Automated Alignment of Chromatographic Data

Thomas Skov*, Frans van den Berg & Rasmus Bro

Quality and Technology, Department of Food Science (IFV), The Royal Veterinary and Agricultural University (KVL)
Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark
*corresponding author, e-mail: thsk@kvl.dk

SUMMARY

This paper focuses on the practical aspects and implications of preprocessing chromatographic data to correct for undesirable time-shifts. An approach to automate the alignment of chromatographic data based on warping is proposed. This approach deals with selection of the essential parameters including selection of reference sample to warp towards, and choose warping settings based on a new evaluation criterion for goodness of alignment. The new criterion aims at quantifying goodness of alignment while at the same time penalizing significant shape or area-changes in the warped peaks. The whole selection procedure is automated using a discrete-coordinates simplex-like optimization routine. Examples with simulated chromatographic data, GC-FID and HPLC-Fluorescence measurement series illustrate the benefit of using this automated alignment tool.

KEYWORDS: Automated alignment, Correlation Optimized Warping, Peak area preservation, Chromatographic data, Optimization

1. INTRODUCTION

Preprocessing of chromatographic data to correct for undesirable phenomena is often a crucial step in the proper data analysis chain. This holds especially if the data are to be used for multivariate data analysis either in the form of peak areas or raw chromatograms. Some of these "artifacts" can be taken care of using traditional chromatographic procedures such as correcting the signal by internals standards or normalization. Other, more challenging artifacts such as peak shifting and baseline variations need more advanced preprocessing techniques to remove their undesired contribution in the subsequent data analysis steps like Principal Components Analysis (PCA) or PARAllel FACtor analysis (PARAFAC). The reason is that if data are not brought to a form where elements in the matrix or data cube for individual objects or samples describe the same phenomena, the required assumption of bi- or tri-linearity in the data is no longer valid. Several preprocessing methods have been put forward in literature to correct for shifted peaks in chromatographic data [1-7].

The Correlation Optimized Warping (COW) algorithm has shown great potential for alignment correction in chromatographic data, due to its peak shape and area preserving properties [1-4]. The COW algorithm is based on aligning a sample chromatogram in the form of a digitized vector towards a target chromatogram (i.e. a reference sample vector) by piecewise linear stretching or compression in combination with interpolation, optimizing the correlation coefficients between corresponding segments in reference and sample [3]. The same reference sample is used for correcting/aligning the entire data set. As the chromatograms are split in segments and all boundaries between segments are allowed to move a certain number of data points in either direction - the local flexibility of alignment - the COW algorithm requires two user input parameters: the *segment length* and the flexibility (so-called *slack size*). These two parameters are typically selected on a trial and error basis by visual inspection of the chromatic profiles after preprocessing (peak shape, width, etc.). An automated method to investigate whether these parameters are optimal has not yet been proposed in the literature. This paper introduces a new concept for this purpose that calculates a so-called *simplicity* value for each combination of input parameters. It can be used as a measure of the similarity of the shapes of the aligned chromatograms.

As shown in the result section of this paper one parameter combination of segment and slack size will give the highest simplicity value, but more combinations will provide simplicity values very close to the optimum. Since an exhaustive search will typically be too time consuming for representative research questions, the latter observation can be utilized in a stratified optimization procedure that investigates fewer combinations and provides a satisfying alignment in considerably less time. However, when aligning peaks consisting of only a small number of sample points, the interpolation step can cause a significant change in peak shapes and areas. This paper includes a second optimization criterion of minimizing area alterations in the optimization routine leading to a more conservative and reasonable measure of a correct choice of segment length and slack size.

An obstacle in alignment is the selection of reference sample, and so far no method is put forward in the literature that can suggest a good choice. The ideal reference sample should be as representative as possible for all phenomena of interest in the data set. This could e.g. be the chromatogram in the middle of the run sequence [2], the chromatogram containing the highest number of common chemical constituents (i.e. peaks) [3-4], a composite artificial sample, or the chromatogram that is most similar to the loading of the first principal component in a PCA model on the un-aligned data set. This paper discusses and demonstrates a simple and quick way of selecting a reference in a given data set based on the product of correlation coefficients.

In this paper we focus on the practical aspects and implications of preprocessing chromatographic data to correct for undesired time-shifts. Optimization of the alignment of chromatographic data will be demonstrated on three separate data sets. A simulated set (Figure 1) will be used to explain the principles of the simplicity value and the importance of preserving the peak area as quantitative measure for an optimal alignment. Next, a GC-FID data set of ground coffee samples (results from this are NOT included in this presentation) will be used to illustrate the above on real data and to show the effect of selecting a suitable reference vector [3]. And a HPLC-Fluorescence data set is also included to show the possibilities of the ideas presented here for chromatographic method with broad-peak features (results from this are NOT included in this presentation). For further information about GC and HPLC the reader is referred to:

Skov, T., van den Berg, F., Bro, R. (2006). *Automated Alignment of Chromatographic Data*, Submitted for publication.

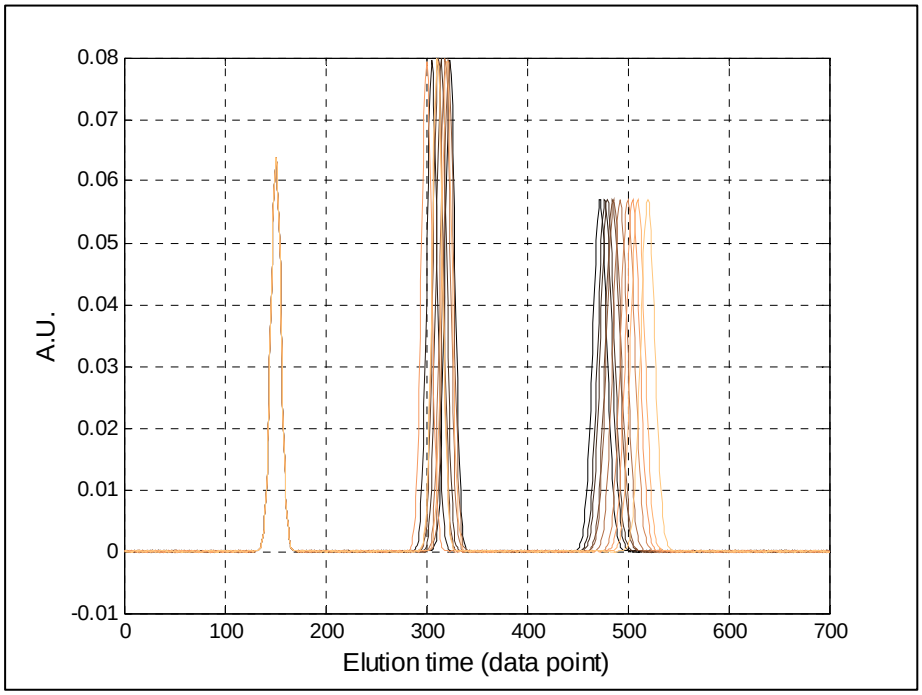

Figure 1. Illustration of ten simulated chromatograms, each containing three Gaussian peaks. First peak (left): no shifts, second peak (middle): random shifts and third peak (right): systematic peak shifts where a higher sample number is related to a later eluting time.

2. THEORY

2.1. Nomenclature and terminology

All samples will be referred to as sample chromatograms or simply *chromatograms*, independent of the analytical technique used. The direction along which the chemical constituents elute and where warping/alignment will be performed is referred to as *time*. Throughout this work, lowercase italics are used for scalars (i.e. x) and lowercase bold for column vectors (i.e. $\mathbf{x}$). T in superscript is the transpose operation (i.e. $\mathbf{x}^T$ is a row vector). Data matrices will be denoted with bold capital letters (i.e. $\mathbf{X}$). The ijth element of a $\mathbf{X}$ is thus denoted $x(i,j)$, where the indices run as $i = 1,\ldots,I$ and $j = 1,\ldots,J$.

2.2. Correlation optimized warping (COW)

The correlation optimized warping algorithm (COW) was introduced by Nielsen *et al*. [1] as a method to correct for shifts in discrete data signals. It is a piecewise or segmented data

preprocessing technique that aligns a sample chromatogram towards a reference chromatogram by stretching or compression of sample segments using linear interpolation. The theory of COW will not be explained further here but the reader is referred to the literature for more details [1,3-4].

2.3. Reference chromatogram selection

The selection of reference sample (i.e. reference chromatogram) is often made from *a priori* knowledge on the data set. This could e.g. be the chromatogram in the middle of the run sequence [2] or the chromatogram containing the highest number of common chemical constituents (i.e. peaks) [3-4]. However, to make sure that the most appropriate reference sample is selected in a given data set, a more objective method is needed. One solution could be to choose the chromatogram that is most similar to the loading of the first principal component in a PCA model on the unaligned data or simply to the mean of all chromatograms. Such a generic approach can be problematic because the mean chromatogram as well as the first loading from a PCA of the raw data will have too many or heavily distorted/broadened peaks due to the original problem at hand: the shifts present in the data set.

In this paper a method is presented, which is based on the product of the correlation coefficients between all individual chromatograms. For a given chromatogram $\mathbf{x}_t$, this *similarity index* (0 < similarity index ≤ 1) can be calculated as:

$$Similarity\ \ index = \prod_{i=1}^{I} \left| r\left(\mathbf{x}_t, \mathbf{x}_i\right) \right| \tag{1}$$

where $r(\mathbf{x}_t,\mathbf{x}_i)$ is the conventional correlation coefficient between two chromatograms in the data set calculated as:

$$r\left(\mathbf{x}_t, \mathbf{x}_i\right) = \frac{\sum_{j=1}^{J} \left(x_t(j) - \bar{x}_t\right)\left(x_i(j) - \bar{x}_i\right)}{\sqrt{\sum_{j=1}^{J} \left(x_t(j) - \bar{x}_t\right)^2 \sum_{j=1}^{J} \left(x_i(j) - \bar{x}_i\right)^2}} \tag{2}$$

with *J* being the number of data points in the chromatograms. Taken the absolute value in Equation (1) will safeguard the similarity index selection from the situation where strongly deviating samples in the data set will have a low correlation coefficient with arbitrary sign. But like in all data processing operations such samples should preferably be caught and removed before computations.

The similarity index for each sample in the set will be less then or equal to one (in case of perfectly aligned and identical chromatograms). The chromatogram that is most similar to all others will have the largest similarity index and is selected to be the most suitable reference chromatogram to use within the given data set.

2.4. The simplicity value

The overall goal when aligning chromatograms is to make the chromatographic profiles as similar in appearance as possible while preserving the peak shape and area. Stated differently, with the right preprocessing, the numerical rank of the data set, disregarding random noise, will be lowered to the chemical rank [3].

The *simplicity* value is used to measure how well aligned a set of chromatograms are. The principle of the simplicity value is related to the properties of the Singular Value Decomposition (SVD), where the size of the squared singular values is directly related to the variation explained in the data matrix. Any data matrix, **X** can be decomposed as:

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T \tag{3}$$

where **S** is a diagonal matrix containing the singular values equal to the eigenvalues of $\mathbf{X}^T\mathbf{X}$. **U** and **V** are both orthogonal matrices, where the columns in **U** are the eigenvectors of $\mathbf{X}\mathbf{X}^T$ and the columns of **V** the eigenvectors of $\mathbf{X}^T\mathbf{X}$. The sum of the eigenvalues equals the total sum of squares of all the original data entries in **X**. Thus, the ratio of each individual eigenvalue divided by the sum of all eigenvalues can be regarded as a measure of how much of the variation in **X** is partitioned into each eigenvalue.

In unaligned data the chromatographic profile will differ among samples and thus less of the total variation will be explained by the first few eigenvalues. This is due to the deviation from bilinearity, which causes more significant eigenvalues in the decomposition of the data matrix, and thus a higher rank. If data are aligned and all chromatographic profiles only differ in the magnitude from a common profile (**V**), then the above ratio for the first eigenvalue will be equal to one. Generally, better aligned chromatograms will result in fewer and larger significant eigenvalues, which would represent the (true) chemical information we are looking for.

The sum of the first *R* eigenvalues is a measure of how much of the variation is explained by these eigenvalues:

$$\sum_{r=1}^{R}\left(SVD\left(\mathbf{X} / \sqrt{\sum_{i=1}^{I}\sum_{j=1}^{J} x(i,j)^2} \right)\right)^2 = 1 \qquad (4)$$

where SVD(**M**) implies the vector of singular values – the diagonal of matrix **S** in (3) - and where the data are scaled to a total sum of squares of one for convenience. The above expression will always be one if all eigenvalues are retained (the mathematical rank), and as such this sum cannot be used to evaluate preprocessing and the effect of alignment. However, for finding the optimal combination of segment and slack size the following expression is put forward as the *simplicity* value ($0 \leq$ simplicity ≤ 1), where the principle of simplicity is adapted from Henrion & Andersson [8] and Christensen *et al*. [9]:

$$Simplicity = \sum_{r=1}^{R}\left(SVD\left(\mathbf{X} / \sqrt{\sum_{i=1}^{I}\sum_{j=1}^{J} x(i,j)^2} \right)\right)^4 \qquad (5)$$

This sum of the squared eigenvalues or singular values taken to the fourth power will be higher the more of the variation that is explained in the first components. As a simple example consider the two alternative series of singular values SVD(M_1) = $[1\ 0\ 0\ 0]^T$ and SVD(M_2) = $[½\ ½\ ½\ ½]^T$. Although the squared sum of the series are the same, the fourth-power sum is 1 and 1/8, respectively, indicating that in the first case, the data are more similar in shape as

seen from the lower numerical rank; they can be explained with one component only. In general the simplicity value will be noticeable smaller if the chromatograms are not well aligned. Having achieved perfect alignment the simplicity value will be closer to though not necessarily equal to one.

In COW alignment it is often possible to achieve high simplicity values with several combinations of segment and slack parameters. This is illustrated in Figure 4A for the simulated chromatograms. It is obvious that combining a small segment length with a large slack size (i.e. high flexibility) will result in interpolation steps over many data point and thus the possibility to align peaks efficiently, but this also carries the danger to undesirably change both shape and area of peaks. To avoid this potential pitfall we include a criterion in the optimization of simplicity that takes into account this "change in area" effect and can guide the selection of the optimal combination of segment and slack via an additional penalty term.

2.5. The peak factor

When aligning chromatograms the peak area and shape should ideally be the same before and after the procedure. One prerequisite for success is that the reference chromatogram has been carefully selected, but this alone cannot guarantee that peak shapes and areas do not change. We quantify the change by a measure called *peak factor* ($0 \leq$ peak factor ≤ 1). It indicates how much the sample set is changed when preprocessed by a certain combination of segment length and slack size:

$$Peak\ \ factor = \frac{\sum_{i=1}^{I}\left(1 - \text{minimum}(c(i),1)^2\right)}{I} \tag{6}$$

where,

$$c(i) = \frac{\|\mathbf{x}_w(i)\| - \|\mathbf{x}(i)\|}{\|\mathbf{x}(i)\|} \tag{7}$$

and $\|\mathbf{x}(i)\| = \sqrt{\sum_{j=1}^{J} x(i,j)^2}$ is the Euclidian length or norm for **x**(i); **x**(i) is the chromatogram before warping while $\mathbf{x}_w$(i) is the same sample after alignment. In this criterion (7), if the norm stays the same, the absolute term (or relative change) is 0, and the overall contribution for that sample is 1 in Equation (6). If a sample is almost the same, the absolute term will be between 0 and 1, and the overall contribution will be smaller than 1. When the warped sample is very distorted the absolute term will grow, and its overall contribution will be 0.

Values of the peak factor measure are shown in Figure 4B together with the simplicity values for the simulated data. Notice that some combinations of segment length and slack size provide high simplicity values but "low" peak factor values, and thus should not be considered as suitable alignment parameters.

2.6. The warping effect

The new quantitative measure combining the simplicity and the peak factor value is called the *warping effect* ($0 \leq$ warping effect ≤ 2):

$$Warping\ \ effect = simplicity + peak\ \ factor \tag{8}$$

The relation between the three measures - simplicity, peak factor and warping effect - is illustrated in Figure 2 for the simulated chromatograms.

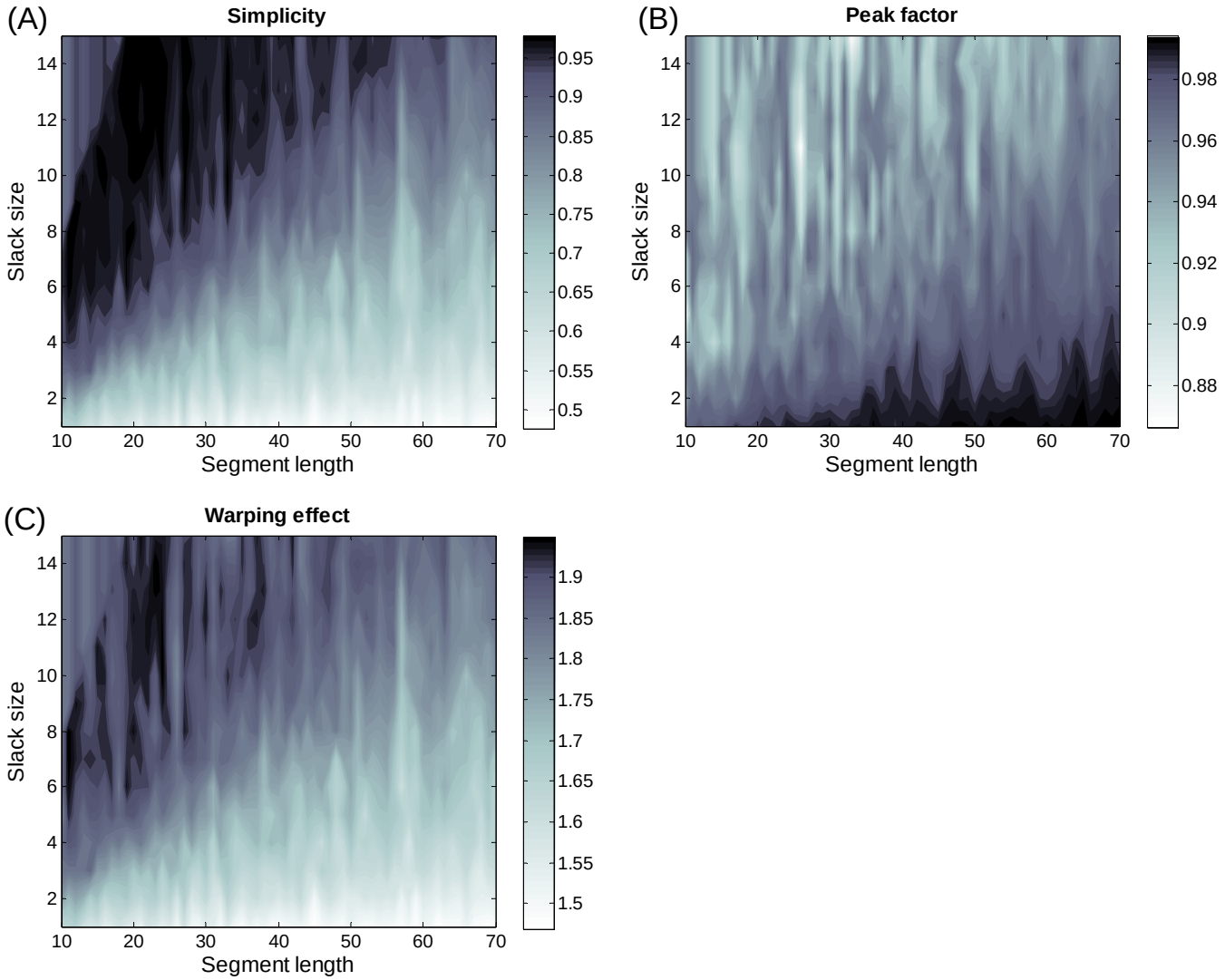

Figure 2. Simplicity (A), peak factor (B) and warping effect (C) values for all combinations of segment length and slack size using the simulated data. For plots (A) and (B) a value close to 1 indicates that data are well aligned and that the area has changed insignificantly, respectively. For plot (C) a value close to 2 means that peaks are both aligned and that the change in the area is minimal.

Figure 2 shows that some combinations of segment length and slack size result in a significant distortion of the profiles and thus are inappropriate choices for the alignment parameters (e.g. the dark colored band around segment length twenty five). The warping effect overall gives a smoother, less pronounced optimization landscape with more gradual changes as compared to the simplicity plot and thus an automated optimization routine will have a better chance of finding (near) optimal parameters of segment length and slack size in reduced computation time.

2.7. Optimization

An exhaustive search for the optimal combination of segment length and slack size will be rather time consuming for realistic problems, and as illustrated in Figure 2 is not rational as (near) optimal solution(s) (i.e. combinations) are present over a considerable, well defined area/corner of all segment length and slack size combinations. For the situation depicted in Figure 2 an exhaustive search involves the calculation of 71 × 15 = 1065 combinations. Many of these combination-areas are suboptimal, and thus, it would be more rational to limit the number of calculations to find a good starting point for further optimization. One way this can be done is by using a predefined sparse search grid, equally spaced over the search region combinations, as illustrated in Figure 3 for the simulated chromatograms.

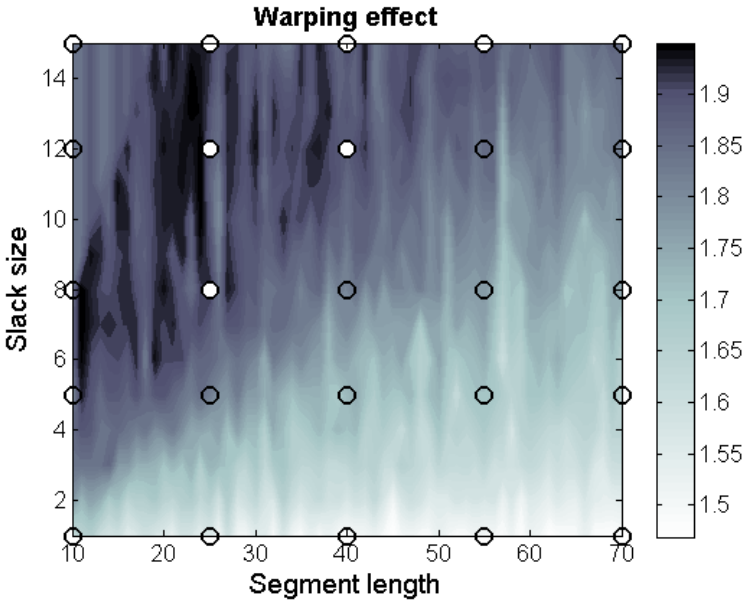

Figure 3. 5x5 sparse search grid (circles) for the global search space of segment length 10 to 70 and slack size 1 to 15 using the simulated data as an example. The six largest 'winning' parameter sets of the warping effect in this search grid are indicated with solid dots.

The optimization of the warping effect values is done from a discrete-coordinates simplex-like optimization routine carried out in three steps [10].

First step in our optimization is to establish global search space boundaries from the combination of all segment length and slack sizes of interest. In the second step a sparse global search grid is determined where we default select a 5×5 grid in both the segment and slack direction, as indicated Figure 3, and the warping effect for these 25 points is determined.

The six best (default choice) combinations, providing the highest warping effect scores, are then selected and used as starting points in the discrete-coordinates simplex optimization part. This routine is depicted in Figure 4. It works by establishing a triangle with the base (the intersection of the two legs) on the grid point (triangle I). The three points (combinations of segment and slack) are then calculated for as far as they were not known a priori, and the triangle is flipped over the axis formed by the two points possessing the smallest warping effect (I to II). If the value in the corner of the new triangle after evaluation is lower than the one found in the previous step the flip is not carried out (II to III). The routine instead flips the triangle over the second lowest axis and makes a new evaluation (II to IV). The optimization stops when no further flips are possible according to the rules explained above and visualized in Figure 4.

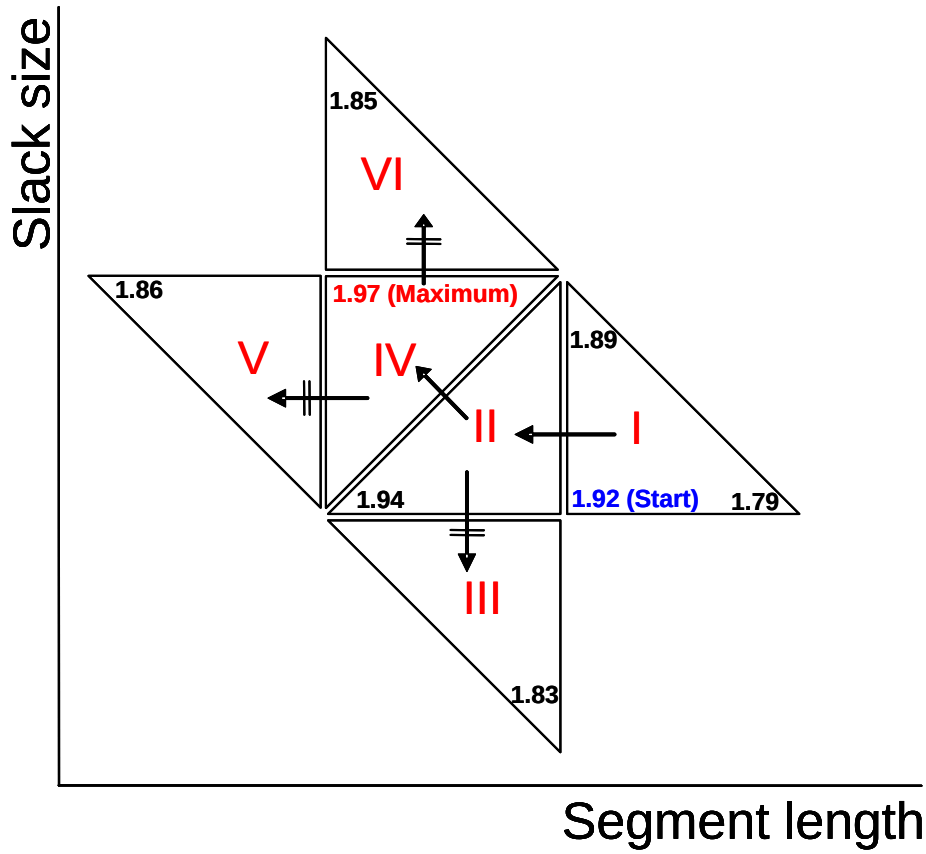

Figure 4. Illustration of the discrete-coordinates simplex-style optimization routine using warping effect values. See text for explanation of optimization steps. The roman letters indicate the consecutive steps of the optimization routine.

The main characteristics of the optimization routine for the simulated data are highlighted in Table 1. The end point in the path of each of the six optimization steps can be seen in Figure 5.

Table 1. Optimization of warping effect values: six starting points for the discrete-coordinates simplex optimization routine to find the optimal alignment parameters for the simulated chromatograms.

Optimization	Starting point (segment/slack)	Warping effect	Optimized to (segment/slack)	Steps	Warping effect	Simplicity Value	Peak factor
1	25/8	1.9264	25/8	4	1.9264	0.9663	0.9601
2	25/15	1.9218	24/16	7	1.9432	0.9786	0.9646
3	40/15	1.9075	40/16	3	1.9226	0.9586	0.9640
4	55/15	1.8949	55/16	3	1.9001	0.9569	0.9432
5	40/12	1.8841	41/12	3	1.9022	0.9631	0.9390
6	25/12	1.8817	**23/13**	6	**1.9537**	**0.9853**	**0.9684**

As seen in Table 1 the best combination of segment length and slack size (23/13) holds the largest simplicity and peak factor value of the six end points. However, for real data the situation is often more complex as the warping effect value is a compromise between alignment and peak area preservation. For some experiments the alignment might be most important and in others the preservation of the peak area more essential. By default the two quantitative measures are weighted equal in the warping effect value (a simple addition as shown in Equation (8)), but this can be changed dependent on the purpose of an experiment.

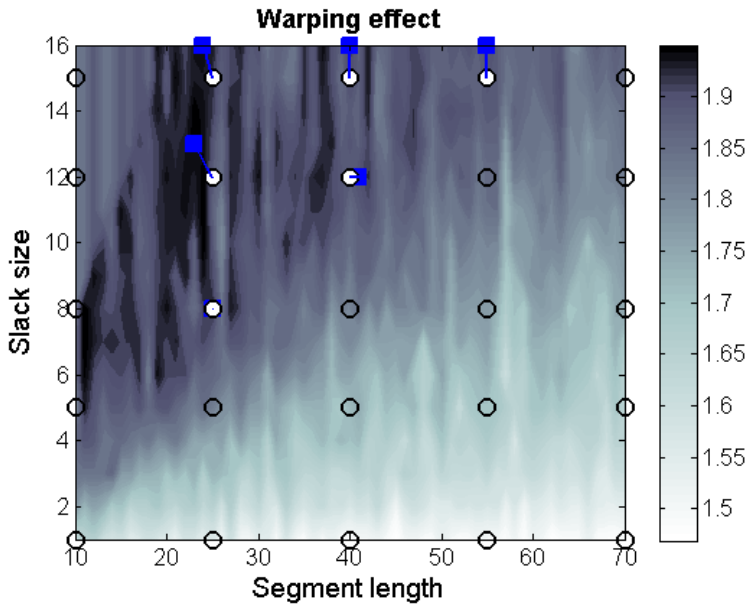

Figure 5. Warping effect: start points (filled circles) and end points (filled squares) in the optimization path for each of the six starting values. Notice that in contrast to Figure 2 and 3 the slack dimension had been extended to include size sixteen as three of the optimization end points include this slack size.

The optimization space (and sparse search grid) in Figure 5 includes combinations of segment length and slack size close to the dark area located in the upper left corner. This means that only a few steps are required for the simplex-style optimization routine to find these combinations assuming that no local maxima are found on the path. The distance in optimization steps to (near) optimal solutions and the smoothness (less high values – i.e. less local maxima) of the global search space are the two main issues that affect the optimization routine. The first issue is taken care of using a predefined sparse search grid as shown in Figure 3, whereas the other part is accomplished by using the warping effect instead of simplicity values in the optimization routine. The latter can be seen by comparing the plots in Figure 2A and 2C, where less (near) optimal solutions are found for the warping effect values.

The contour plot of the warping effect in Figures 2, 3 and 5 might give the impression that optimal solutions must be found in the regions that appear darkest visually. However, very local combinations can give warping effect values just as high, but they can be difficult to see in this kind of graphical presentation. Thus, the use of multiple starting points on the sparse search grid can lead the optimization path in other directions, as will be shown for the GC and HPLC data. The overall routine still ends up in points (combinations of segment length and slack size) with near optimal alignment characteristics.

To illustrate the effect of including the change in area (the peak factor) in the optimization routine, four examples of aligned peaks are shown in Figure 6. Notice that all peaks in the simulated data are Gaussian peaks and thus, any deviations from this are due to the alignment process.

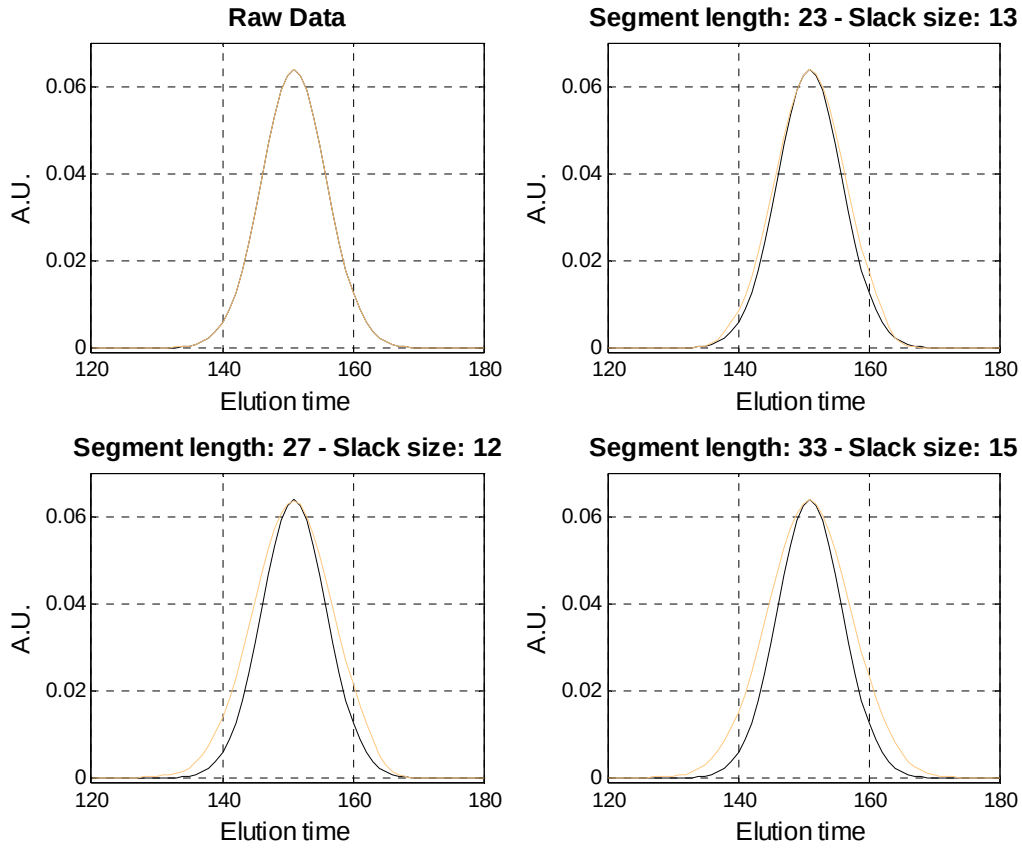

Figure 6. Illustrations of alignment and peak area characteristics for key combinations of segment length and slack size for two samples in the simulated chromatograms on single peak eluting between 120 to 180 data points.

Figure 6 shows that even though the samples in the first peak region (from 120 to 180 data points – see Figure 1) are perfectly aligned in the raw data, significant changes can be introduced as a result of the alignment process because it works on the total chromatogram. All three combinations give simplicity values that would indicate an efficient alignment, but the change in area results in the combination of segment length 23 and slack size 13 as the outcome of the optimization routine. Another measure to use when evaluating these combinations of segment and slack is the area change under the chromatograms in relation to the area of the chromatograms before alignment. For the combinations shown in Figure 6 an average change in area (calculated as (1-peak factor) × 100) of 3.5%, 6.4% and 14.7% are

found for the combination 23/13, 27/12 and 33/15, respectively. This measure is based on the absolute change in area (se Equation (6) and (7)) and thus describes the fact that some peaks might get broader and some more narrow.

2.8. Defining the optimization space

As shown in Figure 2 and 3, the search space for the segment lengths includes several feasible choices as long as the flexibility (slack size) is large enough. In general, longer segment lengths require more flexibility to give good alignments. However, this depends significantly on the chromatographic data at hand and the following guidelines should be used with some care.

Segment length

A rule of thumb for selecting the segment length optimization space is:

$$PW_A \pm \frac{PW_A}{2}$$

where PW_A is the approximate peak width average at the base over all peaks in the reference chromatogram. By this rule, the segment lengths will contain both peak fragments and entire peaks. In the result section this rule of thumb will be used to set the upper limit for the segment length, whereas the lower limit is set to ten points for all calculations. This is done here for illustrative purposes, to show the effect of combining a low segment length and a high flexibility (large slack size) on the preservation of the peak areas.

Slack size

The right slack size search space is more difficult to define as features such as different local peak shifts, data points before and after the first and last peak, and increased flexibility of the COW algorithm in the middle of the chromatogram will have an effect on the outcome of the alignment procedure. However, a rule of thumb is that if the number of data points before and after the first and last peak, respectively, is approximately the same as the peak widths (ensuring enough flexibility), then a slack size search space ranging from 1 to 15 should do the job, also because higher values will increase the computation cost considerably. In the

following the lower limit will always be set to one, whereas the upper limit is set higher than the suggested if peaks are severely shifted (e.g. HPLC data) and lower when only small shifts (e.g. GC data) are observed.

REFERENCES

[1] Nielsen NPV, Carstensen JM, Smedsgaard J. Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *J. Chromatogr. A* 1998; 805: 17–35.

[2] Bylund D, Danielsson R, Malmquist G, Markides KE. Chromatographic alignment by warping and dynamic programming as a pre-processing tool for PARAFAC modelling of liquid chromatography-mass spectrometry data. *J. Chromatogr. A* 2002; 961: 237–244.

[3] Tomasi G, van den Berg F, Andersson C. Correlation Optimized Warping and Dynamic Time Warping as preprocessing methods for chromatographic data. *J. Chemometrics*. 2004; 18: 231-241.

[4] Pravdova V, Walczak B, Massart DL. A comparison of two algorithms for warping of analytical signals. *Anal. Chim. Acta* 2002; 456: 77–92.

[5] Wang CP, Isenhour TL. Time-warping algorithm applied to chromatographic peak matching gas-chromatography Fourier-transform infrared mass-spectrometry. *Anal. Chem*. 1987; 59: 649–654.

[6] Grung B, Kvalheim OM. Retention time shift adjustments of two-way chromatograms using Bessel's inequality. *Anal. Chim. Acta* 1995; 304: 57-66.

[7] Wong JWH, Durante C, Cartwright HM. Application of fast Fourier transform cross-correlation for the alignment of large chromatographic and spectral datasets, *Anal. Chem*. 2005; 17: 5655-5661.

[8] Henrion R, Andersson CA. A new criterion for simple-structure core transformations in N-way principal components analysis, *Chemom Intell Lab Syst*. 1999; 47: 189-204.

[9] Christensen J, Tomasi G, Hansen AB. Chemical Fingerprinting of Petroleum Biomarkers Using Time Warping and PCA, *Environ. Sci. Technol*. 2005; 39: 255-260.

[10] Spendley W, Hext GR, Himsworth FR. Sequential Application of Simplex Designs in Optimisation and Evolutionary Operation, *Technometrics* 1962; 4: 441-461.

[11] The Quality & Technology Website: http://www.models.kvl.dk [August 2006]