Analytical Methods

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/c7ay00268h

Discrete wavelet assisted correlation optimised warping of chromatograms: optimizing the computational time for correcting the drifts in peak positions†

Keshav Kumar

Received 28th January 2017
Accepted 24th February 2017

DOI: 10.1039/c7ay00268h

rsc.li/methods

Correlation optimised warping (COW) has been the most favourite chromatographic peak alignment approach in recent years. After optimization of the two parameters, slack and segment length, COW works well in aligning the chromatograms. However, one of the serious disadvantages of COW is that it is computationally time consuming. Often several segment lengths and slack parameters need to be tested to find the optimum combination for achieving the alignment that makes the whole analysis take several hours. In the present work, it has been shown that with the application of wavelet analysis prior to alignment it is possible to provide the necessary computational economy to the COW algorithm.

1. Introduction

Data analysis workflow requires constant improvement to ensure fast and efficient analysis of a large amount of chromatographic datasets. For a successful application of any data analysis workflow the same chromatographic peak must appear at the same retention time point across the sample set.[1–3] Thus, correction of the retention time drifts becomes the bottleneck for the successful implementation of the data analysis workflow. There are different alignment approaches available in the literature that can broadly be classified into two classes: (i) insertion and deletion and (ii) compression and expansion.[3–8] Of the two, the compression and expansion based alignment approach has been found to work best for chromatograms consisting of several peaks with complex (*i.e.* irregular) retention time drifts. Correlation optimised warping (COW)[1–6] has been the most commonly used technique belonging to the compression and expansion model. COW ensures that the shape and area of the peaks are preserved.[3–8] Successful application of the COW algorithm also requires optimization of the two parameters, slack (t) and size of segments (m). The COW algorithm also suffers from a great disadvantage of being computationally time consuming. For a given slack and segment length, the alignment for few chromatograms consisting of a few thousands of data points can take hundreds of minutes. Often several combinations of slack and segments need to be tested to achieve the optimum alignment of the chromatograms and hence the whole data analysis workflow can take several days to obtain any meaningful results for the analysed peptidoglycans. Implementing a pre-processing step prior to COW that can reduce the dimension of the dataset while preserving all chemically relevant information can really catalyse the COW analysis and hence in result can speed up the entire data analysis workflow.

Modern chromatographic instruments are quite sensitive and can be used to analyse the constituents even at the trace levels. However, the sensitivity of the instruments tends to drop down with use and the signals tend to have strong interferences from the background noises originating due to the depletion or malfunction of different electrical, optical and mechanical components. Thus, in order to either push the sensitivity limits further or to help in maintaining the constant sensitivity of the instrument, or to enhance the weak signals it is also important to introduce signal denoising steps in the data analysis workflow to ensure unbiased chromatographic data analysis.

In recent years, wavelet analysis has gained its ground in analytical chemistry.[9–24] Wavelets are oscillating waves with null moments.[9–13] Similar to Fourier analysis, it also takes the signal from one domain (*e.g.* time domain) and represents it as another (*i.e.* frequency). However, compared to FFT it ensures simultaneous localization of the signal in both frequency and time domains.[9–25] Wavelet analysis essentially decomposes a signal into its high and low frequency components at different levels. There are different bases in the wavelet analysis algorithm that can be used to approximate the signals. Wavelet analysis can be classified as discrete wavelet transformation (DWT) analysis and continuous wavelet transformation (CWT) analysis.[9–13] Most of the analytical processes including chromatography generate datasets that are discrete in nature. For

Department of Molecular Biology, Umeå University, 90187 Umeå, Sweden. E-mail: keshavkumar29@gmail.com

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c7ay00268h

example, a chromatogram represents the intensity at a discrete time point and each time point is further stepped up with a constant discrete time interval. DWT analysis is ideal for processing such discrete datasets. DWT has been mostly applied for pre-processing the raw datasets belonging to MALDI imaging and similar fields.[26–29] However, the potential of DWT analysis lies towards achieving two important goals in the chromatographic data analysis workflow: (i) optimising the computational time for the COW algorithm and simultaneously (ii) improving the signal to noise ratio; these goals have not been tested so far. The present work deals with these two aspects and proposes a DWT assisted COW alignment approach to achieve a fast and sensitive data analysis workflow for aligning the chromatographic datasets.

2. Theory

2.1 Wavelet analysis

Wavelet analysis[9–24] allows the mapping of a signal from the time domain to the frequency domain retaining both frequency and time information. The wavelet analysis essentially measures the degree of similarity between the original signal and the mother wavelets. The wavelet family consists of several classes of wavelets such as Haar, Daubechies, Symlets, and Coiflets.[9–13,16,18,23] The following wavelets, haar, db2–db10, sym1–sym8 and coif1–coif5 belonging to these wavelet families are shown in Fig. S1 of the ESI.† The prefixes haar, db, sym and coif indicate that they belong to Haar, Daubechies, Symlet and Coiflets wavelet families. The numerical suffix defines the number of vanishing moments, for example db8 indicates that this wavelet has 8 vanishing moments. These wavelets differ in their unique properties such as the number of vanishing moments, symmetry, regularity, support size, presence of scaling function *etc.* Each wavelet is appropriate for a certain range of applications. The wavelet analysis can be divided into continuous wavelet transform (CWT) analysis[9–13] and discrete wavelet transform (DWT) analysis.[9–24] As discussed earlier, between these two classes of wavelet transform, DWT is preferred because the analytical data obtained with modern analytical instruments are discrete in nature. DWT analysis[9–24] involves the decomposition of the signal into two sets of coefficients called approximation (A) and detail (D) coefficients. The A and D coefficients represent the signal and noise components of the original signal, respectively. These coefficients are obtained by convolving the signal with low and high pass filters. The convolution of the signal with the low pass filter and high pass filter followed by dyadic decimation (*i.e.* down sampling) generates approximation and detail coefficients, respectively. In the next step the approximation coefficients are further split into approximation and detail coefficients by convoluting the signal with high and low pass filters followed by the downsampling. The process is repeated to the desired level of decomposition. It is to be noted that with each level of decomposition the size of the approximation and detail coefficients reduces by a factor of 2. Fig. S2 in the ESI† shows the general scheme of DWT based signal decomposition. A more detailed theoretical discussion on wavelet analysis can be seen elsewhere in the literature.[9–24]

2.2 Correlation optimised warping (COW) algorithm

The COW algorithm[3–5,30–32] aligns the two chromatograms by piecewise expansion and compression. The quality of alignment is evaluated using the Pearson correlation coefficient (ρ) between the aligned segments of the chromatograms. In order to understand the working mechanism of COW, two chromatograms of length L are considered. Of these two chromatograms, one is selected as the target (T) and the second chromatogram (R) needs to be aligned with it. Both the chromatograms are divided into N segments of length m ($=L/N$). Each segment (w) is warped (*i.e.* expanded or compressed) using linear interpolation. The amount of warping for each segment is controlled by the parameter called slack (t). Except the first and last segments the remaining N-2 segments can be warped on both sides in the range $[-t, +t]$. The start and end points of the chromatograms are fixed in the COW algorithm, therefore the first and last segments can only be warped on the right and left ends, respectively. The scheme of COW algorithm is shown in Fig. S3 of the ESI.† The correlation coefficient between the segment of target chromatogram (T) and aligned chromatogram (R_1) can be calculated using eqn (1):

$$\rho = \frac{(w_{\mathrm{T}} - \mathrm{mean}(W_{\mathrm{T}}))^{\mathrm{T}}(w_{\mathrm{R}_1} - \mathrm{mean}(W_{\mathrm{R}_1}))}{\mathrm{std}(w_{\mathrm{T}})\mathrm{std}(w_{\mathrm{R}_1})} \tag{1}$$

where w_T and w_{R_1} are the segments of target and aligned chromatograms; mean (w_T) and mean (w_{R_1}) are the mean values of target and aligned chromatograms, respectively; std (w_T) and std (w_{R_1}) are the standard deviation values of target and aligned chromatograms, respectively. The dynamic programming approach is used to find the optimum warping combination for each of the N segments. This approach involves the calculation of two matrices F and U of equal dimensions. The matrix F contains benefit function values and the matrix U contains the control input. The matrix U containing optimal warping for each segment is used to reconstruct the signal. The elaborate explanation of the theoretical aspects of COW can be found elsewhere.[3–5,30–32]

One of the severe limitations of COW is the significantly high computational time that depends primarily on two parameters, segment length (m) and slack (t). For a given segment length the computational time increases by a factor equal to $t^{1/2}$. Therefore, alignment of the two chromatograms consisting of a large number of data points will require a significantly large computational time compared to the alignment of the chromatograms containing a smaller number of data points. As discussed above, the application of DWT analysis to the chromatogram at an appropriate decomposition level can reduce the size of the dataset, while retaining all the information of the original chromatogram can really be useful in reducing or optimising the computational time of the COW algorithm.

3. Materials and methods

3.1 Simulation of the chromatograms

In total, 21 chromatograms, labelled as S1–S21, are simulated using the Gaussian functions, shown in eqn (2). The simulated chromatograms differ in some or all of these three aspects: (i) number of peaks, (ii) peak positions and (iii) amplitude of the peaks. There is deliberate choice in simulating the chromatograms with complex features. A random noise of significant intensity is also added to the simulated chromatograms. It is necessary to have these complexities in chromatograms to evaluate the proposed DWT assisted COW alignment of the chromatograms.

$$f(x)=\sum_{i=1}^{n}a_i\exp\left(-\left[\frac{(x-p_i)}{w_i}\right]^2\right)\tag{2}$$

In the above equation, a_i, p_i, w_i and n are the amplitude, peak position, width and number of simulated peaks.

3.2 Software and codes used

All the analyses are carried out on the MATLAB platform (version 2008 b). The wavelet analyses are performed using the wavelet toolbox of MATLAB. The COW analysis is carried out using the codes available on the website http://www.models.life.ku.dk/DTW_COW. The chromatographic data of the nine coffee samples are also taken from the above website.

4. Results and discussion

4.1 COW based alignment of the chromatograms

The simulated chromatograms with randomly added noises are shown in Fig. 1(A) and (B). It is clear that chromatograms have complex retention drifts in the peak position and it is indeed a computational challenge to fix it. As it is known, the application of COW alignment requires the selection of the reference chromatograms. An ideal reference chromatogram is the one that has all the features present in the other chromatograms. In principal, any of the following chromatograms S3, S7, S10, S11, S14 and S15 can be taken as the reference. In the present work a random choice is made and S10 is selected among the available five chromatograms. Each of the remaining fourteen chromatograms has different shifts in the peak position, thus they will need a different slack and segment length for achieving the optimal alignment with the reference chromatogram. To start with, an attempt is made to align the sample S1 with S10 with a slack of 250 and segment length of 350. The obtained COW retrieved S1 is compared with S10 and shown in Fig. 1(C). The COW analysis with these parameters apparently did not work out. The most critical thing that needs to be seriously considered is the computational time invested to test the applicability of the selected segment length and slack parameters. For example, here in this analysis it consumed more than 2700 seconds to test and get the negative results. Often, several combinations of the slack and segment length need to be tested to achieve the optimal alignment. Testing the different combinations of slack and segment length for each of the chromatograms can take several hours making the whole analysis seriously laborious and computationally challenging. In order to speed up the alignment approach it is necessary that a suitable pre-processing step is added prior to COW analysis. The pre-processing step should be able to filter out the noise and reduce the dimension of the dataset while preserving all the essential features of the chromatograms.

4.2 DWT assisted COW (DWT-COW) alignment on the chromatograms

Towards optimising the computational time, the present work proposes discrete wavelet transform (DWT) analysis as a pre-processing technique prior to COW analysis. In order to proceed, a mother wavelet needs to be selected from the available wavelet families. The selection needs to be done in a judicious way because it can affect the accuracy of the wavelet analysis of the chromatograms. In the present work, the wavelet that maximises the energy to Shannon entropy ratio of the approximation coefficients is taken as the mother wavelet for analysing the chromatographic datasets.[16] In other words, a wavelet is considered as the mother or base wavelet if it captures the maximum amount of energy from the chromatogram and also minimises the entropy of the wavelet coefficients. The energy to Shannon entropy ratio represented as R can be calculated using eqn (3) given below.

$$R=\frac{\text{energy}}{\text{Shannon entropy}}=\frac{\sum_{i=1}^{N}|w(s,i)|^2}{-\sum_{i=1}^{N}p_i\log 2\,p_i}\tag{3}$$

In the above equation, N is the number of wavelet coefficients, $w(s,i)$ represents the wavelet coefficients and p_i is the energy probability distribution of the wavelet coefficients.

The energy and Shannon entropy are calculated for each of the 21 chromatograms with all the commonly used 22 wavelets *i.e.* Haar, db2–db10, sym2–sym8 and coif1–coif5. The energy and Shannon entropy values for all the chromatograms obtained with a specific wavelet are averaged to obtain the mean energy and mean Shannon-entropy of the datasets using eqn (4) and (5) given below.

$$\text{Mean energy}=\frac{\sum_{k=1}^{K}(\text{energy})_k}{K}\tag{4}$$

$$\text{Mean Shannon entropy}=\frac{\sum_{k=1}^{K}(\text{Shannon entropy})_k}{K}\tag{5}$$

In the above equations, $(\text{energy})_k$ and $(\text{Shannon entropy})_k$ are the energy and Shannon entropy of the k^{th} chromatogram. The mean energy and mean Shannon-entropy of the datasets are divided as shown in eqn (6) to obtain the energy to Shannon entropy ratio (R)

$$R=\frac{\text{mean energy}}{\text{mean Shannon entropy}}\tag{6}$$

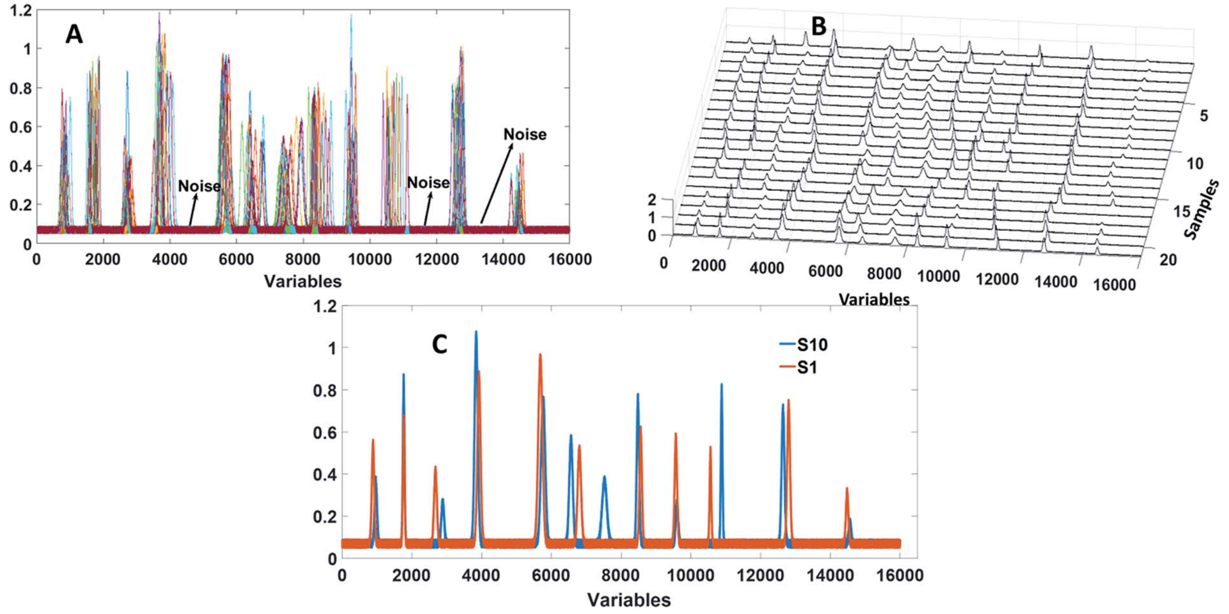

Fig. 1 (A) Simulated chromatograms with a significant amount of noise, (B) complex different in the peak position in the three dimensional plot of the simulated chromatograms, (C) COW retrieved S1 chromatogram and reference S10 chromatogram, the COW is carried out with a segment length of 350 and slack of 250.

It is also important to mention that analyses of the simulated chromatograms with all the selected 22 wavelets are found to be optimum at level 4. At higher levels (>4) the shapes of the wavelet retrieved chromatograms are found to be distorted whereas at lower levels (<4) the denoising of the chromatogram could not be achieved to a satisfactory level. The mean energy and mean Shannon entropy values and their ratio R for each of the 22 wavelets at level 4 are summarised in Table 1. From the reported values, it can be seen that except the Haar wavelet, all the remaining wavelets are found to have similar energy to Shannon entropy ratios. However, comparatively db5 is found to maximise the energy and minimise the entropy and as a result have a relatively higher energy to Shannon entropy ratio. Therefore, db5 is selected as the mother wavelet to analyse the chromatograms.

The db5 wavelet at level 4 is applied for each of the 21 chromatograms. The db5 wavelet decomposes the chromatogram in four levels. Each level consists of two types of coefficients, A (approximation) and D (details). In the first step, db5 takes the chromatogram and decomposes it into $A1$ and $D1$, in the second step it decomposes $A1$ into $A2$ and $D2$, in the third step it decomposes $A2$ into $A3$ and $D3$, and in the fourth and last step it decomposes $A3$ into $A4$ and $D4$. Each level reduces the dimension of the datasets by half. Therefore, with the db5 wavelet at level 4, it is possible to represent a chromatogram consisting of 16 000 points with $A4$ comprising just 1000 data points. The db5 decomposition of the S1 chromatogram is shown in Fig. 2. It can be seen that application of DWT allows simultaneous reduction of the data size and denoising of the chromatograms.

The db5 pre-processed chromatograms of all the 21 samples are shown in Fig. 3(A) that are further subjected to the COW analysis. It is reported in the literature that for a given segment length, calculation time for COW is directly proportional to the square of the slack parameter.[3–5,30–32] A chromatogram consisting of few data points would require a smaller magnitude of slack while performing the COW analysis and thus providing the reduction of the calculation time. *In a nutshell, there is a golden principle that the smaller the data points the* shorter *the*

Table 1 The mean energy, mean Shannon entropy and mean energy to mean Shannon-entropy ratio for 22 different wavelets for the S1–S21 chromatograms

Wavelet	Mean energy	Mean Shannon entropy	R = mean energy/(mean Shannon entropy)
Haar	99.1079	−481.2649	−0.205932
db2	99.5654	−489.8053	−0.203275
db3	99.5696	−489.8945	−0.203247
db4	99.5731	−489.9772	−0.203220
db5	99.5744	−490.0408	−0.203196
db6	99.5730	−490.0087	−0.203207
db7	99.5739	−490.0267	−0.203201
db8	99.5736	−490.0175	−0.203204
db9	99.5735	−490.0086	−0.203208
db10	99.5737	−490.0051	−0.203209
sym2	99.5654	−489.8053	−0.203275
sym3	99.5696	−489.8945	−0.203247
sym4	99.5740	−489.9878	−0.203217
sym5	99.5748	−489.9761	−0.203224
sym6	99.5736	−489.9878	−0.203216
sym7	99.5738	−489.9323	−0.203240
sym8	99.5733	−490.0060	−0.203208
coif1	99.5620	−489.7411	−0.203295
coif2	99.5710	−489.9447	−0.203229
coif3	99.5731	−490.0129	−0.203205
coif4	99.5736	−490.0310	−0.203199
coif5	99.5735	−490.0076	−0.203208

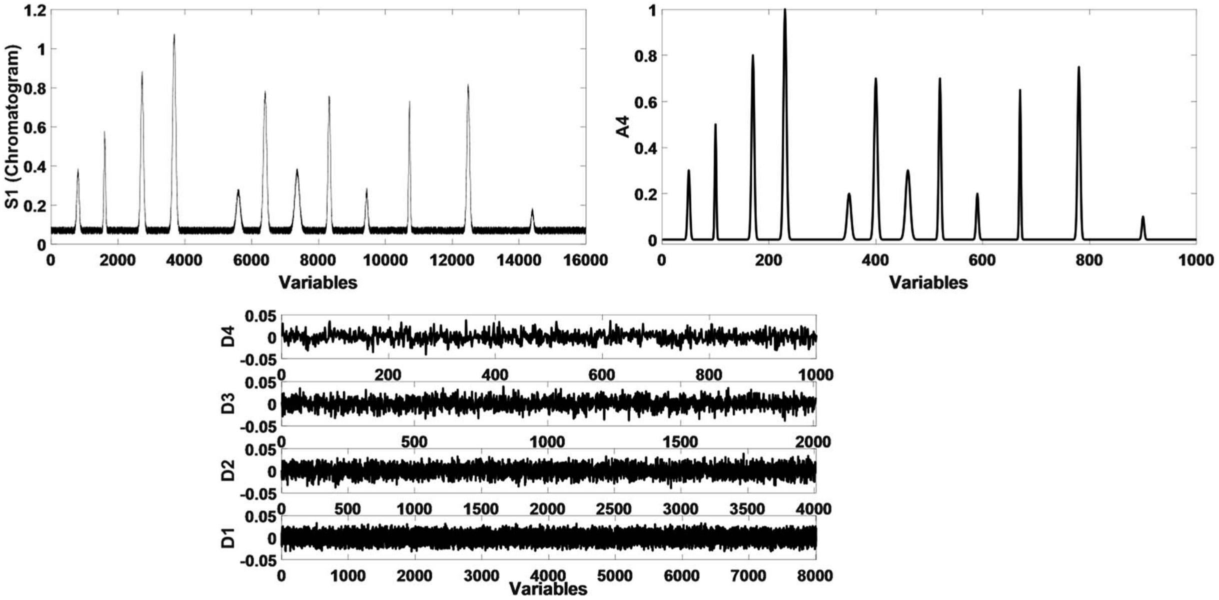

Fig. 2 The chromatogram of S1 and its decomposition with the db5 wavelet at level 4. *A*4 is the approximation coefficient and *D*1, *D*2, *D*3 and *D*4 are the detail coefficients of the analysed chromatogram.

calculation time required. Therefore, with wavelet pre-processed chromatograms that are 1/16 times smaller in size it would be expected that the calculation time can be reduced significantly.

As discussed above, the average calculation time for evaluating different combinations of segments and slack parameters for all the chromatograms are found to be reduced significantly from >2500 s to < 25 s. For each of the 20 chromatograms, several combinations of slack and segment lengths could be evaluated in a very short amount of time to achieve their optimum alignment with reference sample S10. The optimised slack and segment length for each of the remaining 20 chromatograms are summarised in Table 2. The COW aligned chromatograms are shown in Fig. 3(B). From the obtained results, it can be seen that application of DWT as a pre-processing technique prior to COW can really be useful in providing the required computational economy. It is to be noted that one can also use different interpolation approaches to reduce the dimension of the dataset. Therefore, in order to further justify the proposed approach a comparison between

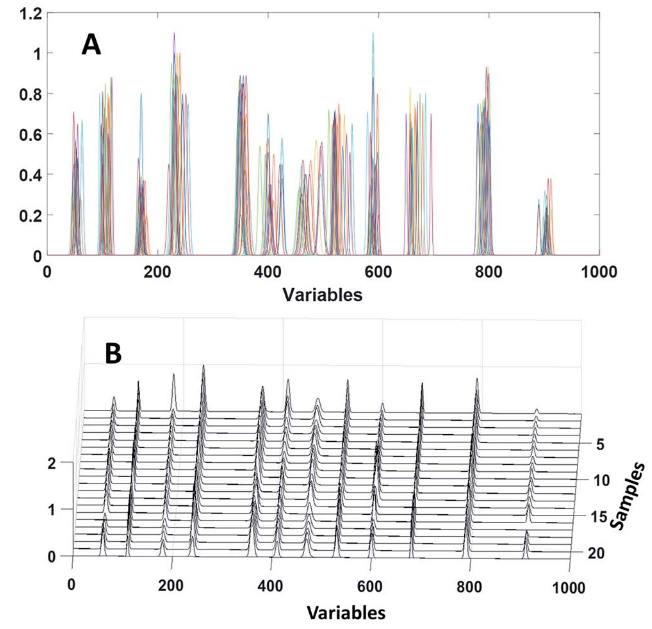

Fig. 3 (A) The db5 analysed (reduced size and denoised) chromatograms and (B) DWT-COW aligned chromatograms.

Table 2 The segment length, slack used and computational time spent in DWT assisted COW analysis for aligning different chromatograms to reference sample S10

Sample	Segment length	Slack	Time (seconds)
S1	35	20	1.01
S2	37	21	1.03
S3	32	18	0.95
S4	28	21	1.00
S5	35	19	1.00
S6	32	21	1.02
S7	35	20	1.01
S8	37	22	1.07
S9	29	19	0.95
S11	30	25	1.15
S12	31	23	1.09
S13	27	21	1.00
S14	29	22	1.05
S15	30	25	1.14
S16	37	21	1.06
S17	32	19	0.99
S18	29	21	1.04
S19	28	22	1.03
S20	35	20	1.00
S21	32	23	1.03

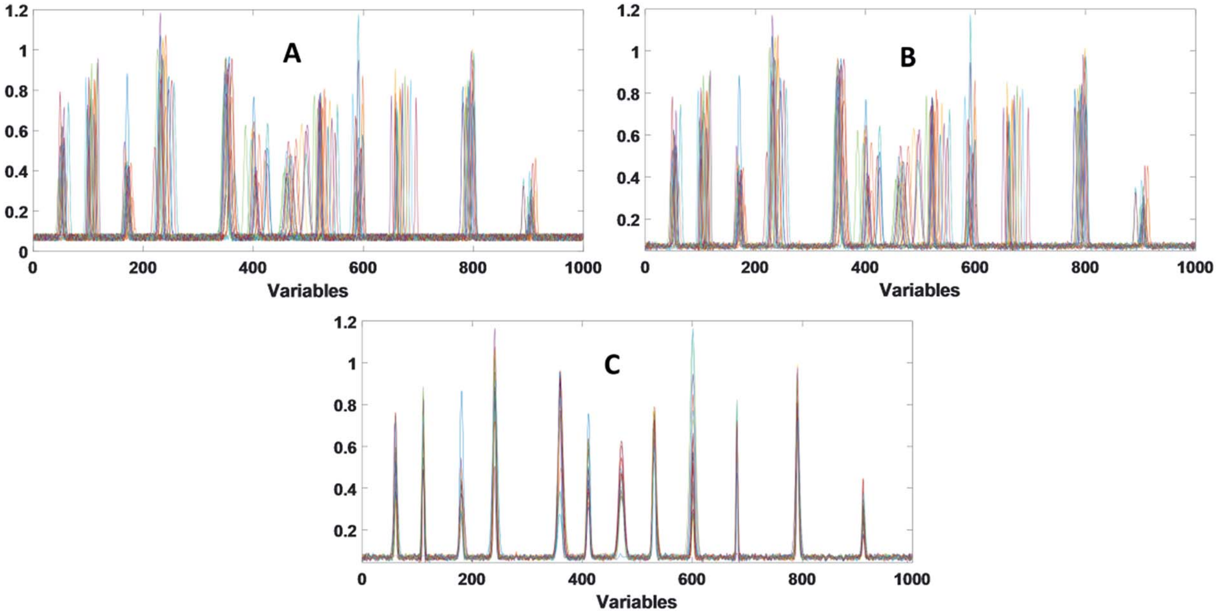

Fig. 4 (A) Spline interpolated chromatograms; interpolation is carried out to reduce the size of the dataset, (B) Savitky–Golay based denoised chromatograms and (C) interpolated-COW chromatograms.

DWT-COW and interpolation assisted COW needs to be performed which is presented below.

4.3 Interpolation assisted COW (interpolation-COW)

As discussed above, in principle, the interpolation (linear or spline) approach can provide the necessary reduction in data size. In the present work, the spline interpolation approach is used to reduce the sizes of the chromatograms from 16 000 data points to 1000 data points. The shapes of chromatograms are retained in this procedure. The spline interpolated data are shown in Fig. 4(A) and it can be seen that before applying COW algorithm a further pre-processing step is required to remove the high frequency noise present in the reduced chromatogram. It can be achieved with application of a smoothing technique such as the Savitky–Golay (SG) algorithm.[33] SG assumes that variables adjacent to each other contain similar information and can be averaged to reduce the noise without losing any important information present in the data. The SG algorithm involves fitting individual polynomials to windows around each data point in the chromatogram. The data are smoothed by moving these polynomials. Successful application requires optimisation of the two parameters (i) size of the window and (ii) polynomial order. With the optimum window size and polynomial order the chromatograms are denoised and are shown in Fig. 4(B). It is to be noted that finding the optimum parameters for the SG algorithm is also a time consuming step. The pre-processed chromatograms are further subjected to COW analysis. The slack and segment length used for the db5 wavelet pre-processed chromatograms are found to work well. The aligned chromatograms are shown in Fig. 4(C). The obtained results show that it is possible to provide the necessary computational economy to COW using the combination of interpolation and SG as the pre-processing step. However, it is also important to compare the denoising efficiency of the db5 wavelet and SG algorithm. The comparison is carried out by calculating the parameter called smoothing index (SI).[34] The smaller the SI value the greater the denoising achieved for the chromatogram. The SI can be calculated using eqn (7) given below:

$$\mathrm{SI} = \frac{\sum_{i=1}^{n} \left|\mathrm{DS}_{i+1} - \mathrm{DS}_i\right|}{\sum_{i=1}^{n} \left|S_{i+1} - S_i\right|} \tag{7}$$

In the above equation, DS represents the denoised chromatogram and S is the raw chromatogram, n is the number of data points in the chromatogram. The SI for each chromatogram is calculated and reported in Table 3. The obtained results clearly indicate that the DWT based approach provides a greater extent of chromatogram denoising. Therefore, DWT as pre-processing not only provides the necessary computational economy to COW but it also ensures that analysed datasets are of high quality. Moreover, the application of wavelets separates noise from the signal and provides the pure signal and noise profile that can be used for further analysis, whereas, SG does not separate noise from the signal, rather it removes the noise by averaging the data point, therefore it is difficult to have a noise profile for further analysis. Overall, it can be inferred that the DWT-COW combination provides a better way of pre-processing the chromatograms. In the literature, a cross-correlation based approach is available that can also provide fast alignment of the chromatograms. Therefore, it is also important that a comparison should be carried out between one such cross-correlation based approach and DWT and spline-SG assisted COW analysis.

Table 3 Smoothing index (SI) obtained for S1–S21 with the wavelet and Savitky–Golay approach

Chromatograms	Smoothing index (SI) of db5 wavelet at level 4	Smoothing index (SI) of Savitky–Golay
S1	0.21	0.68
S2	0.21	0.69
S3	0.23	0.70
S4	0.21	0.69
S5	0.22	0.69
S6	0.23	0.70
S7	0.23	0.70
S8	0.23	0.70
S9	0.25	0.73
S10	0.26	0.74
S11	0.27	0.74
S12	0.26	0.73
S13	0.24	0.71
S14	0.22	0.71
S15	0.21	0.69
S16	0.23	0.69
S17	0.22	0.69
S18	0.24	0.71
S19	0.23	0.71
S20	0.26	0.72
S21	0.24	0.71

4.4 Interval correlation shifting (Icoshift) algorithm: fast Fourier transform (FFT) cross-correlation based alignment approach

Interval correlation shifting (Icoshift) is one such FFT cross-correlation based alignment technique that allows the calculation of all the cross-correlation values in a short period of time.[6–8] Theoretical details of Icoshift can be found elsewhere.[6–8] Similar to COW it also requires a reference chromatogram to which the rest of the chromatograms are aligned. However, there is a fundamental difference in the approach of COW and Icoshift, the former technique uses an expansion and compression approach for achieving the alignment whereas the latter technique involves an insertion and deletion approach for aligning the chromatograms. The FFT engine makes the Icoshift a computationally economical approach for achieving the alignment of the chromatograms. Therefore, it is worth comparing its performance with the DWT-COW approach. The Icoshift analysis is performed on the dataset using S10 as the reference. The Icoshift analysis is performed on the entire length, despite that computational time is found to be less than 25 seconds and hence is comparable to that of DWT-COW. The obtained Icoshift analysed chromatograms are given in Fig. 5(A). It can be seen that application of Icoshift to the entire length of the chromatograms failed to provide the alignment. Therefore, the entire chromatograms for each of the samples are divided into 11 segments and Icoshift is applied separately to each of them. The obtained results are shown in Fig. 5(B). It can be seen that most of the peaks of the chromatograms are aligned to the reference, however they also contain certain artefacts, *i.e.* horizontal lines in the middle of the chromatograms. These artefacts appear due to the insertion and deletion approach. These artefacts can severely affect the outcomes of the entire data analysis workflow. The simulated chromatograms are highly complex with uneven separation of the peaks that test the efficiency of the Icoshift alignment algorithm.

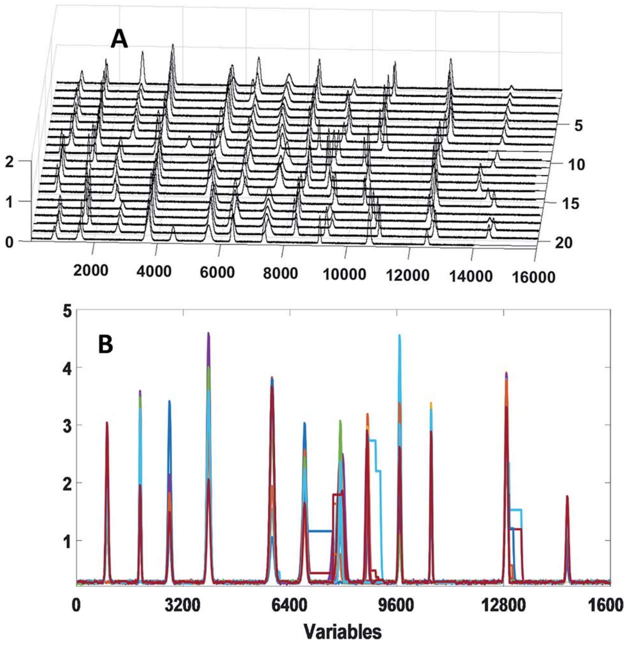

Fig. 5 (A) Icoshift analysed chromatograms where analysis is carried out on the entire length of the chromatogram. The results indicate that it is not possible to achieve the alignment by applying the Icoshift on full length and (B) Icoshift analysed on different segments provides the alignment for majority of the peaks, however it also generates the artefacts, which could be seen as the horizontal segments.

Table 4 Comparison of the different statistical parameters mcc, msv, and mpfv for evaluating the quality of alignment achieved using DWT-COW, spline-SG assisted COW and Icoshift approach for the simulated dataset

Approach	mcc	msv	mpfv
DWT-COW	0.925	0.966	0.999
Spline-SG assisted COW	0.823	0.888	0.912
Icoshift	0.789	0.812	0.757

The optimisation of the computational time is an important issue in the chromatographic data analysis workflow but at the same time one cannot compromise the quality of the alignment. Therefore, it is also important that the quality of alignment achieved with different approaches is compared. In order to achieve this, three statistical parameters, namely (i) mean correlation coefficient (mcc) (ii), mean simplicity value (msv) and (iii) mean peak factor value (mpfv), are calculated for DWT assisted COW, spline-SG assisted COW and Icoshift approaches and reported in Table 4. These statistical parameters were proposed by Liang and co-workers.[35] A perfect alignment is achieved if these statistical parameters are close to unity.[35] The reported statistical parameters in Table 4 clearly show that DWT-assisted COW provides the best alignment followed by spline-SG assisted COW and Icoshift approach.

The overall results obtained so far clearly show that COW suffers due to the fact that it is computationally time

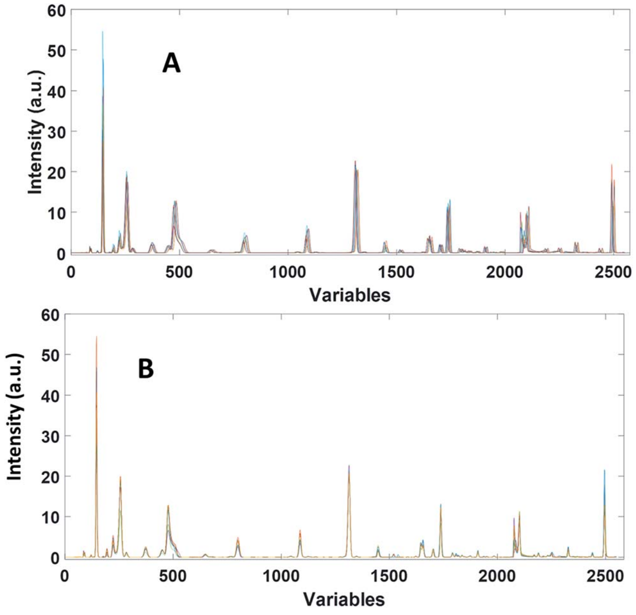

Fig. 6 (A) The unaligned chromatograms of the nine coffee samples and (B) the COW aligned chromatograms where the alignment process is found to be computationally demanding.

Table 6 Mean energy, mean Shannon entropy and mean energy to mean Shannon-entropy ratio for 22 different wavelets for the C1–C9 chromatograms

Wavelets	Mean energy	Mean Shannon entropy × 10^5	R = mean energy/ (mean Shannon entropy)
haar	98.9880	−1.1337	−87.3141
db2	99.8615	−1.1501	−86.8285
db3	99.9609	−1.1524	−86.7415
db4	99.9830	−1.1530	−86.7155
db5	99.9905	−1.1532	−86.7070
db6	99.9942	−1.1533	−86.7027
db7	99.9942	−1.1533	−86.7027
db8	99.9980	−1.1532	−86.7135
db9	99.9986	−1.1532	−86.7140
db10	99.9985	−1.1532	−86.7139
sym2	99.8615	−1.1501	−86.8285
sym3	99.9609	−1.1524	−86.7415
sym4	99.9832	−1.1530	−86.7157
sym5	99.9926	−1.1533	−86.7013
sym6	99.9943	−1.1532	−86.7103
sym7	99.9968	−1.1534	−86.6974
sym8	99.9968	−1.1533	−86.7049
coif1	99.8674	−1.1504	−86.8110
coif2	99.9847	−1.1532	−86.7020
coif3	99.9946	−1.1535	−86.6880
coif4	99.9969	−1.1536	−86.6825
coif5	99.9977	−1.1536	−86.6832

consuming; spline-SG assisted COW is computationally economical but the quality of the data being analysed still needs improvement; Icoshift based on FFT cross-correlation is computationally economical but tends to fail in aligning the chromatograms when applied to the entire length, moreover applying Icoshift to segments introduces artefacts. These issues can easily be taken care of by applying the DWT-COW approach. DWT not only provides computational economy to COW but at the same time also removes the noise part from the signals making sure that analysed data are of the best quality. In order to validate the proposed approach it is important that an application of DWT assisted COW for aligning real chromatograms should be presented. To achieve this, chromatograms of 9 different coffee samples are taken as the test case.

4.5 Application of DWT-COW on a real life dataset

The chromatograms of the coffee samples are shown in Fig. 6(A). Each chromatogram contains 2550 data points collected over the 0 to 17 minute with a step size of 0.0069 minutes. In order to achieve their alignment, sample C5 is chosen as the reference chromatogram in a random fashion

Table 5 The segment length and slack parameter used, computational time spent, mcc, msv and mpfv parameters obtained from COW and DWT assisted COW analysis while aligning different chromatograms to reference sample C5

Approach	Sample	Segment length	Slack	Time (seconds)	mcc	msv	mpfv
COW	C1	60	35	6.79	0.884	0.927	0.987
	C2	45	15	6.05			
	C3	30	25	8.57			
	C4	35	21	6.14			
	C6	35	21	5.21			
	C7	45	20	7.23			
	C8	45	23	8.32			
	C9	35	28	8.21			
DWT-COW	C1	13	7	1.14	0.933	0.977	0.999
	C2	11	5	1.32			
	C3	8	3	1.48			
	C4	8	3	1.48			
	C6	9	4	1.54			
	C7	11	5	1.31			
	C8	11	3	1.31			
	C9	9	5	1.45			

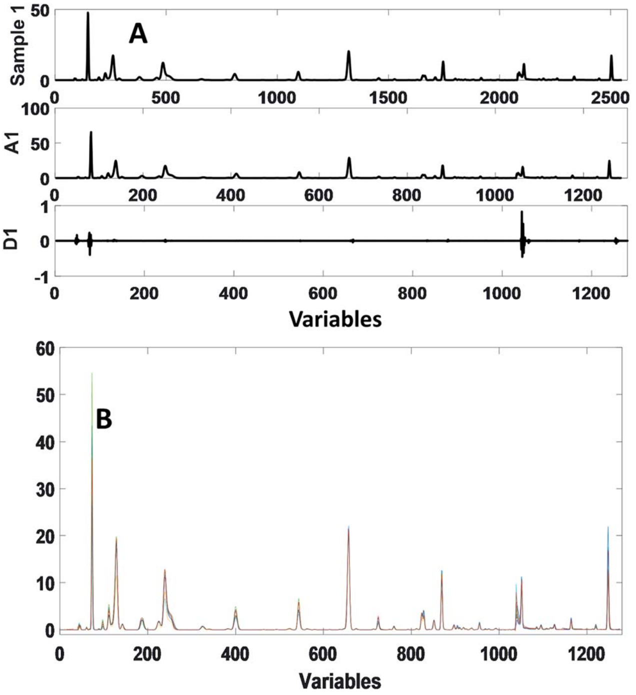

Fig. 7 (A) Chromatogram of the C5 (fifth coffee) sample and its decomposition with coif4 wavelet at level 1 and (B) DWT-COW aligned chromatograms of the coffee samples. The application of COW is found to reduce the computational time by a significant amount.

(the random approach is used because all the samples have similar features). The evaluation for a particular combination of slack and segment length towards aligning a chromatogram to the reference sample with COW is found to take roughly 6–12 seconds. Several combinations need to be evaluated for finding the optimum combination of the slack and segment length. The COW computational time spent aligning each chromatogram to the reference sample with optimum slack and segment length is summarised in Table 5. The COW aligned chromatograms are shown in Fig. 6(B). In order to reduce the computational time the DWT-COW combination is applied. All the 22 selected wavelets are applied to the chromatograms at the level 1. The wavelet analyses at higher levels are found to distort the shape of the chromatogram. As discussed above, the wavelet that maximises the energy to Shannon-entropy ratio is chosen as the mother wavelet. Table 6 containing the energy, Shannon-entropy, and energy to Shannon-entropy ratio for all the 22 wavelets indicates that coif4 should be considered as the mother wavelet. The coif4 decomposition of sample C5 at level 1 is shown in Fig. 7(A). The coif4 pre-processed chromatograms consisting of 1250 data points are further subjected to COW and it is found that the evaluation time for a particular combination of slack and segment length is reduced from 6–12 seconds to 0.5–2 seconds. For the optimised combination of segment length and slack, the time involved in COW analysis for each coif4 pre-processed chromatogram is also summarized in Table 5. Compared to COW analysis on unprocessed chromatograms, the computation for DWT assisted COW analysis is found to be 3–4 times faster. DWT-COW aligned chromatograms are shown in Fig. 7(B). The statistical parameters, mcc, msv and mpfv are also calculated and reported in Table 5 to evaluate the quality of alignments achieved with COW and DWT-COW approaches. The obtained results of DWT-COW combination clearly show that DWT as a pre-processing step prior to COW can significantly reduce the calculation time, denoise the chromatograms, and enhance the quality of alignment.

The proposed DWT-COW combination can also be really useful for the analysis of chromatography coupled with mass spectrometry datasets that are much bigger in size and require alignment not just along the time axis but along the *m*/*z* axis. However, a practical demonstration would still be required.

5. Conclusions

In the present work, it has been shown that DWT analysis provides twin benefits: (i) it denoises the chromatograms and (ii) reduces the dimension of the dataset while retaining all the information of the original chromatograms. Despite being computationally time consuming, COW is the most commonly used chromatographic peak alignment approach. COW involves optimisation of slack and segment length for achieving the optimum alignment. Often the optimisation takes several hours. In the present work, it is also successfully shown that application of DWT as a pre-processing technique prior to alignment can provide the required computational economy to the COW algorithm.

Acknowledgements

The author thanks the reviewers for their suggestions that improved the quality of the work being reported. The author acknowledges and thanks R. Bro, University of Copenhagen, Denmark for making the useful codes available to the scientific community through his website: http://www.models.life.ku.dk/DTW_COW.

References

1 G. Malmquist and R. Danielsson, *J. Chromatogr. A*, 1994, **687**, 71–88.

2 S. Liu, Y. Z. Liang and H. Liu, *J. Chromatogr. B*, 2016, **1015–1016**, 82–91.

3 N. P. V. Nielsen, J. M. Carstensen and J. Smedsgaard, *J. Chromatogr. A*, 1998, **805**, 17–35.

4 G. Tomasi, F. van den Berg and C. Andersson, *J. Chemom.*, 2004, **18**, 231–241.

5 T. Skov, F. van den Berg, G. Tomasi and R. Bro, *J. Chemom.*, 2006, **20**, 484–497.

6 F. Savorani, G. Tomasi and S. B. Engelsen, *J. Magn. Reson.*, 2010, **202**, 190–202.

7 G. Tomasi, F. Savorani and S. B. Engelsen, *J. Chromatogr. A*, 2001, **1218**, 7832–7840.

8 F. Savorani, G. Tomasi and S. B. Engelsen, Alignment of 1D NMR data using the Icoshift tool: A tutorial, in *Magnetic resonance in food science, Food for thought*, ed. J. van

Duynhoven, H. V. As, P. S. Belton and G. A. Webb, Royal Society of Chemistry, Cambridge, 2013, pp. 14–24.

9 B. Walczak and D. L. Massart, *TrAC, Trends Anal. Chem.*, 1997, **16**, 451–463.

10 X. Shao, W. Cai and Z. Pan, *Chemom. Intell. Lab. Syst.*, 1999, **45**, 249–256.

11 X. Shao, W. Cai, P. Sun, M. Zhang and G. Zhao, *Anal. Chem.*, 1997, **69**, 1722–1725.

12 D. Labat, *J. Hydrol.*, 2005, **314**, 275–288.

13 B. K. Alsberg, A. M. Woodward and D. B. Kell, *Chemom. Intell. Lab. Syst.*, 1997, **37**, 215–239.

14 C. Perrin, B. Walczak and D. L. Massart, *Anal. Chem.*, 2001, **73**, 4903–4917.

15 L. Pasti, B. Walczak, D. L. Massart and P. Reschiglian, *Chemom. Intell. Lab. Syst.*, 1992, **48**, 21–34.

16 R. X. Gao and R. Yan, *Wavelets: Theory and application for manufacturing*, Springer, New York, 2011.

17 X. G. Shao, A. K. M. Leung and F. T. Chau, *Acc. Chem. Res.*, 2003, **36**, 276–283.

18 F. T. Chau, Y. Z. Liang, J. Gao and X. G. Shao, *Chrmometrics: From basic to wavelet transform*, John Wiley & Sons, Inc., Hoboken, New Jersey, 2004.

19 F. Ehrentreich, *Anal. Bioanal. Chem.*, 2002, **372**, 115–121.

20 D. Barache, J. P. Antoine and J.-M. Dereppe, *J. Magn. Reson.*, 1997, **128**, 1–11.

21 V. J. Barclay and R. F. Bonner, *Anal. Chem.*, 1997, **69**, 78–90.

22 Z. M. Zhang, S. Chen and Y. Z. Liang, *Talanta*, 2011, **83**, 1108–1117.

23 V. S. Chourasia and A. K. Mittra, *J. Med. Technol.*, 2009, **33**, 442–448.

24 B. Walczak and D. L. Massart, *Chemom. Intell. Lab. Syst.*, 1997, **36**, 81–94.

25 M. Sifuzzaman, M. R. Islam and M. Z. Ali, *J. Phys. Sci.*, 2009, **13**, 121–134.

26 M. Schwartz, B. Mayer, B. Wirnitzer and C. Hopf, *Anal. Bioanal. Chem.*, 2015, **407**, 2255–2264.

27 M. Lagarrigue, T. Alexandrov, G. Dieuset, A. Perrin, R. Lavigne, S. Baulac, H. Thiele, B. Martin and C. Pineau, *J. Proteome Res.*, 2012, **11**, 5453–5463.

28 S. Cappadona, F. Levander, M. Jansson, P. James, S. Cerutti and L. Pattini, *Anal. Chem.*, 2008, **80**, 4960–4968.

29 S. Chen, D. Hong and Y. Shyr, *Comput. Stat. Data Anal.*, 2007, **52**, 211–220.

30 A. M. van Nederkassel, M. Daszykowski, P. H. C. Eilers and Y. V. Heyden, *J. Chromatogr. A*, 2006, **1118**, 199–210.

31 A. M. van Nederkassel, M. Daszykowski, D. L. Massart and Y. V. Heyden, *J. Chromatogr. A*, 2005, **1096**, 177–186.

32 V. Pravdova, B. Walczak and D. L. Massart, *Anal. Chim. Acta*, 2002, **456**, 77–92.

33 A. Savitzky and M. J. E. Golay, *Anal. Chem.*, 1964, **36**, 1627–1639.

34 H. Yang, D. Zhang, J. Zhao and L. Huang, *Int. J. Agric. Biol. Eng.*, 2014, **7**, 36–42.

35 W. Jiang, Z. M. Zhang, Y. H. Yum, D. J. Zhan, Y. B. Zheng, Y. Z. Liang, Z. Y. Yang and L. Yu, *Chromatographia*, 2013, **76**, 1067–1078.