Available online at www.sciencedirect.com
ScienceDirect
Journal of Chromatography B, 850 (2007) 236–240
JOURNAL OF CHROMATOGRAPHY B
www.elsevier.com/locate/chromb

A GC-based metabonomics investigation of type 2 diabetes by organic acids metabolic profile

Kailong Yuan [a], Hongwei Kong [a], Yufeng Guan [b], Jun Yang [a], Guowang Xu [a,*]

[a] *National Chromatographic R&A Center, Dalian Institute of Chemical Physics, Chinese Academy of Science, Dalian 116023, PR China*
[b] *Dalian Medical University, Dalian 116023, China*

Received 2 April 2006; accepted 20 November 2006
Available online 22 December 2006

Abstract

"Metabonomics" method requires the development of rapid, advanced analytical tools and GC will play an important role for its special advantage. In this study we show the application of GC-based metabonomics to investigate the control and type 2 diabetes (DM2) patients by urinary organic acids metabolic profile. After peak matching, multivariate statistical analysis methods: principal components analysis (PCA) and partial least squares-discriminant analysis (PLS-DA) were used. The results showed that there was a relationship between organic acids metabolic profiles and DM2, and PLS-DA can distinguish the DM2 patients from the control. Five organic acids as potential biomarkers were identified.

Keywords: Metabonomics; GC; Organic acids; Peak matching algorithm; Type 2 diabetes; Biomarkers

1. Introduction

In the last decade "metabonomics" has demonstrated enormous potential in furthering the understanding of, for example, disease processes, toxicological mechanism, and biomarker discovery [1,2]. Metabonomics is a holistic approach for measuring time-related biochemical responses in key intermediary biochemical pathways as a result of physiological, pathological, or interventional genetic events, and this has been achieved principally through the use of 1H NMR spectroscopy on biofluids such as urine or plasma [3,4].

However, "metabonomics" method requires the development of rapid, advanced analytical tools to comprehensively profile biofluid metabolites. The key point of the methodology is to have in disposition a generic analytical method for rapid biofluid sample profiling together with a chemometric method for data evaluation. As NMR spectroscopy is non-destructive, not selective and high throughput it has widely been used for metabonomics. But NMR has its drawbacks: poor sensitivity and resolution. Chromatography has been used mainly in biofluid analysis, especially for target component analysis and not for whole sample profiling combined with chemometrics [5–8]. The high separation power and the ability to achieve high sensitivity are strong incentives for the consideration of its use in biofluid fingerprinting as well. So chromatography would provide additional and complementary information that cannot be achieved with NMR [9]. Capillary GC provides the highly sensitive analyses required for multivariate statistical processing and has a greater chance to find biomarkers of disease or toxicity.

Type 2 diabetes (DM2) was a typical metabolism disorder disease with the very high frequency of 1.5–2.5% in the Western hemisphere [10]. Some of the changes in the concentrations of the metabolites are detectable in blood serum and in urine even when the diabetic patients are well controlled by therapy. These alterations are therefore inherently associated with the disease [10]. GC analysis has revealed a number of pathophysiological changes to accompany diabetes mellitus.

In order to reduce the complexity of biofluid GC data and facilitate analysis, automatic data-reduction followed by chemometric methods, for example, principal components analysis (PCA) and partial least squares-discriminant analysis (PLS-DA), can be applied. Herein, an efficient GC-based

* Corresponding author. Tel.: +86 411 84379530; fax: +86 411 84379559.
E-mail address: xugw@dicp.ac.cn (G. Xu).

1570-0232/$ – see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.jchromb.2006.11.035

metabonomic approach to understand pathophysiological process has been developed; FID was used for the quantitative analysis of the profiling, and MS for the qualitative analysis. A metabonomic strategy including peak matching algorithm has been applied to investigate whether there is a relationship between the concentration of organic acids and DM2, then to distinguish the DM2 patients from the control and to discover potential biomarkers.

2. Experimental

2.1. Sample collection and preparation

Urine samples were collected from 26 healthy adults and 28 patients with DM2. The age range was 28–63 years (43 ± 15, 46 ± 17, respectively) and roughly age-matched. All patients were from the Second Affiliated Hospital of Dalian Medical University (Dalian, China) with the fasting plasma glucose concentration above 7.0 mmol/L. The patients were enrolled into the study via "informed consent". Samples were stored at $-20\,^{\circ}C$ until assayed.

Urine samples were individually processed for organic acid analysis. The first step is the preparation of SPE column: the SAX column was washed successively with 2 mL methanol, 2 mL water, 2 mL phosphate buffer at pH 7 (0.336 mol/L potassium dihydrogenphosphate and 0.665 mol/L disodium hydrogenphosphate), and finally 2 mL of diluted phosphate buffer at pH 7 (0.013 mol/L potassium dihydrogenphosphate and 0.020 mol/L sodium hydrogenphosphate). After addition of ferulic acid as internal standard (I.S) at 50 μg/L, an aliquot (2 mL) of centrifuged urine was passed through the cartridge. The cartridge was dried by sucking air through them and elution of the adsorbed organic acids was performed with 1.5 mL of aqueous HCl in methanol (1 mL of concentrated HCl adjusted with methanol to 25 mL).

Each sample was dried by rotary evaporator and placed with 140 μL of 6:1 *N*-(*tert*-butyldimethylsilyl)-*N*-methyltrifluoroacetamide (MTBSTFA)/*N*,*N*-dimethylformamide (DMF). The vial was airproof and heated at 50 °C for 1 h, then cooled to room temperature before being placed on the autosampler vial for GC analysis.

2.2. GC–FID and GC–MS

After derivatized, the extracts were analyzed using an Agilent 6890N gas chromatograph equipped with FID. The column used was a DB 5 ms (30 m × 0.25 mm × 0.25 μm) (J&W, USA). One microliter sample was injected in the split mode (30:1). The carrier gas was He with the flow rate 30 cm/min. The split/splitless injection port was at 280 °C. The oven temperature was programmed: initially 40 °C, then ramped to 260 °C at 2.5 °C/min, and held for 2 min.

The mass (MS) detector was used for qualitative analysis. All mass spectra were acquired in the electron impact (EI) mode at 70 eV and scanned in the range of 40–500 amu. The ion source and the interface temperatures are 200 and 290 °C, respectively.

2.3. Processing and pattern recognition of GC chromatogram data

2.3.1. Statistical analysis

For quantitative analysis, all of the peaks exceeding a signal-to-noise (S/N) of 10 were selected. Their relative peak areas (RPA) to the internal standard were evaluated by multivariate data analysis using the software program SIMCA-P (Umea, Sweden). The data were subjected to PCA and PLS-DA.

2.3.2. Principal components analysis (PCA)

In PCA the multivariate data set is projected down to a lower dimensional plane formed by the principal components (PCs) which approximate the data as well as possible in the least square sense. Principal components analysis (PCA) is a bilinear decomposition method used for overviewing 'cluster' within multivariate data. The GC data (X) were represented in K-dimensional space (where K is equal to the number of chemical shift regions) and reduced to a few principal components (or latent variables) which described the maximum variation within the data, independent of any knowledge of class membership. The principal components were displayed as a set of 'scores' (t) that highlighted clustering or presence of outliers and a set of "loading" (p) that described the influence of input variables on t.

The PCs are the uncorrelated (orthogonal) variables, obtained by multiplying the original correlated variables with the eigenvector (loadings or weightings). Thus, the PCs weighted linear combinations of the original variables. PC provides information on the most meaningful parameters, which describe whole data set affording data reduction with minimum loss of original information [11,12]. It is a powerful technique for pattern recognition that attempts to explain the variance of the large set of inter-correlated variables and transforming into a smaller set of independent (uncorrelated) variables (principal components). PCA performed on correlation matrix of individually rearranged data explains the structure of the underlying data set. The correlation coefficient matrix measures how well the variance of each constituent can be explained by relationship with each of the others [13].

2.3.3. Partial least square (PLS)

PLS can be described as the regression extension of PCA. Instead of describing the maximum variation in the measured data (X), which is the case for PCA, PLS attempts to derive latent variables, analogues to PCs, which maximize the co-variation between the measured data (X) and the response variable (Y) regressed against.

Partial least squares-discriminant analysis (PLS-DA) which discriminates the known classes in calibration set is a special form of PLS modeling aims to find the variables and directions in multivariate space. In PLS-DA, an indicator Y matrix of category variables is constructed which contains as many columns as there are known classes in the calibration set, i.e., each class has a column in Y. Each class variable is assigned a value 1 or 0 depending into which class a subject belongs. In the current work, PLS-DA was used to generate models that could distinguish between the control and the DM2.

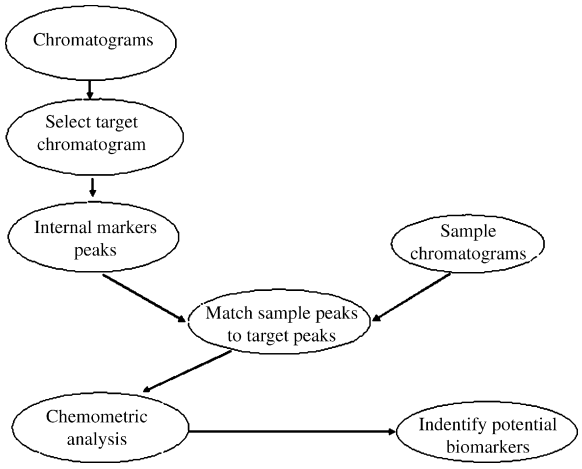

Fig. 1. A flow chart diagram of the peak matching algorithm and metabonomics data analysis.

3. Results and discussion

Metabonomics is based on large amount of samples. So high throughput profiling requires the development of a GC method allowing highly detailed fingerprinting of complex samples containing a wide range of compounds. Compared with NMR, GC has some problems to solve; the most significant one is the retention time shifting of chromatographic peaks due to phenomena related to the instrument itself as well as to the chemical interactions between different samples and the instrument. So a peak matching method should be developed to solve the problem of time shift.

3.1. Peak matching algorithm

The flow chart is given in Fig. 1. Firstly, a target chromatogram was selected (Fig. 2) that is typical of the whole set of compound analyses. An alignment target chromatogram must be chosen carefully. It is important for a chromatographic peak in the target chromatogram to be located as close as possible to the center of the distribution of peak positions. If a target with peaks on the edge of the distributions of peak locations (i.e., a chromatogram that is significantly shifted relative to the others in the set) is chosen, the peak mismatch will increase [14].

Secondly, all of the peaks that can be detected exceeding a signal-to-noise (S/N) of 10 were identified in all of the chromatograms by using the software from GC instrument, 195 peaks with the S/N > 10 were found in the target chromatogram (Fig. 2). All other chromatograms were then matched against this peak list of target chromatogram.

The main intention of the algorithm is to define a series of internal marker peaks firstly which could be easily identified by selecting a wider retention window. A chromatogram was then divided into several zones, which were residing between the internal marker peaks. Compared with other peaks, 10#, 23#, 33#, 54#, 74#, 102#, 133# and 158# which were higher than the others and occurred in most of the other sample chromatograms were selected as the internal marker peaks. The internal standard was also added to the internal marker peak set and numbered as 17#. While the internal marker peaks' retention values have been assigned to a series of pre-defined values, all peaks' retention values were scaled based on the adjusted retention index (ari) that was similar to Kovats retention index and calculated according to Eq. (1):

$$\mathrm{ari}_i = \frac{\mathrm{tr}_i - \mathrm{tr}_j}{\mathrm{tr}_{j+1} - \mathrm{tr}_j} \times (\mathrm{ARI}_{j+1} - \mathrm{ARI}_j) + \mathrm{ARI}_j$$
$$\mathrm{tr}_j < t_i \leq \mathrm{tr}_{j+1} \quad (1)$$

where ari_i is the adjusted retention index of the peak i (i is the peak no., equals to 1, 2, 3, ..., 195); tr_i and tr_j are the retention time of peaks i and j, respectively; $tr_0 = 0$. j ($j = 0, \ldots, C$) is the number of internal marker peaks, in this study, C is equal to 9. ARI_j is the internal marker peak's adjusted retention index, equal to $j \times 100$. Based on Eq. (1), the ari values of the peaks would be very stable in different chromatograms.

Thirdly, the alignment was carried out on the basis of the ari value. The sample and the target chromatograms are compared by stepping through each of the peaks in the target chromatogram and finding the sample chromatogram peak that most closely matches it in the ari value. If the closest match is within a selected distance from the ari of the target peak, then the peaks are matched. The matching window of the ari can be adjusted on the basis of the distribution density of the peaks in the chromatograms. The region crowded with more target peaks was assigned to a smaller ari difference. This allowed the peaks

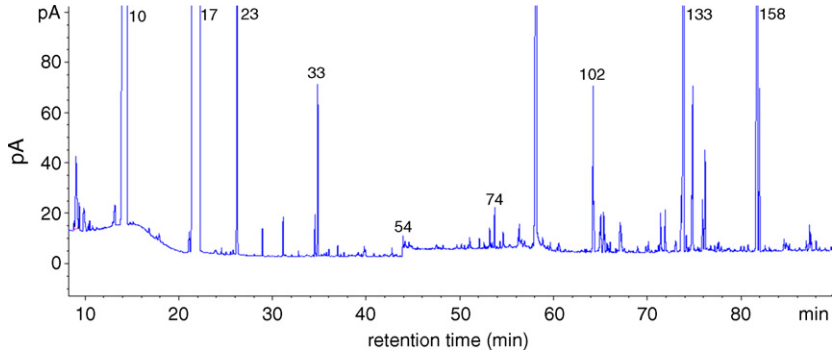

Fig. 2. A target chromatogram from a typical analysis of the whole set of compound analyses.

to be correctly aligned and transformed into the data array for multivariate statistical analysis.

3.2. Multivariate analysis

The purpose of applying multivariate statistical methods to the analysis of organic acids data was to identify the profiling characteristic of DM2 comparing with control and discover potential biomarkers.

Here, there were 54 samples from two classes (the control and the DM2) including 195 variables (chromatographic peak no.). On the basis of their relative peak areas to the internal standard, the PCA analysis of urinary organic acids from the control and DM2 was carried out. As overlapping happened (Fig. 3a), the DM2 patients could not be separated from the controls.

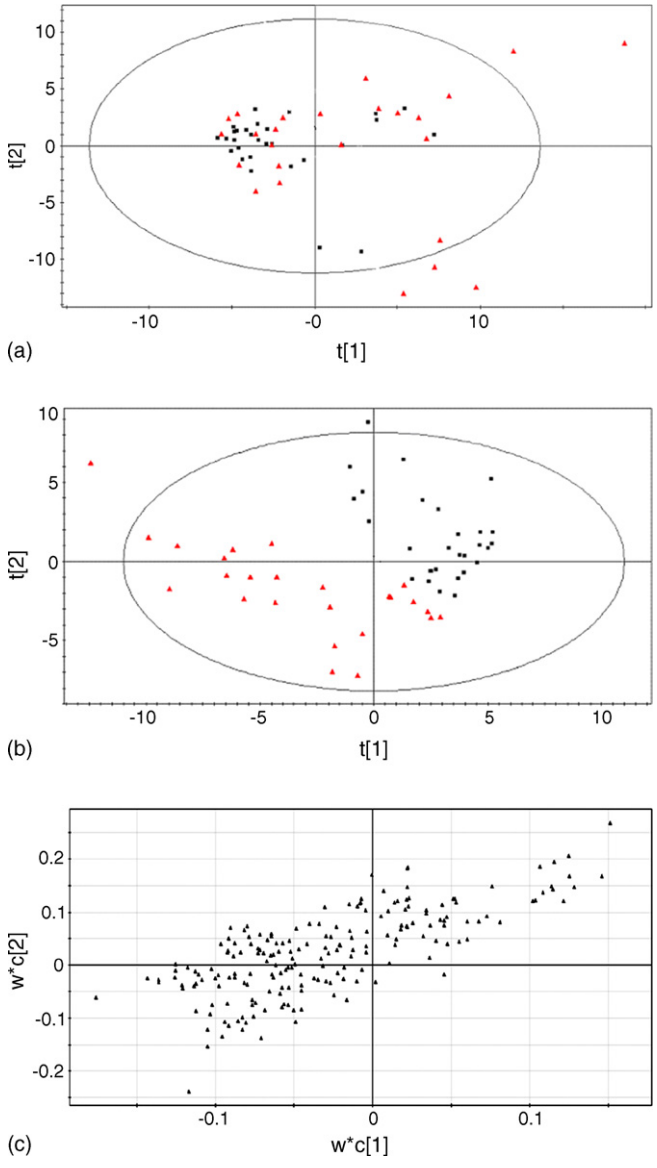

Fig. 3. (a) Score plot from principal component analysis (PCA). (b) Score plot from Partial least squares-discriminant analysis (PLS-DA) based on relative amounts of 195 constituents of urine samples ((black ■) the DM2 patient; (red ▲) the control). (c) PLS-DA loadings plot.

To improve the classification of the DM2 and the control, the PLS-DA was used. In PLS-DA the data set is modeled in way similar to PCA, but in combination with a discriminant analysis. The objective of PLS-DA is to find a model that separates (discriminates) the *X* data according to above described treatments as well as possible. Therefore an additional *Y* matrix was made up as a dummy variable, containing the values 1 and 0 for each treatment, respectively. The number of significant principal components was determined by cross-validation. It can be found that two groups according to the treatments were fairly discriminated. Nearly all samples were clearly separated from the other class; the DM2 patients appeared in the up-right zone, and the control in the area of the down-left zone (Fig. 3b).

The PCA shows a dissatisfactory separation of the two classes, but the separation can be greatly improved by the use of PLS-DA. The reason is that PCA is a technique that finds a lower dimensional space capturing the maximum amount of variance in an input data matrix, *X*, without losing any useful information. PLS is a similar approach to PCA except it reduces the dimension of both input and output data matrices, *X* and *Y*, by capturing the maximum amount of covariance between *X* and *Y*, to best predict *Y*.

3.3. Potential biomarkers

The PLS-DA loading plot given in Fig. 3c shows which variables contribute strongly to the separation of classes. From the loading, we can know potential biomarkers that were the furthest one from the origin in the loading plot.

The structures of potential biomarkers were identified by mass spectrometry. Table 1 gives their molecular formula and the quantitative analysis result. The compounds detected are mostly organic acids and they are derivated to *tert*-bu-TMS derivatives. CI and SIM scan modes by using $m/z = 73$ and 115 were combined, the data were then submitted to a NIST library search (NIST147, NIST27) which resulted in a hit for *tert*-bu-TMS-derivatized organic acids (similarity more than 85%). The agreement between the exact mass measurement and library search results is excellent and is supported by the CI data.

The potential biomarkers identified are mostly organic acids. Organic acids play an important role in nearly all the metabolic processes and take part in many different physiological and pathophysiological functions such as nutrient deficiencies, mitochondrial energy production, intestinal dysbiosis, free radical overload, and so on [15]. And they are also considered as biomarkers for many diseases, such as the inborn errors of organic acidurias, diabetes, central nervous system diseases, etc. [16–18]. Four biomarkers identified are all short-chain organic acids interrelated with the metabolic disorder as reported before [19,20]. And these newly found biomarkers probably imply the underlying metabolic disorder of the patients, or the trends to acidemia. This information provides important clues for the understanding and monitoring of the disease. For its high sensitivity, selectivity, and identification capability of GC–MS, the organic acid profiling method we established can further be used for the treatment of type 2 diabetes to prevent the metabolic disorders. It is very important that early and accurate diagnosis of

Table 1
Identification result of the potential biomarkers

Compounds	Molecular formula	Average (I.S)		*T*-test
		DM2	Control	
Maleic acid, dimethyl ester	$C_6H_8O_4$	2.80E−5	8.20E−5	0.03
Oxyl acetic acid	$C_2H_4O_3$	4.31E−5	1.67E−4	0.02
4-Aminobenzoic acid	$C_7H_7NO_2$	4.24E−4	9.78E−5	0.0008
2,5-Bisoxy-benzeneacetic acid	$C_8H_8O_4$	2.61E−4	9.53E−4	0.0002

metabolic disorders is made. The treatments of these disorders perhaps are simple, yet when undiagnosed and untreated, they will result in serious syndrome or even worse.

4. Conclusions

Metabonomics is now recognized as an independently and widely used technique for evaluating the toxicity of drug-candidate compounds, deriving new biochemically based assays for disease diagnosis, understanding the relationships between gene function and metabolic control in health and disease, and identifying combination biomarkers for disease. We have demonstrated the utility of GC for urinary metabonomics studies after sample preparation. The results have shown that peak matching of the spectra followed by multivariate techniques is an good method for screening DM2 biomarkers. One of the significant advantages using GC–MS is that the interesting compounds in the loading plot can be identified based on the NIST database. So we can allege that GC will play an important role in the future metabonomics research. It should be emphasized that the main aim of this study is to develop a peak matching algorithm based on GC and to apply the method to the DM2 biomarker discovery for distinguishing healthy adult controls from DM2 patients, the patients with other non-DM2 metabolic disorders will be included in the future.

Acknowledgements

The studies have been supported by the foundation (No. 20425516) for Distinguished Young Scholars from National Natural Science Foundation of China and the Knowledge Innovation Program of the Chinese Academy of Sciences (KSCX2-SW-329).

References

[1] J.K. Nicholson, I.D. Wilson, Prog. Nucl. Magn. Reson. Spectrosc. 21 (1989) 449.
[2] J.K. Nicholson, J. Connelly, J.C. Lindon, E. Holmes, Nat. Rev. Drug. Discov. 1 (2002) 153.
[3] J.K. Nicholson, J.C. Lindon, E. Holmes, Xenobiotica 29 (1999) 1181.
[4] J.C. Lindon, J.K. Nicholson, E. Holmes, J.R. Everett, Prog. NMR Spectrosc. 12 (2002) 89.
[5] J. Yang, G.W. Xu, H.W. Kong, Y.F. Zheng, T. Pang, Q. Yang, J. Chromatogr. B 780 (2002) 27.
[6] J. Yang, G.W. Xu, H.W. Kong, H.M. Liebich, K. Lutz, R.M. Schmulling, H.G. Wahl, J. Chromatogr. B 813 (2004) 53.
[7] J. Yang, G.W. Xu, H.W. Kong, T. Pang, S. Lv, Q. Yang, J. Chromatogr. B 813 (2004) 59.
[8] J. Yang, G.W. Xu, Y.F. Zheng, H.W. Kong, T. Pang, S. Lv, Q. Yang, J. Chromatogr. B 813 (2004) 59.
[9] H. Pham-Tuan, L. Kaskavelis, C.A. Daykin, H. Janssen, J. Chromatogr. B 789 (2003) 283.
[10] H.M. Liebich, J HRC & CC 6 (1983) 640.
[11] D.A. Wunderlin, M.P. Diaz, M.V. Ame, S.F. Pesce, A.C. Hued, M.A. Bistoni, Water Res. 35 (2001) 2881.
[12] B. Helena, R. Pardo, M. Vega, E. Barrado, J.M. Fernandez, L. Fernandez, Water Res. 34 (2000) 807.
[13] C.W. Liu, K.H. Lin, Y.M. Kuo, Sci. Total Environ. 313 (2003) 77.
[14] K.J. Johnson, B.W. Wright, K.H. Jarman, R.E. Synovec, J. Chromatogr. A 996 (2003) 141.
[15] B. Baena, A. Cifuentes, C. Barbas, Electrophoresis 26 (2005) 2622.
[16] K. Tanaka, D.G. Hine, A. West-Dull, T.B. Lynn, Clin. Chem. 26 (1980) 1839.
[17] T. Niwa, K. Meada, T. Ohki, A. Saito, I. Tsuchida, J. Chromatogr. 225 (1) (1981) 1.
[18] S. Kolker, E. Mayatepek, G.F. Hoffmann, Neuropediatrics 33 (5) (2002) 225.
[19] H. Chen, Y. Xu, F.V. Lente, J. Chromatogr. B 679 (1996) 49.
[20] K.B. Elgstoen, J.Y. Zhao, J.F. Anacleto, E. Jellum, J. Chromatogr. A 914 (2001) 265.