Talanta 83 (2011) 1108–1117

Contents lists available at ScienceDirect

Talanta

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

Peak alignment using wavelet pattern matching and differential evolution

Zhi-Min Zhang, Shan Chen, Yi-Zeng Liang*

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

ARTICLE INFO

Article history:
Available online 14 August 2010

Keywords:
Alignment
Chromatography
Wavelet
Differential evolution

ABSTRACT

Retention time shifts badly impair qualitative or quantitative results of chemometric analyses when entire chromatographic data are used. Hence, chromatograms should be aligned to perform further analysis. Being inspired and motivated by this purpose, a practical and handy peak alignment method (alignDE) is proposed, implemented in this research for one-way chromatograms, which basically consists of five steps: (1) chromatogram lengths equalization using linear interpolation; (2) accurate peak pattern matching by continuous wavelet transform (CWT) with the Mexican Hat and Haar wavelets as its mother wavelets; (3) flexible baseline fitting utilizing penalized least squares; (4) peak clustering when gap of two peaks is smaller than a certain threshold; (5) peak alignment using differential evolution (DE) to maximize linear correlation coefficient between reference signal and signal to be aligned. This method is demonstrated with both simulated chromatograms and real chromatograms, for example, chromatograms of fungal extracts and Red Peony Root obtained by HPLC-DAD. It is implemented in R language and available as open source software to a broad range of chromatograph users (http://code.google.com/p/alignde).

1. Introduction

As a most commonly used laboratory technique for separating mixtures, chromatography plays a crucial role in analytical chemistry. The visual output of chromatography is chromatograms, which consist of peaks corresponding to components of the separated mixtures. Ideally, same component of different samples should have an equal retention time. But in real analysis, retention time shifts occur frequently due to instrumental drift, poor mixing of mobile phase, stationary phase decomposition, columns changes during usage, interaction between analytes and etc. It influences foundational chemometric algorithms heavily, since principal component regression (PCR) [1] and partial least squares (PLS) [2] require assumption as follows: data matrix are bilinear factor models. The bilinear factor model means that alike peaks represent the same components in all samples [3–8]. In order to analyze entire chromatograms accurately and effectively using chemometrics, the same peaks in different chromatograms should be aligned at the same matrix column.

Dozens of alignment techniques have been developed and published in literatures to align retention time shifts in chromatography and spectroscopy. Time warping was first introduced by Wang and Isenhour [9] to eliminate the time axis stretch and squeeze in the chromatograms. After 10 years since Wang introduced time warping for chromatograms, two practical methods, namely dynamic time warping (DTW) [10] and correlation optimized warping (COW) [11], were proposed for aligning retention time shift in chromatograms. Eilers made effort to model the warping function, then presented the parametric time warping (PTW) [12]. These three methods were popular at the time, which could be seen from both discussion about and improvement in them [13–17]. The fuzzy warping method, proposed by Walczak and Wu [18], firstly detected of a few more intense peaks in individual chromatograms and aligned signals to avoid time consuming or requiring time-consuming optimization of input parameters. Then Wu et al. [6] revised the fuzzy warping into an iterative version, and it was used on-line for an alignment of the NMR spectra. Meanwhile some methods, relied on peak detection, were proposed and published, and they are partial linear fit [19], PARS (Peak Alignment using Reduced Set mapping) [20] and Landmark Selection [21]. The algorithm PAGA (Peak Alignment by a Genetic Algorithm) of Forshed et al. [22] is worth mentioning here, which takes each segment of the spectrum out individually, and uses genetic algorithm to apply a separate shift to each segment for alignment until the best correlation. It is so different and novel from the above mentioned kinds and seems to be fast and request less memory.

Generally speaking, dynamic programming [9–12,16,18], simple peak detection [6,18–21] and evolutionary algorithms [22] are three mainstream algorithms for retention time shift aligning. All the three methods have been extensively applied in chromatography and spectroscopy, because of their desirable property in certain circumstance. But each of them has drawbacks in certain aspects: (1) warping methods using dynamic programming are time consuming or require time-consuming optimization of input

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

0039-9140/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.talanta.2010.08.008

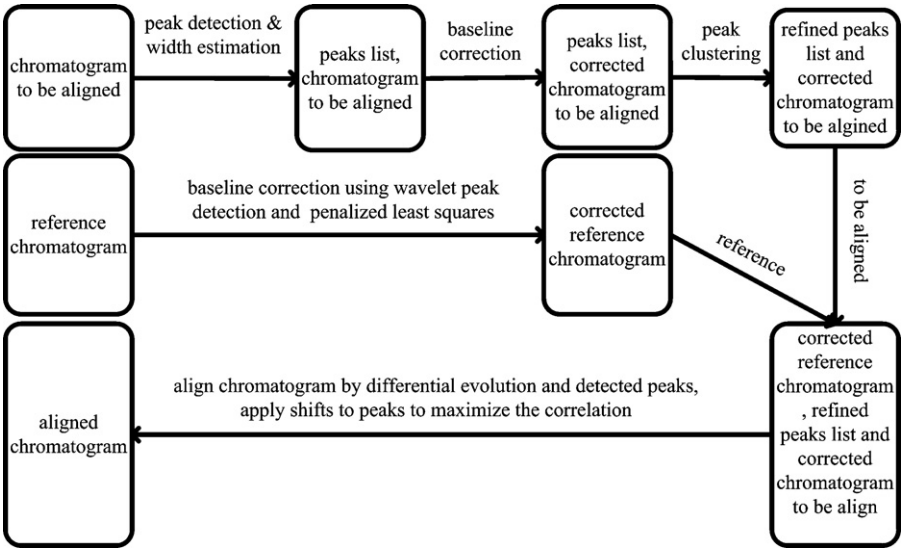

Fig. 1. The flow chart describing the framework of the differential evolution alignment method.

parameters [18], and PTW is fast but not flexible enough to align complex chromatographic shifts [17]; (2) PARS and fuzzy warping, relied on simple peak detection, align major peaks at the expense of minor peak shift accuracy [5]; (3) shortcoming of PAGA is difficult to optimize the segment size, which means that large segment size can result in nonalignment of small peaks and small segment can result in distortions in peaks due to a peak being cut at segment boundaries [5].

Warping, peak detection and evolutionary algorithms are complementary to some extent when they are used in shifting alignment. The success of COW can be attributed to the application of maximization of the correlation, and the time consuming shortcoming can be overcome using peak detection or evolutionary algorithm. Here, the fuzzy warping is an example, whose success can be attributed to the combination of warping and peak detection together. If evolutionary algorithms are used to apply a separate shift to peaks but not segment, there are not longer any needs to optimize the segment size. Consequently, the accurate peak detection can help PAGA methods to overcome their shortcoming. So one can intuitively draw the conclusion that the combination of the three mainstream algorithms together might overcome shortcomings of each other, and further develop a more general-purpose method.

The alignDE method described in this paper aligns chromatograms by differential evolution (DE) [23]. It moves the accurate detected peaks in the range of (−slack, slack), and warps the non-peak parts between two detected peaks using linear interpolation. Maximizing the correlation between reference chromatogram and chromatogram to be aligned is the objective function of the moving and warping steps using DE. By transforming the chromatogram into the wavelet space, the peaks can be detected accurately and robustly [24]. The DE optimization method is used to adjust the position of the accurately detected peak to maximize the correlation coefficient between the reference chromatogram and chromatogram to be aligned. The differential evolution guarantees the fast execution speed of the proposed method. Accurate detected peaks and warping in non-peak regions preserve the area and shape of peaks. Finally, the correction of baseline can improve the quality in calculation of linear correlation coefficient.

The paper is organized as follows. Principles of lengths equalization using linear interpolation, peak detection relies on wavelet, penalized least squares for background correction, peak clustering by gap threshold and alignDE are detailedly investigated in the method section. Then experiments of the used data are briefly introduced. In the following section, results of the alignment are presented and also accompany with discussions about them. The result section starts with simulated chromatograms. Then real experimental chromatograms are used to demonstrate the proposed methods. Finally, some conclusions and outlooks are drawn at the end of this paper.

2. Materials and methods

Aligning using DE involves the application of separate shifts to accurate detected peaks, which are outputs of wavelet pattern matching. In order to eliminate the influence of baseline on linear correlation coefficient and treat some special peak regions, such as peaks with shoulders, overlapping peaks and etc, baseline correction and peak clustering are added in the framework of proposed method. If sample intervals are different in reference chromatogram (R) and chromatogram to be aligned (C), data points of these two chromatograms are not equal. Therefore, linear interpolation is also integrated into the framework to equate length of chromatograms. The flow chart describing the architecture of the proposed alignment method is shown in Fig. 1.

2.1. Chromatograms length equalization

Difference in sample intervals leads unequal data points of chromatograms. In this work, linear interpolation is used to equate data points. Let's start with the simplest case: two points are given by the coordinate (x_1, y_1) and (x_2, y_2). The linear interpolation is a straight line between these two points. Since this straight line is across these two points, it can be given in the equation:

$$y = y_1 + (x - x_1)\frac{y_2 - y_1}{x_2 - x_1} \quad (1)$$

Generalization above equation to a set of data points means concatenation of linear interpolation between each pair of data points together. So assuming that number of data points of R is N_R and number of data points of C is N_C. The retention time of C is divided into N_R equally spaced points spanning, and interpolation takes place at these N_R points using points in C adjacent to them.

2.2. Peak matching using wavelet

After comparing many peak detection algorithm [25], the signal-to-noise ratio (SNR) and ridge lines method was applied in this study. The main idea of this approach was described in Ref. [26] and improved in estimation of peak width in Ref. [24]. It is based on wavelet, which is one of the most powerful tools in signal analysis.

Wavelet transformation can be categorized into discrete wavelet transform (DWT) and continuous wavelet transform (CWT). Unlike the non-redundant, more efficient DWT, CWT allows wavelet transforms at every scale with continuous shifting, which makes the information available in peak scale and peak position of chromatograms. The chromatograms are transformed into wavelet space using CWT, and peak is identified in wavelet space. CWT is defined as the sum over all time of the signal multiplied by scaled, shifted versions of the wavelet function Ψ. Mathematically, this process of CWT is represented as follows:

$$C(a,b) = \int_{-\infty}^{+\infty} s(t)\psi_{a,b}(t)\ dt, \qquad \psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right),$$
$$a \in R^{+}, \quad b \in R \tag{2}$$

Here $s(t)$ is the signal, a the scale, b the shifting, $\psi(t)$ the mother wavelet, $\psi_{a,b}(t)$ is the scaled and shifted wavelet and C is the 2D matrix of wavelet coefficients.

CWT can be regarded as kind of convolution between the signal and mother wavelet. If one would like to pick out peaks of chromatograms, the mother wavelet must be similar to the peaks. The Mexican Hat wavelet is chosen from a set of mother wavelets, which can be mathematically described by:

$$\psi(x) = \left(\frac{2}{\sqrt{3}}\pi^{-1/4}\right)(1-x^2)e^{-x^2/2} \tag{3}$$

Here $\psi(t)$ means the Mexican Hat mother wavelet, π represents the ratio of the circumference of a circle to its diameter, and e is Euler's constant, which is an irrational constant approximately equal to 2.718281828.

Observing the 2D CWT coefficients carefully after application CWT on chromatogram, one can find that the corresponding CWT coefficients at each scale have a local maximum around the peak center and the local maximum increases corresponding to the CWT scale, which reaches its maximum when the scale best match the peak width. So the peak detection problem has been transformed into finding ridges over the 2D CWT coefficient matrix. The peak can find in three steps in the 2D CWT coefficient matrix of chromatogram: (1) identify the ridges by linking the local maxima; (2) identify the peaks based on the ridges lines; (3) refine the peak parameter estimation.

Fig. 2 is an illustration about the peak matching steps. Fig. 2(a) is simulated chromatogram with the circles representing the detected peak positions. CWT with the scale parameter from 1 to 56 was applied to the simulated chromatogram and the 2D CWT coefficients are displayed in Fig. 2(b). The ridges lines identified from the 2D CWT coefficients. One can clearly see the correspondence between the major peaks and the ridge lines from Fig. 2(c).

The usage of Mexican Hat wavelet for peak detection can lead to underestimate scale of the peak. In order to accurately estimate the peak width, Zhang et al. [24] proposed signal-to-noise ratio (SNR) enhancing derivative calculation based on CWT using Haar wavelet function, which can start and end points of detected peaks accurately. The procedure for peak-width estimation basically consists of four steps: (1) CWT with Haar wavelet is performed at the same scales as in the peak detection step to obtain the 2D CWT coefficients; (2) replace each value in the matrix by its absolute value; (3) take out the row of best scale from the 2D Haar CWT coefficients, searching for the local minima from the peak index on both

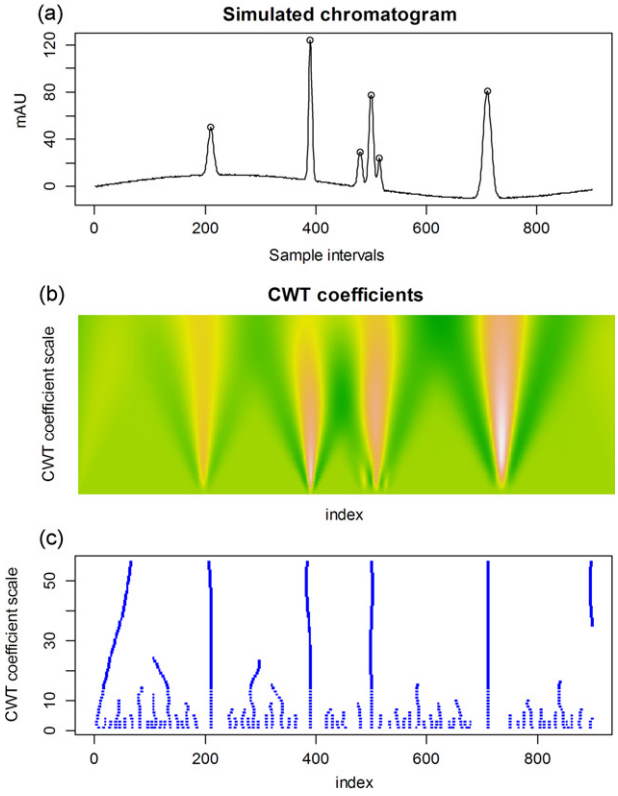

Fig. 2. Application of peak matching algorithm proposed by Du to chromatogram. (a) Chromatogram with the circles representing the detected peak positions. (b) Image of the 2D CWT coefficients. (c) The ridges lines identified from the 2D CWT coefficients.

sides within the range of three times its peak scale or the next peak index; (4) if local minimums exist, choose the minimum within the range of three times of its peak, else choose the local minimums between the peaks. This procedure is illustrated in Fig. 3. CWT with Haar wavelet was also performed from 1 to 56. The row of best scale from the 2D Haar CWT coefficients are taken out and plotted in Fig. 3(a). The taken out vectors are replaced with the absolute values of themselves, and the start and end points of the peaks were found according to steps 3 and 4, which were marked out with asterisks and circles.

2.3. Baseline correction

In this study, the used baseline correction algorithm is proposed by Zhang et al. [24,27]. The procedure for baseline fitting basically consists of three steps: (1) fit an rough background using penalized least squares algorithm with proper value of λ; (2) using rough background instead of original chromatogram to fit a newer rough background; (3) refining the newer rough background based on the peak position and peak width to obtain the final baseline.

One should fit out baseline and correct it, since baseline affects accuracy during calculating linear correlation coefficient. Utilizing peak detection results from previous section, an ideal fitting algorithm should match the baseline well at non-peak segment and could also just link the start and end points directly at peak segment. Here penalized least squares algorithm [27–30] is chosen for baseline fitting, which adapts to boundaries automatically, handles missing values, even in large stretches, automatically by introducing a vector of 0–1 weights, and is very efficient when sparse matrices are used.

Penalized least squares algorithm balanced between fidelity to the data and roughness of the fitting data by penalties on roughness

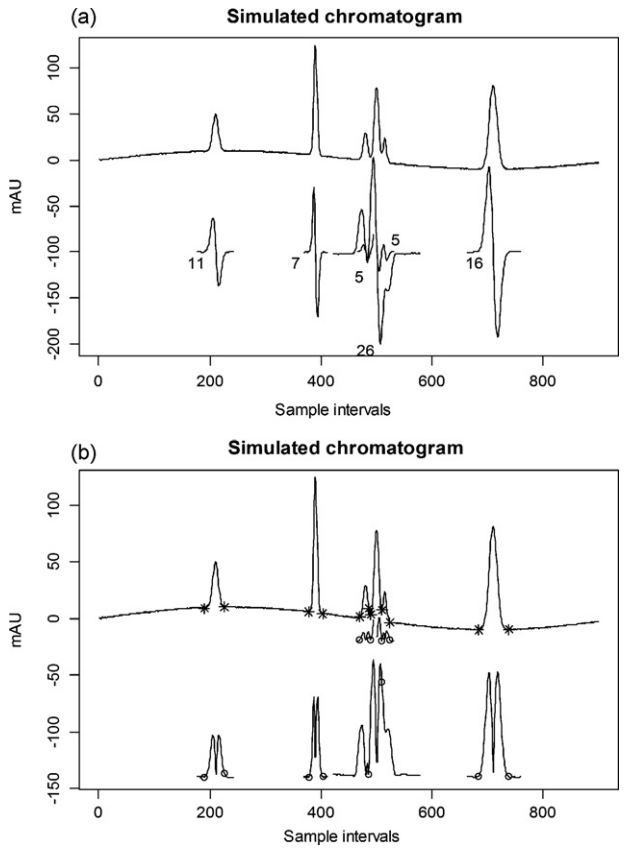

Fig. 3. Zhang's peak width estimation steps with application to chromatogram. (a) Detected peak derivative calculation using CWT with Haar wavelet, and plot of their optimum scales in the peak detection step. (b) Searching for the local minima to determine the start and end points of each peak. The nearest pairs of the local minimums to the corresponding peaks index are marked out with circles, and the estimated start and end points of peaks with asterisks.

of the fitting data. Fidelity of fitting vector z to chromatogram c can be measured using sum of squares of their differences using the formula:

$$F = \sum_{i=1}^{m} (c_i - z_i)^2 \tag{4}$$

Roughness of the fitting data z can be measured as squaring and summing its differences between neighbors, which can be expressed as:

$$R = \sum_{i=2}^{m} (z_i - z_{i-1})^2 = \sum_{i=1}^{m-1} (\Delta z_i)^2 \tag{5}$$

The balanced between these two goals can be measured as the fidelity plus with penalties on the roughness, and it can be given by:

$$Q = F + \lambda R = |\mathbf{c} - \mathbf{z}|^2 + \lambda |\mathbf{D}\mathbf{z}|^2 \tag{6}$$

where λ is an adjustable parameter, it can be adjusted by user. The larger λ is, the stronger the influence of R on the goal Q and the smoother $\mathbf{z}$ will be, but at the cost of a deterioration of the fit to the data (which is the penalization concept). $\mathbf{D}$ is the derivative of the identity matrix such that $\mathbf{D}\mathbf{z} = \Delta\mathbf{z}$. For example, when $m = 5$, $\mathbf{D}$ would be

$$\begin{bmatrix} -1 & 1 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & -1 & 1 \end{bmatrix} \tag{7}$$

By finding for the vector of partial derivatives and equating it to 0 ($\partial Q/\partial \mathbf{z}$) = $\mathbf{0}$, we get the linear system of equations that can be easily solved:

$$(\mathbf{I} + \lambda \mathbf{D}'\mathbf{D})\mathbf{z} = \mathbf{c} \tag{8}$$

Consider the peak segment as the missing value. This smooth algorithm can easily be modified to handle this situation. The peak segments of $\mathbf{x}$ set to an arbitrary value, say 0, and a vector $\mathbf{w}$ of weights is introduced, with $w_i = 0$ for missing observations and $w_i = 1$ otherwise. The fidelity of $\mathbf{z}$ to the chromatogram $\mathbf{x}$ is changed to

$$F = \sum_{i=1}^{m} w_i (c_i - z_i)^2 = (\mathbf{c} - \mathbf{z})'\mathbf{W}(\mathbf{c} - \mathbf{z}) \tag{9}$$

Here $\mathbf{W}$ is a diagonal matrix with $\mathbf{w}$ on its diagonal.

Eq. (8) changes to

$$(\mathbf{W} + \lambda \mathbf{D}'\mathbf{D})\mathbf{z} = \mathbf{W}\mathbf{c} \tag{10}$$

Solve Eq. (10), and one can obtain the fitting vector easily:

$$\mathbf{z} = (\mathbf{W} + \lambda \mathbf{D}'\mathbf{D})^{-1}\mathbf{W}\mathbf{c} \tag{11}$$

2.4. Peak clustering

Special peak regions, such as peaks with shoulder, overlapping peaks etc., needs to do some special treatments. One could see clearly from Fig. 3 that peak 3, peak 4 and peak 5 overlap heavily and are difficult to determine start point and end point individually. But the start point of peak 3 and end point of peak 5 can be determined accurately. So these peaks, whose distances to its neighbors are smaller than a certain threshold, are defined as a peak cluster. Peaks in a peak cluster are aligned simultaneously as one peak. Fig. 4 is an example of peak cluster. Three peaks in the rectangle in Fig. 4(a) change into a peak cluster in Fig. 4(b) with a gap threshold equals 3.

2.5. Peak alignment using differential evolution

Alignment problem can be transformed into optimization problem by maximizing linear correlation coefficient between reference chromatogram and chromatogram to be aligned. Accordingly, one should choose an optimization method simple to implement, easy to use, reliable and fast, which means that the selected method is reliably converge to the true optimum, spends acceptable time on searching for a solution and should be easy for an analyst but not an expert in optimization to solve a difficult optimization problem in analytical chemistry. Differential Evolution (DE) [23,31–33] is such a general-purpose method. It is a population-based optimizer, intelligent evolves to the true optimum with good probability using of differences between individuals.

DE algorithm was first proposed by Storn and Price [23]. It is a parallel direct search method utilizing N_D dimension parameter vector as a population for each generation G.

$$x_{j,G}, \quad j = 1, 2, \ldots, N_D \tag{12}$$

The basic structure of DE is initialization, mutation, crossover and selection. Once initialized, the mutation, crossover and selection steps are repeated to search the optimum.

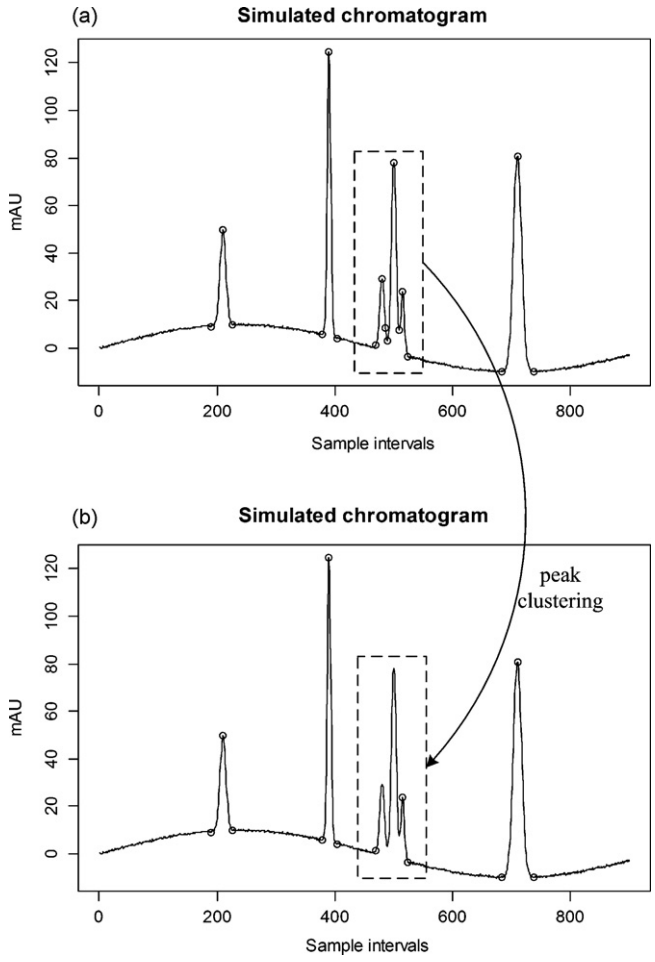

Fig. 4. Example of peak cluster with gap threshold equals 3. (a) Peaks before clustering, which is marked out with a rectangle. (b) Peak cluster after clustering, which is also marked out with a rectangle.

Initialization: The population should be initialized using upper and lower bounds for each parameter and a random number generator:

$$x_{j,i,0} = rand_j(0,1)\cdot(b_{j,U} - b_{j,L}) + b_{j,L}, \quad j = 1, 2, \ldots, N_D \quad \text{and} \quad i = 1, 2, \ldots, N_P \tag{13}$$

Here $x_{j,i,0}$ is the initial value of the jth parameter of the ith vector, $b_{j,U}$ and $b_{j,L}$ are the lower and upper bounds of the jth parameter, and N_P is the population size.

Mutation: Differential mutation adds a scaled, randomly sampled, vector difference to a third vector. A mutant vector is generated for each target vector according to

$$\mathbf{v}_{i,g} = \mathbf{x}_{r0,g} + F\cdot(\mathbf{x}_{r1,g} - \mathbf{x}_{r2,g}) \tag{14}$$

The scale factor, $F \in (0,1+)$, is a positive real number that controls the rate at which the population evolves. $r_0, r_1, r_2 \in \{1, 2, \ldots, N_P\}$ are random indexes.

Crossover: The uniform crossover is complement to the differential mutation search strategy. DE crosses each vector with a mutant vector to generate a trail vector:

$$\mathbf{u}_{i,g} = u_{j,i,g} = \begin{cases} v_{j,i,g} & \text{if } (rand_j(0,1) \le \text{Cr} \text{ or } j = j_{rand}) \\ x_{j,i,g} & \text{otherwise.} \end{cases} \tag{15}$$

$\text{Cr} \in [0,1]$ is crossover probability, which is defined by user to controls the fraction of parameter values that are copied from the mutant.

Selection: If $f(\mathbf{u}_{i,g})$ is lower than $f(\mathbf{x}_{i,g})$, $\mathbf{u}_{i,g}$ replace $\mathbf{x}_{i,g}$ in next generation; otherwise $\mathbf{x}_{i,g}$ retains.

$$\mathbf{x}_{i,g} + 1 = \begin{cases} \mathbf{u}_{i,g} & \text{if } f(\mathbf{u}_{i,g}) \le f(\mathbf{x}_{i,g}) \\ \mathbf{x}_{i,g} & \text{otherwise.} \end{cases} \tag{16}$$

The process of mutation, crossover and selection is repeated until the optimum is located, or one of the termination criterions is satisfied, such as the number of generations reaches the preset maximum generation (itermax).

On account of reducing the searching space in optimization [34], the peaks are grouped into 4 peaks a group by their order, and positions of peaks in the same group are optimized using DE one group per time. DE is used to adjust the n peaks' position of ith group with in the offset [−slack, +slack] simultaneously; warping the non-peak parts between peaks with linear interpolation; memorize peak positions that maximize linear correlation coefficient between corresponding parts of chromatogram to be align and corresponding parts of reference chromatogram. After dealing with all the peak groups, the memorized peak positions and linear interpolation warping are applied to reconstruct the aligned chromatogram.

The proposed method was tested both on simulated and real chromatograms. Firstly simulated data with known shifts were used to evaluate the accuracy of the proposed algorithm. Then chromatograms of fungal extracts obtained with HPLC-DAD at 202 nm were used to test the method's performance on real chromatograms.

2.6. Simulated chromatograms

Simulated chromatogram is the sum of Gaussian peaks, sinus curve baseline and random noise. Reference one is denoted as R, whose noise is normally distributed, with variance 0.2. The one, to be aligned, is denoted as C, but normally distributed noise with variance 1. The peaks of C were shifted by 50 positions from peaks of R except the second peak. Both R and C are created in R language, and shown in Fig. 5(a).

2.7. Real chromatograms

Real chromatograms are available from Refs. [11,35]. The chromatograms were analyses of extracts from fungal cultivated on Yeast Extract Sucrose agar (Penicillium cyclopium, denoted by IBT 11415 and 15670) using ultrasonic extraction and HPLC, which collected at the Department of Biotechnology (IBT), Technical University of Denmark. IBT11415 and IBT15670 were downloaded as an MATLAB™ 6 MAT-file from website mentioned in Ref. [11]. 2 UV spectra per second from 200 nm to 600 nm with a bandwidth of 4 nm resulted in 100 data points in each UV spectrum. The first column, representing the 202 ± 2 nm was selected to demonstrate the performance of aligned. IBT11415 was used as reference chromatogram, and IBT15670 was used as chromatogram to be aligned. Figure of these chromatograms is illustrated in Fig. 6(a).

Chromatograms, analyses of the Red Peony Root using HPLC, were selected to show how much peak heights and peak areas change due to this linear interpolation and quantify the retention time before and after alignment. 7 of Red Peony Roots were collected from different producing areas in China. Then a standard sample was purchased from the National Institute for control of Pharmaceutical and Biological Products. The HPLC experiments were performed at Chromap Co., Ltd., Zhuhai, China. The data were transformed into ASCII format using HP chemstations

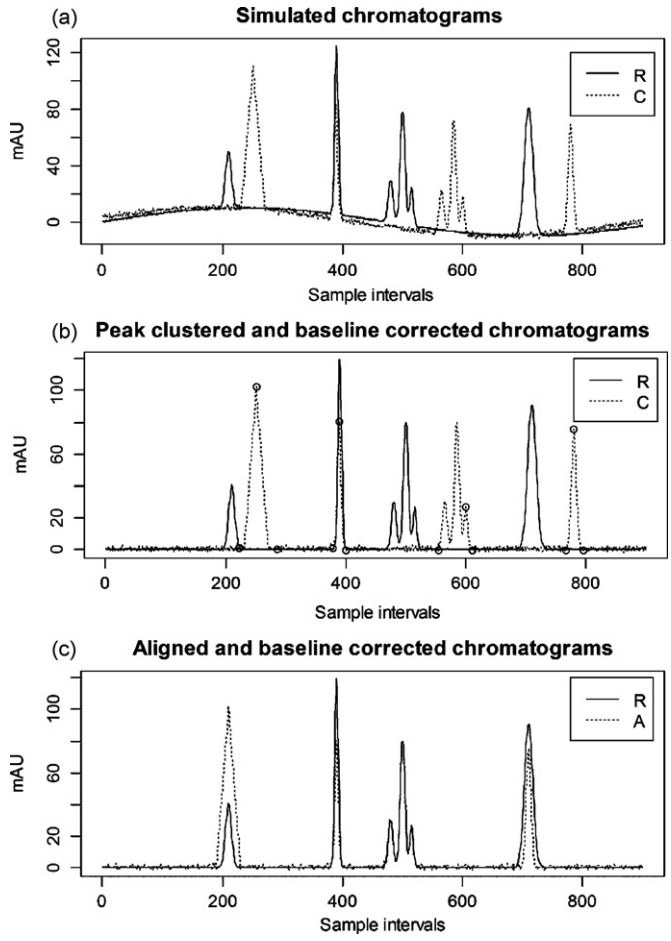

Fig. 5. Simulated chromatograms (R, reference chromatogram; C, chromatogram to be aligned). (a) Plots of simulated chromatograms before aligning, (b) after baseline correction and peak clustering, and (c) after aligning using alignDE.

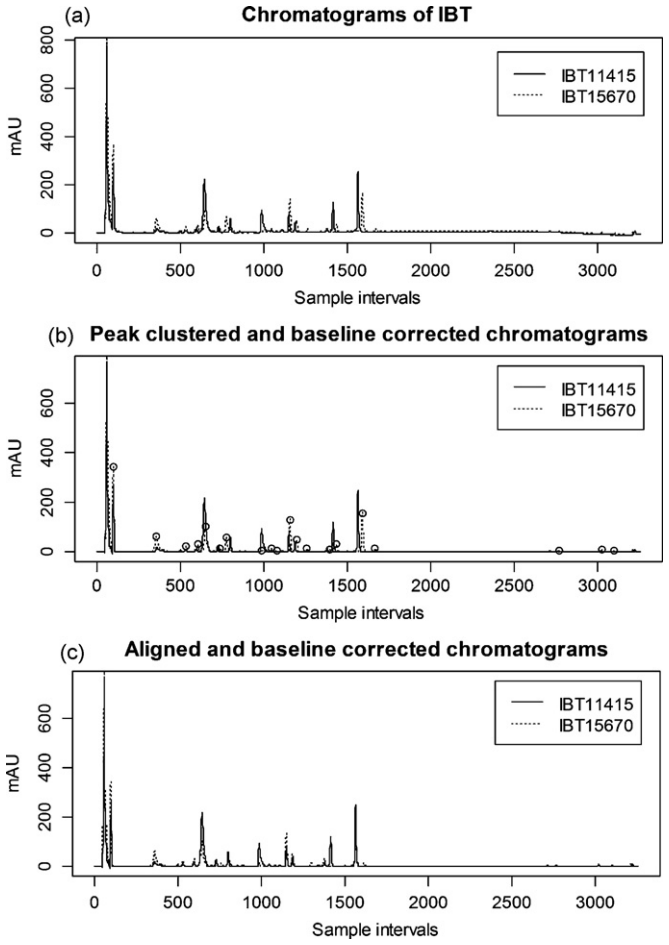

Fig. 6. Real chromatograms (IBT11415, reference chromatogram; IBT15670, chromatogram to be aligned). (a) Plots of real chromatograms before aligning, (b) after baseline correction and peak clustering, and (c) after aligning using alignDE.

(version A.09.01) for peak heights calculation and peak areas integral.

3. Results and discussion

Results on the alignment of both simulated and real life chromatograms, using the newly proposed method, will be presented to evaluate performance of the proposed method in this section. Then the results are discussed and the properties are explored in this section. In the end, effects of some parameters are also discussed to guide chromatograph users to use this method.

3.1. Result of simulated chromatograms

The parameters used in peak detection step were: SNR.Th = 3, ridgeLength = 5, $\lambda = 100$ and peak shape threshold = 0.3, and parameters used in peak clustering and peak aligning steps were: Gap = 5, slack = 100, $N_P = 200$, itermax = 150. Peak clustering and baseline correction results are shown in Fig. 5(b), and peak aligning results are shown in Fig. 5(c). One could obviously discover from Fig. 5 that both sinus baselines in R and C have been corrected satisfactorily; all the peaks positions, start point, and end point of them were detected with wavelet pattern matching exactly; all the peaks have been well aligned, which have demonstrated the method's ability to align chromatograms no matter special peak regions (such as peaks with shoulder, overlapping peaks, etc.), various baselines and different level noises.

3.2. Result of real chromatograms

The proposed method was also applied to real chromatograms of fungal extracts. Parameters are as follows: SNR = 3, ridge length = 5, $\lambda = 100$, peak shape threshold = 0.3, Gap = 5, slack = 100, $N_P = 200$, itermax = 150. It is seen from Fig. 6 that IBT11415 and IBT15670 are not well aligned originally, but after aligning using alignDE the alignment is much improved and visually good.

3.3. Comparison of alignDE with COW

Performance of the proposed method was also compared with performance of COW, which is evaluated by the form of linear correlation coefficient of two chromatograms, namely reference chromatogram and chromatogram to be aligned. The segment and slack parameters of COW are optimized using the "grid-search" method. The results are presented in Table 1. One could see from the table that alignDE for both simulated and real chromatograms gives results comparable with COW. The linear correlation coefficients of real chromatograms of COW are a little larger than alignDE. The reason is that COW will change shapes of some peaks during the interpolation, and it is prone to stretch or compress the peak, when there are some differences between corresponding peaks of the reference chromatogram and chromatogram to be aligned. This phenomenon could be seen in Fig. 7. Fig. 7(c) zooms in the first two peaks aligned using COW, and one can see that there are significant differences of peak 1 before and after aligning. It means that the high correlation coefficients obtained using COW is not

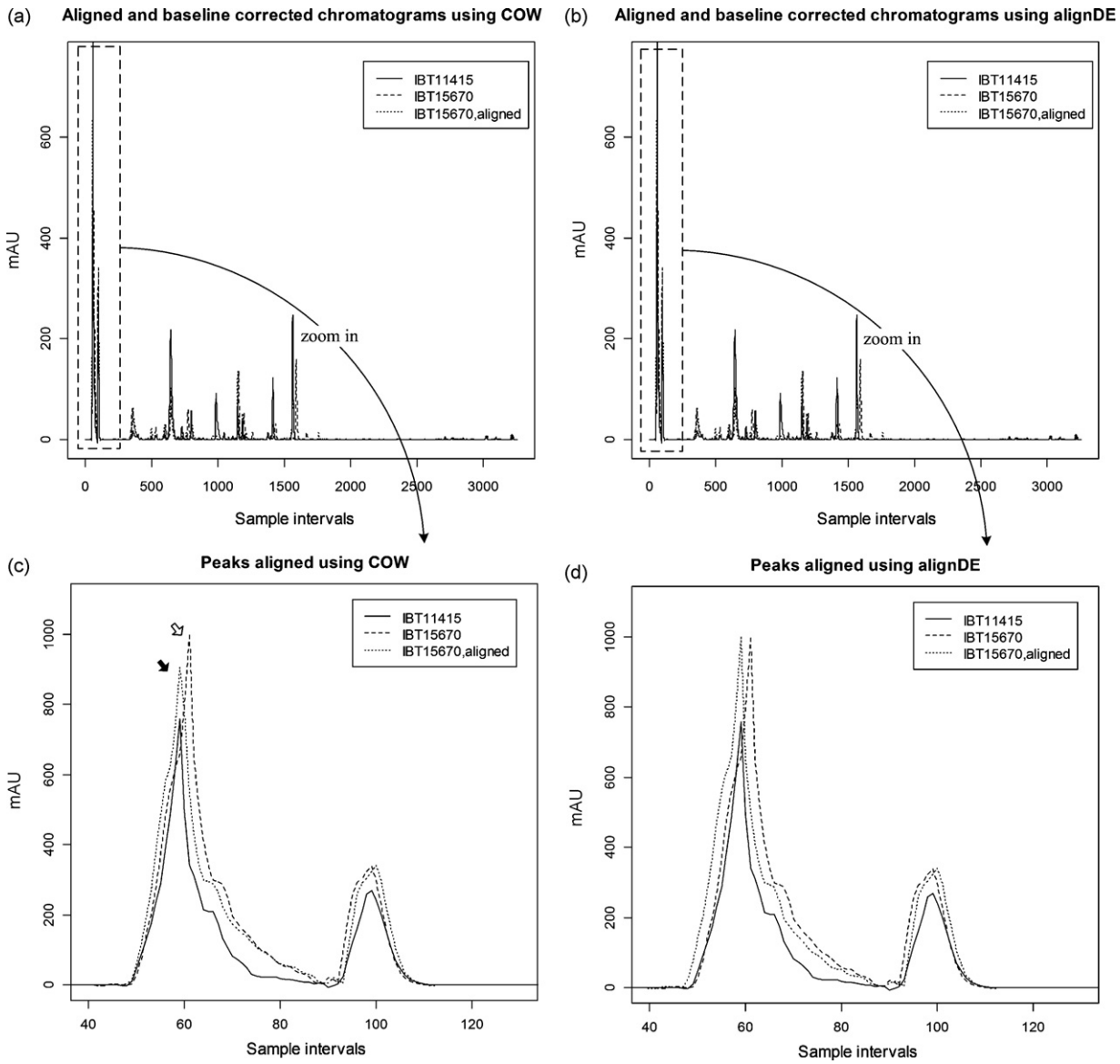

Fig. 7. Comparison the property of preserving area and shape of peaks using different method. (a) Full chromatograms aligned by COW, (b) full chromatograms aligned by alignDE, (c) zoom in the first two peaks of (a), and (d) zoom in the first two peaks of (b).

reliable. The detail of aligned peaks using alignDE can be seen in Fig. 7(d), which is aligned satisfactorily. The accurate and robust wavelet peak matching can guarantee no alteration in peak shape, peak height and peak area authentically.

3.4. Tuning parameters for peak matching and baseline correction

When comparing peak alignment parameters, i.e. peak detection, width estimation and baseline correction, one can infer that they are all insensitive to chromatograms with various baselines and noises. It is very easy to tune the parameters for peak matching and baseline correction, and key points are as followings: SNR, ridge length should be small enough to detect as many as possible peaks for clustering and aligning, and the ranges of SNR and ridge length should be within (0,3] and [5,10] respectively; λ controls smoothness of the fitted baseline, the lager the smoother, and the range of λ should be within [10,1000]; the gap parameter in peak clustering step merges close peaks in special peak regions together, whose value should be from 3 to 5. One can read Ref. [24] for detail.

Table 1
Comparison results of different methods of alignment.

Chromatograms	Baseline	Linear correlation coefficient		
		Initials	COW[a]	alignDE[b]
Simulated	Uncorrected	0.2401	0.6966	0.8059
	Corrected	0.1181	0.7133	0.7821
Real(IBT)	Uncorrected	0.8049	0.9287	0.9223
	Corrected	0.7862	0.9222	0.9189

[a] For all the datasets, segment = 22 and slack = 16.
[b] For all the datasets, slack = 100, N_P = 200, itermax = 150.

3.5. Tuning parameters for alignDE

The alignDE algorithm is not sensitive to the parameters, because the used peak detection, baseline correction and optimization algorithms are robust enough. The 7 user parameters are offered to the user for the complex situations. This aligning method can be tuned to obtain more accurate result and faster speed by several parameters. Three of them are of great importance, namely slack (number of points to shift peaks), N_P (number of populations of DE) and itermax (preset maximum generation of DE). Both slack

Table 2
Peak heights and peak areas changes due to this linear interpolation.

Before length equalizing			After length equalizing		
RT	Area	Height	RT	Area	Height
2.431167	5.201775e+002	6.025600e+001	2.428704	5.078457e+002	6.003727e+001
3.271167	7.180195e+003	1.547172e+003	3.274002	7.176067e+003	1.519094e+003
4.337833	1.626241e+003	4.179645e+002	4.34688	1.635648e+003	3.974140e+002
5.077833	1.714932e+003	2.341580e+002	5.078388	1.717186e+003	2.340413e+002
5.777833	1.954913e+002	1.882935e+001	5.777385	1.946253e+002	1.877934e+001
6.317833	2.791530e+002	2.549124e+001	6.313824	2.754173e+002	2.541697e+001
7.451166	3.098633e+002	2.961684e+001	7.451725	3.149424e+002	2.960835e+001
7.6045	1.383203e+002	1.873398e+001	7.614282	1.346856e+002	1.859885e+001
8.4245	3.234238e+002	3.192044e+001	8.427069	3.333269e+002	3.190882e+001
9.497833	4.364499e+002	4.310560e+001	9.499947	4.208180e+002	4.305978e+001
9.6645	3.115619e+002	2.573776e+001	9.662504	3.225781e+002	2.564828e+001
10.0645	2.189177e+002	2.352190e+001	10.0689	2.171675e+002	2.338090e+001
10.6445	6.423835e+002	6.551838e+001	10.6541	6.399862e+002	6.527614e+001
11.8045	5.518335e+002	4.334975e+001	11.80826	5.554847e+002	4.331932e+001
13.57117	4.398410e+002	2.602434e+001	13.58014	4.380027e+002	2.587457e+001
14.1245	2.704960e+002	1.906443e+001	14.11657	2.724714e+002	1.896465e+001
14.85783	4.825620e+002	2.619410e+001	14.86434	4.817871e+002	2.610096e+001
16.7245	2.509267e+004	1.795248e+003	16.73375	2.509593e+004	1.793913e+003
18.85117	2.863715e+003	1.394339e+002	18.84699	2.861342e+003	1.392666e+002
21.47783	1.654988e+003	1.058879e+002	21.48042	1.654171e+003	1.058568e+002
24.39117	1.255558e+003	7.891274e+001	24.3902	1.255913e+003	7.890025e+001
26.25783	5.283034e+003	3.340859e+002	26.2596	5.283775e+003	3.340504e+002
32.23783	6.414516e+003	2.849059e+002	32.24171	6.414657e+003	2.848560e+002
34.9645	4.613905e+002	2.240944e+001	34.97268	4.612419e+002	2.240658e+001
48.5045	5.483888e+002	5.049467e+001	48.5137	5.487869e+002	5.046863e+001
55.15783	4.539447e+002	2.483320e+001	55.16229	4.539947e+002	2.483257e+001

and N_P are easy to choose. To select the value of slack, one can observe the chromatograms and estimate the largest shifts number points as slack. The proper N_P is doubled slack. The reason is that values of initial population are generated by a random number generator, which uniformly distributed within the range [−slack, slack]. The itermax parameter is affiliated with convergence speed of DE. To investigate the convergence speed of alignDE, both simulated and real chromatograms are used with different itermaxs (from 30 to 300). The results are shown in Fig. 8. Fig. 8(a) is the convergent test of alignDE using simulated data, and one could see that alignDE converges swiftly within 80 iterations. Fig. 8(b) is the convergent test using real chromatograms, and alignDE also converges swiftly within 150 iterations. It is concluded that parameters of alignDE can be determined using following procedure: (1) observe the chromatograms and estimate the largest shifts number points as the slack; (2) number of populations of DE (N_P) is initialized with the doubled slack; (3) itermax parameter should set the value to 150 or higher to assure convergence according to the convergent test.

3.6. Peak heights and areas change due to linear interpolation

There are 8 chromatograms of Red Peony Root. The number of chromatogram with 9751 points is 7, and the third one is chromatogram with 9752 points. Since there are unequal data points among chromatograms, the chromatogram lengths were equalized to 4000 points by linear interpolation. C-09-02 was chosen to study that how much peak heights and peak areas change due to this linear interpolation. The used peak detection method is local maximum method with threshold values of peak height and peak area. The peak heights and peak areas were listed in Table 2. The changes in peak heights and peak areas due to linear interpolation are acceptable.

3.7. Quantitative analysis of alignment results

The lengths of 8 chromatograms of Red Peony Root were equalized by linear interpolation to 4000 firstly. Then alignDE algorithm

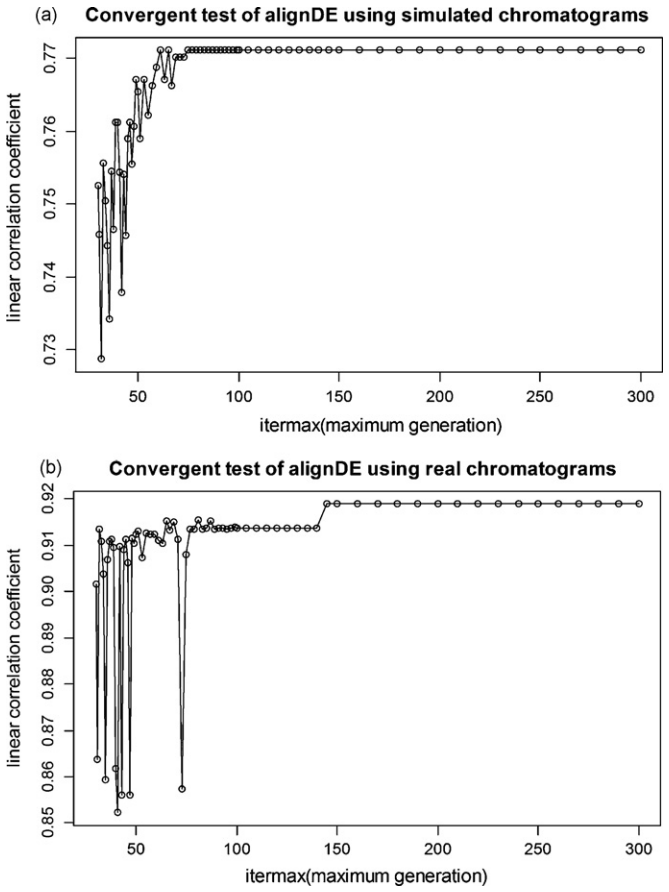

Fig. 8. Convergent test of alignDE using both simulated and real chromatograms. (a) Convergent test using simulated data and (b) convergent test using real chromatograms.

Table 3
Quantitative results of the retention times of Red Peony Root chromatograms before and after alignment using alignDE.

Sample number	Alignment	Peaks retention times (min)				Correlation coefficient
		P1	P2	P3	P4	
Reference	No	3.275996	16.669344	26.242987	32.208229	1.0000
C-09-001	No	3.259742	16.685598	26.210479	32.208229	0.9743
	Yes	3.259742	16.701852	26.242987	32.208229	0.9809
C-09-002	No	3.275996	16.718106	26.259241	32.208229	0.9797
	Yes	3.275996	16.701852	26.242987	32.208229	0.9836
C-09-003	No	3.275996	16.701852	26.242987	32.191975	0.9794
	Yes	3.275996	16.701852	26.242987	32.208229	0.9795
C-09-004	No	3.275996	16.636836	26.291750	32.191975	0.9856
	Yes	3.275996	16.653090	26.242987	32.208229	0.9955
C-09-005	No	3.259742	16.685598	26.308004	32.191975	0.9868
	Yes	3.259742	16.701852	26.242987	32.224483	0.9901
C-09-006	No	3.275996	16.685598	26.291750	32.191975	0.9811
	Yes	3.275996	16.685598	26.242987	32.208229	0.9834
C-09-007	No	3.259742	16.685598	26.438036	32.256991	0.9625
	Yes	3.275996	16.669344	26.356766	32.224483	0.9788

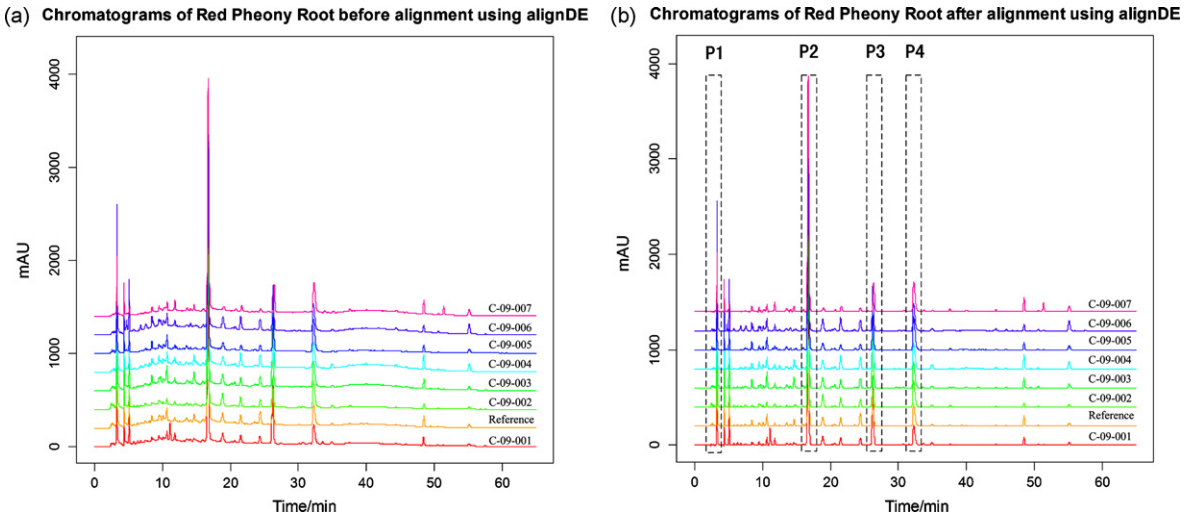

Fig. 9. Chromatograms of Red Peony Root before and after alignment using alignDE.

was applied with parameters as follows: SNR = 1, ridge length = 10, λ = 100, peak shape threshold = 0.3, Gap = 5, slack = 5, N_P = 30, itermax = 300. The positions before and after alignment of four large peaks are detected and listed in Table 3, and four large peaks are detected using wavelet peak matching and marked out with red rectangle in Fig. 9. By comparison the retention times of matched peaks before and after alignment with peaks of the reference chromatogram, Table 3 indicates that peak1, peak3 and peak4 are well aligned. The retention times of peak2 are not exactly as it of reference chromatogram. The reason is that the vertex of the peak2 consists of several points and is not as sharp as other peaks. When matching the peak using wavelet, the detected positions vary from sample to sample. (Notice: Since the peak detection methods are different, the positions of the peaks are not the same in Tables 2 and 3.)

4. Conclusion

In this study, a practical and handy peak alignment method is proposed with wavelet pattern matching and differential evolution optimizing, which can provide user a baseline corrected, aligned chromatogram with its peaks list simultaneously. The wavelet peak matching algorithm is robust and accurate to detect peak position, peak's start point and end points. Attributing to the differential evolution method, alignDE converges swiftly. The testing with simulated and real chromatogram demonstrates its performance. The parameters are intuitional and easy to adjust. It is the off-the-shelf aligning tool for all chromatograph users.

Acknowledgments

This work is financially supported by the National Nature Foundation Committee of P.R. China (Grants No. 20875104 and Grants No. 10771217), the international cooperation project on traditional Chinese medicines of ministry of science and technology of China (Grant No. 2007DFA40680), the Academic Scholarship of Ministry of Education for Doctoral Postgraduate, Hunan Provincial Innovation Foundation For Postgraduate, the Graduate degree thesis Innovation Foundation of Central South University and Special Funds of Central South University for Fostering Outstanding Doctoral Degree Thesis (Grant No. 2009yb0XX). 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. Also, the authors want to thank Peishan Xie of Chromap Co., Ltd., Zhuhai, China for providing the chromatograms dataset.

References

[1] S. Wold, K. Esbensen, P. Geladi, Chemom. Intell. Lab. Syst. 2 (1987) 37–52.
[2] S. Wold, M. Sjotrom, L. Eriksson, Chemom. Intell. Lab. Syst. 58 (2001) 109–130.
[3] G. Malmquist, R. Danielsson, J. Chromatogr. A 687 (1994) 71–88.
[4] A.M. van Nederkassel, C.J. Xu, P. Lancelin, M. Sarraf, D.A. MacKenzie, N.J. Walton, F. Bensaid, M. Lees, G.J. Martin, J.R. Desmurs, D.L. Massart, J. Smeyers-Verbeke, Y. Vander Heyden, J. Chromatogr. A 1120 (2006) 291–298.
[5] 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) 56–66.
[6] 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. Inform. Model. 46 (2006) 863–875.
[7] R. Siuda, G. Balcerowska, D. Aberdam, Chemom. Intell. Lab. Syst. 40 (1998) 193–201.
[8] R.H. Jellema, in: D.B. Stephen, T. Roma, W. Beata (Eds.), Comprehensive Chemometrics, Elsevier, Oxford, 2009, pp. 85–108.
[9] C.P. Wang, T.L. Isenhour, Anal. Chem. 59 (1987) 649–654.
[10] K. Athanassios, F.M. John, A.T. Paul, AIChE J. 44 (1998) 864–875.
[11] N.-P.V. Nielsen, J.M. Carstensen, J. Smedsgaard, J. Chromatogr. A 805 (1998) 17–35.
[12] P.H.C. Eilers, Anal. Chem. 76 (2004) 404–411.
[13] D. Bylund, R. Danielsson, G. Malmquist, K.E. Markides, J. Chromatogr. A 961 (2002) 237–244.
[14] V. Pravdova, B. Walczak, D.L. Massart, Anal. Chim. Acta 456 (2002) 77–92.
[15] G. Tomasi, F. van den Berg, C. Andersson, J. Chemometr. 18 (2004) 231–241.
[16] T. Skov, F. van den Berg, G. Tomasi, R. Bro, J. Chemometr. 20 (2006) 484–497.
[17] A.M. van Nederkassel, M. Daszykowski, P.H.C. Eilers, Y.V. Heyden, J. Chromatogr. A 1118 (2006) 199–210.
[18] B. Walczak, W. Wu, Chemom. Intell. Lab. Syst. 77 (2005) 173–180.
[19] J.T.W.E. Vogels, A.C. Tas, J.v.d. Venekamp, J. Greef, J. Chemometr. 10 (1996) 425–438.
[20] R.J.O. Torgrip, M. Aberg, B. Karlberg, S.P. Jacobsson, J. Chemometr. 17 (2003) 573–582.
[21] M.D. Krebs, R.D. Tingley, J.E. Zeskind, M.E. Holmboe, J.-M. Kang, C.E. Davis, Chemom. Intell. Lab. Syst. 81 (2006) 74–81.
[22] J. Forshed, I. Schuppe-Koistinen, S.P. Jacobsson, Anal. Chim. Acta 487 (2003) 189–199.
[23] R. Storn, K. Price, J. Global Optim. 11 (1997) 341–359.
[24] Z.M. Zhang, S. Chen, Y.Z. Liang, Z.X. Liu, Q.M. Zhang, L.X. Ding, F. Ye, H. Zhou, J. Raman Spectrosc. 41 (2010) 659–669.
[25] C. Yang, Z.Y. He, W.C. Yu, BMC Bioinformatics 10 (2009), Article No.: 4.
[26] P. Du, W.A. Kibbe, S.M. Lin, Bioinformatics 22 (2006) 2059–2065.
[27] Z.M. Zhang, S. Chen, Y.Z. Liang, Analyst 135 (2010) 1138–1146.
[28] E.T. Whittaker, P. Edinburgh, Math. Soc. 41 (1922) 63–75.
[29] P.H.C. Eilers, Anal. Chem. 75 (2003) 3631–3636.
[30] J. Carlos Cobas, M.A. Bernstein, M. Mart-Pastor, P.G. Tahoces, J. Magn. Reson. 183 (2006) 145–151.
[31] K.V. Price, R.M. Storn, J.A. Lampinen, Differential Evolution: A Practical Approach to Global Optimization, Springer, New York, 2005.
[32] V. Feoktistov, Differential Evolution: in Search of Solutions, Springer, New York, 2006.
[33] D. Ardia, DEoptim: Differential Evolution Optimization in R, 2007, (http://cran.r-project.org/package=DEoptim).
[34] J. Friedman, T. Hastie, R. Tibshirani, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, New York, 2001.
[35] J. Smedsgaard, J. Chromatogr. A 760 (1997) 264–270.