Comparative study of the alignment method on experimental and simulated chromatographic data

Rabia Korifi, Yveline Le Dréau, Nathalie Dupuy

Rabia Korifi
Yveline Le Dréau
Nathalie Dupuy

Laboratoire LISA, EA 4672
Equipe METICA, Case 451,
Aix-Marseille Université,
Marseille Cedex, France

Comparative study of the alignment method on experimental and simulated chromatographic data

One of the major problems in the signal comparison of chromatographic data is the variability of response caused by instrumental drifts and others instabilities. Measures of quality control and evaluation of conformity are inherently sensitive to shift. It is essential to be able to compare test samples to reference samples in an evolutionary analytical environment by offsetting the inevitable drift. Therefore, prior to any multivariate analysis, the alignment of analytical signals is a compulsory preprocessing step. During recent years, many researchers have taken a greater interest in the study of the alignment. The present paper is an updated review on the alignment algorithms, methods, and improvements used in chromatography. The study is dedicated to one-dimensional signals. Several of the exposed methods have common theoretical bases and can differ through their optimization methods. The main issue for the operator is to choose the appropriate method according to the type of signals to be processed.

Keywords: Alignment methods / Chromatographic data / Preprocessing / Signal comparison / Warping

1 Introduction

A large amount of data can be produced by modern chromatographic instrumentation. To extract a maximum of qualitative and quantitative information about the analyzed samples and to easily compare signals, chemometric analysis is necessary. However, instrumental drifts and other instabilities provide unwanted shifts in chromatographic peak positions. More precisely, peak components from the same molecule have different spatial positions on the measurement axis in different samples. These variations are an impediment against the use of chemometric techniques [1–3]. Some current chemometric techniques are unable to correctly model information that shifts from variable to variable within a dataset.

Chemometric tools for visualization and data mining such as principal component analysis and for quantitative analysis by predictive model such as partial least squares regression require uniform presentation of data. All signals must be adjusted to the same length and corresponding variables have to be placed in proper columns of the data matrix. The effect of the alignment procedure on the principal component analysis has been demonstrated on chromatographic signals by Malmquist et al. [4].

Signals obtained by chromatography are prone to shift and do not fulfill this condition. More precisely, to accurately and effectively analyze data using chemometrics, the same peaks in different signals should be aligned at the same matrix column. Even minor shifts introduce extra factors in multivariate analysis, therefore data alignment is a prerequisite before any multivariate chemometric analysis [5]. To overcome the problem of misaligned peaks, several approaches have been proposed in the literature. Examples of this misalignment problem have previously been treated for several types of chromatographic and spectroscopic data. Jellema [6] and more recently Bloemberg et al. [7] introduced several of the alignment techniques presented in our comparative study. The *intersample correspondence* problem, defined by recent reviews [8, 9], is based on the difficulty to evaluate which peaks correspond to which in datasets from measurements on different sample, the fact is that features from different samples are nonsynchronized in the measured data. While the majority of peaks can be assigned obviously, some peaks are problematic at this level, so Åberg et al. [8] defined the concept of ambiguous correspondence that identifies three basic cases of ambiguous correspondence. (i) Two peaks in the second sample can match one peak in the first sample, second ambiguity. (ii) Two peaks in the first sample can match two

Correspondence: Dr. Rabia Korifi, Aix-Marseille Université, Laboratoire LISA Metica, case 451, Avenue escadrille Normandie Niemen, Marseille 13397, France
E-mail: rabia.korifi@yahoo.fr

Abbreviations: **COW**, correlation optimized warping; **DP**, dynamic programming; **DTW**, dynamic time warping; **FID**, flame ionization detection; **PAGA**, peak alignment by genetic algorithm; **PAFFT**, peak alignment by fast Fourier transform; **PARS**, peak alignment using reduced set mapping; **PTW**, parametric time warping; **RAFFT**, recursive alignment by fast Fourier transform; **STW**, semiparametric time warping

peaks in the second sample but with the last peak of the first sample matching the first peak of the second sample with few or no shifting; in this case it is not possible to determine if there are two or just one corresponding peak. (iii) Peaks change order between the two samples.

The aim of this paper is to test and compare all the freely available alignment techniques able to treat 1D chromatographic signals. In the first part, the sources of shifts are exposed. Then, the alignment techniques are briefly introduced. The alignment algorithms are tested on real and simulated datasets for different types of shift. Finally, the validation of alignment results are evaluated through quality criteria and to give hints for the choice of algorithm, a summary of the practical aspects in the form of table is provided, then, to highlight issues a drawbacks are discussed.

2 Sources of retention time shifts

Instrumental drift is identified and principally associated to ageing, and deterioration of the chromatographic column drifts in detection conditions, and slight variations of the analytical conditions. More precisely, variations in the stationary phase, film thickness, column head pressure, split ratio, and so on [10,11]. In HPLC, the column can be exposed to frequent mobile phase conditions changes (pH value variations), varying temperatures, and changing flow rates. Those situations are able to promote column degradation and so provide a decrease in resolution and retention factors responsible for peak shifts in time [12]. In GC, temperature and carrier gas flow fluctuations from run-to-run, stationary phase, composition changes from column-to-column, and matrix effects from sample-to-sample lead to variations of retention time for a given component over a series of several chromatographic runs [13, 14]. The correction of shifts is a major issue and requires the development of alignment algorithms.

3 Alignment methods

The source of all the warping alignment procedures involves the transformation of the measurement axis of the sample signal to match the best with the reference signal. The following notation will be used: the sample signal will be denoted $S(x)$, the reference signal $R(x)$, with x representing the abscissa axis in data points. Alignment techniques differ in several basic aspects. First, the way the warping path or warping function denoted by F is defined, the measurement of quality of alignment (similarity between signals) and the optimization algorithm used to find the optimal F (e.g. dynamic programming (DP), evolutionary algorithm, etc.)

For each studied method, the operating principles and the conditions of use (form of shifts, type of processed data tested) are outlined. The methods are classified depending on the kind of optimization algorithms and techniques.

3.1 Dynamic programming

3.1.1 Correlation optimized warping

The $R(x)$ and $S(x)$ signals, defined by their respective length on abscissa axis L_R and L_S expressed into data points, are uniformly divided into a number of sections N of length m [15].

If the length of the two signals is equal, the number of sections is then given by Eq. (1):

$$N = \frac{L_R}{m} = \frac{L_S}{m} \tag{1}$$

To treat unequal lengths of two signals, the difference δ in section length in $S(x)$ and $R(x)$ is calculated by Eq. (2):

$$\delta = \frac{L_R}{N_S} - m_S \tag{2}$$

The alignment of two signals is done by an iterative piecewise linear stretching and compression of the $S(x)$ signal. Aligning by piecewise linear warping involves dividing signals into a user-specified number of sections N. Each section is warped by linear interpolation, meaning that it is stretched or shortened by shifting the position of its end points by a finite number of possible warping magnitudes, the so-called flexibility or slack parameter s. s is the maximum shifting degree of boundary allowed in the section to align. Each section boundary can be shifted from $-s$ to $+s$ points, except for the start and end points of the signal that are fixed. For example, if the slack equals 1, there are three possible shifting of the boundary, -1, 0, and $+1$. For unequal lengths of signals, warpings permitted are included in the interval ($\delta - s$; $\delta + s$) [16]. For every possibility, a linear interpolation of the sample segment computes the number of data points corresponding to the reference section. Then, a correlation coefficient is calculated.

Nielsen et al. [15] developed a method based on this principle, which corrects shifts in vectorized data signals, the correlation optimized warping (COW). The DP is used as the optimization procedure for the COW. Determining the optimal alignment is a question of finding the optimal combination of warpings of the N sections warped by the slack parameter from $-s$ to $+s$. DP was first introduced by Bellman in 1954 in the Bulletin of the American Mathematical Society. The algorithm of DP has since been used in several alignment methods to solve these issues. DP [17] solves combinatorial optimization problems by examining all possible combinations of the variables. For each section i ($i = 1$: N), the optimal warping u_i is calculated for each possible position of the node x_i (starting point of section i after warping). Suboptimal combinations of warping are found for every section, and when all sections have been treated only the optimal combination is kept, providing a global optimum. A meaningful example is given by Tomasi et al. [18] to show the implementation of the DP. An example with $s = 1$ is given Fig. 1. The current implementation starts at the end of the

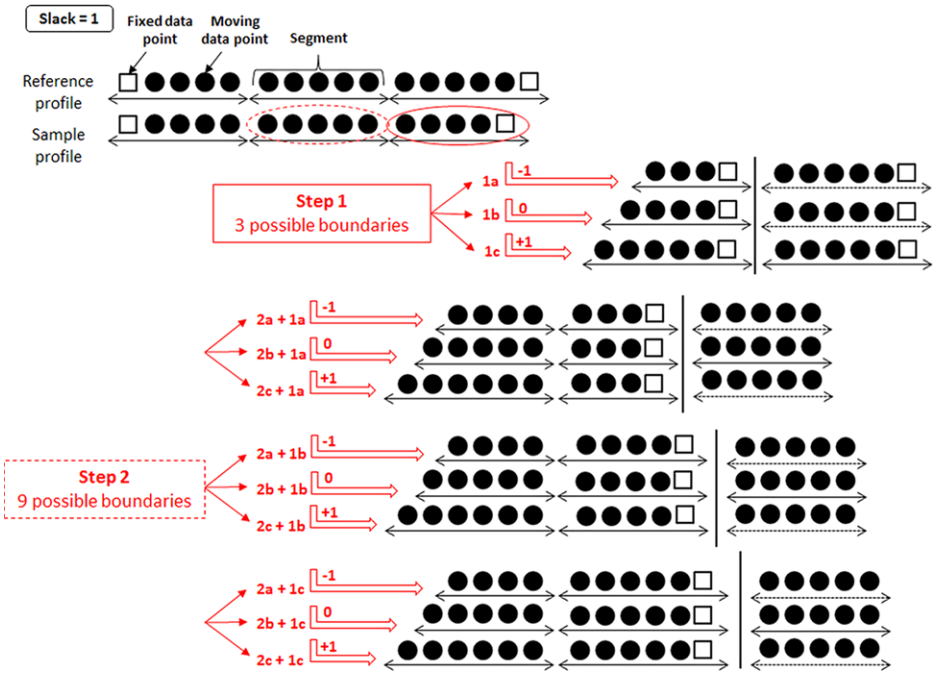

Figure 1. Illustration of the first two steps of dynamic programming algorithm used in correlation optimized warping.

chromatogram and progresses toward the beginning. The external boundaries of the segments are kept at fixed positions. As the slack is equal to 1, it means that three possible boundaries exist.

In the first step, the first sample signal segment (circled) is compared to the first reference signal segment. The left boundary of this sample segment is moved one data point to the right (1a), not moved (1b), and moved one data point to the left (1c). Each of these segments is compared to the reference sample and each correlation coefficient is calculated and stored. If the reference segment is a solid line no interpolation is needed, while a segment in dotted line means that the sample segment has been interpolated to fewer or more data points.

In the second step, the second sample segment is included. The second segment (circled by dotted lines) is five data points, three new segments of length four, five, and six data points, respectively, are created. The performance of the total warping is then the sum of the two correlation coefficients available in each of the now nine possible combinations $(2a + 1a, 2b + 1a, 2c + 1a, 2a + 1b, \ldots)$. Figure 1 only describes the first two steps of the DP algorithm. The number of possibilities increases with the next steps. The implementation keeps going until all boundaries have been moved, except for the start and finish positions that are fixed. At this time, the optimal warping path is found by constructing a global function P that is the cumulative sum of maximum correlation coefficients of the previous sections. Once this function is known, the handling of the axis, explicitly, the right positioning of the segments boundaries can be done. Finally, sample segments are linearly interpolated to be aligned to reference segment. The COW was demonstrated on data collected from GC with flame ionization detection (FID) [19–21].

3.1.2 Dynamic time warping

Initially, the time warping technique came from the speech recognition. The idea of time warping exploited as a method to align signals was first introduced by Wang and Isenhour [22]. About 10 years later, Kassidas et al. [23] developed a method called dynamic time warping (DTW). This algorithm is based on nonlinear warping of signals to align similar peaks with a minimal distance between them. A plot is constructed with the reference signal in the x-axis and the sample signal in the y-axis. The algorithm creates a path, as shown in Fig. 2a, such that corresponding peaks in $S(x)$ and $R(x)$ signals are linked. When this path is known, it can be used to align the whole signals. Each point $p(i)$ in the grid is described by a pair of indices and indicates a position in the grid: $p(i) = [n(i), m(i)]$, where n and m are the indices of, respectively, reference and sample signals. The path is found by constructing a grid with size $L_R \times L_S$, respectively, the lengths of reference and sample signals, and a sequence through the grid is noted. Once the path through the grid is found, it has to be optimized. The DP algorithm enables to reach the path in the grid, which minimizes the total cumulative distances between signals. To avoid excessive compression or expansion, constraints must be set up, Pravdova et al. [19] in their example had recourse to allowable predecessors. Each point (i,j) has three local constraints: $(i-1, j)$, $(i-1, j-1)$, $(i, j-1)$. For each of these three predecessors (Fig. 2b), local cumulative distances are calculated using the Euclidean distance formula as seen in Eq. (3):

$$d(i, j) = \sqrt{(R_i - S_j)^2} \quad \text{for } i = 1,\ 2,\ \ldots,\ L_R \text{ and } j = 1,\ 2,\ \ldots,\ L_S \tag{3}$$

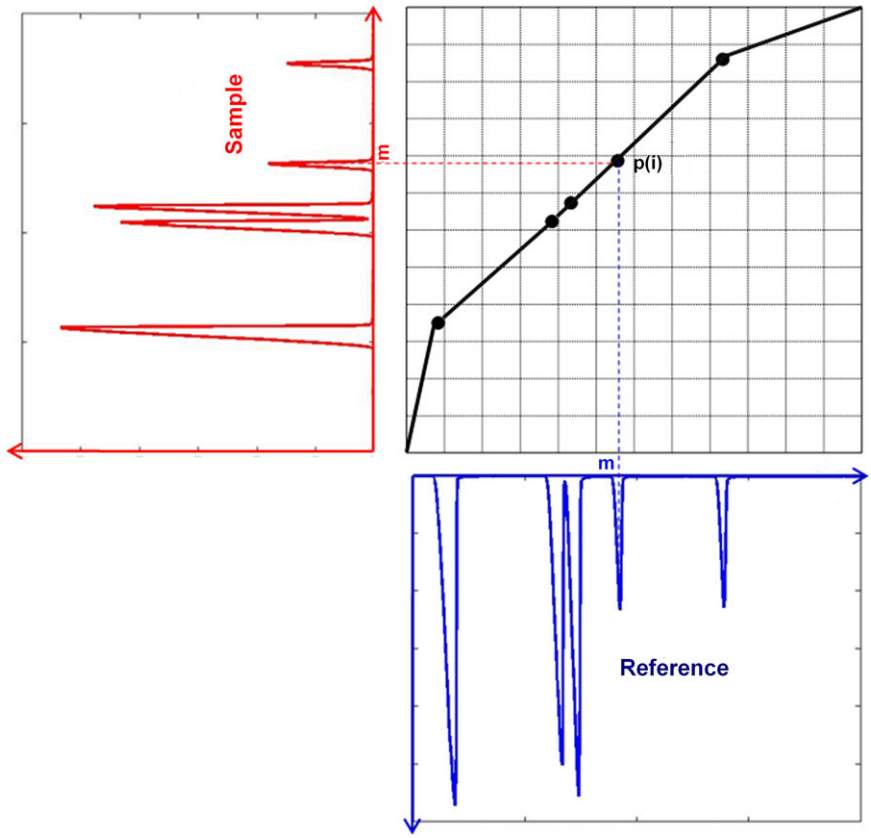

Figure 2. The warping path construction in dynamic time warping algorithm (a) and the process to calculate the local cumulative distances using allowable predecessors (b).

It allows us to define the cell (*i,j*) recursively shown by Eq. (4):

$$\text{cell}(i, j) = \text{local distance}(i, j) + \text{MIN}(\text{cell}(i-1, j), \text{cell}(i-1, j-1), \text{cell}(i, j-1)) \quad (4)$$

Only the shortest distance is kept for point (*i,j*). When all the points in the search area are filled with cumulative distances, the optimal path can be constructed. The DTW has been used by Wang et al. to eliminate the time shift in data collected from GC–FTIR.

3.1.3 Peak alignment using reduced set mapping

In 2003, Torgrip et al. [24] developed a global alignment method based on mapped data depending on peak positions termed peak alignment using reduced set mapping (PARS). This technique is based on the detection of peak maxima in reference and sample signals, data are mapped by means of the peak positions. The representation of the peaks is transformed into a sparse vector of zeros with integers located at the abscissa axis data corresponding to the peak maxima. PARS is a sequence alignment method that differs from the time warping method by the way the search map is created. A typical sequence alignment method is the algorithm of Smith and Waterman [25] that involves the calculation of a match map of similarity/dissimilarity, i.e. a matrix with entries of ones for matches and negative numbers for mismatches. Then the maximum alignment score is searched by DP and the optimal route for alignment, in the form of inserts/deletions at optimal positions, can be established by the DP solution. The PARS is a modified form of this sequence alignment method, a new mapping scheme is proposed by tracing sparse match maps with optimization algorithm. In their study, three of them have been tested. The first one is DP, the second one is the complexity reduced DP. Based on the same principle as DP, it is proposed to reduce memory handling. The matrix sizes are reduced by adding a Sakoe–Chiba constraint for the search of the maximum distance allowed between peaks [26]. The third is the breadth first search algorithm. It is a fast search-tree algorithm, which treats a sparse representation of peak maxima, the peak locations of the sample and reference signals are put in a list. A search space *s* is defined. A graph is constructed with vertices representing possible matches between peaks in sample to align and reference. The number of possible matches is bounded in a window of size $2s + 1$. In Fig. 3, vertices are depicted as circles at the possible match positions between peaks in *Rx* and *Sx*, the search space is $\pm s$ from the diagonal. The breadth first search algorithm operates on a list of the graph vertices sorted by peak index in the reference. Another list, sorted by peak index in the sample contains the vertex positions in the reference list. Each vertex holds fields with information about the previous vertex and the matching score. The search stops when the vertex holds the best combination.

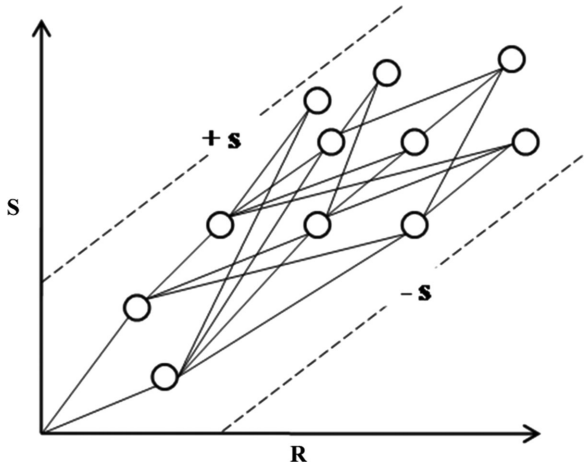

Figure 3. An example of breadth first search algorithm tree.

3.2 Evolutionary algorithms

3.2.1 Peak alignment by genetic algorithm

Forshed et al. [27–29] developed the peak alignment by genetic algorithm (PAGA). The sample signal $S(x)$ is divided into a number of segments N. Each segment is aligned to the corresponding segment in a reference signal $R(x)$. The alignment is achieved by taking each segment of the signal individually, shifting it sidewise, and linearly interpolating it to stretch or shrink until the best correlation with a corresponding reference signal segment is obtained. Cutting in peaks are avoided based on a routine that finds the minimum intensities of both signals in a spectral window. As shown in Fig. 4, each section is shifted sidewise m points and/or stretched or shrunk with linear interpolation i points. By shifting and interpolation, z points are removed from the segment and $p = (i + m)$ points are added to fill out the segments to make the length equal with s.

The genetic algorithm is an optimization method introduced by Holland. It can be seen as an evolutionary process in which a population of candidate solutions to a problem evolves over a sequence of generations. Based on the survival of the fittest strategy, a better solution will have a higher probability of surviving as the genetic algorithm proceeds. Genetic algorithm can solve linear and nonlinear problems. In the case of the alignment problem tackled by Forshed et al. [28] in the PAGA, the candidate solutions are comparable to the parameters of the shifting, stretching, and shrinking of the sample segments. These parameters are allowed by the genetic algorithm to evolve and find the best fit to the reference segment. Then, after some linear interpolation, the fit is evaluated as the correlation coefficient between the sample and the reference. The benefit is that not every possible solution has to be evaluated, but the best solution in a set of candidate solutions is favored while the algorithm proceeds. As soon as all the sample segments are treated, the signal is reconstructed to form the aligned signal. Kaya et al. [30] developed a signal alignment method based on genetic algorithm (SAGA), a warping function is modeled with an ordinary differential equation and the parameters of the function are optimized by using a genetic algorithm.

3.2.2 Peak alignment by beam search

Following Forshed et al., Lee et al. [31] developed the peak alignment by beam search. On the same approach as the PAGA, the aim is to identify the parameters that maximize correlation coefficient between sample and reference segments. They implemented a beam search algorithm, which contrary to the PAGA, finds the shift position without interpolation. Beam search algorithm, originally used for speech recognition and image processing, is a heuristic algorithm that progresses layer by layer to build the search tree shown in Fig. 5. In each layer, a heuristic evaluation function is used to estimate the promising solutions. This number of solutions is called beam width k, only the k most promising solutions are selected, the others are removed. The k value is fixed at the start, in the example (Fig. 5) below $k = 2$, so two nodes by layer are analyzed. For the first layer, the two nodes on the solution path are B and D and the node C is cutting back. At each level, k solutions are tested. Each solution is used to align the sample segment to its reference, and evaluated by a correlation coefficient calculus. At each layer, only the best solution is kept, meaning the one providing the higher correlation coefficient.

3.2.3 Differential evolution

A peak alignment method using wavelet pattern matching has been proposed by Zhang et al. in 2011 [32]. It moves the accurate detected peaks in the range of a value of flexibility called slack denoted s, so from a range of $-s$ to $+s$, and warps the nonpeak parts between two detected peaks using linear interpolation. To maximize linear correlation coefficient between reference signal and sample signal, an evolutionary algorithm has been put forward as a population-based optimizer called differential evolution.

First introduced by Storn and Pierce [33], differential evolution algorithm is a parallel direct search method using N_D dimension parameter vectors as a population for each generation G. Four steps constitute the differential evolution algorithm: initialization, mutation, crossover, and selection, the three last steps are repeated to find the optimum. Parameter vectors of great importance used in peak aligning steps are slack (number of points to shift peaks), N_P (number of populations), and itermax (predetermined maximum generation). Differential evolution generates new parameter vectors by adding the weighted difference vector between two population members (reference and sample) to a third member. It works by having a population of candidate solutions that are moved around in the search space to combine the positions of existing peaks from the population. If the new position of a peak is an improvement, it is accepted and forms part of the population, otherwise the new position is simply discarded.

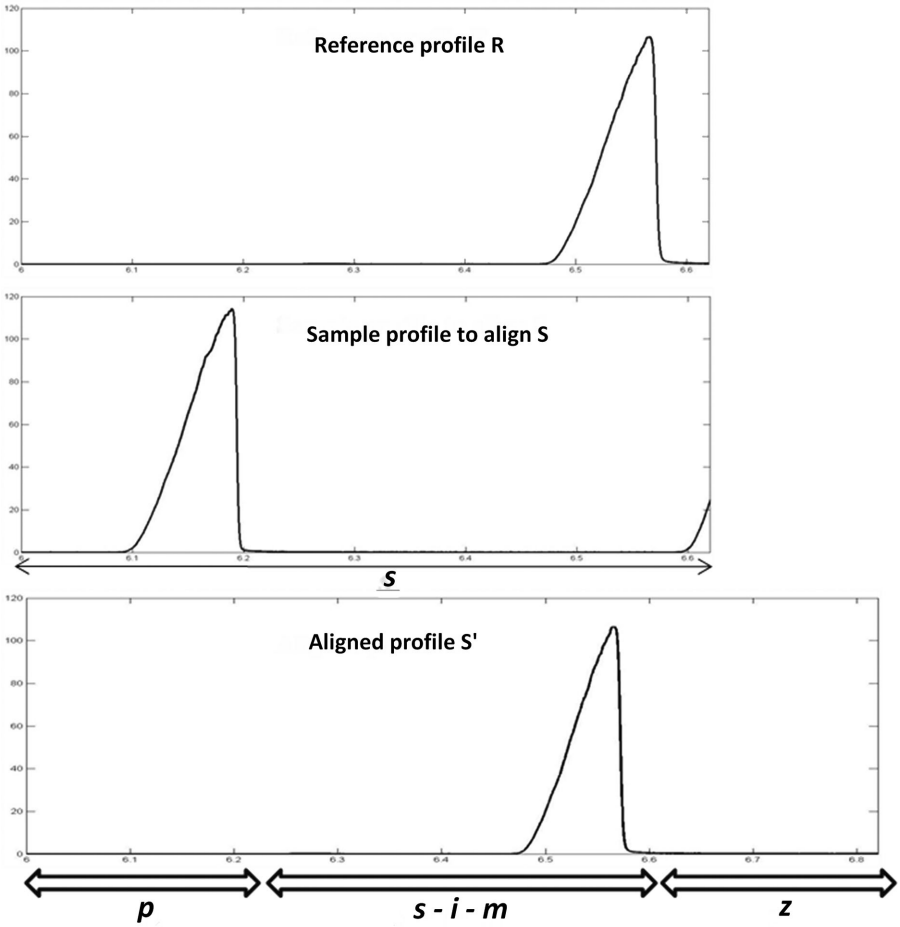

Figure 4. Scheme of the principle of the peak alignment by genetic algorithm.

To sum up, the algorithm intelligently evolves to the true optimum with good probability using of differences between individuals.

3.3 Cross-correlation function by fast Fourier transform

Two techniques also based on spectral segmentation are introduced by Wong et al. [34], the peak alignment by fast Fourier transform (PAFFT) and the recursive alignment by fast Fourier transform (RAFFT). The signals are divided into an arbitrary number of segments such that the shift in each signal can be corrected independently. An optimal shift size is found and the segment is shifted by this amount. The Icoshift [35], based on PAFFT and RAFFT and the recursive segment-wise peak alignment [36], uses the same segmentation principle to improve local spectral alignment. More recently, a method also based on the cross-correlation function by fast Fourier transform has been developed [37] to deal with nonlinear retention time shift by a moving window procedure. A well-known method for efficiently measuring correlation and signal shift is the calculation of the cross-correlation function using fast Fourier transform. The

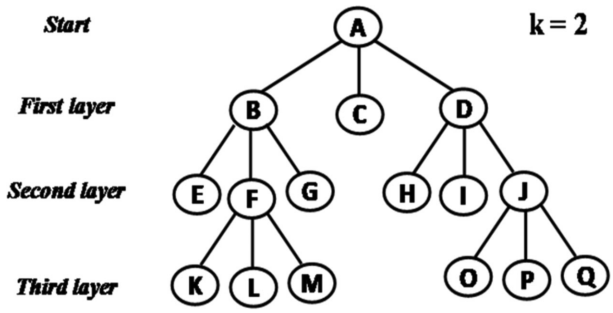

Figure 5. An example of a beam search tree used in the beam search algorithm.

two algorithms proposed by Wong et al. are based on this optimization method. The cross-correlation function by fast Fourier transform ensures rapid calculation of the correlation between two datasets where one of the signals is shifted compared to the other. The cross-correlation corresponds to a similarity degree between the two signals to be aligned. Besides, it is also able to provide an accurate estimation of the shift between two datasets. The cross-correlation function, denoted $\mathrm{Corr}(r,s)_u$, for two functions reference $R(x)$ and sample $S(x)$ at any shift position u, is written in Eq. (5):

$$\mathrm{Corr}(r, s)_u = \int_{-\infty}^{\infty} R(x)\, S(x+u)\, dx \tag{5}$$

To generate this function, the first step is to apply the Fourier transform on the functions $R(x)$, Eq. (6) and $S(x)$, Eq. (7):

$$R(\lambda) = \int_{-\infty}^{\infty} R(x) e^{2\pi\lambda x} dx \tag{6}$$

$$S(\lambda) = \int_{-\infty}^{\infty} S(x) e^{2\pi\lambda x} dx \tag{7}$$

where $R(\lambda)$ and $S(\lambda)$ are the Fourier-transformed functions in the inverse wavelength λ domain. The complex conjugate of the function $S(\lambda)$, denoted as $S^*(\lambda)$, is multiplied to the function $R(\lambda)$, and the last step is to perform a reverse Fourier transform on this product. The cross-correlation function is now generated and the optimal shift u_{op} between $R(x)$ and $S(x)$ can be found at the maximum of $\mathrm{Corr}(r,s)_u$, that is Eq. (8):

$$u_{op} = \max_u (\mathrm{Corr}(r, s)_u) \tag{8}$$

Once this optimal shift is found, the sample segment to align is shifted by that amount. This optimization method is used in the same way by Veselkov et al. in a recursive segment-wise peak alignment technique, by Savorani et al. for the Icoshift, which allows both the entire signal and user-defined intervals to be worked on.

3.4 Parametric alignment methods

3.4.1 Parametric time warping

A warping function is defined as an appropriate transformation of the abscissa axis useful to align two signals. Eilers [38] proposed a parametric model for the warping of chromatograms, the parametric time warping (PTW). It aligns a sample signal $S(x)$ to a reference signal $R(x)$ using a warping function. The warping function is a second degree polynomial of the abscissa axis in data points (x). As explained by Nederkassel et al. [5], the algorithm interpolates the sample signal $S(x)$ to the points in the warping function $W(x)$, to obtain the aligned signal $S(W(x))$, which is supposed to match as good as possible to the reference signal $R(x)$. The warping function $W(x)$ has to be optimized for each warping problem and is defined by Eq. (9):

$$W(x) = \sum_{k=0}^{2} a_k x^2 = a_0 + a_1 x + a_2 x^2 \tag{9}$$

K symbolizes the polynomial degree ($K = 2$), x the abscissa axis, then, a_k (a_0, a_1, a_2) are the warping coefficients.

3.4.2 Semiparametric time warping

Analogous to PTW, a method called semiparametric time warping (STW), also developed by Eilers, uses a warping function to align a sample signal with a reference signal. However, the difference between STW and PTW is that the STW warping function is now made up with a series of B-splines, constructed from polynomial parts that are joined at certain points of the abscissa axis, then, the input parameters are the number of splines chosen. The warping function for the alignment in the STW technique can be written as Eq. (10):

$$w(t_i) = \sum_{j=1}^{n} a_j b_j(t_i) \tag{10}$$

In Eq. (4), n represents the number of B-splines, a_j the warping coefficient for the jth B-spline, the latter indicated as b_j, and t_i represents the abscissa axis computed as $t_i = ih$ with $i = 1, \ldots, m$, and m is the length of the signal, h represents the sampling interval.

Related to the STW approach, Daszykowski et al. [39] modeled an alignment function for the development of an automated alignment method. The alignment function is modeled explicitly; it is continuous and smooth and can be approximated by several spline functions. To treat complex retention shifts of chromatographic fingerprints, the use of a large number of spline functions is required.

In the case of warping functions like in the PTW algorithm, once the polynomial equation for the alignment is found, the warping coefficients must be optimized. The aim is to find the warping function that minimizes the sum of squared residuals S between the reference signal and the interpolated signal. An objective function S is defined by Eq. (11):

$$S = \sum [y_i - x(w(t_i))]^2 \tag{11}$$

Similar to PTW, the aim of STW algorithm is to optimize the warping function, so the sum of squared residuals S is minimized.

3.5 Peak matching algorithm

There are two types of warping path; one based on the warping of the whole signal and another based on peak detection followed by a warping of the regions between those peaks. For the method involving peak detection [40–42], on both signals, reference and sample, peaks are identified and a list of peak positions is generated. Peaks can be automatically detected by finding zero crossings in an estimate of the signal's first derivative. Reference and sample signals are compared. For each reference peak, the algorithm is looking for the sample peak, which most closely matches it in a user-defined window width. Two cases are possible. Either the distance between the peaks is lower than the threshold distance, then the peaks are

matched. Or there is no matching in the window width, which means that the peak is missing in the sample signal. For the first case, the regions between peaks in the sample signal are stretched and shrunk by interpolating more or fewer points to align them with the peaks in the reference signal. Therefore, the selection of the optimum peak matching window width is crucial for the alignment. Based on the comparison between the retention times of each detected compound in a sample, a free alignment program called Gcaligner 1.0 [43] has been developed to compare GC data matrix. The peak detection system being problematic, it is more appropriate to align all the points of signal, not just the peaks.

3.5.1 Piecewise alignment

Piecewise alignment [44] operates like the COW by subdividing the data into local windows of length W, but then each window is iteratively shifted along the sample signal with a specified limit l, and no stretching or compressing interpolation step is done.

3.6 Validation of alignment results

Even if all the methods have proved their ability to align different signals, it should be important to measure the quality of the alignment.

Many alignment methods use the Pearson's correlation coefficients to evaluate the quality of the alignment. Indeed, the measure of the similarity between two datasets R, the match reference, and S, the warping sample section with N points each, is well expressed by the correlation coefficient defined by Eq. (12):

$$\rho(R, S) = \frac{\sum RS - \frac{\sum R \sum S}{N}}{\sqrt{\left(\sum R^2 - \frac{\left(\sum R\right)^2}{N}\right)\left(\sum S^2 - \frac{\left(\sum S\right)^2}{N}\right)}} = \frac{\mathrm{cov}(R, S)}{\sqrt{\mathrm{var}(R)\mathrm{var}(S)}} \quad (12)$$

This correlation coefficient indicates the degree to which the two signals are linearly related. Two identical signals have a correlation coefficient that equals 1. The higher the correlation coefficient, the higher the quality of the alignment between the two signals.

One of the first to use this quality criterion is Nielsen et al. [15], with the development of the COW. The measure is influenced by the type of peaks that are shifted (smaller or larger on the abscissa axis). Chronologically, these methods are peak matching, PAGA, PARS, peak alignment by beam search, RSPA, Icoshift, aligneDE.

4 Practical aspects and drawbacks

4.1 Data and experimental conditions

All the alignment techniques publicly and freely available were tested on simulated datasets as well as real chromatographic datasets to assess the extent of the alignment algorithms capabilities.

Simulated datasets have the advantage that important features can be controlled; two major categories of shifts are simulated, small shifts and large shifts. The concept of small shift corresponds to a shift less or equal to the peak width to align. In contrast, the concept of large shift corresponds to a shift greater or equal to the peak width to align. In these categories, four types of shift are simulated, one-direction shift systematic and unsystematic, two-direction shift systematic and unsystematic. As explained in Section 2.2, the term systematic means that the shift is the same throughout the signal, i.e. when all peaks move by a constant offset. The term "two-direction shift" means that peaks in the signal are shifted in both directions. Tests were achieved on three real large chromatographic datasets that differ in the complexity of the signal, i.e. the different types of shifts described as simple, intermediate, and complex. Three samples corresponding to the three levels of complexity were analyzed twice with the same chromatographic method at an interval of one month, the first signal corresponds to the reference signal and the second to the sample signal.

4.2 Software

The computations were run on a computer equipped with an Intel Core 2 Duo (2.10 GHz) processor, 4.00 Go RAM, and the Microsoft Windows 7 operating system. GC analyses were performed using Agilent Model 6890 gas chromatograph with FID. Each chromatographic run was 20 min long with FID readings acquired at a rate of 200 Hz, yielding 240 000 points per chromatogram. The chromatograms were imported from the Chemstation (Agilent Technologies) into MatLab 7.14 software (MathWorks, Natick, MA, USA), where the data processing was done. Table 1 summarizes the sources of all freely available MatLab routines used in data processing.

4.3 Aligning of simulated and real data

Two quality criteria were investigated: one to measure the accuracy of alignment and the other to measure the performance of the technique. The alignment accuracy is established by the Pearson's correlation coefficient (R), which is the most popular measure in the literature to quantify the similarity between two datasets. The performance is characterized by the computation time (T). The optimal alignment technique should require a short computation time because an algorithm has to be fast to be useful and valuable, and

Table 1. Sources of the MatLab routines

Alignment algorithm	Author(s)	Website
COW	***Giorgio Tomasi, Frans van den Berg, Thomas Skov*** University of Copenhagen, Faculty of Life Sciences, Department of Food Science, Quality and Technology, Section Spectroscopy and Chemometrics, Denmark E-mail: gt@kvl.dk, fb@kvl.dk, thsk@kvl.dk	www.models.kvl.dk/DTW_COW
DTW		
PAGA	***Jenny Forshed*** Department of Analytical Chemistry, Stockholm University, Stockholm, Sweden E-mail: jenny@forshed.se	http://www.forshed.se/jenny/index.php?n = Research.SoftwareAmpCode
PAFFT	***Jason W. H. Wong*** Physical and Theoretical Chemistry Laboratory, Chemistry Department, Oxford University, England E-mail: Jason.wong@chem.ox.ac.uk	http://powcs.med.unsw.edu.au/research/adult-cancer-program/services-resources/specalign
RAFFT		
Icoshift	***Francesco Savorani*** University of Copenhagen, Faculty of Life Sciences, Department of Food Science, Quality and Technology, Section Spectroscopy and Chemometrics, Denmark E-mail: frsa@life.ku.dk	http://www.models.life.ku.dk/icoshift
Piecewise alignment	***Jérémy Nadeau*** Université de Washington, The Synovec Research Group E-mail: nadio@uw.edu	http://synoveclab.chem.washington.edu

only a minimal or no operator intervention. The usefulness of seven algorithms was studied in this section using both simulated and real data: COW, DTW, PAGA, PAFFT, RAFFT, Icoshift, piecewise alignment. For space reasons the single implementations will not be discussed in detail, the MatLab 7.14 guidelines for improving performances were followed.

The results obtained through each alignment algorithm arise from the optimization of the input parameters of each algorithm because the running time of COW, DTW, and PAGA algorithms depends on the parameterization. This is particularly true in the case of alignment with the COW, parameters, number of segments N, and flexibility s, which must be optimized to the extent that the computation time increases with the increase of these values. The aim is to determine the parameters with the lowest values that provide the best alignment signals. At each test on simulated and real datasets, several combinations of N and s were tested and for each of them, the correlation coefficients between the reference signal and the sample signal and the calculation time have been evaluated. By analogy, in the case of the piecewise alignment, window width W and offset L settings were optimized in the same way. For the PAGA, parameters required are three times more than normal parameters involved; the segmentation as well as the genetic algorithm are the parameters that have to be optimized.

It was found that several different combinations of parameters can provide similar correlation coefficients, the computation time being different. There is therefore an optimal combination of parameters for which a high correlation coefficient is obtained in a minimum time.

4.3.1 Simulated datasets

Simulated signals were computer-generated with 800 data points in the form of three Gaussian peaks centered at $t1 = 100$, $t2 = 300$, and $t3 = 500$, respectively. Shifts were artificially induced to test the potential alignment of the algorithms. Tables 2 and 3 sum up the results of alignment for each algorithm tested for different types of shift on simulated data.

Figure 6 shows the alignment of one-direction systematic shift with the COW algorithm, the dashed line corresponds to the reference signal, the dotted line to the sample signal, and the solid line to the aligned signal. This figure shows that the choice of the input parameters is crucial for the accuracy of the alignment. Indeed, for a slack of 20, a segment width of 50 provides a correct alignment while a segment width of 40 results in a broadening of the first peak.

Results obtained show that the tested algorithms are able to align the simulated data with varying degrees of success and computation times shorter or longer. DTW and PAGA are the greatest time consumers, but DTW is much more interesting given the quality of alignment for all types of small and large shifts. The simulated signals being composed of few data points, in the case of small shifts (Table 2) except the PAGA, all algorithms are capable of achieving proper alignment (correlation coefficients very close to 1). In the case of large shifts (Table 3), the PAGA, the COW, and the piecewise alignment do not allow proper alignment signals (correlation coefficients inferior to 0.98). The bad performances of the PAGA and the COW could be explained by a poor

Table 2. Alignment algorithms behavior for different type of small shifts (noncomplex) on simulated datasets

Type of shifts	ODSS		TDSS		ODSU		TDSU	
Alignment method	*R*	*T*	*R*	*T*	*R*	*T*	*R*	*T*
COW	0.995	Inst.	0.976	Inst.	0.997	Inst.	0.997	Inst.
DTW	1.000	60 s	1.000	60 s	1.000	60 s	1.000	60 s
PAGA	0.919	40 s	0.967	50 s	0.928	40 s	0.926	40 s
PAFFT	1.000	Inst.	0.999	Inst.	1.000	Inst.	0.999	Inst.
RAFFT	1.000	Inst.	1.000	Inst.	1.000	Inst.	1.000	Inst.
Icoshift	1.000	Inst.	0.989	Inst.	0.984	Inst.	0.990	Inst.
Piecewise alignment	1.000	2 s	1.000	2 s	1.000	2 s	1.000	2 s

ODSS, one-direction shift systematic; TDSS, two-direction shift systematic; ODSU, one-direction shift unsystematic; TDSU, two-direction shift unsystematic; *R*, Pearson's correlation coefficient; *T*, computation time; Inst., instantly.

Table 3. Alignment algorithms behavior for different type of large shifts (complex) on simulated datasets

Type of shifts	ODSS		TDSS		ODSU		TDSU	
Alignment method	*R*	*T*	*R*	*T*	*R*	*T*	*R*	*T*
COW	0.868	Inst.	0.970	3 s	0.992	1 s	0.976	3 s
DTW	1.000	60 s	0.999	60 s	1.000	60 s	1.000	60 s
PAGA	0.897	55 s	0.899	80 s	0.974	55 s	0.840	55 s
PAFFT	1.000	Inst.	0.999	Inst.	1.000	Inst.	0.995	Inst.
RAFFT	1.000	Inst.	1.000	Inst.	1.000	Inst.	1.000	Inst.
Icoshift	1.000	Inst.	0.999	Inst.	0.999	Inst.	0.999	Inst.
Piecewise alignment	0.995	2 s	0.557	2 s	0.989	2 s	0.778	2 s

ODSS, one-direction shift systematic; TDSS, two-direction shift systematic; ODSU, one-direction shift unsystematic; TDSU, two-direction shift unsystematic; *R*, Pearson's correlation coefficient; *T*, computation time; Inst., instantly.

optimization of the input parameters. Concerning the Icoshift algorithm, alignment of one-direction systematic shift is correct with the "whole signal" alignment mode. However, for the three others shift types, the correlation coefficients obtained with the alignment mode "whole signal" are $R_{\mathrm{DDDS}} = 0.299$, $R_{\mathrm{DDDN}} = 0.387$, and $R_{\mathrm{DUSN}} = 0.750$, it is therefore necessary to use another alignment mode. In this case, the one that yielded the values in Table 3 and thus a correct alignment is to specify intervals of definition.

4.3.2 Real datasets

Each chromatogram tested was located into a MatLab workspace as a vector composed of the time series of FID detector readings over the duration of that particular GC run. Table 4 recaps the results of alignment for each algorithm tested for signal complexity on real large chromatographic datasets.

The computation time required for alignment of large datasets with the COW algorithm being much too high, the results have not been classified in Table 4. Given its configuration, the computing capabilities of the computer cannot reach the end of the calculation. Indeed, the computer freezes after 90 min computing. In the case of the DTW, the computer crashes immediately after launching the calculation.

The alignment module of piecewise alignment is provided with a parameters optimization module, which in the case of large datasets reached its limits and becomes unsatisfactory. The computation times are much lower than other methods of alignment using DP optimization, as the COW, but remain high for routine use (7 min for complex signals alignment).

The alignment algorithm PAGA is very difficult to handle because of the many input parameters to optimize, the alignment results were obtained with empirical values considered optimal. Better results can probably be obtained with a better parameter optimization, the use of design of experiments could be envisaged to improve more rationally the potential of the alignment algorithm.

It is clear from the results that the algorithms that enable proper alignment instantly are those based on the calculation of the cross-correlation coefficient by Fourier transform. Comparing these three algorithms, the one that provides the highest quality of alignment in a reasonable amount of computation time is the RAFFT.

Table 5 summarizes the characteristics, recommendations, advantages, disadvantages, and the resulting alignment of the studied methods. The methods that have not been tested are mostly paying methods; those proposals were evaluated based on studies in the literature.

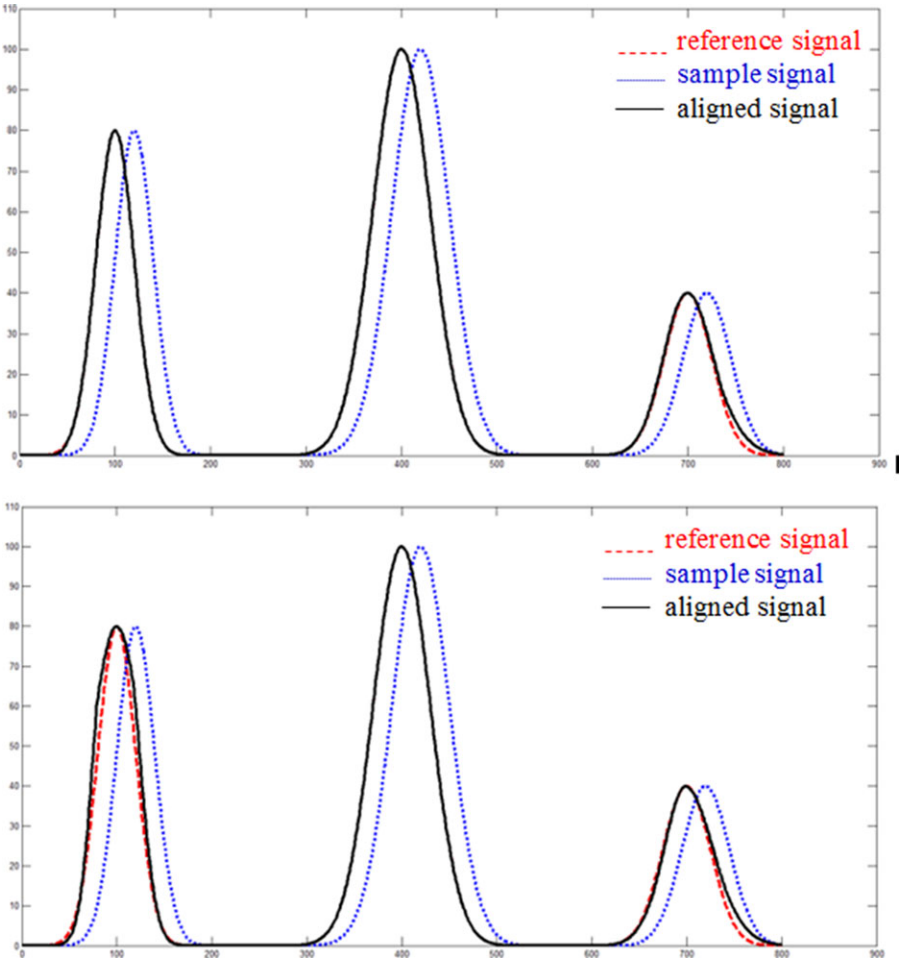

Figure 6. Alignment treatment of ODSS by COW (a) $N = 50$, $s = 20$, (b) $N = 40$, $s = 20$.

Table 4. Alignment algorithms behavior for different signal complexity on real datasets

Signal complexity	Simple		Intermediate		Complex	
Alignment method	*R*	*T*	*R*	*T*	*R*	*T*
Piecewise alignment	0.980	8 s	0.988	2 min	0.925	7 min
PAGA	0.988	2 min15 s	0.987	10 min	0.911	18 min
PAFFT	0.977	Inst.	0.805	Inst.	0.878	Inst.
RAFFT	0.996	Inst.	0.987	3 s	0.981	5 s
Icoshift	0.977	Inst.	0.967	Inst.	0.893	Inst.

4.4 Drawbacks

If so many alignment techniques have been developed, it is because each of them has drawbacks in certain aspects. All alignment techniques need a reference signal to process the alignment, so the reference selection appears to be a crucial concern. Daszykowski et al. [43] analyzed through several datasets of chromatographic signals various proposals for reference selection. Their study established that the optimal reference is the one that provides the highest mean correlation coefficient with respect to the remaining chromatograms.

Warping methods using DP are time consuming or require time-consuming optimization of input parameters [5]. As many interpolation steps are required, the computational complexity is the major impediment of the COW, the DTW, and the piecewise alignment technique. Indeed, the processing time of calculation is proportional to N^2, where N is the number of data. The choice of the input parameters can affect the peak's features and lead to peak deformation. Indeed, a segment length too short or not positioned correctly and a slack too large can significantly change peak shapes. Therefore, there is an optimal segment length and slack parameter for which a high correlation is obtained in the shortest time.

Table 5. Summary of the discussed alignment methods

Alignment method	Input parameters	Peak picking	Application areas	Form of shifts	Ease of use	Speed of execution	Advantages	Disadvantages
COW	Segment size, slack size	No	Chromatography, spectral data	Noncomplex, two-direction shifts	−−	−	- Efficiency of the alignment	- Time consuming on large datasets - Optimization of input parameters - Too small segments and too large flexibility can cause distortion of peaks
DTW	Constraints	No	Chromatography, spectral data	Complex, two-direction shifts	+−	−−	- Efficiency of the alignment	- Time consuming - Inability to handle more data than 1000 points - Sensitive to peak intensities - Synchronization step can produce artifacts
PARS (untested)	Constraints	Yes	Chromatography, NMR	Complex two-direction shifts	+	+	- No alteration in peak shapes - Fast algorithm	- Lack of accuracy for the alignment of minor peaks - Misalignment of overlapped signals regions - Imprecise matching criterion (it assumes the closest peak is the right match)
PAGA	Segment size, maximum range of sideways movements, maximum range of interpolation, genetic algorithm parameters	No	NMR	Noncomplex, one-direction shifts	−−	−−	- Avoid cutting in a peak	- Time consuming - Optimization of the numerous input parameters
PABS	Beam width, maximum range of sideways movements, maximum range of interpolation	No	NMR	Complex, two-direction shifts	−	+	- Faster than the PAGA	- Optimization of the input parameters - Low alignment performance in complex cases
DE (untested)	Slack size	No	Chromatography	Complex, two-direction shifts	+	+	- Combination of warping, peak detection, and evolutionary algorithm to overcome shortcomings of each algorithm - No alteration in peak shape, peak height, and peak area	None

Table 5. Continued

Alignment method	Input parameters	Peak picking	Application areas	Form of shifts	Ease of use	Speed of execution	Advantages	Disadvantages
PAFFT	Segment size, shift freedom	No	Chromatography	Complex, two-direction shifts	++	++	- Fast algorithm	- Selection of the optimal segment length
RAFFT	None	No	Chromatography	Complex, two-direction shifts	++	++	- No operator intervention - Fast algorithm	- Rare alteration in peak shapes
Icoshift	Interval definition	No	NMR	Complex, two-direction shifts	+	++	- Fast algorithm - Possibility of defining intervals for better alignment	- Definition of intervals is sometimes required for the alignment hence a waste of time
PTW (untested)	Polynomial warping function	No	Chromatography	Noncomplex, one-direction shifts	+	+	- Fast algorithm	- Lack of flexibility for complex shifts alignment
STW (untested)	Number of B-splines	No	Chromatography	Complex, two-direction shifts	+	+	- More flexible than the PTW	- Selecting the number of B-splines can cause too much flexibility and distort peaks
Peak matching (untested)	Window size, threshold matching distance	Yes	Chromatography	Noncomplex, one-direction shifts (typical distance between two adjacent peaks)	−	+	- Easy to understand	- Sensitive to peak intensities - Dependence about the way the signal is derived - Imprecise matching criterion (it assumes the closest peak is the right match) - No optimization method

Table 5. Continued

Alignment method	Input parameters	Peak picking	Application areas	Form of shifts	Ease of use	Speed of execution	Advantages	Disadvantages
Fuzzy warping (untested)	Threshold for peak picking, width of Gaussian peaks	Yes	Chromatography, NMR	Noncomplex, one-direction shifts	+	+	- Combination of warping and peak detection	- Require an estimation of the threshold value for the first derivative - Imprecise matching criterion (it assumes the closest peak is the right match) - Lack of accuracy for the alignment of minor peaks
Piecewise alignment	Window size, maximum scalar shift	No	Chromatography, NMR	Complex, one-direction shifts	++	+	- Efficiency of the alignment for one-direction shifts	- Time consuming - Optimization of input parameters
PLF (untested)	Segment size, shift freedom	Yes	NMR	Complex, two-direction shifts	++	+	- Easy to understand	- Not powerful enough to handle complex shifts
GFHT (untested)	Locations of a few user-specified peaks clearly assigned	Yes	NMR	Complex, two-direction shifts, peak shifts described by single parameter model	+	+	- No target selection	- The algorithm converges to local maximum instead of global maximum for complex shifts

PABS, peak alignment by beam search.

As a matter of fact, Skov et al. [20] introduced a routine that automatically optimizes the segment length and slack size for the COW alignment. The selection procedure of the input parameters uses a discrete-coordinates simplex-like optimization routine.

In the PAFFT, the selection of the optimal segment length can be a matter, but this problem has been fixed with the recursive approach of the RAFFT that minimizes the operator intervention. Recursive alignment is achieved by aligning the full signal on a global scale to progressively smaller segments until no further alignment is required [34].

For long or complex signals, the optimization of the input parameters (segments size and slack) can be quite time consuming in the implementation of the PAGA. Large segment size can result in nonalignment of small peaks and short segment can result in distortions in peaks due to peak being cut at segment boundaries [27].

The DTW has been widely used in many scientific fields and the conclusion put forward is that, like the peak matching technique, the algorithm is sensitive to peak intensities [19]. Explicitly, the location of peaks with a small S/N is problematic because a small peak can be missed due to the low detection threshold. Another drawback of the peak matching technique is that the achievement of the approach depends much about the way the signal is derived. No specific shoulder peak detection is used for larger peaks, some are detected others are not depending on their sizes in comparison to the parent peak [13]. To return to the DTW, the step following the construction of the optimal path is called a synchronization step. This step is also an issue because it produces artifacts. There are two versions, symmetric and the asymmetric synchronization. The one used by Pravdova et al. is the asymmetric one: when a vertical transition occurs (more than one point of the warped signal aligned with the same point in the reference), the average of these points is calculated and aligned with the corresponding point in the reference signal. In the case of the symmetric version, when a vertical or horizontal transition occurs, the response of the signal for the index involved is not averaged but taken twice.

The PTW is certainly a fast algorithm, but its lack of flexibility does not allow the alignment of complex chromatographic shifts [37]. In fact, for PTW the flexibility is restricted as only a global second-order polynomial (quadratic) warping function. In consequence, only systematic shifts (shifts in one direction) or noncomplex shifts in both directions can be treated. The STW is able to correct complex and unsystematic shifts in two directions, but it requires an optimization of the penalty term λ and the number of B-splines to avoid peak shape changes due to a too large flexibility. The quality of the alignment depends on the S function, the sum of square residuals. The more the number of B-splines is high, the more the warping function will produce variation, which is not necessarily valuable for the alignment, and which will cause a useless increase of time of calculation [5]. To reduce this flexibility, a penalty term is added to the initial equation introduced in the paragraph entitled STW.

$$S = \sum [y_i - x(w(t_i))]^2 + \lambda \sum_{j=k+1}^{n} (\Delta^k a_j)^2 + \kappa \tag{13}$$

The first term is the sum of square residuals with y_i reference signal and $x(w(t_i))$ the interpolated signal. The second term is the penalty term, n number of B-splines, a_j warping coefficients of the jth B-splines, the parameter λ controls the smoothing, and the parameter κ works out the flexibility (if κ increases, the flexibility decreases). Too many B-splines will result in too much flexibility and deformations in peak shape, so this number of B-splines has to be optimized to the best compromise, enough flexibility with the shortest computation time.

The PARS relied on simple peak detection that allows the alignment of major peaks but causes a lack of accuracy concerning the alignment of minor peaks [32]. The correspondence of peaks in overlapped signal regions is difficult and can also result in misalignment [36].

5 Conclusion

All the methods introduced in this paper are able to properly correct the simplest signals, i.e. those with small shifts. Therefore, to not waste time unnecessarily, it seems wise to treat the simplest signals with the less time-consuming methods that are often the basic ones. As regards the more complex cases, the operator has to rely on elaborated methods that can be more time consuming. The aim is to avoid the alignment of major peaks at the expense of minor peak shift precision to get a suitable alignment quality. In consequence, a compromise between the quality of the alignment and the computation time has to be set up according to the type of purpose to reach. All alignment techniques tested need a reference signal to process the alignment, so the reference selection appears to be a crucial concern. Daszykowski et al. [45] analyzed several datasets of chromatographic signals various proposals for reference selection. Their study established that the optimal reference is the one that provides the highest mean correlation coefficient with respect to the remaining chromatograms. Indeed, the quality of the alignment depends on the reference selection that must be the most representative signal among all signals.

Another major issue is the number of data points per signal. If this number is small (i.e. ≤10 data points per peak), the interpolation steps can result in considerable changes in area and thus increase the ambiguity of the area estimation. This can be solved by interpolating to a higher resolution before the alignment procedure. Though, increasing the number of data points leads to a more time-consuming alignment process. Therefore, a less-restrictive means would be to integrate the signal before alignment rather than the aligned signal that may have undergone changes and distortions that could

alter the results of integration. However, the area preservation issue depends on what you want to do with the data after alignment. This issue is much more problematic in quantitative analysis than in qualitative analysis.

Ideally, the research aims to create highly automated software that allows, regardless of the complexity and type of shifts of the signals, an alignment in the shortest time and that does not require intervention from the operator.

The authors have declared no conflict of interest.

6 References

[1] Geladi, P., *Spectrochim. Acta Part B* 2003, *58*, 767–782.

[2] Duarte, A. C., Capelo, S., *J. Liq. Chromatogr. Relat. Technol.* 2006, *29*, 1143–1176.

[3] Daszykowski, M., Walczak, B., *TrAC, Trends Anal. Chem.* 2006, *25*, 1081–1096.

[4] Malmquist, G., Danielsson, R., *J. Chromatogr. A* 1994, *687*, 71–88.

[5] Van Nederkassel, A. M., Xu, C. J., Lancelin, P., Sarraf, M., MacKenzie, D. A., Walton, N. J., Bensaid, F., Lees, M., Martin, G. J., Desmurs, J. R., Massart, D. L., Smeyers-Verbeke, J., Vander Heyden, Y., *J. Chromatogr. A* 2006, *1120*, 291–298.

[6] Jellema, R. H., *Comprehensive Chemometrics*, Elsevier, Oxford, 2009, pp. 85–108.

[7] Bloemberg, T. G., Gerretzen, J., Wouters, H. J., Gloerich, J., van Dael, M., Wessels, H. J. C., van den Heuvel, L. P., Eilers, P. H., Buydens, L., Wehrens, R., *Chemom. Intell. Lab. Syst.* 2010, *104*, 65–74.

[8] Åberg, K. M., Alm, E., Torgrip, R. J. O., *Anal. Bioanal. Chem.* 2009, *394*, 151–162.

[9] Torgrip, R. J. O., Alm, E., Åberg, K. M., *Bioanal. Rev.* 2010, *1*, 105–116.

[10] Guiochon, G., Guillemin, C. L., *Quantitative Gas Chromatography: For Laboratory Analyses and On-Line Process Control*, Elsevier, The Netherlands 1998.

[11] Tranchant, J., Chromatographie en Phase Gazeuse (Ed. Techniques Ingénieur) 1995.

[12] Van Nederkassel, A. M., Aerts, A., Dierick, A., Massart, D. L., Vander Heyden, Y., *J. Pharm. Biomed. Anal.* 2003, *32*, 233–249.

[13] Johnson, K. J., Wright, B. W., Jarman, K. H., Synovec, R. E., *J. Chromatogr. A* 2003, *996*, 141–155.

[14] Etxebarria, N., Zuloaga, O., Olivares, M., Bartolomé, L. J., Navarro, P., *J. Chromatogr. A* 2009, *1216*, 1624–1629.

[15] Nielsen, N.-P. V., Carstensen, J. M., Smedsgaard, J., *J. Chromatogr. A* 1998, *805*, 17–35.

[16] Ramaker, H.-J., van Sprang, E. N. M., Westerhuis, J. A., Smilde, A. K., *Anal. Chim. Acta* 2003, *498*, 133–153.

[17] Hillier, F. S., Lieberman, G. J., *Introduction to Mathematical Programming*, McGraw-Hill, New York.

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

[19] Pravdova, V., Walczak, B., *Anal. Chim. Acta* 2002, *456*, 77–92.

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

[21] Szymańska, E., Markuszewski, M. J., Capron, X., van Nederkassel, A.-M., Vander Heyden, Y., Markuszewski, M., Krajka, K., Kaliszan, R., *Electrophoresis* 2007, *28*, 2861–2873.

[22] Wang, C. P., Isenhour, T. L., *Anal. Chem.* 1987, *59*, 649–654.

[23] Kassidas, A., Taylor, P. A., MacGregor, J. F., *J. Process Control* 1998, *8*, 381–393.

[24] Torgrip, R. J. O., Åberg, M., Karlberg, B., Jacobsson, S. P., *J. Chemom.* 2003, *17*, 573–582.

[25] Smith, T. F., Waterman, M. S., *J. Mol. Biol.* 1981, *147*, 195–197.

[26] Sakoe, H., Chiba, S., *IEEE Trans. Acoust. Speech Signal Process.* 1978, *26*, 43–49.

[27] Forshed, J., Andersson, F. O., Jacobsson, S. P., *J. Pharm. Biomed. Anal.* 2002, *29*, 495–505.

[28] Forshed J., Schuppe-Koistinen I., Jacobsson S. P., *Anal. Chim. Acta* 2003, *487*, 189–199.

[29] Forshed, J., Torgrip, R. J., Åberg, K. M., Karlberg, B., Lindberg, J., Jacobsson, S. P., *J. Pharm. Biomed. Anal.* 2005, *38*, 824–832.

[30] Kaya, H., Gündüz-Öğüdücü, S., *Inform. Sci.* 2013, *228*, 113–130.

[31] Lee, S.-Y., *Comput. Stat. Data Anal.* 1985, *2*, 279–295.

[32] Zhang, Z. M., Chen, S., Liang, Y. Z., *Talanta* 2011, *83*, 1108–1117.

[33] Storn, R., Price, K., *J. Glob. Optim.* 1997, *11*, 341–359.

[34] Wong, J. W. H., Durante, C., Cartwright, H. M., *Anal. Chem.* 2005, *77*, 5655–5661.

[35] Savorani, F., Tomasi, G., Engelsen, S. B., *J. Magn. Reson.* 2010, *202*, 190–202.

[36] Veselkov, K. A., Lindon, J. C., Ebbels, T. M. D., Crockford, D., Volynkin, V. V., Holmes, E., Davies, D. B., Nicholson, J. K., *Anal. Chem.* 2009, *81*, 56–66.

[37] Zhong, L., Wang, J., Huang, J., Zhang, Z., Lu, H., Zheng, Y., Zhan, D., Liang, Y., *J. Sep. Sci.* 2013, *36*, 1677–1684.

[38] Eilers, P. H. C., *Anal. Chem.* 2004, *76*, 404–411.

[39] Daszykowski, M., Vander Heyden, Y., Boucon, C., Walczak, B., *J. Chromatogr. A* 2010, *1217*, 6127–6133.

[40] Vivó-Truyols, G., Torres-Lapasió, J. R., van Nederkassel, A. M., Vander Heyden, Y., Massart, D. L., *J. Chromatogr. A* 2005, *1096*, 133–145.

[41] Peters, S., Velzen, E., Janssen, H.-G., *Anal. Bioanal. Chem.* 2009, *394*, 1273–1281.

[42] Debrus, B., Lebrun, P., Ceccato, A., Caliaro, G., Govaerts, B., Olsen, B. A., Rozet, E., Boulanger, B., Hubert, P., *Talanta* 2009, *79*, 77–85.

[43] Dellicour, S., Lecocq, T., *J. Sep. Sci.* 2013, *36*, 3206–3209.

[44] Pierce, K. M., Hope, J. L., Johnson, K. J., Wright, B. W., Synovec, R. E., *J. Chromatogr. A* 2005, *1096*, 101–110.

[45] Daszykowski, M., Walczak, B., *J. Chromatogr. A* 2007, *1176*, 1–11.