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 $\psi$. Mathematically, this process can be represented as:

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

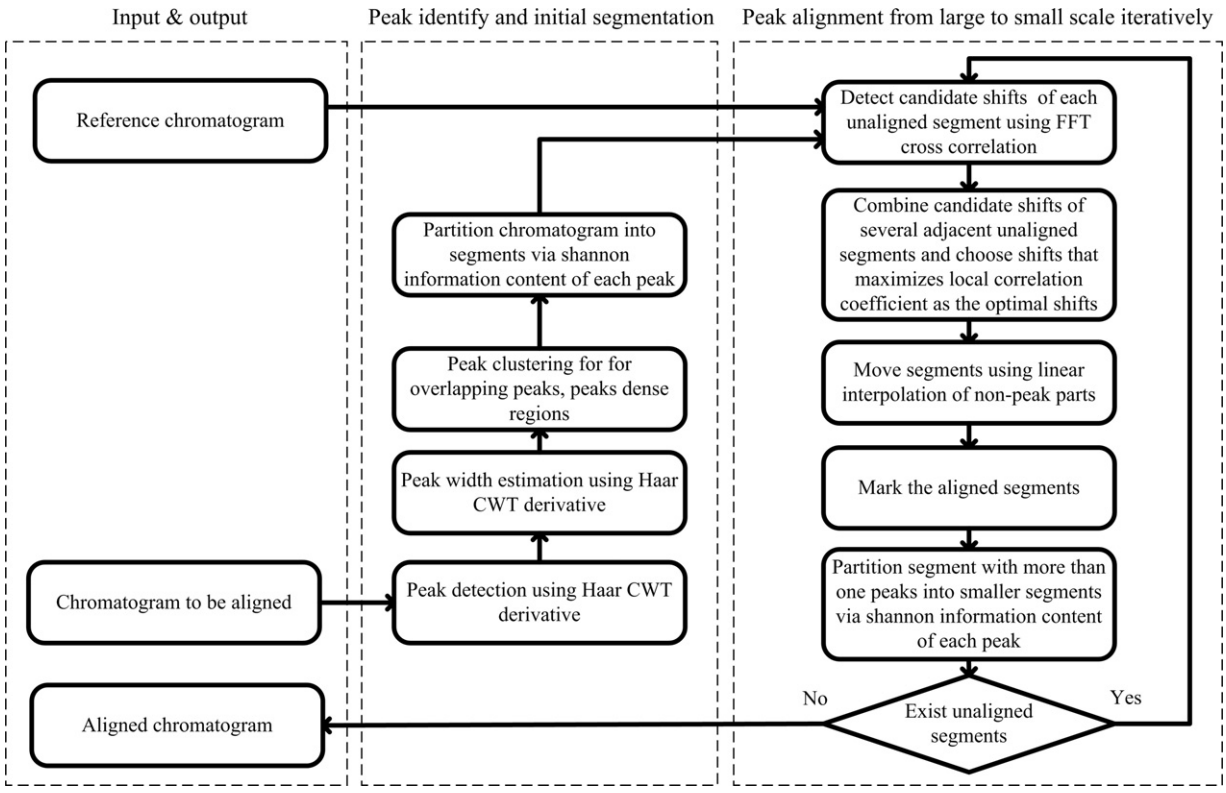


**Fig. 1.** Flow chart for describing the architecture and overview of MSPA method.

here $s(t)$ is the signal, and $\mathbf{C}$ is a 2D matrix of wavelet coefficients.

The approximate $n$th derivative of an analytical signal can be obtained by applying Haar CWT $n$ times to the signal. Haar wavelet is the simplest wavelet function among all the wavelet functions, which can be defined as:

$$\psi(t) = \begin{cases} 1 & 0 \le t < 1/2 \\ -1 & 1/2 \le t < 1 \\ 0 & \text{other} \end{cases} \tag{3}$$

Peak can be defined as a local maximum of $N$ neighboring points, whose intensity is significantly larger than the noise level. The local maximums can be found from the derivative calculation via Haar CWT of the signal. Then, false positive peaks are eliminated from whose SNR is lower than a pre-specified threshold. Similar to finding of the peak position, the start and end points of each peak can also be found by this derivative calculation method. For each peak, its start and end points can be obtained by searching the nearest point from its peak position in the vector of detected peak position and peak width. Principles for peak detection and width estimation are concisely illustrated in Fig. 2. The middle part of the figure depicts derivative calculation via Haar CWT. The top and bottom parts of this figure describe peak detection and width estimation, respectively.

### 2.2. Segmentation based on Shannon information content

There exist different scale peaks in signal. When correlation based alignment methods are used, it is intuitive that alignment of large scale peak is easier than small scale peak. One could also say that the difficulty level of aligning the specific scale peak has direct relationship with the uncertainty of this peak in the entire signal profile. In information theory [56], Shannon information content is a good measurement of uncertainty. Consequently, Shannon information content of peaks can be regarded as a good measurement of the difficulty level of specific scale peak. The Shannon information content is defined to be:

$$\mathbf{h}_i = -\log_2 \mathbf{p}_i \tag{4}$$

where $\mathbf{p}_i$ is the probability of distribution function and $\mathbf{h}_i$ is Shannon information content.

Some reasonable modifications [8] should be employed on the equation above to calculate Shannon information content of peaks in the chromatogram. Firstly, the signal is normalized with its overall peak area equal to one and then its information content is calculated based on:

$$\mathbf{h}_i = -\log_2 \frac{\mathbf{p}_i}{\sum \mathbf{p}_i} \tag{5}$$

where $\mathbf{p}_i$ is the area of the $i$th peak or peak cluster of the real chromatogram, and $\sum \mathbf{p}_i$ is the overall peak area of the real chromatogram.

A small Shannon information content of a peak means a small uncertainty and large scale, and vice versa. It is intuitive that the large scale peak with small uncertainty should have priority over small scale peaks during alignment, so peaks were aligned in MSPA against the reference chromatogram from larger to smaller scale gradually. The iterative segmentation scheme is illustrated in Fig. 3.

### 2.3. Candidate shift detection via FFT cross correlation

Cross-correlation is a standard method to measure the similarity and linear shift of two signals as one is time-lag to the other, which involves shifting one signal and calculating the correlation coefficient between the shifted one and the other. For two continuous functions, $r$ and $s$, the cross-correlation of them at lag $j$ is defined as:

$$c(j) = \int_{-\infty}^{+\infty} r(x)s(x+j)\,dx \tag{6}$$

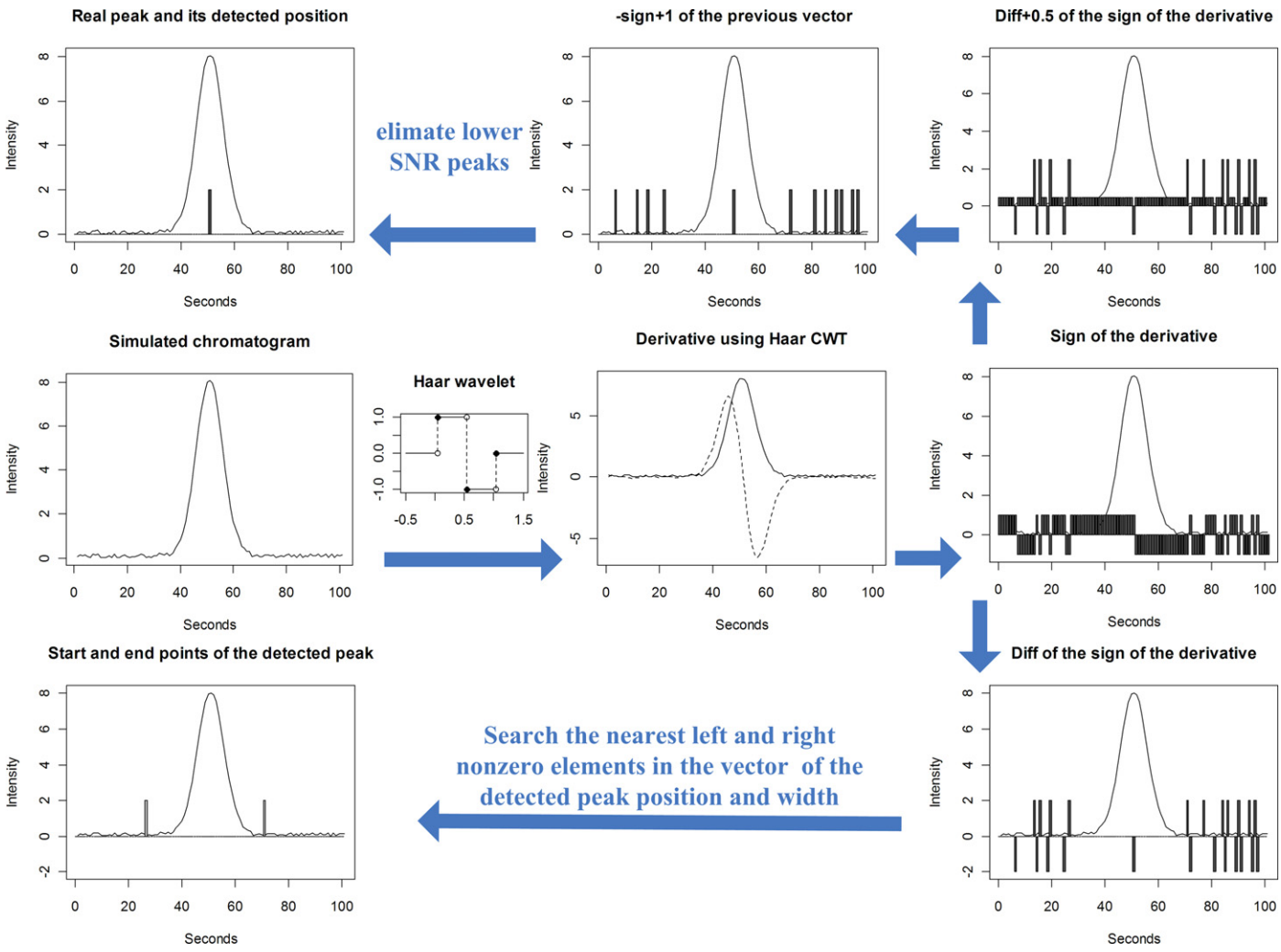


**Fig. 2.** Principles for peak detection and width estimation using derivative calculation based on continuous wavelet transform with Haar wavelet as the mother wavelet.

Similarly, for discrete signal such as chromatograms, the cross-correlation is defined as:

$$\mathbf{c}(j) = \frac{\sum_i (\mathbf{r}(i) - \bar{r})(\mathbf{s}(i+j) - \bar{s})}{\sqrt{\sum_i (\mathbf{r}(i) - \bar{r})^2} \sqrt{\sum_i (\mathbf{s}(i+j) - \bar{s})^2}} \tag{7}$$

where $\mathbf{r}$ is the reference signal, $\mathbf{s}$ the signal to be aligned, $\mathbf{c}$ the cross-correlation values for all lags.

The direct evaluation of cross correlation requires $O(N^2)$ time complexity for a chromatogram of length $N$, which is time-consuming for chromatograms with several thousands of data points. Fortunately, cross correlation can be calculated via FFT to achieve a much better performance, which can dramatically reduce time complexity of cross correlation from $O(N^2)$ to $O(N \log N)$. FFT computes discrete Fourier transform (DFT) and produces the same result as DFT. In order to clarify how to calculate cross correlation via FFT, a brief introduction about DFT is required. The forward and reverse DFT is defined by the formulas

$$\begin{aligned} X_k &= \sum_{n=0}^{N-1} x_n e^{-(2\pi i/N)kn}; \quad k = 0, \ldots, N-1 \\ x_n &= \frac{1}{N} \sum_{n=0}^{N-1} X_k e^{(2\pi i/N)kn}; \quad k = 0, \ldots, N-1 \end{aligned} \tag{8}$$

where $X$ is the discrete Fourier transformed data in the wavelength domain. DFT and inverse DFT are often denoted by $\mathbf{X} = \mathcal{F}\{\mathbf{x}\}$ and $\mathbf{x} = \mathcal{F}^{-1}\{\mathbf{X}\}$ respectively.

If $\mathbf{R}$ and $\mathbf{S}$ are DFTs of $\mathbf{r}$ and $\mathbf{s}$ then circular convolution theorem and cross-correlation theorem [57] for DFT state:

$$\mathbf{c} = \mathcal{F}^{-1}\{\mathbf{R} \cdot \mathbf{S}^*\} \tag{9}$$

here $\mathbf{c}$ is cross correlation between $\mathbf{r}$ and $\mathbf{s}$, and $\mathbf{S}^*$ is the complex conjugate of $\mathbf{S}$.

FFT cross correlation can only estimate linear shift between signals, but retention time shifts are often nonlinear for chromatogram of real sample. In MSPA method, chromatogram to be aligned will be iteratively divided into small segments and FFT cross correlation will be used to estimate candidate shifts for each segment and align peaks from large scale to small scale gradually. This strategy can solve the alignment of nonlinear retention time shifting problem by FFT cross correlation.

Previous alignment methods based on FFT cross correlation only use the maximum of cross correlation as the optimal shift. But the maximum of cross correlation of small segment as the optimal shift may sometimes be the optimal shift locally but not optimal at larger scale or globally. Therefore, all the local maximums of FFT cross correlation should be detected as the candidate shifts via CWT derivative calculation. Then the optimal shift is found by combining

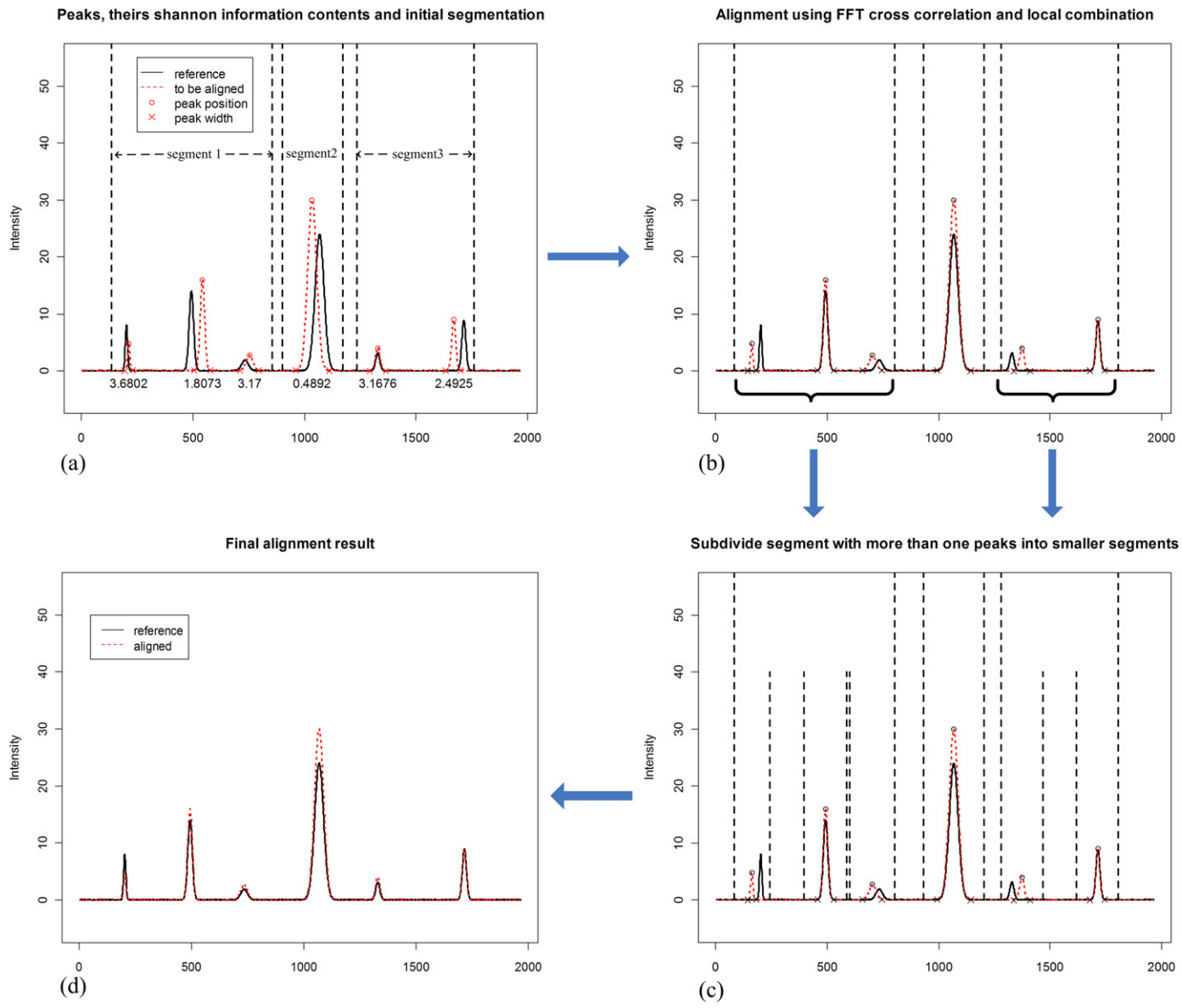


**Fig. 3.** Scheme of iterative segmentation of chromatogram based on Shannon information content.

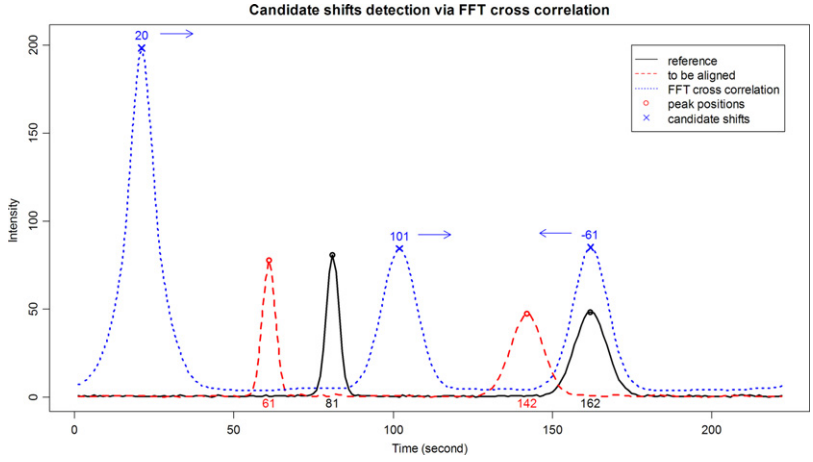


**Fig. 4.** Candidate shift detection of simulated chromatograms by finding the local maximums in FFT cross-correlation.

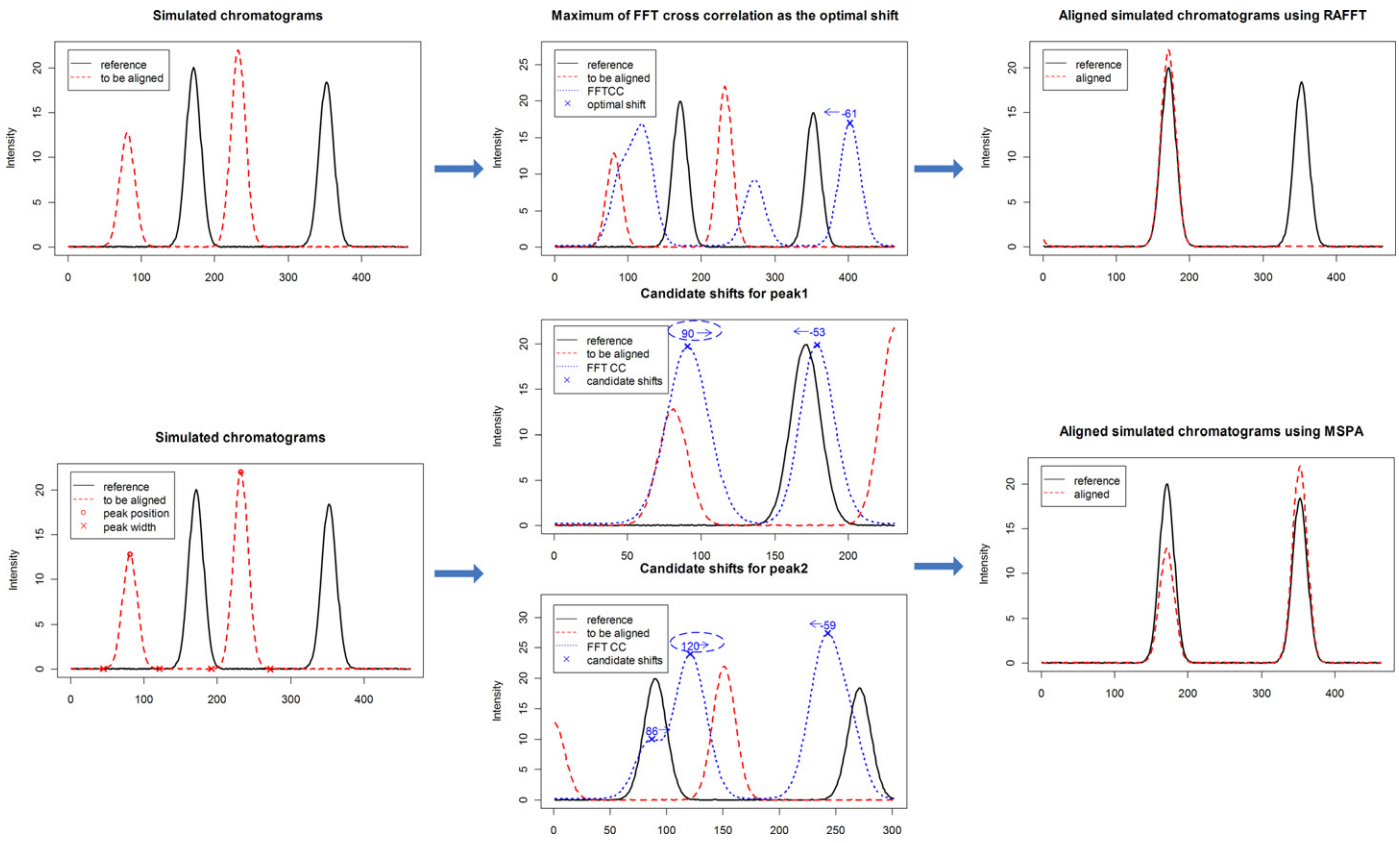


**Fig. 5.** Example shows the collision in the shift with only the largest cross correlation and how to solve this problem by combining candidate shifts of adjacent segments.

the candidate shifts of several adjacent segments to avoid locally optimal problem.

Here is one simple example for candidate shift detection of simulated chromatograms. Consider two chromatograms (the reference one is denoted as **r** and test one is **s**) that differ by an unknown shift along the retention time. One can rapidly calculate cross-correlation **c** between **s** and **r** using FFT cross correlation, and the candidate shifts between **s** and **r** can be found at the local maximums of **c**. Fig. 4 depicts this procedure visually. The number 20 of candidate shifts in Fig. 4 means shift the test profile by 20 points and the maximum cross-correlation between test and reference profile will be obtained.

### 2.4. Optimal shift determination

Sometimes the maximum of cross correlation of small segment as the optimal shift may be not the optimal shift at larger scale or globally. To avoid this problem, the candidate shifts can be detected using FFT cross correlation and CWT derivative according to Section 2.3; then, the optimal shift of each segment can be determined by combining the candidate shifts of several adjacent segments to maximize the correlation coefficient between test profile and reference profile. Fig. 5 is an example of simulated chromatograms with collision in the optimal shift between different segments, when only the largest cross correlation is used as the optimal shift. By combination of the local maximums of adjacent segments, one can see that MSPA can obtain more reasonable aligning result and larger correlation coefficient.

### 2.5. Move segments

By warping the non-peak parts to move the segments with peaks using linear interpolation, the detected peaks in each segment can be aligned without altering their shapes. It can reduce the emergence of artifacts by linear interpolation of the non-peak parts. The linear interpolation of the non-peak parts can also conserve the information of small peaks as much as possible, which are difficult to detect.

### 2.6. Implementation

All these works were done on a Dell Inspiron 530 PC with an Intel® Core™2 Quad Q6600 processor and 2048M memory. This method is implemented and available at http://code.google.com/p/mspa. MSPA can rapidly and accurately align chromatograms. The user is required to provide the dataset with some intuitive parameters such as reference and test profiles, SNR threshold for peak detection, allowed maximum shift parameter for each segment. During detection of the candidate shifts via FFT cross correlation, the candidate shifts should not be excessively large. Here the allowed maximum shift parameter for each segment can be used to prevent overshifting problem.

## 3. Experimental

To illustrate the used techniques in MSPA and demonstrate the effect of MSPA, three datasets have been used. Firstly, simulated chromatograms were constructed to test its basic functions. Then MSPA was applied to two experimental chromatographic datasets, covering from total ion chromatograms (TIC) of plasma metabolites to HPLC fingerprints of herb medicine extractions to evaluate its alignment performance. The summary of these three datasets is presented in Table 1.

**Table 1**
Dataset summary of simulated, free fatty acids (FFAs) in plasma and fructus aurantii immaturus (FAI).

| | Simulated dataset | FFAs dataset | FAI datasets |
|---|---|---|---|
| No. of samples | 1 | 121 | 38 |
| Sample length | 900 | 3900 | 12,000 |

### 3.1. Simulated chromatograms

Simulated chromatograms were created according to literature [24], which consist of Gaussian peaks, sinus curve baseline and random noise. These two simulated chromatograms have different peak position, noise level and baseline drifts, which are shown in Fig. 6(a). The solid one is the reference, whose noise is normally distributed with variance 1. The dashed one is the simulated chromatogram to be aligned, whose noise is normally distributed with variance 0.2.

### 3.2. Total ion chromatograms of free fatty acids in plasma

A total of 121 overnight fasting plasma samples were collected from patients at the Xiangya Hospital of Hunan, Changsha city of China. Each blood sample was centrifuged at 3000 × *g* for 10 min and transferred into a clean Eppendorf tube. The EDTA–$Na_2$ was added as anticoagulant. Aliquots (200 μl) of plasma were spiked with internal standard (I.S.) working solution (25 μl C17:0 and 25 μl C17:0 methyl ester), lipid extraction was carried out using hexane. The methyl esters of free fatty acids (FFAs) were extracted into hexane after the first esterification reaction, and the hexane phase was removed. Then, the methyl esters of free fatty acids (FFAs) were also extracted into hexane after the second esterification reaction, and concentrated under $N_2$ gas. Hexane (100 μl) was added to each tube prior to analysis. Chromatography and mass detection were performed on Shimadzu GC2010A (Kyoto, Japan) coupled to GCMS-QP2010 mass spectrometer. For each run, 1.0 μl plasma extractions were injected into DB-23 capillary column (30 m × 0.25 mm i.d., film thickness 0.25 μm) with split ratio of the injector being 1:10. Helium carrier gas was used at a constant flow rate of 1.0 ml $min^{-1}$. Column temperature was programmed from 70 °C to 220 °C. Mass spectra from 30 to 450 amu were collected at 0.2 s $scan^{-1}$. The ionization voltage was 70 eV and ion source temperature was 200 °C.

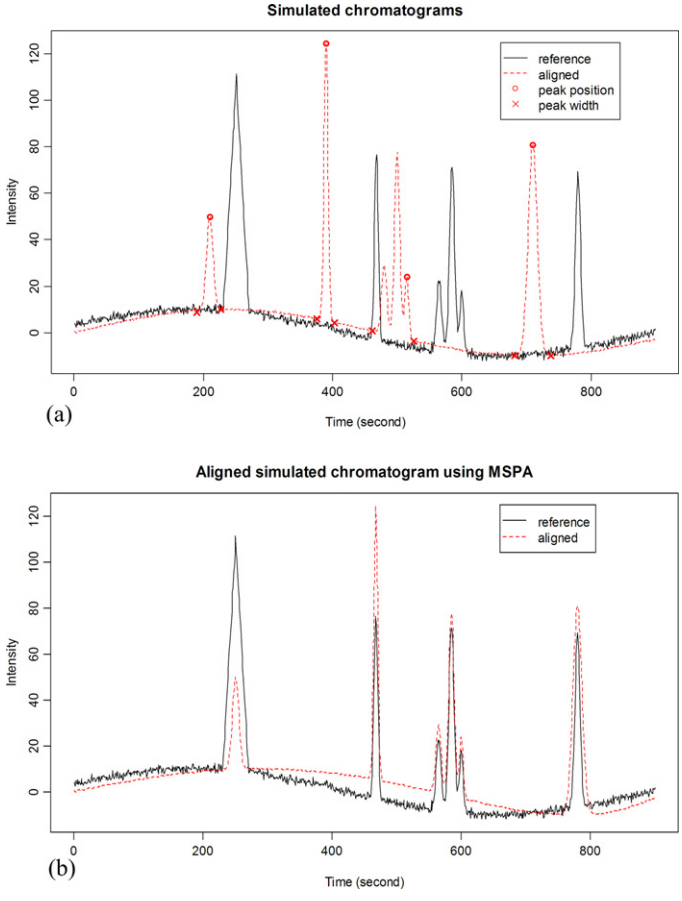


**Fig. 6.** Simulated chromatograms: (a) plots of simulated chromatograms before aligning and (b) after aligning using MSPA.

### 3.3. HPLC fingerprints of fructus aurantii immaturus

HPLC fingerprints of 38 samples of fructus aurantii immaturus (FAI) from nineteen provinces (or municipality) of China and a standard sample from National Institute for Control of Pharmaceutical and Biological Products were measured using Agilent/HP 1100 Series HPLC-DAD system (Agilent, Palo Alto, CA, USA). The pulverized FAI sample (60 mesh, 0.5 g) was placed into 150 ml round bottom flask, then extracted for 10 min at room temperature under ultrasound with 25 ml methanol and finally filtered. For each run, 20 μl herbal medicine extractions were injected into Sepax column (C18, 5 μm, 250 mm × 4.6 mm). The mobile phase consisted of acetonitrile, methanol and 0.05% polyphosphoric acid. The flow rate was 0.8 ml $min^{-1}$, column temperature was maintained at 30 °C. The DAD was set at 284 nm for acquiring chromatograms. The data were exported in netcdf format using HP chemstations (version A.09.01) for further analysis using MSPA method.

## 4. Results and discussion

Alignment results on both simulated and real chromatograms will be presented to evaluate the performance of MSPA method. By comparing MSPA with several widely used alignment methods, one can see the advantages of MSPA. The obtained results will also be discussed as well as some key points about MSPA and the effects of some parameters are discussed to explore the properties of MSPA.

### 4.1. Simulated chromatograms

The effect of MSPA method was firstly benchmarked using simulated chromatographic dataset with overlapped peaks, baseline and noises. Fig. 6 shows the alignment results. It can be seen in Fig. 6(a) that peaks were shifted nonlinearly between the reference and the chromatogram to be aligned, and the position, start and end points of the peaks from the chromatogram to be aligned can be exactly detected with Haar wavelet. All the peaks of chromatogram to be aligned have been synchronized to match the reference chromatogram after alignment by MSPA method in Fig. 6(b), which demonstrated MSPA can align chromatograms with overlapping peaks, baselines and different level noises.

### 4.2. Total ion chromatograms of free fatty acids in plasma

The MSPA method was run on TIC of free fatty acids in plasma to test its performance on metabolomics dataset. Both unaligned chromatograms (left part) and chromatograms aligned by MSPA (right part) are illustrated in Fig. 7, and some peaks are zoomed in to demonstrate the performance of MSPA method in a more clear way. One can see from the zoomed regions of unaligned chromatogram that there exist variations in peak positions. After processing by MSPA, the aligned chromatograms were also amplified at the same peaks' regions to show the aligning results. One can see from the zoomed regions of aligned chromatograms that all the peaks were

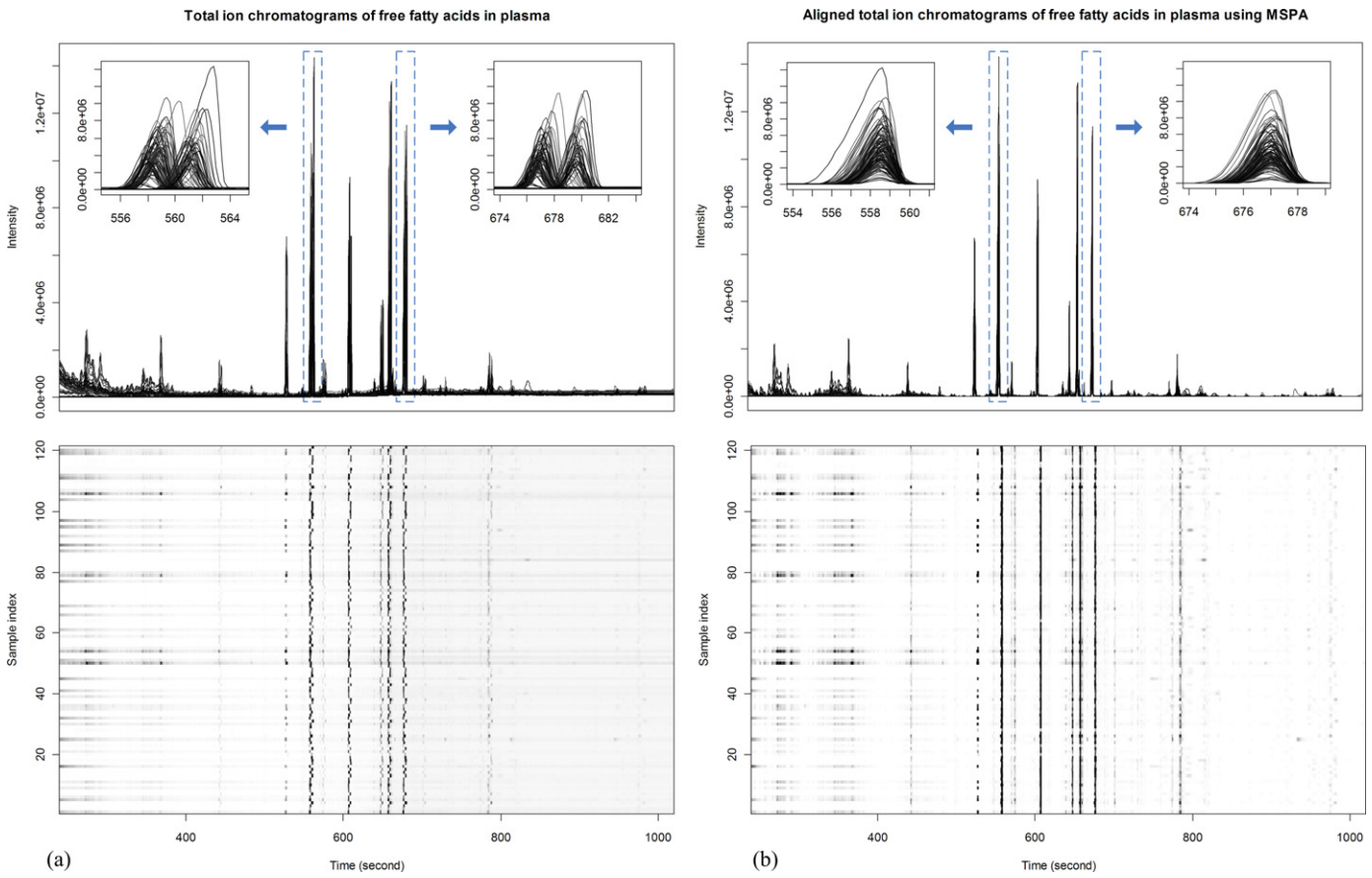


**Fig. 7.** Total ion chromatograms (TIC) of free fatty acids in plasma: (a) plot of TIC before alignment and (b) plot of the aligned TIC using MSPA.

successfully synchronized. Two images were also created to display overall variations in peak positions and alignment results of the entire chromatograms in a global manner. It can be seen obviously from the bottom left image that the lines are in "zigzag" and not straight enough, which means that there are variations in peak positions from sample to sample. But after aligning them with MSPA, all the "zigzag" lines in the bottom left image became the straight lines in the bottom right image, which means that MSPA can effectively eliminate variations in peak positions from sample to sample. By comparing the same zoomed regions of unaligned and aligned by MSPA, it can be seen that all peaks' shapes are intact, which demonstrated that MSPA has the capacity to preserve peaks' shape during alignment.

### 4.3. HPLC fingerprints of fructus aurantii immaturus

Original HPLC fingerprints of fructus aurantii immaturus were adopted from a previous FAI fingerprints study [58] involving 38 samples and a standard. These fingerprints were aligned by MSPA method as an example to help researchers to apply MSPA method in quality control of herbal medicines. The alignment procedure of these fingerprints was initiated by peak detection, width estimation using CWT derivative. By setting threshold for SNR as 500, dozens of peaks were detected in each fingerprint. The fingerprint of standard sample was used as the reference. Then, both the reference and fingerprints to be aligned can be drawn in the same plot, and one can easily estimate the maximum shift between the reference fingerprint and fingerprints to be aligned. In the alignment of these FAI fingerprints, the maximum shift was set as 285. The unaligned fingerprints are shown in the left part of Fig. 8. The zoomed "peaks rich" region and the full image are illustrated in Fig. 8 to provide readers with the detailed view of peak position variations and the overview of peak position variations of the entire chromatograms in a global manner. The aligned fingerprints by MSPA method are also shown in the right part of Fig. 8. The aligned fingerprints were also amplified at the same peak regions to show the aligned results in detail, and one can see that all the peaks were also successfully synchronized and the shapes of all the peaks were preserved by comparing the same zoomed region of unaligned fingerprints and the aligned ones. The aligned fingerprints are also illustrated in the image. It is clear that all the "zigzag" lines in the bottom left image became straight in the bottom right image, which means that MSPA can also effectively eliminate variations in peak positions of herb medicine fingerprints. By combining both the detailed views of zoomed region and the overview of the images, one can come to the conclusion that MSPA can accurately align the fingerprints of herb medicines, meanwhile preserving the shapes of the peaks.

### 4.4. Influence of noise on different derivative calculation methods

Four signals, which consist of the same Gaussian peak but with noises of different variances, were created to investigate influence of noise on Haar wavelet and numeric derivative calculation methods for peak detection and width estimation. For these four signals, their calculated derivatives by Haar wavelet and numeric derivative are shown in Fig. 9 to check that which derivative method was more suitable for peak detection and width estimation. One can clearly observe from the figure that the numeric derivative method is only effective when the noise level is lower. When the noise level

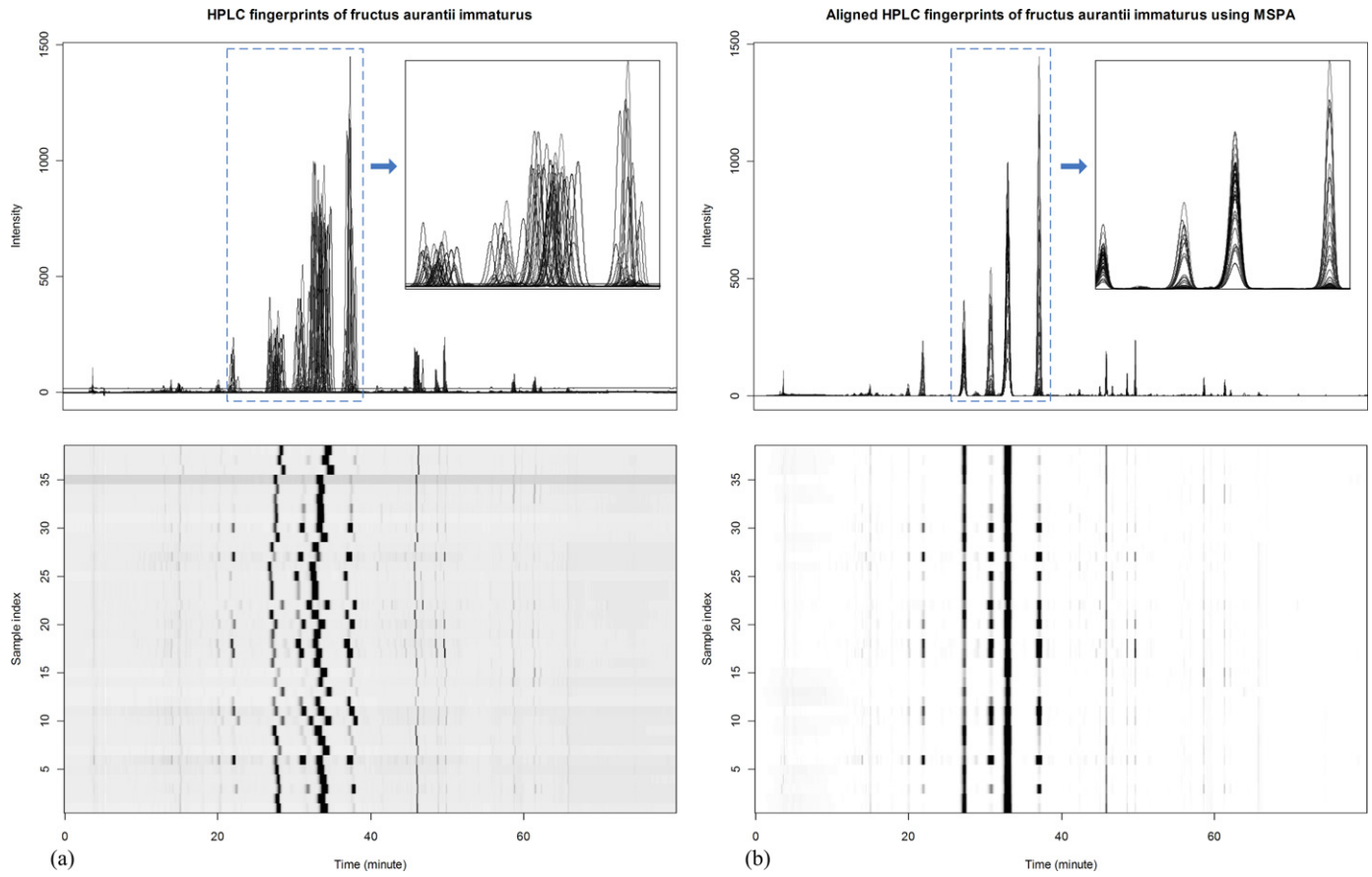


**Fig. 8.** Fingerprints of fructus aurantii immaturus: (a) plot of fingerprints before alignment and (b) plot of the aligned fingerprints using MSPA.

was higher, SNR of the obtained derivative by numeric derivative method was badly degenerated and it can no longer be used for peak position detection and width estimation. By comparing the derivatives by Haar wavelet with numeric derivative methods, one can see that the SNRs of the obtained derivatives by Haar wavelet derivative method are all good enough to be used for accurate detection of the peak position and estimation of the peak width for all the four signals with different noise levels.

### 4.5. *Influence of baseline on different alignment methods*

The simulated chromatograms with and without sinus curve baseline were used to test the influence of baseline on MSPA, RAFFT and COW methods. Simulated chromatograms without sinus curve baseline were obtained by correcting baseline with airPLS [10]. The parameters of airPLS for this correction were: lambda = $10^4$ and order = 2. Baseline correction results can be seen in Fig. 10. The alignment results by MSPA methods of uncorrected and corrected chromatograms are illustrated in Fig. 10(a) and (b) respectively. One can observe that both the uncorrected and corrected chromatograms are properly aligned by MSPA method, which demonstrates that MSPA method is not sensitive to baseline during alignment. The maximum shift parameter of MSPA can be easily set by observing and estimating from the plot of reference and chromatograms to be aligned. In this dataset, the estimated maximum shift using the plot was 85 points, and it worked well with MSPA method. The alignment results by RAFFT are plotted in Fig. 10(c) and (d). The estimated maximum shift from the plot of reference and chromatograms to be aligned did not work well with RAFFT. By enumerating shift from 85 to 10, 70 was chosen as the maximum shift of RAFFT which can provide the best alignment results. This means that the maximum shift parameter of RAFFT is not intuitive enough to be adjusted easily, its optimization being too time-consuming. From Fig. 10(c) it can be observed that without baseline correction, RAFFT method can align the peaks of chromatogram properly. With the same parameter, the first peak of the baseline corrected chromatogram cannot be aligned properly by RAFFT in Fig. 10(d), which means that RAFFT method is sensitive to baseline during alignment. COW method was implemented by Tomasi [59], which is available at website of R. Bro group [60]. The segment and slack parameters of COW were optimized using the "grid-search" method, and the optimized parameters were segment = 22 and slack = 16. Fig. 10(e) and (f) shows alignment results by COW. All peaks have been well aligned, and thus COW method is also not sensitive to baseline too. But if one examines and compares aligned peaks and unaligned peaks carefully, it can be seen that COW method changes the shapes of peaks during alignment procedure. The segment and slack parameters of COW method needed even more time-consuming optimization than RAFFT. Therefore, in this dataset, MSPA outperforms RAFFT and COW in robustness, insensitiveness to baseline, capacity to preserve shapes of peaks and easiness of parameter adjustment.

Concerning the insensitivity to the baseline, in our opinion, it is because that the detection of peaks in signal and movement of the peaks are conducted separately in MSPA. The peak shape, in general, does not change when it was moved in the whole procedure. On the other hand, it had little effect on the FFT correlation coefficient between the detected peak and reference signal when the segment of each detected peak is small and the baseline is slowly changing within a narrow range.

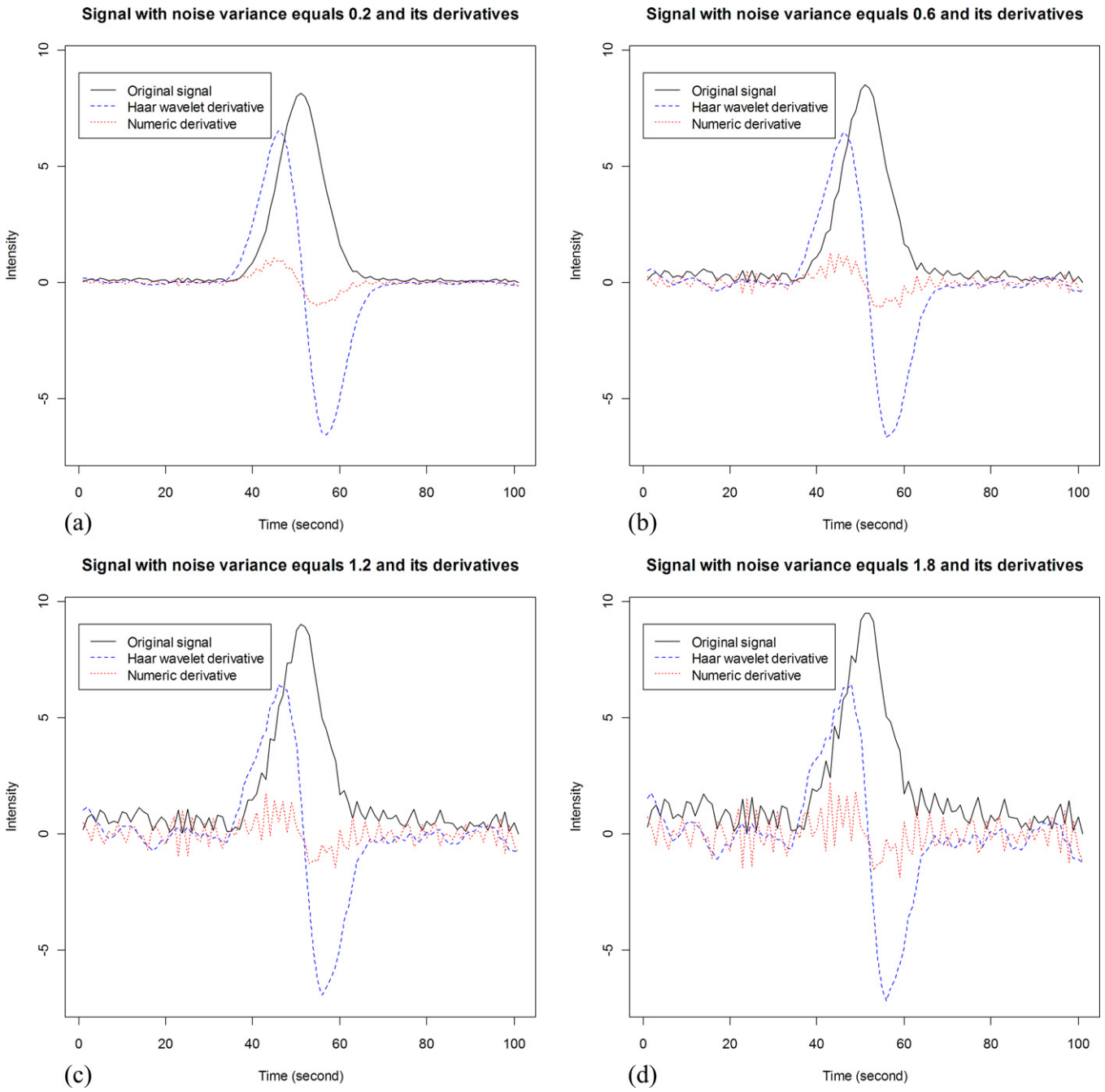


**Fig. 9.** Comparison of the influence of noise on different derivative calculation methods for peak detection and width estimation.

### 4.6. Alignment quality assessment and speed issue

The fructus aurantii immaturus dataset was difficult to align due to the large variation between herbal samples, so this dataset was used to benchmark MSPA, RAFFT and COW methods. The alignment results of fructus aurantii immaturus dataset by MSPA, RAFFT and COW are presented in Fig. 11 via images of correlation maps between samples. The correlation maps in the right part of the figure show that all the three methods can improve similarity between samples. If we carefully observe the correlation maps, it can be seen that correlation coefficients between samples of MSPA are larger than correlation coefficients between samples of RAFFT and COW. The reason is that MSPA can align the peaks more accurately than RAFFT and COW.

The alignment quality can also be assessed by the means of the mean correlation coefficients (*mcc*) between signals to be aligned and reference, which can be expressed by following equation:

$$mcc(\mathbf{r},\mathbf{S}) = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{\sum_{j=1}^{n}(\mathbf{r}_j-\bar{\mathbf{r}})(\mathbf{S}_{i,j}-\bar{\mathbf{S}}_i)}{\sqrt{\sum_{j=1}^{n}(\mathbf{r}_j-\bar{\mathbf{r}})^2}\sqrt{\sum_{j=1}^{n}(\mathbf{S}_{i,j}-\bar{\mathbf{S}}_i)^2}}\right) \quad (10)$$

where $\mathbf{r}$ is a vector of the reference signal and $\mathbf{S}$ is a matrix, and each row of $\mathbf{S}$ is a vector of signal to be aligned.

Since there are peak shape changes during alignment when some alignment methods are used, peak area changes should be also quantified to evaluate the capacity to preserve shapes of peaks during alignment. The mean relative change in area is adopted to

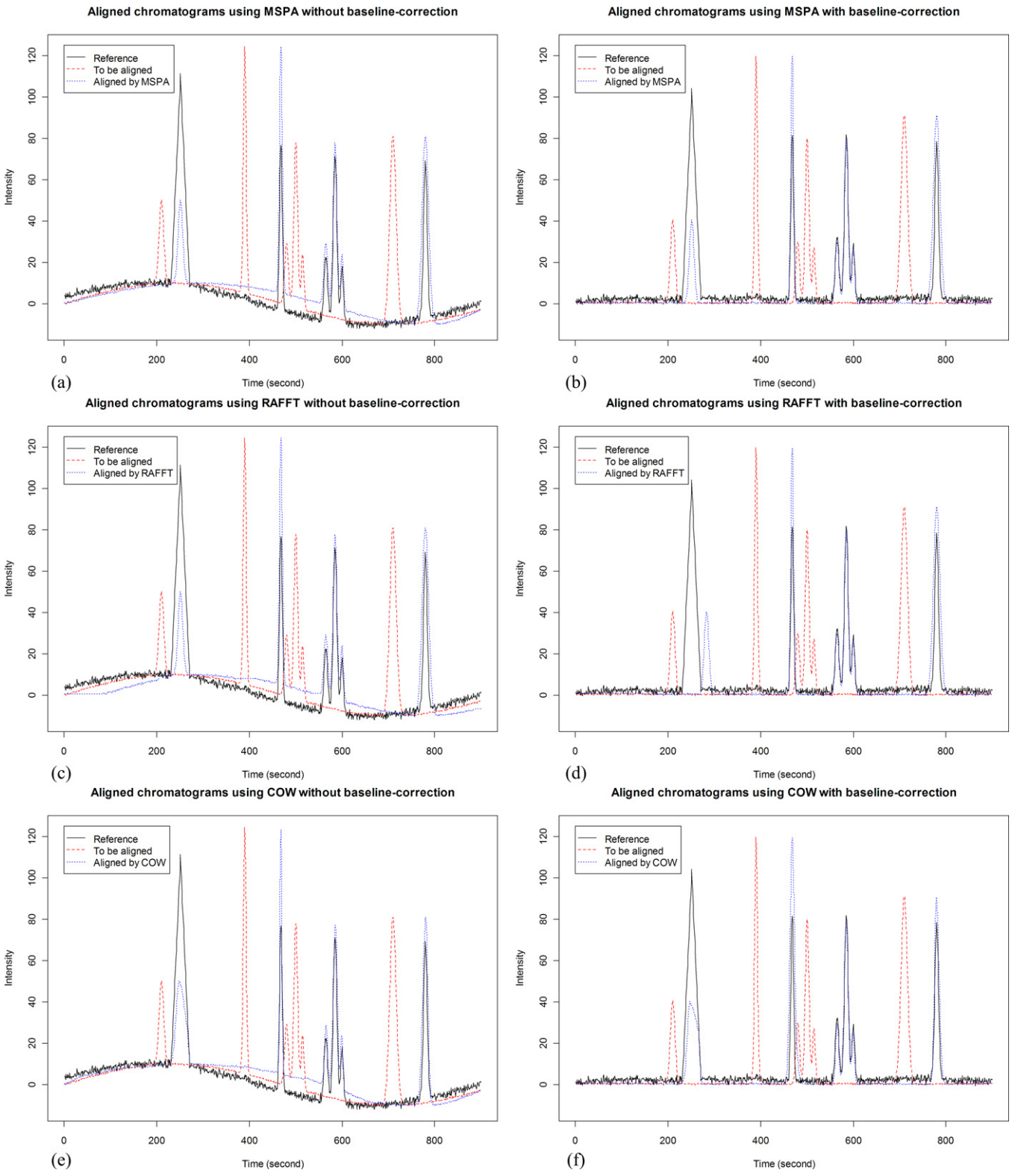


**Fig. 10.** Investigation of the affect of baseline on different alignment methods.

evaluate the area changes during alignment. The definition of mean relative change in area (*mrca*) is

$$mrca = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{\left|\sum_{j=1}^{n}\mathbf{A}_{i,j} - \sum_{j=1}^{n}\mathbf{S}_{i,j}\right|}{\sum_{j=1}^{n}\mathbf{S}_{i,j}}\right) \quad (11)$$

where **S** is a matrix and each row of **S** is a vector of signal to be aligned. **A** is also a matrix and each row of **A** is a vector of aligned signal.

The *mcc*, *mrca* and mean calculation time of simulated, free fatty acids in plasma and fructus aurantii immaturus datasets with MSPA, RAFFT and COW methods are listed in Table 2 to assess alignment quality and speed of these methods. The parameters for

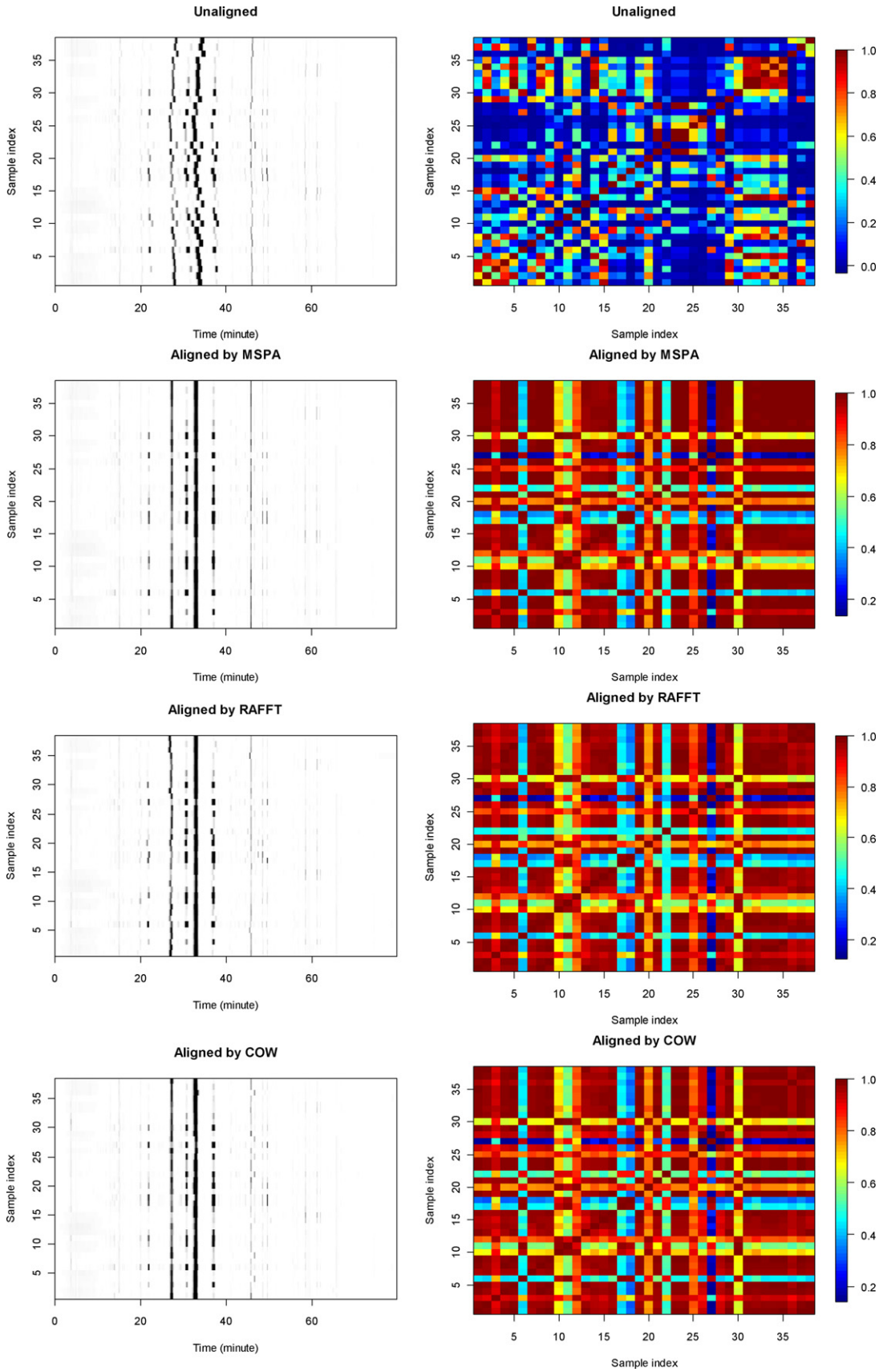


**Fig. 11.** Benchmarks of MSPA, RAFFT and COW methods on fructus aurantii immaturus dataset which is difficult to align because of the large variation between herbal samples.

**Table 2**
*mcc*, *mrca* and mean calculation time of simulated, free fatty acids (FFAs) in plasma and fructus aurantii immaturus (FAI) datasets with MSPA, RAFFT and COW methods.

| Datasets | Alignment methods[a] | *mcc*[b] | *mrca*[c] (%) | Mean calculation time (s) |
|---|---|---|---|---|
| Simulated dataset | Unaligned | 0.0561 | – | – |
| | MSPA | 0.7764 | 0.16 | 0.0540 |
| | RAFFT | 0.6060 | 2.14 | 0.0036 |
| | COW | 0.8420 | 14.04 | 1.5582 |
| FFA dataset | Unaligned | 0.5783 ± 0.4137 | – | – |
| | MSPA | 0.9486 ± 0.0279 | 0.11 ± 0.21 | 0.2215 ± 0.3681 |
| | RAFFT | 0.9382 ± 0.0405 | 0.30 ± 0.48 | 0.0058 ± 0.0022 |
| | COW | 0.9375 ± 0.0607 | 0.96 ± 1.22 | 13.2554 ± 0.0665 |
| FAI dataset | Unaligned | 0.2871 ± 0.3054 | – | – |
| | MSPA | 0.8859 ± 0.1314 | 0.07 ± 0.08 | 1.0428 ± 0.4373 |
| | RAFFT | 0.8631 ± 0.1325 | 0.26 ± 0.40 | 0.0224 ± 0.0057 |
| | COW | 0.8751 ± 0.1285 | 4.46 ± 4.08 | 160.6447 ± 6.2072 |

[a] *Parameters for each alignment method on different datasets*: for simulated dataset, MSPA (max_shift = 85), RAFFT (max_shift = 70) and COW (segment = 22, slack = 16); for FFA dataset, MSPA (max_shift = 300), RAFFT (max_shift = 300) and COW (segment = 80, slack = 30); for FAI dataset, MSPA (max_shift = 285), RAFFT (max_shift = 273) and COW (segment = 75, slack = 30).
[b] *mcc*, mean correlation coefficients, is used to evaluate alignment quality assessment, the larger the better.
[c] *mrca*, mean relative change in area, is used to evaluate peak area changes during alignment, the smaller the better.

each alignment method of different datasets were optimized by the same procedures described in Section 4.5, which are also presented as footnote of Table 2. For the simulated dataset, alignment quality of MSPA was much better than the one of RAFFT. Furthermore, *mrca* of MSPA is only 0.16%, which means that MSPA can better preserve the peak shapes. Although *mcc* of COW method were larger than the *mcc* of MSPA in the simulated dataset, but its *mrca* of COW is 14.04%. We can question the reliability of the large *mcc* of COW, having been obtained at the cost of peak shapes from its large *mrca*. MSPA obtained the best *mcc* and *mrca* in both FFAs and FAI datasets among these three alignment methods. Although the speed of MSPA is slower than RAFFT, it is acceptable that alignment of fingerprint with 12,000 data points by MSPA takes only about 1 s. Both MSPA and RAFFT are much faster than COW. From Table 2, one can deduce the following conclusions about these alignment methods: (1) MSPA can synchronize signals with better alignment quality than RAFFT and COW; (2) RAFFT has the best alignment speed, and alignment speed of MSPA is acceptable even for large dataset with dozens of thousands of data points; (3) COW is too slow to align large dataset and changes the shapes of peaks more seriously than RAFFT and MSPA. It seems that MSPA has the best balance between alignment quality and speed.

## 5. Conclusion

Massive chromatographic datasets can be accurately and rapidly aligned by the proposed MSPA method when two intuitive parameters, namely threshold for SNR and maximum shift, are properly set. By testing with simulated, real chromatograms and comparisons with several widely used alignment methods, it was demonstrated that MSPA method has the capacity to preserve the shapes of peaks, performs well with nonlinear retention time shifts, can avoid locally optimal problem and seems robust and not sensitive to noise and baseline. Furthermore, the availability of source code makes MSPA to be easily customized and optimized for particular applications and more significant to a broad range of chromatography researchers. These advantages guarantee that MSPA may address the challenges of alignment of massive dataset in metabolomics and quality control of herbal medicines, which enables researchers to preprocess, analyze, interpret and extract useful information from these datasets within an acceptable time using statistics and chemometrics.

## Acknowledgments

This work is financially supported by the National Nature Foundation Committee of PR China (Grant No. 20875104, Grant No. 10771217, Grant No. 20975115, Grant No. 21105129 and Grant No. 21175157), the international cooperation project on traditional Chinese medicines of Ministry of Science and Technology of China (Grant No. 2007DFA40680), Hunan Provincial Innovation Foundation For Postgraduate (Grant No. CX2010B058), Scholarship Award for Excellent Doctoral Student granted by Ministry of Education of China (No. 092301020), the Graduate Degree Thesis Innovation Foundation of Central South University (No. 092301020), Special Funds of Central South University for Fostering Outstanding Doctoral Degree Thesis (Grant No. 092301020) and Central South University for special support of the basic scientific research project (No. 2010QZZD007 and No. 2011QNZT053). The studies meet with the approval of the university's review board. We are grateful to all employees of this institute for their encouragement and support of this research.

## References

[1] J.M. Amigo, T. Skov, R. Bro, Chem. Rev. 110 (2010) 4582.
[2] W.B. Dunn, D.I. Ellis, TRAC – Trends Anal. Chem. 24 (2005) 285.
[3] J.K. Nicholson, J.C. Lindon, Nature 455 (2008) 1054.
[4] Y.Z. Liang, P.S. Xie, K. Chan, J. Chromatogr. B – Anal. Technol. Biomed. Life Sci. 812 (2004) 53.
[5] Y.Z. Liang, P.S. Xie, F. Chau, J. Sep. Sci. 33 (2010) 410.
[6] C.A. Smith, E.J. Want, G. O'Maille, R. Abagyan, G. Siuzdak, Anal. Chem. 78 (2006) 779.
[7] J. Trygg, E. Holmes, T. Lundstedt, J. Proteome Res. 6 (2006) 469.
[8] F. Gong, Y.Z. Liang, P.S. Xie, F.T. Chau, J. Chromatogr. A 1002 (2003) 25.
[9] C.J. Xu, Y.Z. Liang, F.T. Chau, Y. Vander Heyden, J. Chromatogr. A 1134 (2006) 253.
[10] Z.M. Zhang, S. Chen, Y.Z. Liang, Analyst 135 (2010) 1138.
[11] F. Gong, Y.Z. Liang, Y.S. Fung, F.T. Chau, J. Chromatogr. A 1029 (2004) 173.
[12] Z.M. Zhang, S. Chen, Y.Z. Liang, Talanta 83 (2010) 1108.
[13] B.Y. Li, Y. Hu, Y.Z. Liang, L.F. Huang, C.J. Xu, P.S. Xie, J. Sep. Sci. 27 (2004) 581.
[14] O.M. Kvalheim, Y.Z. Liang, Anal. Chem. 64 (1992) 936.
[15] Y.Z. Liang, O.M. Kvalheim, H.R. Keller, D.L. Massart, P. Kiechle, F. Erni, Anal. Chem. 64 (1992) 946.
[16] S. Wold, K. Esbensen, P. Geladi, Chemom. Intell. Lab. Syst. 2 (1987) 37.
[17] S. Wold, M. Sjostrom, L. Eriksson, Chemom. Intell. Lab. Syst. 58 (2001) 109.
[18] S. Wold, Pattern Recognit. 8 (1976) 127.
[19] M. Barker, W. Rayens, J. Chemom. 17 (2003) 166.
[20] J. Trygg, S. Wold, J. Chemom. 16 (2002) 119.
[21] R.H. Jellema, in: D.B. Stephen, T. Romà, W. Beata (Eds.), Comprehensive Chemometrics, Elsevier, Oxford, 2009, p. 85.
[22] C.P. Wang, T.L. Isenhour, Anal. Chem. 59 (1987) 649.
[23] K. Athanassios, F.M. John, A.T. Paul, AIChE J. 44 (1998) 864.
[24] N.-P.V. Nielsen, J.M. Carstensen, J. Smedsgaard, J. Chromatogr. A 805 (1998) 17.

[25] D. Clifford, G. Stone, I. Montoliu, S. Rezzi, F.-P. Martin, P. Guy, S. Bruce, S. Kochhar, Anal. Chem. 81 (2009) 1000.
[26] H. Sakoe, S. Chiba, IEEE Trans. Acoust. Speech Signal Process. 26 (1978) 43.
[27] S. Stan, C. Philip, Intell. Data Anal. 11 (2007) 561.
[28] J. Forshed, I. Schuppe-Koistinen, S.P. Jacobsson, Anal. Chim. Acta 487 (2003) 189.
[29] G.-C. Lee, D.L. Woodruff, Anal. Chim. Acta 513 (2004) 413.
[30] P.H.C. Eilers, Anal. Chem. 76 (2004) 404.
[31] V. Pravdova, B. Walczak, D.L. Massart, Anal. Chim. Acta 456 (2002) 77.
[32] A.M. van Nederkassel, M. Daszykowski, P.H.C. Eilers, Y.V. Heyden, J. Chromatogr. A 1118 (2006) 199.
[33] J.W.H. Wong, C. Durante, H.M. Cartwright, Anal. Chem. 77 (2005) 5655.
[34] J.W.H. Wong, G. Cagney, H.M. Cartwright, Bioinformatics 21 (2005) 2088.
[35] F. Savorani, G. Tomasi, S.B. Engelsen, J. Magn. Reson. 202 (2010) 190.
[36] K.A. Veselkov, J.C. Lindon, T.M.D. Ebbels, D. Crockford, V.V. Volynkin, E. Holmes, D.B. Davies, J.K. Nicholson, Anal. Chem. 81 (2009) 56.
[37] G. Tomasi, F. Savorani, S.B. Engelsen, J. Chromatogr. A 1218 (2011) 7832.
[38] M. Daszykowski, Y. Vander Heyden, C. Boucon, B. Walczak, J. Chromatogr. A 1217 (2010) 6127.
[39] B. Walczak, W. Wu, Chemom. Intell. Lab. Syst. 77 (2005) 173.
[40] K.J. Johnson, B.W. Wright, K.H. Jarman, R.E. Synovec, J. Chromatogr. A 996 (2003) 141.
[41] R.J.O. Torgrip, M. Aberg, B. Karlberg, S.P. Jacobsson, J. Chemom. 17 (2003) 573.
[42] Y. Yasui, D. McLerran, B.L. Adam, M. Winget, M. Thornquist, Z.D. Feng, J. Biomed. Biotechnol. (2003) 242.
[43] Z.M. Zhang, S. Chen, Y.Z. Liang, Talanta 83 (2011) 1108.
[44] W. Wu, M. Daszykowski, B. Walczak, B.C. Sweatman, S.C. Connor, J.N. Haseldeo, D.J. Crowther, R.W. Gill, M.W. Lutz, J. Chem. Inf. Model. 46 (2006) 863.
[45] A.L. Duran, J. Yang, L. Wang, L.W. Sumner, Bioinformatics 19 (2003) 2283.
[46] M. Katajamaa, M. Oresic, BMC Bioinformatics 6 (2005) 179.
[47] H.P. Benton, D.M. Wong, S.A. Trauger, G. Siuzdak, Anal. Chem. 80 (2008) 6382.
[48] R. Tautenhahn, G.J. Patti, E. Kalisiak, T. Miyamoto, M. Schmidt, F.Y. Lo, J. McBee, N.S. Baliga, G. Siuzdak, Anal. Chem. 83 (2011) 696.
[49] A. Lommen, Anal. Chem. 81 (2009) 3079.
[50] J. Xia, N. Psychogios, N. Young, D.S. Wishart, Nucleic Acids Res. 37 (2009) W652.
[51] J. Xia, D.S. Wishart, Nat. Protoc. 6 (2011) 743.
[52] C. Yang, Z.Y. He, W.C. Yu, BMC Bioinformatics 10 (2009), Article No. 4.
[53] X. Shao, C. Pang, Q. Su, Fresen. J. Anal. Chem. 367 (2000) 525.
[54] I. Daubechies, Ten Lectures on Wavelets, Society for Industrial Mathematics, 1992.
[55] X.-G. Shao, A.K.-M. Leung, F.-T. Chau, Acc. Chem. Res. 36 (2003) 276.
[56] C.E. Shannon, W. Weaver, The Mathematical Theory of Communication, University of Illinois Press, Urbana, IL, 1949.
[57] http://en.wikipedia.org/wiki/Discrete_Fourier_transform.
[58] X. Xu, J. Jiang, Y. Liang, L. Yi, J. Cheng, Anal. Methods 2 (2010) 2002.
[59] G. Tomasi, F. van den Berg, C. Andersson, J. Chemom. 18 (2004) 231.
[60] http://www.models.life.ku.dk/DTW_COW.