Journal of Chromatography A, 1223 (2012) 93–106

Contents lists available at SciVerse ScienceDirect

Journal of Chromatography A

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

Multiscale peak alignment for chromatographic datasets

Zhi-Min Zhang, Yi-Zeng Liang*, Hong-Mei Lu, Bin-Bin Tan, Xiao-Na Xu, Miguel Ferro

College of Chemistry and Chemical Engineering, Research Center of Modernization of Chinese Medicines, Central South University, Changsha 410083, China

ARTICLE INFO

Article history:
Received 23 October 2011
Received in revised form
10 December 2011
Accepted 12 December 2011
Available online 22 December 2011

Keywords:
Chromatography
Peak alignment
Fast Fourier transform
Cross correlation
Shannon information content

ABSTRACT

Chromatography has been extensively applied in many fields, such as metabolomics and quality control of herbal medicines. Preprocessing, especially peak alignment, is a time-consuming task prior to the extraction of useful information from the datasets by chemometrics and statistics. To accurately and rapidly align shift peaks among one-dimensional chromatograms, multiscale peak alignment (MSPA) is presented in this research. Peaks of each chromatogram were detected based on continuous wavelet transform (CWT) and aligned against a reference chromatogram from large to small scale gradually, and the aligning procedure is accelerated by fast Fourier transform cross correlation. The presented method was compared with two widely used alignment methods on chromatographic dataset, which demonstrates that MSPA can preserve the shapes of peaks and has an excellent speed during alignment. Furthermore, MSPA method is robust and not sensitive to noise and baseline. MSPA was implemented and is available at http://code.google.com/p/mspa.

1. Introduction

Chromatography with various detectors can provide quantification and identification information of complex systems at an unprecedented level [1], which has been extensively applied to metabolomics [2,3], quality control of herbal medicines [4,5] and other fields. For example, gas chromatography (GC) technique can detect, identify and quantify volatile compounds in metabolites and herb medicines' extractions, and liquid chromatography (LC) technique with electrospray ionization (ESI) can detect and quantify nonvolatile compounds complementary to GC [6]. However, both metabolomics and quality control of herbal medicines involve massive experiments and dataset collection, and the datasets usually are generated through experiments performed on different samples. In order to capture differences among samples caused by their composition, the key point of an experiment is to limit experimental variability as much as possible. However, deviations from normal conditions may appear, causing peak shifts observed among signals. For this reason, the acquired datasets are often too complex to easily extract meaningful information. Recently, great efforts have been made by chemometricians to provide researchers in quality control of herbal medicines with chemometrics and chemometrical toolbox to cope, analyze and interpret these complex datasets [4,7]. In order to evaluate the fingerprints of herbal products, several novel chemometric methods have been developed, such as the methods based on information theory [8], pretreatments [9,10], alignment [11,12], spectral correlative chromatogram [13] and multivariate resolution [14,15]. In metabolomics, the usages of more and more variables to characterize samples have driven researchers from traditional statistics to chemometric methods such as principal component analysis (PCA) [16], partial least squares (PLS) [17] and their derivatives [18–20], since they are more efficient and capable of handling collinear datasets.

Chromatograms consist of peaks corresponding to components of the mixtures, and ideally peaks of the same component of different samples should have an equal retention time. But in real analysis, the dataset does not conform to this hypothesis due to retention time shifts between samples. Since the bilinear factor models are the basic requirement of foundational chemometric algorithms such as PCA and PLS [21], peak alignment is necessary to reduce the variation in peak positions, which can improve useful information extraction using chemometrics and statistics. Glancing at literatures, dozens of methods have been proposed to align shifts in peak positions among spectra of different samples when analytical instruments, such as chromatography, nuclear magnetic resonance (NMR) and mass spectrometry, are used. Generally, they can be divided into two major categories: synchronize entire signals and handle only the detected peaks.

Alignment methods that synchronize entire signals usually divide signals into segments, warping these segments by interpolation or transformation to maximize correlation between signal to be aligned and reference. The concept of time warp was initially introduced to align retention time shift of chromatograms by Wang and Isenhour [22]. By then in 1998, two practical alignment methods were introduced, dynamic time warping (DTW) [23],

* Corresponding author. Tel.: +86 731 88830824; fax: +86 731 88822841.
E-mail address: yizeng_liang@263.net (Y.-Z. Liang).

0021-9673/$ – see front matter © 2011 Published by Elsevier B.V.
doi:10.1016/j.chroma.2011.12.047

applied to the analysis and monitoring of batch processes and correlation optimized warping (COW) [24], proposed by Nielsen for chromatograms. Both DTW and COW utilize dynamic programming to search all solutions with respect to all possible combinations of parameters, and they have been demonstrated to be effective on chromatograms at that moment. But currently, chromatogram often contains several thousands of data points, original COW is not suitable for these signals due to large requirements in both execution time and memory, and DTW often "over-warps" signals and introduces artifacts into the aligned profiles when signals were only recorded using a mono-channel detector [25]. Therefore, many heuristic optimization methods, parametric model and fast correlation algorithms have been applied to accelerate this time-consuming procedure and improve the aligning result. In order to improve the computational cost and optimize memory usage of DTW, some global constraints were introduced by Sakoe and Chiba [26]. Stan [27] introduced FastDTW, an approximation of DTW that has a linear time and space complexity. Genetic algorithm [28] and beam search [29] were adopted to align large signals in acceptable time, but it is difficult to optimize the segment size. Eilers proposed a parametric model for the warping function, and presented parametric time warping (PTW) [30], which is fast, stable and consumes little memory. Pravdova [31] and van Nederkassel [32] compared DTW, COW and PTW for chromatogram alignment. Wong [33,34] applied fast Fourier transform (FFT) cross correlation to estimate shift between segments, which is amazingly fast and has solved computational inefficiency problems of alignment. However, both peak alignment by FFT (PAFFT) and recursive alignment by FFT (RAFFT) move segments by insertion and deletion of data points at the start and end of segments without considering peak information, which may change the shapes of peaks by introducing artifacts and removing peak points [35]. Based on RAFFT and PAFFT, recursive segment-wise peak alignment (RSPA) [36] was proposed by Veselkov to improve the accuracy of alignment using peak position information for recursive segmentation and interval correlation shift (icoshift) [35,37]. This method can reduce the artifacts by inserting missing values instead of repeating the value on boundary. Variable penalty dynamic time warping was proposed by Clifford [25] to overcome DTW's "over-warps" shortcomings. Recently, Daszykowski [38] proposed an automatic peak alignment method by explicitly modeling the warping function for chromatographic fingerprints.

Among others, fuzzy warping and reduced set mapping often convert signals into peaks' lists, which can speed up alignment by reducing the dimensions of problems dramatically [39–44]. But they align major peaks at the expense of minor peaks, which are harder to detect. Besides, they are prone to misalignment in special peak regions, such as peaks with shoulder, overlapping peaks and peak dense region.

There are also many mature and competing alignment algorithms or toolbox including alignment algorithms in metabolomics and bioinformatics. MSFACTs [45] can automatically import, reformat and align large chromatographic datasets. MZmine [46] was proposed and implemented by Katajamaa, which contains methods for all data processing stages including spectral filtering, peak detection, alignment and normalization of LC/MS data in proteomics and metabolomics. XCMS [6], XCMS^2 [47] and metaXCMS [48] were developed by Scripps Center for Metabolomics, providing the researchers with a series of tools for preprocessing, analyzing, and visualizing datasets from hyphenated instruments. MetAlign [49] can preprocess and align a broad range of accurate mass and nominal mass datasets. MetaboAnalyst [50,51] provides an integrated web-based platform for data processing, data normalization, statistical analysis and high-level functional interpretation of metabolomics dataset.

In this paper, MSPA method is proposed. MSPA can rapidly align sample signal toward a reference without altering the peaks' shapes. By transforming the chromatogram into the wavelet space using CWT with Haar wavelet as the mother wavelet, peaks can be accurately and robustly detected. Subsequently, we can calculate Shannon information content for each detected peak, and pick out peak of each segment with the smallest Shannon information content value to iteratively divide chromatogram or each segment into smaller segments. Then candidate shifts of each segment can be rapidly found by FFT cross correlation. The optimal shift for each segment can be determined by combining candidate shifts of adjacent segment to maximize the correlation coefficient. Finally, we move the segments via linear interpolation of non-peak parts. This iterative procedure will stop when all the segments are well aligned. One can see that MSPA gradually aligns peaks from small to large scale, which is the reason why the proposed method is named as multiscale peak alignment (MSPA).

This paper is organized as follows. First of all, relevant principles to MSPA are described and dissected in Section 2, including peak detection, width estimation, Shannon information content, FFT cross correlation, candidate shift estimation, optimal shift determination by combining candidate shifts of adjacent segments and segments move via linear interpolation of non-peak parts. Then details of simulated signal and experiments of real chromatograms are introduced, and alignment results will be presented together with discussions about MSPA method. Finally, some conclusions and perspectives are given in Section 5.

2. Theory and implementation

The heart of MSPA is the usages of local maximums in FFT cross correlation as candidate shifts, which can guarantee accuracy and alignment speed. Additionally, it also includes several techniques for peak detection, width estimation, iterative segmentation and optimal shift determination. Fig. 1 describes architecture and overview of MSPA method. The techniques used in MSPA will be explained as thoroughly and clearly as possible in the next sections.

2.1. Peak detection and width estimation

Peak detection and width estimation are universal problems in instrument signal analysis, and various criteria have been proposed such as signal to noise ratio (SNR), intensity threshold, slopes of peaks, local maximum, shape ratio, ridge lines, and model-based criterion [52]. In this study, a derivative calculation method via CWT [53] was used for peak detection and width estimation, and SNR to eliminate false positive peak.

In order to detect peak position and estimate its start and end points, derivative calculation is often applied. However, the simplest numerical differentiation is not very effective for real signal due to the noise increasing drawback, so derivative calculation via Haar CWT was adopted to improve SNR during the calculation. Wavelet transform is one of the most powerful tools in signal analysis [54,55]. Wavelet is a series of functions $\psi_{a,b}(t)$, which are derived from $\psi(t)$ by scaling and shifting, according to the equation:

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right); \quad a \in R^+,\ b \in R \tag{1}$$

where a is the scale parameter to control scaling, b the shift parameter to control shifting, and $\psi(t)$ is the mother wavelet.

Wavelet transform is defined as the projection of signal onto the wavelet function ψ. Mathematically, this process can be represented as:

$$\mathbf{C}(a,b) = \langle s(t), \psi_{a,b}(t)\rangle = \int_{-\infty}^{+\infty} s(t)\psi_{a,b}(t)\,dt \tag{2}$$