Journal of Chromatography A, 1452 (2016) 1–9

Contents lists available at ScienceDirect

Journal of Chromatography A

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

A simple multi-scale Gaussian smoothing-based strategy for automatic chromatographic peak extraction

Hai-Yan Fu[a,*], Jun-Wei Guo[b], Yong-Jie Yu[b,c,d,**], He-Dong Li[a], Hua-Peng Cui[b], Ping-Ping Liu[b], Bing Wang[b], Sheng Wang[b], Peng Lu[b,*]

[a] School of Pharmaceutical Sciences, South-Central University for Nationalities, Wuhan 430074, China
[b] Zhengzhou Tobacco Research Institute of CNTC, Zhengzhou 450001, China
[c] School of Pharmacy, Ningxia Medical University, Yinchuan 750004, China
[d] Key Laboratory of Hui Medicine Modernization, Ministry of Education, Yinchuan 750004, China

ARTICLE INFO

Article history:
Received 2 March 2016
Received in revised form 3 May 2016
Accepted 4 May 2016
Available online 6 May 2016

Keywords:
Chromatographic peak detection
Multi-scale Gaussian smoothing
Complex sample analysis
Quality control
Metabolic profiling

ABSTRACT

Peak detection is a critical step in chromatographic data analysis. In the present work, we developed a multi-scale Gaussian smoothing-based strategy for accurate peak extraction. The strategy consisted of three stages: background drift correction, peak detection, and peak filtration. Background drift correction was implemented using a moving window strategy. The new peak detection method is a variant of the system used by the well-known MassSpecWavelet, i.e., chromatographic peaks are found at local maximum values under various smoothing window scales. Therefore, peaks can be detected through the ridge lines of maximum values under these window scales, and signals that are monotonously increased/decreased around the peak position could be treated as part of the peak. Instrumental noise was estimated after peak elimination, and a peak filtration strategy was performed to remove peaks with signal-to-noise ratios smaller than 3. The performance of our method was evaluated using two complex datasets. These datasets include essential oil samples for quality control obtained from gas chromatography and tobacco plant samples for metabolic profiling analysis obtained from gas chromatography coupled with mass spectrometry. Results confirmed the reasonability of the developed method.

1. Introduction

Data analysis is an important endeavor in many scientific fields, including metabolomics and proteomics, because high-throughput data can be generated for a single sample with the aid of advanced analytical instruments [1–5]. Chromatographic instruments, such as gas chromatography, liquid chromatography, gas/liquid chromatography hyphenated with mass spectrometry like high-resolution time-of-flight spectrometry, have been extensively used to analyze complex samples in metabolic research. Compared with the rapid development of modern chromatographic instruments, data analysis has not been paid much attention by scientists until recently. Analysis for metabolic data mainly involves data mining and multivariate statistical analysis. The methods used for the latter include principal component analysis and partial least-squares, both of which have been comprehensively studied by analysts. By contrast, data mining methods, including chromatographic background drift correction, chromatographic peak detection, and peak alignment, have not been developed to levels able to satisfy practical applications [6].

Chromatographic peak detection is the preliminary and critical stage for subsequent downstream data analysis. In metabolic studies, the recorded chromatographic signal usually contains hundreds of chromatographic peaks, together with complicated useless information, such as background drift and instrumental noise. Thus, peak detection becomes a significantly challenging task. False-positive and false-negative peak detection could lead to unsatisfactory results because false-positive peaks may indicate the loss of significant compounds while false-negative peaks may bring additional complexity to the analysis task. Unfortunately, manual verification of peak detection results by visualization is impractical because this procedure could be highly time consuming. Thus, high-performance automatic peak detection methods are urgently needed for complex sample analysis.

* Corresponding authors.
** ** Corresponding author at: School of Pharmaceutical Sciences, South-Central University for Nationalities, Wuhan 430074, China; Zhengzhou Tobacco Research Institute of CNTC, Zhengzhou 450001, China; School of Pharmacy, Ningxia Medical University, Yinchuan 750004, China.
E-mail addresses: fuhaiyan@mail.scuec.edu.cn (H.-Y. Fu), yongjie.yu@163.com (Y.-J. Yu), penglu2004@hotmail.com (P. Lu).

http://dx.doi.org/10.1016/j.chroma.2016.05.018
0021-9673/

A number of methods have been developed for chromatographic detection [6–27]. These methods can be roughly classified into two categories: methods based on the fundamental physical characteristics of chromatographic signals and methods built upon wavelet analysis. The first category is based on the signal-to-noise ratio (s/n) and usually employs pre-defined peak models or first and/or second-order derivatives of the analyzed signal to detect meaningful chromatographic peaks. Such methods have been extensively employed by commercial instruments, such as Agilent MassHunter, Thermo Fisher Xcalibur, and AB SCIEX Analyst. Peak model-based methods, such as XCMS [14], are valuable for obtaining chromatographic peaks that can be mathematically modeled by Gaussian curves, wherein peak information, including peak width and area, can be obtained accurately. However, in extreme cases of complex sample analysis, the shapes of the chromatographic peaks may vary according to the physical and chemical characteristics of the corresponding compounds and therefore cannot be satisfactorily modeled. Methods based on first and second-order derivatives of the analyzed chromatographic signal are sensitive to pre-estimated noise levels. A high noise level may filter some useful peaks, while a low noise level can introduce a large amount of peaks belonging to instrumental noise. Although some peak detection parameters, such as the threshold value of first-order derivatives and noise level, can be optimized manually, fluctuation of instrumental noise across samples may require an additional optimization procedure. The present authors recently developed an automatic peak detection method by using a robust statistical instrumental noise estimation strategy [2,6]. This method assumes that a chromatographic peak should be increased or decreased monotonously with very few oscillating points, which is unsuitable for noisy chromatographic signals that are recorded under mass spectral channels.

Peak detection methods based on wavelet analysis are gradually becoming a new research hot spot for studies on chromatographic and mass spectral analysis. A significant advantage of these methods lies in the fact that they are insensitive to noisy chromatographic data because wavelet analysis is commonly employed for denoising. Examples of freely available software [28], including MassSpecWavelet [19] and Mzmine [29], adopt a wavelet based peak detection strategy to analyze metabolic and proteomic datasets. The successful applications of these methods have been demonstrated elsewhere. In practical applications, however, wavelet analysis-based peak detection strategies remain problematic. Determination of the peak start and end positions of the analyzed chromatographic signal is an important task. Although wavelet scales corresponding to the maximum wavelet coefficients can be used to estimate the peak width, the chromatographic shapes obtained may be inconsistent with that of the selected mother wavelet and the peak widths may be inadequately estimated. Additionally, wavelet-based methods are sensitive to overlapping peaks and may yield peak positions located far from the maximum point. These phenomena may be attributed to the limitations of the ridge line search algorithm. Aiming to address these limitations, researchers have resort to new methodologies, such as employing the local minimum and zero-crossing wavelet coefficients [8].

In fact, wavelet analysis-based peak detection can be treated as a smoothing strategy wherein the scale of the wavelet is equal to the width of the smoothing window. However, the incremental step of scale in wavelet analysis may be too large for chromatographic peak detection, causing two overlapping chromatographic peaks under the current wavelet scale to combine into a single peak in the successive scale and increasing the difficult of locating the maximum positions of the peaks with accuracy. In the present work, we employ the core ideology of wavelet analysis to develop a novel peak detection strategy. In this strategy, peaks appear as local maximum values under various smoothing window widths, a simple multi-scale Gaussian smoothing-based approach is developed for chromatographic peak extraction, and accurate estimation of the maximum positions of chromatographic peaks, together with reasonable peak start and end positions, is achieved. We employ our proposed method to analyze a chromatographic fingerprint dataset used to monitor quality changes in essential oils and then evaluate our method against an extremely complicated GC–MS metabolite dataset of tobacco plant samples. Results from MassSpecWavelet and XCMS are also provided for comparison.

2. Methodology

2.1. Multi-scale Gaussian smoothing for peak extraction (MGSPE)

There are three stages involved in our multi-scale smoothing peak extraction method: background drift correction, preliminary peak detection, and peak filtration.

2.1.1. Background drift correction

Background drift is an important data treatment stage that greatly influences the following peak detection. According to Filgueira [10], background correction should be properly addressed prior to any data analysis. Although peak positions can be accurately located through multi-scale Gaussian smoothing, the peak start and end positions could be misestimated in the presence of background drift. In this work, a recently developed automatic background correction method by Yaroshchyk et al. [30], which was designed to remove the baseline in NMR spectra, was introduced for chromatographic data analysis for the first time. This method extracts a series of minimum values by moving window strategy with a size of W and adopts a normalized boxcar smoothing function to estimate the background drift. In this equation,

$$b(j) = \Sigma_{j=i-W/2}^{j=i+W/2} Min(j) \times rect(i-j) \quad (1)$$

where b(j) is the estimated baseline, Min(j) reflects W-points moving minima, and $rect(i–j)$ is a rectangular with a constant value of 1/W for the elements in the window and zeros outside. We assume that the estimated background drift is never larger than the original signals, and a window size of 60 points was employed. Finally, a chromatogram with background correction can be obtained. More detailed information can be found in the Ref. [30].

2.1.2. Preliminary peak detection

Unlike the traditional peak detection strategy that employs the threshold value of instrumental noise coupled with first-order derivative to estimate the peak start and end positions, we emphasize the detection of local maximum points corresponding to underlying peaks. A data point, c_i, represents a local maximum as long as the following equation is satisfied:

$$c_{i-1} \langle c_i and c_i \rangle c_{i+1} \quad (2)$$

where c_{i-1} and c_{i+1} represent the $i-1$th and $i+1$th data points, respectively. A large number of data points belonging to instrumental noise may also meet the above equation. However, if a local maximum value is indeed the position of a meaningful peak, it should also be a local maximum point when a data smoothing strategy with a large smoothing window size is employed. By contrast, the local maximum points of instrumental noise will disappear when a large smoothing window is employed. Therefore, if the smoothing window width is gradually increased and the positions of local maximum points are obtained, the ridge lines of instrumental noise peaks will only exist in narrow smoothing widows and significant peaks will appear in longer series of window widths.

A brief illustration of the proposed peak detection strategy is shown in Fig. 1. Fig. 1A provides part of a chromatogram