Journal of Chromatography A, xxx (2018) xxx–xxx

Contents lists available at ScienceDirect

Journal of Chromatography A

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

Automatic untargeted metabolic profiling analysis coupled with Chemometrics for improving metabolite identification quality to enhance geographical origin discrimination capability

Lu Han[a], Yue-Ming Zhang[a], Jing-Jing Song[c], Mei-Juan Fan[d], Yong-Jie Yu[a,e,*], Ping-Ping Liu[d], Qing-Xia Zheng[d], Qian-Si Chen[d], Chang-Cai Bai[a], Tao Sun[a], Yuan-Bin She[b,**]

[a] College of Pharmacy, Ningxia Medical University, Yinchuan 750004, China
[b] College of Chemical Engineering, Zhejiang University of Technology, Hangzhou 310014, China
[c] Ningxia Institute of Cultural Relics and Archeology, Yinchuan 750001, China
[d] Zhengzhou Tobacco Research Institute of CNTC, Zhengzhou 450001, China
[e] Ningxia Engineering and Technology Research Center for Modernization of Hui Medicine, Ningxia Medical University, Yinchuan 750004, China

ARTICLE INFO

Article history:
Received 11 September 2017
Received in revised form 23 January 2018
Accepted 7 February 2018
Available online xxx

Keywords:
Untargeted metabolic profiling analysis
Peak resolution
Two-way encoding PLS model
Geographical origin discrimination
Flaxseed

ABSTRACT

Untargeted metabolic profiling analysis is employed to screen metabolites for specific purposes, such as geographical origin discrimination. However, the data analysis remains a challenging task. In this work, a new automatic untargeted metabolic profiling analysis coupled with a chemometric strategy was developed to improve the metabolite identification results and to enhance the geographical origin discrimination capability. Automatic untargeted metabolic profiling analysis with chemometrics (AuMPAC) was used to screen the total ion chromatographic (TIC) peaks that showed significant differences among the various geographical regions. Then, a chemometric peak resolution strategy is employed for the screened TIC peaks. The retrieved components were further analyzed using ANOVA, and those that showed significant differences were used to build a geographical origin discrimination model by using two-way encoding partial least squares. To demonstrate its performance, a geographical origin discrimination of flaxseed samples from six geographical regions in China was conducted, and 18 TIC peaks were screened. A total of 19 significant different metabolites were obtained after the peak resolution. The accuracy of the geographical origin discrimination was up to 98%. A comparison of the AuMPAC, AMDIS, and XCMS indicated that AuMPACobtained the best geographical origin discrimination results. In conclusion, AuMPAC provided another method for data analysis.

1. Introduction

Untargeted metabolic profiling analysis has been greatly utilized in biomarker discovery and geographical origin discrimination [1–4]. Its successful application largely depends not only on the experimental design but also on the data analysis procedure. Data analysis continues to be a great challenging task to date. Although several well-known methods such as XCMS [5,6], ADAP [7–9], AMIDS, and Agilent MassHunter have been developed for routine analysis, high-quality methods should be developed to address the challenges in peak extraction and time shift correction, such as false-positive and false-negative peak detection results and incorrectly aligned peaks [10–12].

The performances of current methods chiefly depend on both the peak detection and the instrumental noise estimation strategies. Most of the advanced methods recommend users to optimize the calculation parameters by using a manual verification strategy. However, the optimized parameters based on several samples are not always applicable to the others. For example, the instrumental noise can change not only across samples but also across m/z channels [13]. A more applicable strategy is intelligently estimating the calculation parameters in an adaptive manner. Most of the current methods employ a wavelet-based peak extraction strategy, which extracts peak positions based on local maximum values by using a

* Corresponding author at: College of Pharmacy, Ningxia Medical University, Yinchuan 750004, China.
** Corresponding author at: Zhejiang University of Technology, Hangzhou 310014, China.
E-mail addresses: yongjie.yu@163.com (Y.-J. Yu), sheyb@zjut.edu.cn (Y.-B. She).

https://doi.org/10.1016/j.chroma.2018.02.017
0021-9673/

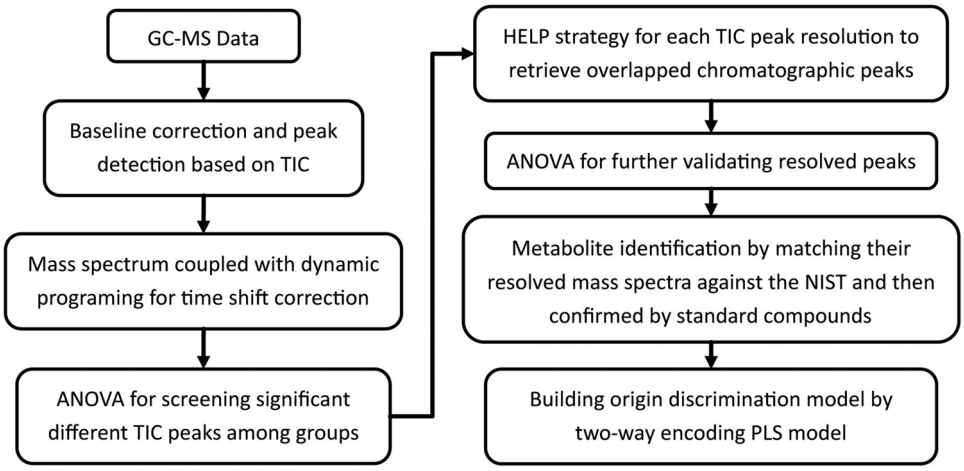

Fig. 1. Workflow of the AuMPAC for accurately screening metabolites with significant differences among the geographical regions and for improving the geographical origin discrimination capability.

Mexican hat wavelet coupled with a ridgeline extraction strategy. However, the wavelet-based peak extraction have several limitations: (i) the number of local maximum values may increase with the increment of wavelet scales and result in a number of false-positive peaks; (ii) overlapped peaks may be mistaken as a single peak and some important metabolites may be lost; and (iii) automatically determining the peak elution ranges is difficult because of a lack of baseline drift correction. Overlapped chromatographic peak is another challenging problem in untargeted metabolic profiling analysis. Only a few untargeted metabolic profiling analysis methods can deal with overlapped peaks [8,9,12]. Unfortunately, false-positive and false-negative results confuse the users in identifying biomarkers.

A primary objective of untargeted metabolic profiling analysis beyond biomarker discovery is to perform specific tasks, such as disease pathogenesis identification and geographical origin discrimination [2]. However, most scientists focus on developing new algorithms that can extract more peaks instead improving the developed algorithms, such as enhancing the geographical origin discrimination capability based on the extracted peaks [14–16].

To address these limitations and to provide a high-quality data analysis method, this study proposed an automatic untargeted metabolic profiling analysis methodology [17] that extracts chromatographic peaks in the total ion chromatogram (TIC) through the Gaussian smoothing strategy coupled with adaptive instrumental estimation and aligns the chromatographic peaks by dynamic programing [18] in combination with mass spectral information. Unfortunately, this method cannot deal with overlapped peaks and thus may lose some important compounds. The present work went a step further by introducing a chemometric overlapped peak resolution strategy into the automatic untargeted metabolic profiling analysis (AuMPAC) to improve metabolite identification quality and to enhance geographical origin discrimination capability. The method was employed to screen markers of flaxseeds from the various growing regions in China for geographical origin discrimination.

Flaxseed or linseed (*Linumusitatissimum*) has been used as food and fiber worldwide. It is rich in fatty acids and has the highest content of linolenic acid among all sources [19–23]. In the northwest provinces of China, such as Ningxia, Gansu, Shananxi, and Inner Mongolia, flaxseed oil occupies a large percentage of the local edible oil market because the flaxseed crop is one of the largest cash crops in these zones. In recent years, consumption of flaxseed and its oil markedly increased in China because of both increased public health consciousness and increased awareness of its rich fatty acid content. The demand for high-quality flaxseed is high in the edible oil industry, and the quality of flaxseed depends on its chemical constituents, which are greatly influenced by genotype, cultural techniques, and especially the climatic elements in the geographical location. Consequently, sales prices of flaxseeds vary depending on the growing zones. The flaxseed from Inner Mongolia is the most expensive, whereas those from Shanxi Province are the cheapest. The price differences are a potential opportunity for flaxseed fraud to increase profit. Unfortunately, methods for the authentication of flaxseed origin in China are not available. To address this circumstance, the present work employed the developed method for the geographical origin prediction of flaxseed samples from distinct growing zones in China.

Given that fatty acid distribution is a major factor in evaluating the value of flaxseed in the edible oil industry, a GC–MS was designed in combination with AuMPAC to accurately identify metabolites that are correlated with geographical origin discrimination. The developed method was compared with AMIDS and XCMS to demonstrate its performance.

2. Methodology

2.1. AuMPAC

The procedure for the data analysis in AuMPAC is shown in Fig. 1. The analysis consisted of two parts: (i) untargeted metabolic profiling analysis procedure to screen TIC peaks that showed significant differences among groups (left column), and (ii) chemometric strategy to resolve the screened peaks and build the geographical origin prediction model (right column). This section provides a brief overview of the involved algorithms.

2.1.1. Screening significantly different TIC peaks

Significantly different peaks were automatically screened using AuMPAC by employing the following substeps, i.e., peak extraction, time shift correction, and statistical analysis.

Peak extraction. Fig. S-1 provides a brief illustration of the peak extraction. First, baseline correction was performed through a local minimum value-based strategy [13]. Then, a multiscale Gaussian smoothing strategy was utilized for peak detection. The ideology behind the peak detection was that chromatographic peaks should be local maximum values and become more evident after smoothing. Peak extraction was performed as follows: (i) the chromatographic signal was smoothed by using a series of Gaussian smoothing functions and simultaneously increasing the smoothing

Please cite this article in press as: L. Han, et al., Automatic untargeted metabolic profiling analysis coupled with Chemometrics for improving metabolite identification quality to enhance geographical origin discrimination capability, J. Chromatogr. A (2018), https://doi.org/10.1016/j.chroma.2018.02.017