. Each sample will yield a chromatographic profile for each resolved component (Figure 1e). The chromatographic profiles for each resolved component were constrained to be unimodal for each sample, and in addition, a resolved component had to elute approximately at the same retention time in every sample. If not, the resolved component was assumed not to exist in the deviating sample. A retention time criterion of ± 0.5 s (from the median retention time (for that component)) was set for accepting the presence of a compound after alignment. These constraints were applied in each iteration step of the AR procedure. The number of components in each time window was found by searching for the solution with the highest rank that fulfills the requirement that components have to elute in the same order in all samples. The search starts from one component ($R = 1$) and then the number of components is continuously increased by one until three subsequent solutions are rejected as not valid (the last valid solution is used). There are no limits of how many resolved components can be found in each time window, but generally there were between 3 and 10 components in each time window in the present investigations. In summary, to be able to resolve components, they had to fulfill either the criterion of having different chromatographic profiles or the component ratio between the samples had to differ in the mass spectral dimension (see also the standard mixtures test case and discussion). For detailed information about the MCR method, see Supporting Information (MCR method).

The integrated peak areas of the chromatographic profiles (all masses in the corresponding mass spectral profile contributes to the area) from all samples in all time windows were thereafter combined to form a data matrix (Figure 1f), which could be subjected to multivariate statistical analysis or other types of statistical analysis. PCA and partial least-squares discriminant analysis (PLS-DA) were used as data analysis tools in the present investigation. PCA describes the largest variation in data using a few orthogonal latent variables. Thus, an overview of the data is obtained, whereby trends, groupings, and outliers can be detected. PCA[34] is an unsupervised method that does not require any prior information about the samples analyzed. PLS-DA[35] is a supervised classification method that can be used if samples are related to different classes or groups, based on prior knowledge. By using these types of supervised classification methods, e.g., PLS-DA, it is possible to efficiently detect the variables (i.e., metabolites) explaining the differences between samples, or rather groups of samples, by interpreting the variable weights. As the mass

(29) Duran, A. L.; Yang, J.; Wang, L. J.; Sumner, L. W. *Bioinformatics* **2003**, *19*, 2283−2293.
(30) Fraga, C. G.; Prazen, B. J.; Synovec, R. E. *Anal. Chem.* **2001**, *73*, 5833−5840.
(31) Malmquist, G.; Danielsson, R. *J, Chromatogr., A* **1994**, *687*, 71−88.
(32) Karjalainen, E. J. *Chemom. Intell. Lab. Syst.* **1989**, *7*, 31−38.
(33) Windig, W.; Guilment, J. *Anal. Chem.* **1991**, *63*, 1425−1432.
(34) Wold, S.; Esbensen, K.; Geladi, P. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37−52.
(35) Wold, S.; Sjöström, M.; Eriksson, L. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109−130.

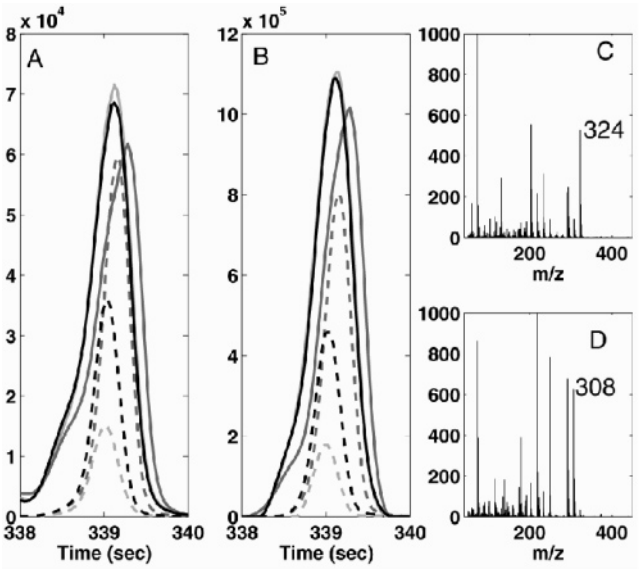

Figure 2. Analysis of three samples containing the standard mixture described in Table 1 and the Supporting Information Table S1. The figure shows how mass spectra from overlapping peaks can be resolved with the MCR method. (a) Ion chromatograms for *m*/*z* 308 (unique for *p*-coumaric acid) and 324 (unique for coniferyl alcohol). Chromatograms are not corrected for retention time drifts between different analyses. (b) Resolved chromatographic profiles obtained using the MCR method. (a) and (b) Solid lines correspond to *p*-coumaric acid and dashed lines to coniferyl alcohol. Lines with the same gray scale correspond to the same sample. (c) and (d) Resolved mass spectra for *p*-coumaric acid and coniferyl alcohol, respectively. Mass spectra identities were confirmed by mass spectra library searches.

Table 1. Results of the Analysis of Mixtures Containing 101 Compounds,[a] 18 of Which Were Varied in Concentration, as Listed in the Table

name	levels (ng μL^{-1})	r^{2b}	match[d]	reverse match[d]
proline (*N*,*O*-TMS)	1/15/30	0.996	940	943
4-hydroxybenzoic acid (2TMS)	1/15/30	0.994	928	937
gluconic acid lactone (4TMS)	1/15/30	0.969	782	815
inositol (6TMS)	1/15/30	0.995	930	933
gluconic acid (6TMS)	1/15/30	0.982	956	957
norleucine (*N*,*O*-TMS)	1/3/5	0.962	905	907
xylose (MeOX 4TMS)	1/3/5	0.903	924	925
aspartic acid (3TMS)	1/3/5	0.996	834	867
citric acid (4TMS)	1/3/5	0.995	919	946
campesterol (TMS)	1/3/5	0.965	820	866
serine (3TMS)	5/15/30	0.993[c]	904	929
asparagine (*N*,*N*,*O*-TMS)	5/15/30	0.997	765	847
coniferyl alcohol (2TMS)	5/15/30	0.951	879	907
trehalose (8TMS)	5/15/30	0.995	917	958
erythritol (4TMS)	0/3/5	0.993	903	905
glutamine (3TMS)	0/3/5	0.996	720	835
tyrosine (3TMS)	0/3/5	0.997	700	867
cholesterol (TMS)	0/3/5	0.981	916	928

[a] See Supporting Information Table S2. [b] r^2 is the correlation between the true amount injected and the detected amount of the compounds ($n = 62$). [c] Due to retention shift drifts for two samples, the amount of serine was incorrectly estimated as zero ($r^2 = 0.691$). Correction increases the r^2 value to 0.993. [d] NIST settings: m/z 50−1000, minimum abundance threshold 5%, identity and normal search mode "match" and "reverse-match" are similarity indices calculated by NIST MS Search 2.0 between the resolved mass spectrum and the user library mass spectrum.

spectrum and the corresponding retention time of each component (Figure 1e; see Supporting Information Table S1 for format) are listed and exported to NIST MS Search 2.0 software (or other mass spectra library search software), the compounds that differ between samples are rapidly identified by means of library comparisons (Figure 1g).

Standard Mixture Test Case. To validate the processing strategy (Figure 1), 19 samples consisting of a standard mixture, were analyzed by GC/TOFMS after methoxymation and trimethylsilylation. The mixture consisted of 101 compounds, including 11 internal standards. The compounds and their respective retention indices, based on the *n*-alkane (C_{12}−C_{40}) series, are listed in the Supporting Information (Table S2). The compounds represent a wide range of different compound classes occurring as endogenous metabolites in biological tissues, and the retention times are well spread along the 11-min analysis time axis. Among the 101 compounds, 84 compounds were kept at constant levels (15 ng μL^{-1}) in the different samples, and for the other 18 compounds, the concentrations were varied at 0, 1, 3, 5, 15, and 30 ng μL^{-1}. Each sample was analyzed three times. The purpose of this experiment was to compare the mass spectra and evaluate the scope, to identify mass spectra of overlapping peaks (deconvolution), and to determine the variation in concentration of compounds. Figure 2 shows the ion chromatograms for m/z 308 and 324 (Figure 2a) between retention time 338−340 s, resolved chromatographic profiles (Figure 2b), and mass spectra obtained after processing the 17 different mixtures analyzed by GC/TOFMS (Figure 2c,d). Two mass spectra with good library matches (Table 1) were obtained, corresponding to *p*-coumaric acid (Figure 2c) and coniferyl alcohol (Figure 2d). These two compounds have almost identical retention times in the GC/MS system used in the present investigation. Using common deconvolution techniques, e.g., AMDIS[20] or Leco Chromatof software, it is not possible to obtain resolved mass spectra for the two compounds (see also Supporting Information Figure S1). However, with the MCR processing strategy described here, it is possible to obtain well-resolved mass spectra for both of these compounds, since the ratio between coniferyl alcohol and *p*-coumaric acid differs between samples. Coniferyl alcohol was one of the compounds that varied in the standard mixtures. The MCR method[27] processes all samples simultaneously, unlike all other published methods, e.g., AMDIS,[20] MS-resolver,[23] or Leco Chromatof software. If only one sample at a time is resolved (deconvoluted), completely overlapping peaks will be regarded as representing a single compound, and consequently, the corresponding mass spectrum will represent a mixture of the overlapping compounds. In addition, if the ratio between the overlapping peaks differs too much, they will appear to represent different compounds in the different samples. However, with the hierarchical MCR method, where all samples are processed simultaneously, completely overlapping peaks can be resolved if there are differences in concentration between the samples and the mass spectra are not identical (see Supporting Information, Figure S2).

The quality of the hierarchical MCR method for analyzing the standard mixtures was validated by comparing the resolved mass spectra with standard mass spectra and by calculating the correlation between estimated amounts of the metabolites and the

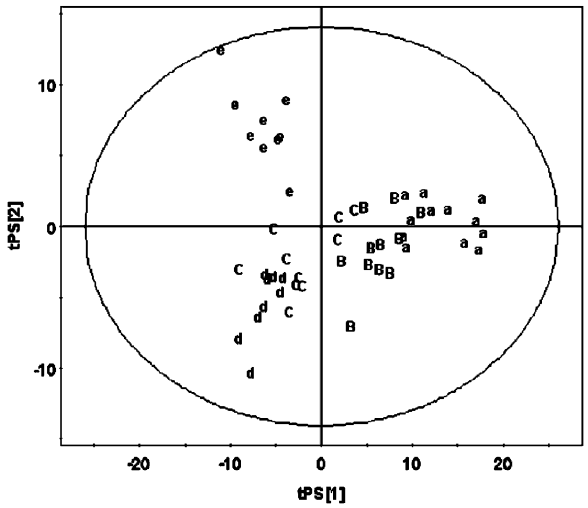

Figure 3. PLS-DA score plot from the analysis of rosette leaves from different *Arabidopsis* genotypes. Key: a = ga1-3; B = gai-t6 ga1-3; C = rga-24 gai-t6 ga1-3; d = rga-24 ga1-3; e = WT.

injected amounts (Table 1). The results show good agreement with the expected values, with r^2 values usually above 0.99, and all compounds that varied in concentration between samples were detected using the described method. The only compound with a lower degree of correlation between the expected and observed values was serine. This was because the retention time drift was greater than the preset maximum allowed during the analysis of two of the samples, and consequently, serine was not detected as a "true" component in these samples. When this was corrected, the correlation was of the same magnitude as for the other compounds studied. The library searches also generally generated high matches, with values above 800. Values between 700 and 800 were mainly attributed to differences in intensity of m/z 73 between the resolved mass spectrum and the user library mass spectrum (data not shown). By omitting comparisons with m/z values below 85, a higher degree of matching can be obtained. Furthermore, including retention index comparisons also increases the strength of the identifications.

***Arabidopsis* Gibberellin-Signaling Mutants Test Case.** To further test the hierarchical MCR method (Figure 1), the metabolite content in rosette leaves from five *A. thaliana* genotypes, four with mutations affected in the plant hormone gibberellin signaling, biosynthesis, or both, were analyzed with GC/TOFMS following extraction and derivatization.[25] Using the hierarchical MCR method, 497 resolved components in 66 time windows were obtained. The data were filtered using orthogonal signal correction[36] (OSC), centered and pareto-scaled prior to PLS-DA classification of three of the genotypes, ga1-3, rga-24 ga1-3, and WT. The obtained PLS-DA score plot for the first two components (latent variables) showed a clear separation of the genotypes (Figure 3; symbols a, d, and e). This is not surprising since, for instance, ga1-3 (a) is a severe dwarf plant that lacks active gibberellins and hence clearly different from the WT plants (e). To validate the model results, predictions were made for the two genotypes gai-t6 ga1-3 and gai-t6 rga-24 ga1-3, using the calculated OSC filter and PLS-DA model based on the other sample-set. The results, shown in the obtained PLS-DA score plot (Figure 3), predicted that the gai-t6 ga1-3 genotype (B) is closer to ga1-3 (a) and less like WT (e) than the gai-t6 rga-24 ga1-3 genotype (C). This is consistent with the facts that gai-t6 ga1-3 is a severe dwarf, similar to the ga1-3 genotype, while gai-t6 rga-24 ga1-3 is tall and WT-like.[37]

Comparison between the dwarf ga1-3 and the semidwarf rga-24 ga1-3 mutants revealed that levels of ~50 metabolites differ in relative concentration between these genotypes. The identification of differences was performed by interpretation of the loadings (as described[38] from the PLS-DA model including only ga1-3 and rga-24 ga1-3, together with the 99% confidence intervals calculated using jackknifing.[39] For example, glycerol 3-phosphate, glutamate, glycerol, and sterols such as campesterol and stigmasterol were found in higher concentrations in the ga1-3 mutant. In contrast, the semidwarfed genotype contained higher amounts of nicotinic acid, threonic acid-1,4-lactone, and erythronic acid (see Supporting Information Figure S3 for resolved mass spectra). As it is beyond the scope of the present work to discuss the biology related to the presented results, no other comparisons between genotypes were performed.

The automatic quantification of metabolites using the present method was also compared with "normal" quantification of metabolites using the software provided with the GC/MS instrument (Figure 4). Quantification with the Leco Chromatof software was done by integrating specific target masses for each compound and normalizing the data against an internal standard. The data showed a high correlation between the two different methods (r^2 > 0.8 for all performed comparisons), suggesting that the automatic integration method is accurate and useful for rapid evaluation of GC/TOFMS data (see Supporting Information, Figure S4 and S5 for additional examples from analysis of human plasma). However, there were two types of cases where the two methods did not show a high degree of correlation. The first was in areas of the chromatograms with many overlapping peaks. In such cases, the poor correlation arose from the difficulties involved in finding unique target masses for quantification with the Chromatof software, which in turn made finding masses that could be integrated with high accuracy problematic. The second type of case was when very low abundance metabolites were to be quantified. In such cases, the poor correlation was usually due to difficulties in finding quantification masses for the Chromatof software with high enough signal-to-noise (S/N) ratios. Too low S/N ratios will result in low-precision integrations (data not shown). However, it must also be emphasized that low-abundant peaks can also be a problem to quantify with the MCR method. This is especially pronounced when the corresponding deconvoluted mass spectrum contains large portions of noise.

DISCUSSION

The presented strategy, which facilitates the rapid identification of metabolites in metabolomic analyses, is a semiautomated method providing both qualitative and quantitative information about GC/MS peaks. The output data should preferentially be analyzed using multivariate statistical tools such as PCA or PLS and be based on the model loadings. The peaks (compounds)

(36) Wold, S.; Antti, H.; Lindgren, F.; Ohman, J. *Chemom. Intell. Lab. Syst.* **1998**, *44*, 175−185.

(37) King, K. E.; Moritz, T.; Harberd, N. P. *Genetics* **2001**, *159*, 767−776.

(38) Trygg, J.; Wold, S. *J. Chemom.* **2002**, *16*, 119−128.

(39) Martens, H.; Martens, M. *Food Qual. Preference* **2000**, *11*, 5−16.

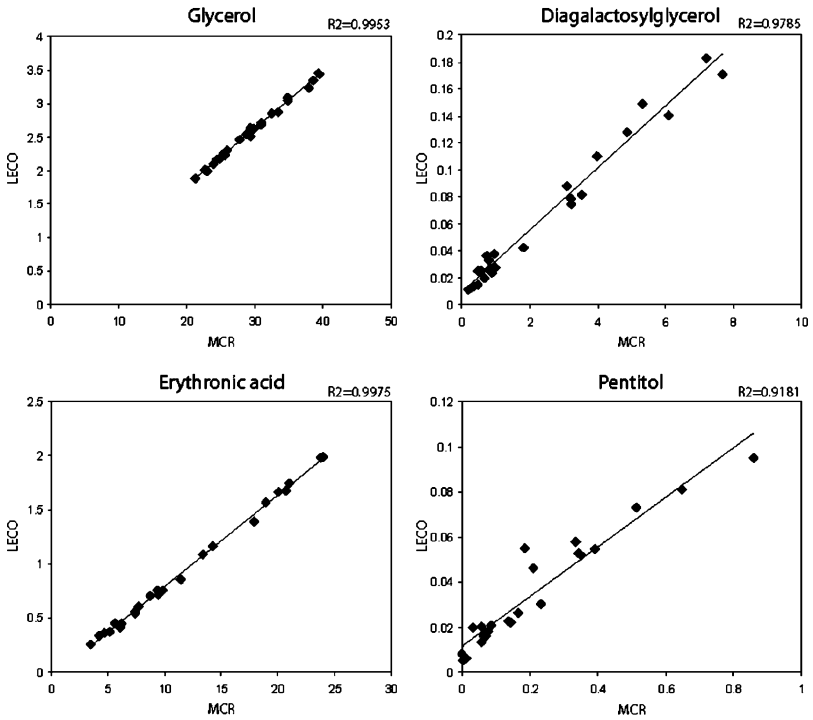

Figure 4. Regression curves for comparing the quantification of metabolites from the *Arabidopsis* (rga-24 ga1-3 mutant line) by the hierarchical MCR method and the traditional method used by the Leco Chromatof software, integration of specific quantification masses.

that differ between sample groups can be identified by exporting the obtained mass spectra to mass spectra library search software, like NIST MS SEARCH 2.0. The method provides fast identification of compounds by comparing mass spectra and retention times (or retention indices) with those in spectra library (user or commercially available). The main advantage of the described method is that it is unbiased, requiring no target sample. In traditional GC/MS peak finding, a master sample is first processed, after which all other samples are processed and peak-matched to fit the master sample (see, for instance, Jonsson et al.[15]). This can result in some peaks being neglected if the master sample is not ideal. The presented method does not use a master sample; instead, multivariate curve resolution[27] is used, which is a multiprocessing approach whereby all samples of interest are analyzed and deconvoluted simultaneously. This allows mass spectra from two completely overlapping peaks to be resolved as long as the intensity of the two peaks varies between samples. The usage of this completely unique feature of the MCR method in complex GC/MS metabolomics data extends the capacity to identify compounds and differences between samples and sample groups. The analysis of all samples simultaneously also eliminates the problems with peak matching, which is a difficult task to automate when deconvoluting one sample at a time. A number of peak-matching procedures have been developed,[40] but problems generally arise using these methods when retention times differ between samples in the analysis of complex GC/MS data sets (such as those generated in metabolomic investigations). For comparison of samples using statistical tools such as PCA and PLS, it is important to ensure that the values in each column of the **X** matrix correspond to the same compound. The hierarchical MCR method automatically generates such an **X** matrix, including values for overlapping peaks. However, it should be pointed out that as with all deconvolution methods sometimes one mass spectrum can be found in two or three adjacent variables. This is mainly due to peak noise or in some cases split peaks during time-window settings. However, in such cases, the peak areas for identical variables can be summarized after the processing.

One of the bottlenecks in the "omics" era is the time-consuming processing and evaluation of the enormous amounts of data generated, for instance, in metabolomic analyses.[3,41] Usually an important issue is to identify the differences between samples, e.g., samples from diseased versus nondiseased or mutant versus wild-type tissues. Therefore, it is essential to develop fast data processing methods that do not slow the whole process from sampling, through extraction, to MS analysis. With the presented method, the whole processing of the *Arabidopsis* test case samples, from exporting the GC/MS data to obtaining the **X** matrix with peak areas for all compounds detected and the mass spectra in text format, using an ordinary PC, took less time than the GC/MS analyses (70 samples in ~14 h). The present method is essentially automatic, the only manual processing stage being the time-window setting. This is, however, still a very fast process, requiring just 10−15 min of manual work, which probably can be automated using algorithms finding low-intensity points. The export of mass spectra to the library search software can also be automated, and although many users will base their final decision of mass spectral identity on visual inspection, together with MS identities and similarities calculated by the library software, these decisions can also be automated, by setting acceptance criteria for similarities and identity. However, since large proportions of the metabolites detected in analyses of complex samples are often

(40) Eide, I.; Neverdal, G.; Thorvaldsen, B.; Shen, H. L.; Grung, B.; Kvalheim, O. *Environ. Sci. Technol.* **2001**, *35*, 2314−2318.

(41) Kell, D. B. *Curr. Opin. Microbiol.* **2004**, *7*, 296−307.

unknown,[41] for instance ~80% of the metabolites in the GC/TOFMS analysis of *Arabidopsis* samples presented here (data not shown), manual operation is very important for classification of mass spectra into compound classes. A system that automatically generates structural information from mass spectra would, of course, be helpful, but little work has been done as yet in this important area. However, nonsupervised mass spectral and retention time index databases can be helpful for the classification and identification of mass spectra.[42]

Although quality control of the data should ideally refer back to the original data, both test cases presented here show that both the qualitative and quantitative data obtained by the hierarchical MCR method are of similar, or better, quality than those obtained by traditional data processing methods. The improvements are mainly due to the opportunity the method provides to identify and resolve overlapping mass spectra and the increased accuracy in integrating low-abundance peaks. Alternative methods for chromatogram alignment and automated time-window settings might further improve the strategy. Other algorithms for deconvolution might also improve the strategy. Applying an automated retention index calculation system instead of relying solely on the retention time provides an even more efficient means for rapidly identifying differences between samples. Furthermore, in contrast to the previously described method for rapidly comparing metabolomic samples,[15] the hierarchical MCR method generates information that can be analyzed using any type of statistical tool or test.

CONCLUSION

This work presents an efficient strategy for high-throughput analysis of metabolomic data generated by GC/MS. In recent years, GC/TOFMS analysis has become one of the main technological platforms for metabolomic analysis due to its very rapid spectral accumulation capacities, providing scope for fast analysis in combination with data deconvolution. To date, metabolomic data processing and evaluation using GC/MS has been a time-related problem. The present strategy for data processing and identification of differences between samples eliminates an analytical bottleneck and, thus, should contribute to the development of high-throughput metabolomic techniques that can play important roles in the functional genomics era, fully complementing the other "omic" technologies. For instance, it is likely to be an important tool for the detection and identification of biomarkers in relation to diseases or developmental processes in biological systems. In combination with GC/MS, it is also an excellent tool for fast predictions of differences in specific biochemical pathways between genotypes, which can then be investigated in detail using other methods.

ACKNOWLEDGMENT

This work was supported by grants from the Wallenberg Consortium North (WCN), the Kempe Foundation, EU-strategic funding, Strategic Research Foundation (SSF), and the Swedish Research Council. Krister Lundgren is acknowledged for help with the GC/MS analysis, IngaBritt Carlsson with preparing the standard mixtures and growing the plants, and Dr. Nicholas Harberd, John Innes Centre, Norwich, U.K., for his gift of *Arabidopsis* seeds. Script is available for academic and noncommercial users.

SUPPORTING INFORMATION AVAILABLE

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

Received for review April 8, 2005. Accepted July 6, 2005.

AC050601E

(42) Wagner, C.; Sefkow, M.; Kopka, J. *Phytochemistry* **2003**, *62*, 887−900.