Journal of Chromatography A, 1399 (2015) 65–73

Contents lists available at ScienceDirect

Journal of Chromatography A

journal homepage: www.elsevier.com/locate/chroma

A chemometric-assisted method based on gas chromatography–mass spectrometry for metabolic profiling analysis

Yong-Jie Yu[a,*,1], Hai-Yan Fu[b,1], Li Zhang[c], Xiao-Yu Wang[a], Pei-Jian Sun[a], Xiao-Bing Zhang[a], Fu-Wei Xie[a,*]

[a] Zhengzhou Tobacco Research Institute of CNTC, Zhengzhou 450001, China
[b] College of Pharmacy, South-Central University for Nationalities, Wuhan 430074, China
[c] Technology Center of China Tobacco Guizhou Industrial Co. Ltd., Guiyang 550009, China

ARTICLE INFO

Article history:
Received 5 December 2014
Received in revised form 23 March 2015
Accepted 16 April 2015
Available online 22 April 2015

Keywords:
Metabolic profiling analysis
Plant sample
Automatic chromatographic peak detection
Peak-shift alignment

ABSTRACT

An automatic and efficient data analysis method for comprehensive metabolic profiling analysis is urgently required. In this study, a new chemometric-assisted method for metabolic profiling analysis (CAMMPA) was developed to discover potentially valuable metabolites automatically and efficiently. The proposed method mainly consists of three stages. First, automatic chromatographic peak detection is performed based on the total ion chromatograms of samples to extract chromatographic peaks that can be accurately quantified. Second, a novel peak-shift alignment technique based on peak detection results is implemented to resolve time-shift problems across samples. Consequently, aligned results, including aligned chromatograms, and peak area tables, among others, can be successfully obtained. Third, statistical analysis using results from unsupervised and supervised classification results, together with ANOVA and partial least square-discriminate analysis, is performed to extract potential metabolites. To demonstrate the proposed technique, a complex GC–MS metabolic profiling dataset was measured to identify potential metabolites in tobacco plants of different growth stages as well as different plant tissues after maturation. Results indicated that the efficiency of the routine metabolic profiling analysis procedure can be significantly improved and potential metabolites can be accurately identified with the aid of CAMMPA.

1. Introduction

Metabolomics aims to achieve the large-scale identification and quantification of metabolites to study metabolite changes and is widely applied in various scientific fields [1–5]. Metabolic analysis can usually be classified as either target or untargeted analysis [2,3,6]. Untargeted analysis detects all metabolites species simultaneously and is commonly achieved by using techniques such as GC–MS and LC–MS. These techniques provide large amounts of data containing information on a broad range of compound classes and are thus valuable for metabolic profiling analysis [7,8]. In contrast to instrumental improvements, data analysis methods for metabolic profiling analysis lag behind in recent years. The problem of hundreds of chromatographic peaks eluting in a single total ion chromatogram (TIC) is commonly encountered in most practical applications. Moreover, manual verification of large-scale datasets is cumbersome, inefficient, prone to errors, and even irreproducible during complex sample analysis [2,5,9]. As such, a data analysis method for the large-scale, high-throughput conversion of raw datasets into organized data for data visualization and potential metabolite extraction must be developed.

A number of methods and software tools are used to process GC–MS or LC–MS data, and these techniques are frequently applied in metabolic studies. Some of these tools include ADAP-GC 1.0 and 2.0 [10,11], MRMPROBS [12], MET-IDEA [13], CAMERA [14], MetSign [15], MetAlign [16], and XCMS [17–20], among others [21–36]. When processing metabolic profiling sets based on data obtained through chromatography coupled to MS, the performance of a data-mining method mainly depends on two critical stages: (1) peak detection and integration [11,23] and (2) peak-shift alignment [31,34]. Various algorithms have been individually proposed for either chromatographic peak detection or time peak alignment [37–51]. For instance, widely used time-shift alignment methods are often represented by correlation-optimized warping and dynamic-time warping [37,52]. However, in practical application, analysis of complicated samples continues to challenge

* Corresponding authors. Tel.: +86 371 67672502; fax: +86 731 67670260.
E-mail addresses: yongjie.yu@163.com (Y.-J. Yu), xiefuwei@sina.com (F.-W. Xie).
[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.chroma.2015.04.029
0021-9673/

researchers, especially when large-scale chromatograms are analyzed for metabolic profiling.

An automatic chromatographic peak detection (ACPD) method was proposed recently in our previous study [53]. This method can efficiently and accurately detect valuable chromatographic peaks with signal-to-noise ratios larger than 10 using an automatic instrumental noise estimation strategy and provide reasonable background-drift correction results. Using peak detection results as bases, a novel peak-shift alignment strategy was further established to align chromatographic peaks across samples. In the present study, a chemometric-assisted method based on GC–MS was developed for metabolic profiling analysis. This method involves three phases: (1) The TICs of samples were analyzed to extract underlying chromatographic peaks; (2) peak-shift alignment based on ACPD results was performed; and (3) a peak area list table that could be directly used for subsequent statistical, such as through ANOVA coupled with partial least square-discriminate analysis (PLS-DA), was constructed to identify potentially valuable metabolites.

To demonstrate the performance of the proposed method, 30 plant samples, including 12 samples from different growth periods and 18 samples from various issues of mature tobacco plants, were analyzed by a GC–MS system. In each TIC obtained, more than 240 chromatographic peaks were detected, and a large number of peaks overlapped. Analysis of such datasets is a laborious and challenging endeavor for most analysts. The new method proposed in the present work allows automatic and efficient data processing with satisfactory classification results. Finally, a number of potential metabolites showing significant differences over three growth stages and plant tissues after maturation were extracted.

2. Experiment

2.1. Sample preparation

Thirty plant samples used for metabolic profiling analysis were collected from Zunyi, Guizhou Province, China. Of these samples, 18 were obtained from the lower, middle, and upper leaf surfaces of tobacco after maturation (six samples were collected per stage). Twelve other samples were collected at different growth stages with six samples per stage. All samples were rapidly frozen in liquid nitrogen and stored before analysis. The samples were then freeze dried over a 24 h period and ground to powder.

Metabolites were extracted using a solvent mixture of MeOH (JT Baker, USA), $CHCl_3$ (JT Baker, USA), and H_2O (5/2/2, v/v/v). About 20 mg of each sample was placed in a 5 mL tube, and then mixed with 2 mL of solvent. The mixture was sonicated for 40 min at room temperature, after which approximately 400 μL of supernatant was transferred to a 2.5 mL Eppendorf tube and dried by a gentle flow of helium gas. Derivatization was performed to reduce the influences of sugars as extensively as possible and improve the volatility of non-volatile compounds. Finally, 1 μL of supernatant was analyzed on an Agilent 7890-5975 GC–MS system. Details of the sample preparation for GC–MS are discussed in our previous work [54].

2.2. GC–MS analysis

The chromatographic column used in this work was a DB-5MS column (30 m × 0.25 mm, 0.25 μm). Analysis was carried out at 290 °C and the injector featured a 1:5 split ratio. Helium, as the carrier gas, was flowed through the system at a constant rate of 1.0 mL min^{-1}. The column temperature was kept constant for 4 min at 70 °C and then increased to 310 °C at a rate of 5 °C min^{-1}. This temperature was maintained for 10 min. The transfer line and ion source temperatures were set to 280 and 230 °C, respectively. Mass spectra were acquired in full scan monitoring mode within the scan range of 40–510. Finally, a TIC with over 10,000 elution channels was collected for each sample. The GC–MS datasets were acquired in netcdf format and converted by Matlab 2013b (Mathworks, USA).

3. Methodology

The proposed CAMMPA method comprised three stages: ACPD of TIC, a novel peak-shift alignment, and statistical analysis (Fig. 1). Each of these stages is discussed in this section in detail.

3.1. Data preprocessing

A denoising strategy was applied to the TICs prior to data analysis. Our investigation indicated that three-point moving window averaging is suitable for most cases.

3.2. ACPD

As the ACPD strategy was explained in our previous work [53], some basic details are provided in the present work. The TIC of a sample was analyzed to obtain chromatographic peaks that can be accurately quantified, i.e. chromatographic peaks with signal-to-noise ratios over 10. Signals that continuously increased or decreased more by over three elution channels were temporally recognized as evidence of analytes:

$$x_i < x_{i+1} < x_{i+2} < x_{i+3} < \dots \quad (1)$$

$$x_j > x_{j+1} > x_{j+2} > x_{j+3} > \dots \quad (2)$$

where x is the recorded signal and i and j represent the ith and jth elution channels, respectively. After eliminating these signals, the instrumental noise level can be immediately estimated by robust statistical analysis. A pseudo-peak elimination step was then used to separate artifacts from real peaks. Using a chromatographic peak clustering strategy, the peak baseline can be accurately estimated, and peak information, including retention time, peak area, elution time, and peak height, can be simultaneously obtained; these data are very helpful for subsequent peak-shift alignment analysis. Small peaks with average signal-to-noise ratios less than 10 were eliminated because these compounds cannot be accurately quantified.

3.3. Peak-shift alignment

A peak-shift alignment procedure was performed based on the TICs of the samples. Fig. 1 describes the peak alignment procedure in detail, and a brief discussion of peak-shift alignment is provided in Fig. 2. The reference sample must be determined prior to analysis. In general, the sample with the most number of chromatographic peaks were separated could be selected as the reference. In this work, the TIC of the 10th sample was selected as the reference.

The chromatographic peaks of the reference and test sample were aligned individually. When a chromatographic peak is selected, its maximum time-shift range should be pre-estimated. According to our experience, a time-shift ±0.5 min can be used for most cases in metabolic profiling analysis based on GC–MS. Fig. 2A shows the elution segment of t_2, which is a chromatographic peak that must be aligned to the reference sample. The corresponding elution segment in the reference sample, r_2, is also provided in Fig. 2A, where three chromatographic peaks are detected by ACPD.

A rough alignment procedure was used in our peak alignment method to align the time-shifts between the segment in the reference sample and that in the test sample. A peak-to-peak alignment strategy was used to improve data analysis efficiency. Three chromatographic peaks in the reference sample were separately aligned to each of the chromatographic peaks in the test sample, and the