· Research Papers ·

Visualization of Multivariate Metabolomic Data

ZHOU Jun[1], Aa Jiye[1], WANG Guang-ji[1*], ZHANG Feng-yi[1], GU Rong-rong[1], WANG Xin-wen[1], ZHAO Chun-yan[1], LI Meng-jie[1], SHI Jian[1], CAO Bei[1], ZHENG Tian[1], LIU Lin-sheng[1], GUO Sheng[2], DUAN Jin-ao[2]

1. Lab of Metabolomics, Key Laboratory of Drug Metabolism and Pharmacokinetics, China Pharmaceutical University, Nanjing 210009, China
2. Key Laboratory of Chinese Medicine, Nanjing University of Chinese Medicine, Nanjing 210029, China

Abstract: **Objective** Although principal components analysis profiles greatly facilitate the visualization and interpretation of the multivariate data, the quantitative concepts in both scores plot and loading plot are rather obscure. This article introduced three profiles that assisted the better understanding of metabolomic data. **Methods** The discriminatory profile, heat map, and statistic profile were developed to visualize the multivariate data obtained from high-throughput GC-TOF-MS analysis. **Results** The discriminatory profile and heat map obviously showed the discriminatory metabolites between the two groups, while the statistic profile showed the potential markers of statistic significance. **Conclusion** The three types of profiles greatly facilitate our understanding of the metabolomic data and the identification of the potential markers.

Key words: discriminatory profile; heat map; metabolomic data; principal components analysis; statistic profile
DOI: 10.3969/j.issn.1674-6384.2011.04.008

Introduction

Metabolomics has been widely applied to studying herbal medicines and their biological responses. High-throughput metabolomic analyses, such as NMR- and MS-based techniques, could generate a large data set containing hundreds even thousands of signal variables. A data matrix of the signal intensity values (i.e., peak areas or peak heights in LC- or GC-MS analysis) is therefore constructed with two vectors: sample names as observations in the first column, and retention times (LC-MS- and GC-MS-based metabolomic data) or chemical shifts (NMR-based metabolomic data) as the response variables in the first row (Aa, 2010). Conventionally chemometrical approaches are applied to analyzing and evaluating the data, which are called multivariate statistical (MVS) analysis (Aa, 2010; Eriksson *et al*, 2001; Trygg, Holmes, and Lundstedt, 2007). Although many approaches are applicable, principal components analysis (PCA) is most popular, and greatly facilitates the visualization and interpretation of the data. The scores plot shows the clustering of groups or samples, suggesting their differences or similarities; While the loading plot shows the distribution of the detected variables which are positively correlated to the samples positioning in scores plot. However, the quantitative concept in both scores plot and loading plot is rather obscure, since the location of each plot is determined by a summary of variables of one sample (observation). In other words, PCA plots are rather abstract and do not show the data in a direct way, especially for a specific metabolite; And they provided little quantitative information for the identification of potential marks among a large number of detected peaks/variables. To assist our understanding of the metabolomic data and facilitate the identification of the potential markers, in this article we introduce a series of techniques to analyze and visualize the multivariate data based on our previous studies using GC-TOF-MS as the

* Corresponding author: Wang GJ Address: Key Laboratory of Drug Metabolism and Pharmacokinetics, China Pharmaceutical University, Nanjing 210009, China Tel: +86-25-8327 1192 Fax: +86-25-8530 6750 E-mail: guangjiwang_cpu@hotmail.com
Received: December 15, 2010; Accepted: March 20, 2011
Fund: the National Key New Drug Creation Special Programs (2009ZX09304-001 and 2009ZX09502-004); National Natural Science Foundation of the People's Republic of China (81072692); National Key Fundamental Research "973" Projects (2011CB505300 and 2011CB505303)

analytical tool (Liu *et al*, 2010; Aa *et al*, 2005).

Methods

Clinic samples collection

Access to human samples complied with the *Guidelines of the First Affiliated Hospital of Nanjing Medical University Ethics Committee* in accordance with the *Declaration of Helsinki*. Written informed consent was obtained from all patients. Altogether, 17 patients with gastric cancer (GC) and 20 control patients with chronic superficial gastritis (CSG) were included. After overnight fasting and a thorough rinse of gastrointestinal tract, tissue samples were collected from CSG and GC patients when undergoing gastroscopic examination, and the biopsy tissue was immediately frozen in liquid nitrogen and stored at −80 ℃. The diagnosis of GC and CSG was confirmed in all patients by histopathologic and pathological examinations.

Sample preparation and GC-TOF-MS analysis

Metabolites in tissue samples were extracted as previously reported (Zhang *et al*, 2009). Approximately 20 mg tissue sample was weighed, and 800 μL monophasic mixture of water-methanol (1:4) containing [1,2-$^{13}C_2$]-myristic acid as internal standard (IS, 2.5 μg/mL) was added to the sample and then manually homogenized in a glass grinder carefully. Subsequently the homogenate was transferred to an Eppendorf tube and vigorously extracted at a frequency of 30 Hz for 3 min using an MM400 Vibration Mill (Retsch GmbH, Haan, Germany) with a 3 mm steel bead in each tube to increase the extraction efficiency. After removal of the beads, the tubes were centrifuged at 20 000 × *g* for 10 min and 200 μL of the supernatant was added into a GC vial. The supernatant was dried and the analytes were trimethylsilylated. Finally 1 μL derivatized sample was injected and analyzed in GC-TOF-MS in the same way as previously described (Zhang *et al*, 2009).

GC-TOF-MS data acquisition and identification of metabolites

Automatic peak detection and calculation of peak areas for IS and detected compounds were made using the ChromaTOF 3.20 software of the Leco Corporation (USA). Peak widths in automatic peak detection and mass spectrum deconvolution were all set to 2 s. Peaks of signal-to-noise (S/N) ratios lower than 10 were rejected and the peak areas were obtained as previously reported (Aa *et al*, 2005; Jonsson *et al*, 2005). The retention index of each peak was calculated by comparison of its retention time with those of the standard alkane series C_8 — C_{40}. Identification of compounds was achieved by comparing the mass spectra and retention index of all detected compounds with authentic reference standards and those available in the National Institute of Standards and Technology (NIST) library 2.0 (2005).

Multivariate data analysis

Multivariate data analysis and modeling were performed with SIMCA-P 11 software (Umetrics, Umeå, Sweden). The data matrix was constructed with the sample names as observations and the peak areas normalized by IS as the response variables. GC-TOF-MS data were analyzed using PCA, partial least squares projection to latent structures and discriminant analysis (PLS-DA), and orthogonal partial least squares projection to latent structures (OPLS) (Trygg, Holmes, and Lundstedt, 2007).

Discriminatory profile

Deconvolution of the GC-TOF-MS profiles (both GC and CSG samples) resulted in a data matrix of peak areas; For each peak the mean values of GC and CSG groups were averaged, respectively, and a relative ratio of the mean values was calculated to indicate the relative intensity. Of the two groups, when the averaged value of a peak in GC is higher than that in CSG, the relative ratio has a positive value, vice versa, when the averaged value of a peak in CSG is higher than that in GC, the relative ratio has a negative value. Statistic significance was assessed for each peak between groups using analysis of variance, and *P* values were evaluated. Finally, the retention times were designated as *X* values, and the relative intensities were determined as *Y* values. The Netcdf files containing total intensity and retention time data sets were exported from ChromaTOF 3.20 software. Then an in-house Matlab script read the Netcdf files, calculated *X*, *Y*, and *P* values, plotted *X* and *Y* values, and colored the plots according to *P* values. Each plot was colored to indicate the statistic significance by red or blue according to the statistic result. Red indicates statistic significance by calculating a value of 1—p. With so many plots on a 2-dimensional plane, a continual colored curve appears as the discriminatory profile.

Creation of heat maps

Based on the data set of GC and CSG, the relative intensity value of one peak in a certain sample was calculated among the samples, and then each value was replaced by a small colored block, i.e., warm color means higher abundance while cool color means the lower. Consequently, the data matrix was transformed into a color panel, and a heat map was created by Heatmap Builder which was freely available (Ashley Labs, Stanford University, CA, USA). The excel sheet consisting of compound ID, compound name, sample name, and peak areas was saved as a tab-delimited text file to be imported to Heatmap Builder, and then it made customized heat map images from the input file. With the help of "heatmap" command in the BioConductor package (Gentleman *et al*, 2004), the heat map can also be ordered after hierarchical clustering of data and a dendrogram is aligned and together with the heat map.

Statistic profile

Based on the same data set as above, the folder change and the statistic significance were evaluated for each metabolite in the two sample classes, i.e., GC and CSG. Briefly, we performed One-way ANOVA for direct comparisons of the peak areas normalized against the IS, and the *P* values were obtained. With the calculated folder change of GC against CSG samples, the *P* values, a statistic profile, could be made and the discriminatory metabolites could be identified easily.

Results and discussion

Discriminatory profile

Discriminatory profile could show the differences of spectra between groups or samples. Spectrum is the primal data from metabolomic analysis, regardless of the analytical tools. As the output of GC-TOF-MS analysis of metabolites in bio-samples, the total ion current (TIC) chromatogram shows the retention times and signal intensities of detected peaks. Direct comparison of the peaks in TIC gave an overview on the whole spectrum and the detected peaks, facilitated the marking of discriminatory metabolites, and retained the original information as much as possible. Therefore, a discriminatory profile directly reflects interclass metabome variance. The different intensity of a metabolite at certain retention time (LC-MS- or GC-MS-based metabolomics) or chemical shift (NMR-based metabolomics) indicates the varied levels of certain metabolite between groups. In comparison with the scores plot and the loading plot (Figs. 1 and 2), the discriminatory profile of GC tissue and CSG tissue clearly showed their metabolomic difference (Fig. 3) and the discriminatory metabolites.

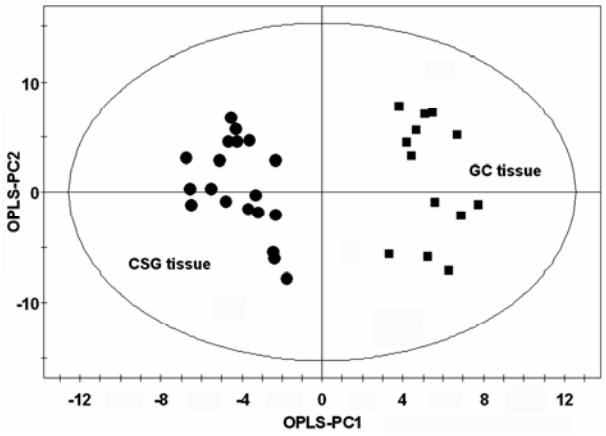

Fig. 1 Overview of GC and CSG tissue samples by scores plots

GC and CSG tissue samples cluster closely within each group, respectively, but show the difference between the two groups

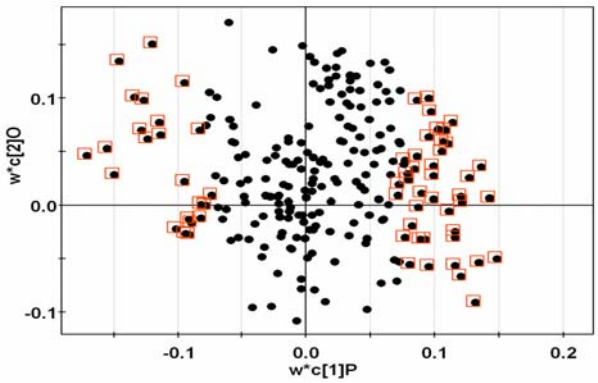

Fig. 2 Loadings plot roughly suggested discriminatory variables/metabolites (marked plots) between GC and CSG tissue samples

Heat maps

Exclusively, metabolomic data contain a data matrix of the signal intensity values with two vectors, peaks or variables, and samples or observations. Usually there are hundreds even thousands of variables and dozens or even hundreds of observations so that the data sheet is very large. Visual inspection of the data set is meaningless. For overview of the data set, heat map is useful. Heat map is typically used in genomics to represent the expression level of many genes across a number of comparable samples as they were obtained from DNA microarrays. In metabolomics, the relative concentration of the endogenous metabolites in experimental group has to be compared with those in

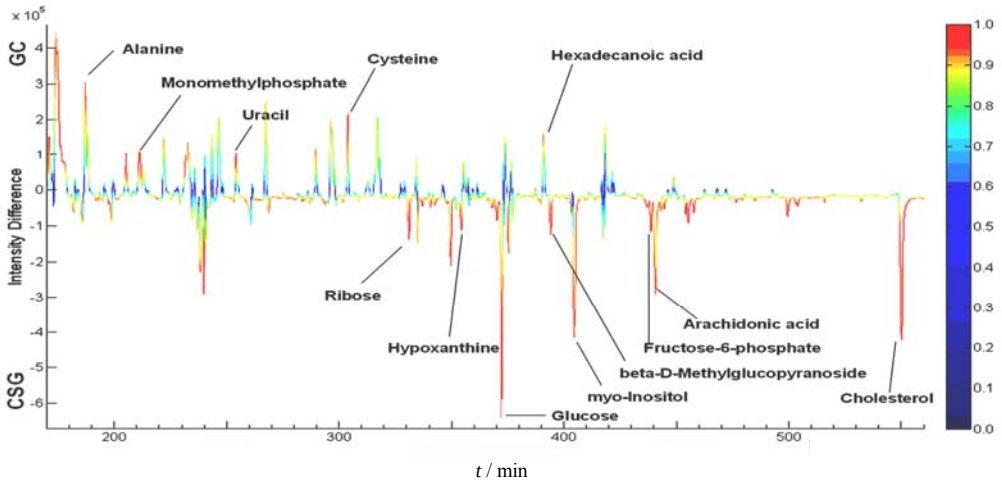

Fig. 3 Discriminatory profile clearly showed the metabolites contributing most to the differentiation of GC and CSG tissue

Relative intensity in GC groups has a positive value; while relative intensity in CSG group has a negative value. Each of the peak was colored by red or blue according to the statistic result. Red rather than blue color indicates statistic significance by calculating a value of 1－p

the control group to characterize the metabolic features and to identify the potential markers. Rapid identification of the discriminatory metabolites is necessary. Heat map just has the advantage of facilitating overview the data and quick screening of discriminatory metabolites by showing the relative abundance in different colors. The heat map gives a direct overview of data matrix analyzed by GC-TOF-MS of GC tissue and CSG tissue (Fig. 4). Each column reflects the levels of the measured metabolites (parts) of a particular sample, while each row reflects the different levels of a metabolite in the samples. Further inspection of the graph shows the differences of samples (i.e., samples A and C) and the discriminatory metabolites of the two groups (e.g., higher level of lactate in sample C and lower level of 2-ketoisovaleric acid in sample A).

Statistic profile

Identification of the potential markers is one of the most important goals in metabolomic research. Although the discriminatory profile, the heat map, and the loading plot in PCA may indicate some of the potential markers, the statistic significance is not clearly shown. To avoid false identification of potential markers in metabolomics, the candidate markers must be strictly validated using appropriate statistical approaches (Broadhurst and Kell, 2006). In the above study on identifying the potential markers of GC, the statistical significances of the discriminatory metabolites were assessed by one way analysis of variance, and *P* values were calculated. This profile

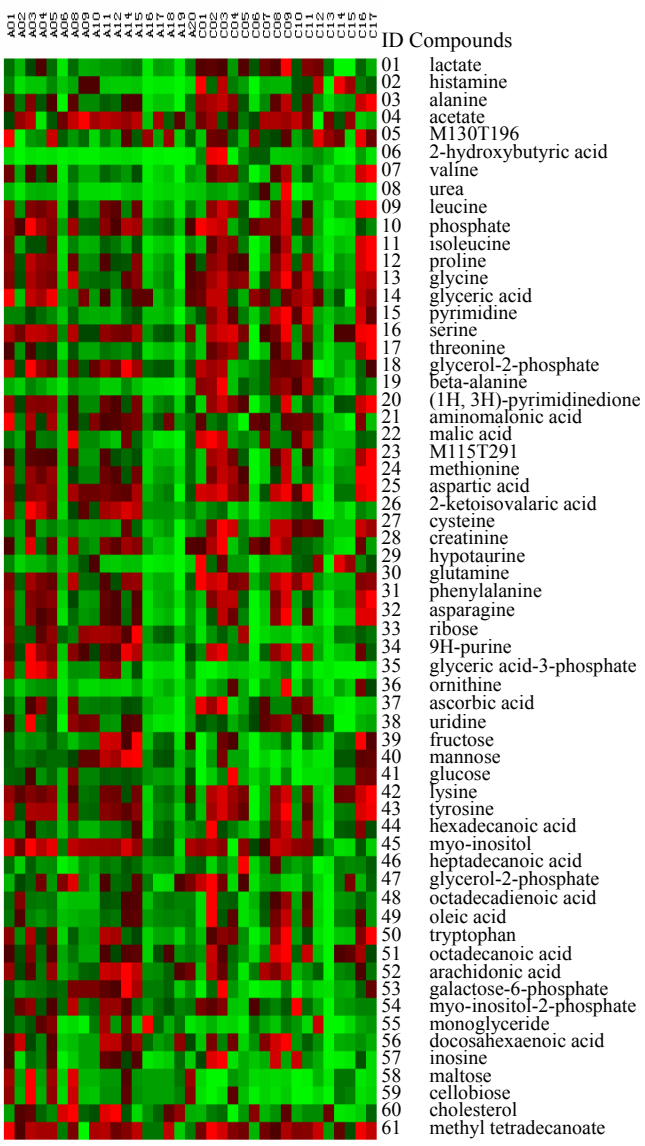

Fig. 4 Heat map gives a direct overview of data matrix analyzed by GC-TOF-MS of GC tissue and CSG tissue

shows distinct significances of the discriminatory metabolites (Fig. 5). It is shown that metabolites below the line (P = 0.01) are of statistical difference between these two groups. On the contrary, metabolites above this line are of no statistic significance (P > 0.01).

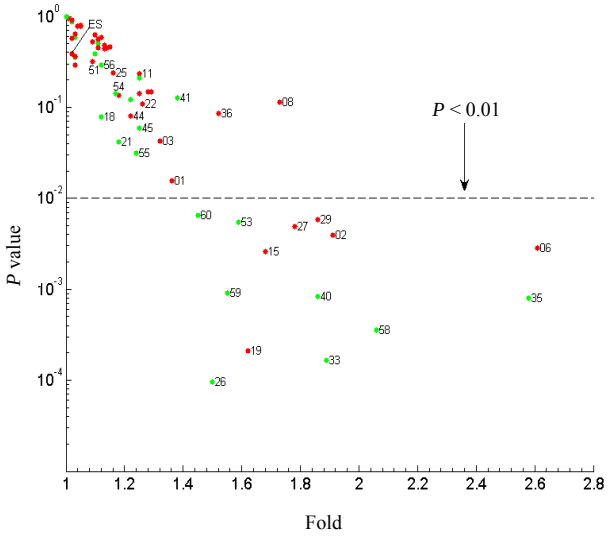

Fig. 5 Statistic results of part of potential markers

Each dot reflects *P*-value of two different statistical analyses of a certain metabolite. Red plot indicates the metabolite is more abundant in CSG tissue compared to GC tissue, and green plot is less concentrated in CSG. Metabolite 26 is identified as 2-ketoisovaleric acid, which is of extremely significant difference between GC and CSG, and is indicated as the potential markers of GC

References

Aa J, 2010. Analysis of metabolomic data: Principal component analysis. *Chin J Clin Pharmacol Ther* 15(5): 481-489.

Aa J, Trygg J, Gullberg J, Jonsson P, Antti H, Marklund SL, Moritz T, 2005. Extraction and GC/MS analysis of the human blood plasma metabolome. *Anal Chem* 77: 8086-8094.

Broadhurst D, Kell D, 2006. Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics* 2: 171-196.

Eriksson L, Johansson E, Kettaneh-Wold N, Wold S, 2001. *Multi- and Megavariate Data Analysis Principles and Applications*. Umeatrics Academy: Umetrics AB, Sweden.

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J, 2004. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biol* 5(10): R80.

Jonsson P, Johansson AI, Gullberg J, Ytygg JAJ, Grung B, Marklund S, Sjöström M, Antti H, Moritz T, 2005. High-throughput data analysis for detecting and identifying differences between samples in GC/MS-based metabolomic analyses. *Anal Chem* 77: 5635-5642.

Liu L, Aa J, Wang G, Yan B, Zhang Y, Wang X, Zhao C, Cao B, Shi J, Li M, Zheng T, Zheng Y, Hao G, Zhou F, Sun J, Wu Z, 2010. Differences in metabolite profile between blood plasma and serum. *Anal Biochem* 406: 105-112.

Trygg J, Holmes E, Lundstedt T, 2007. Chemometrics in metabolomics. *J Proteome Res* 6: 469-479.

Zhang Y, Aa J, Wang G, Huang Q, Yan B, Zha W, Gu S, Liu L, Ren H, Ren M, Sheng L, 2009. Organic solvent extraction and metabonomic profiling of the metabolites in erythrocytes. *J Chromatogr B Analyt Technol Biomed Life Sci* 877: 1751-1757.