ELSEVIER

Journal of Chromatography A, 805 (1998) 17–35

JOURNAL OF CHROMATOGRAPHY A

Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping

Niels-Peter Vest Nielsen[a], Jens Michael Carstensen[b], Jørn Smedsgaard[a,*]

[a]*Department of Biotechnology, Building 221, Technical University of Denmark, DK-2800 Lyngby, Denmark*
[b]*Department of Mathematical Modelling, Building 321, Technical University of Denmark, DK-2800 Lyngby, Denmark*

Received 4 December 1997; accepted 9 January 1998

Abstract

The use of chemometric data processing is becoming an important part of modern chromatography. Most chemometric analyses are performed on reduced data sets using areas of selected peaks detected in the chromatograms, which means a loss of data and introduces the problem of extracting peak data from the chromatographic profiles. These disadvantages can be overcome by using the entire chromatographic data matrix in chemometric analyses, but it is necessary to align the chromatograms, as small unavoidable differences in experimental conditions causes minor changes and drift. Previous aligning methods either fail to utilise the entire data matrix or rely on peak detection, thus having the same limitations as the commonly used chemometric procedures. The method presented uses the entire chromatographic data matrices and does not require any preprocessing e.g., peak detection. It relies on piecewise linear correlation optimised warping (COW) using two input parameters which can be estimated from the observed peak width. COW is demonstrated on constructed single trace chromatograms and on single and multiple wavelength chromatograms obtained from HPLC diode detection analyses of fungal extracts[1]. © 1998 Elsevier Science B.V.

Keywords: Chemometrics; Correlation optimised warping

1. Introduction

Chromatography (high-performance liquid chromatography, HPLC and gas chromatography, GC) has been established as one of the most important analytical methods in the modern laboratory. Developments in instrumentation have greatly increased the capability of the instruments, through-put of samples and the amount of data that can be collected. Chemometrics is therefore often used for processing the large amount of data collected.

An important part of chromatography involves comparing of chromatographic profiles using some sort of pattern recognition routines, for example fingerprinting of flavour components in coffee [1], of oil components in forensic investigations [2], or taxonomy of microorganisms [3–5]. Several chemometric methods are used for comparison, calculation of correlation and ranking. These are, among others,

*Corresponding author.

[1]A copy of the C program containing the COW implementation used in this work may be obtained at http://www.imm.dtu.dk/~jmc/papers/cow/cow.html

0021-9673/98/$19.00
PII S0021-9673(98)00021-1

factor analysis (FA), principal component analysis (PCA) and cluster analysis (CA) [1–4,6,7].

Prior to a chemometric analysis, chromatographic data are in most studies transformed to retention time-peak area data matrices including only selected peaks, whether the identity of the peaks are known or not. These data matrices are usually constructed from one chromatographic trace only. The quality of the data rely on peak detection (integration) and on how the peaks are selected for the data analysis. It can be very difficult to select an optimal set of integration parameters for chromatograms obtained from analysis of complex samples which easily can contain more than 100 peaks. Furthermore, the selection and extraction of peaks to include in the data analysis is difficult, partly subjective and a large amount of the data in the chromatograms are discarded.

The disadvantages of peak detection and integration, and of the introduction of a subjective peak selection can be avoided by using all collected data points in the chemometric analysis. A further advantage is that the peak shape can be included in data analysis. In order to perform direct chemometric analysis of entire chromatographic data matrices, thus utilising all collected data, it is however a prerequisite that the chromatographic profiles are properly aligned to compensate for minor drifts in retention times, either global or in small sections of the chromatogram. These small retention time shifts are known to all chromatographers and is due to changes in the columns during use, minor changes in mobile phase composition, drift in the instrument, interaction between analytes etc.

Previous work in this field has included peak tracking methods, in which the UV spectra of all peaks are determined, and the corrected retention time for the individual peak is then determined by the best fit of its spectrum to a standard profile [8]. Another approach was used by Grung and Kvalheim [9] who used Bessel's inequality to determine the optimal retention time adjustment within a given tolerance for individual peaks, by comparing with pure spectra. Both methods require the detection of the significant peaks in the chromatographic profile, and Grung and Kvalheim's method also requires knowledge of the components in the sample. In complex mixtures with large variations in concentrations, these requirements are rarely met. Malmquist and Danielsson [10] discuss some of the problems in detail, and propose a method, that does not have these requirements. But their method only incorporates information from a single wavelength, and thus does not utilise the spectral information. The method also involves selection of a considerable number of parameters. Finally, Wang and Isenhour [11] use dynamic programming to align two profiles, using a distance measure to optimise the aligning. Their method is, however, limited to deleting or duplicating points, thus resulting in rather coarse aligned profiles. The distance measure used relies on similar peak heights in the two chromatograms in order to give a reliable alignment, thus requiring a number of pretreatment steps in order to make the method effective.

The method described in this paper aims to align two chromatographic profiles by piecewise linear stretching and compression, also known as warping, of the time axis of one of the profiles. The optimal alignment will be determined by calculations of correlation, and so no knowledge of the compounds present is required, and individual peaks need not be detected and integrated. The method will be referred to as correlation optimised warping (COW) and when applying the method only two parameters must be set. Aligning by warping can be used on single chromatographic profiles [two-dimensional (2-D) data from e.g., standard GC–flame ionisation detection (FID) analysis or single wavelength UV detection in HPLC] or multi-trace chromatographic matrices [3-D e.g., from a HPLC with diode array detection (DAD) or GC–Fourier transform infra-red (FTIR) spectroscopy].

COW is demonstrated on HPLC–DAD data collected from analyses of fungal culture extracts using both the multiple wavelength data matrix and single wavelength data extracted from it. The output data matrix from the COW can be used directly in chemometric analysis (e.g., PCA and CA), for pattern recognition and image analysis (to be published).

The method developed is generalised, and can easily be customised to incorporate a priori knowledge of the problem class at hand and take advantage

of special features in the profiles. In the general form, only two parameters must be set in order to use COW.

2. Theory and implementation

Aligning by piecewise linear warping involves dividing chromatographic profiles into a number of sections that are each warped linearly. If the number of sections, and the number of ways each section may be warped, is finite then the number of possible solutions is also finite. The problem of finding the optimal solution among a finite number of possible solutions is called a combinatorial optimisation problem. In this case the feasible domain is constrained by ordering of the variables on a time axis i.e., the sections should remain in the same order and not overlap. Combinatorial optimisation problems subject to this type of constraints may be solved by a technique known as dynamic programming [12].

As warping is only applied in the time direction, all lengths and positions in the chromatograms refer to the time axis.

2.1. Problem formulation

Consider two chromatographic profiles to be aligned. One of the profiles is chosen as the target (T), and the other profile (P) is then aligned with it, forming the aligned profile (P′). The chromatographic profile P, having L_P+1 data points on the time axis and thus a length of L_P sample intervals, is divided into sections of length m, as shown in Fig. 1. It is possible to formulate the problem using varying section lengths, e.g., so that the sections correspond to features in the chromatographic profiles. This might yield better results, but would involve some sort of feature detection. It was therefore decided to use the more general, and thus more easily automated, uniform-length segmentation in this work. The number of sections N is given by

$$N = \frac{L_P}{m} \tag{1}$$

Each section may be warped to a smaller or greater length. In this work, linear interpolation is used to perform the warping. A section having starting point at position x_s and end point at position x_e is warped to starting position x'_s and end position x'_e by calculating

$$p_j = \frac{j}{x'_e - x'_s}(x_e - x_s) + x_s; \quad j = 0, ..., x'_e - x'_s \tag{2}$$

and then calculating the value of $P'(x'_s + j)$ by linear interpolation between the points in P adjacent to p_j.

The end points of the sections are referred to as nodes, and the position of the starting point of section i after warping is denoted x_i. Node 0 is the starting point of the first section ($x_0 = 0$) and node N is the end point of the entire profile after warping. This position corresponds to the length L_T of the target chromatographic profile T, thus $x_N = L_T$. Using this definition, differences in chromatographic profile length are also corrected by the warpings.

For each section only a finite number of possible warping magnitudes can be examined. In order to keep the problem formulation general the same magnitudes are used for all segments. In this work, the warpings examined consist of the integer values from 0 to t, where t will be referred to as ''the slack''. Non-integer and non-uniformly spaced values may also be used, but uniform spacing is more generally suitable, and the time resolution of modern

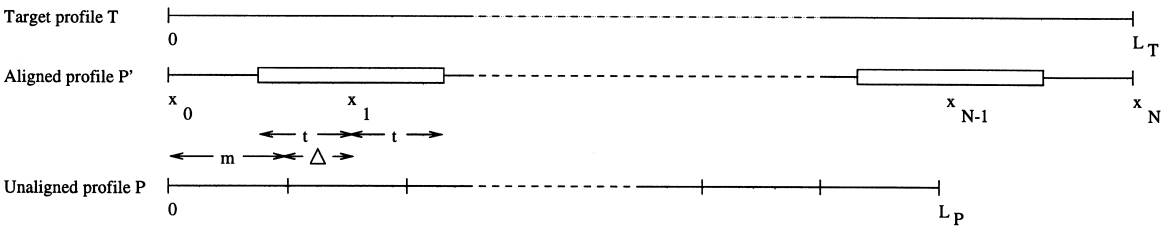

Fig. 1. Schematic presentation of the structure of the optimal warping problem.

laboratory equipment means that warpings smaller than the sample interval are rarely necessary.

Determining the optimal alignment is now a question of finding the optimal combination of warpings of the N sections, where no section may be warped by more than t sample intervals. The warping of section i is called u_i.

If there is a large difference between L_P and L_T a given value of t would not allow an equal degree of stretching and compression. As an example, consider a chromatographic matrix with a length of 10 sample intervals ($L_P = 10$) that is to be aligned with a matrix of length 16 ($L_T = 16$), using a section length of five sample intervals ($m = 5$) and a slack of 4 ($t = 4$). N would then equal 2 and the mean section length after aligning would be 8. With the given values for m and t, this would allow each section to be stretched 1 position or be compressed up to 7 positions relative to the mean length, obviously not an equal amount of slack. Therefore, warpings are allowed to fall in the interval $(\Delta - t;\ \Delta + t)$ where Δ is the difference in section length in P and T, calculated as

$$\Delta = \frac{L_T}{N} - m \tag{3}$$

The quality of the alignment is determined separately for each section i, by calculating the correlation coefficient ρ between section i after warping and the corresponding segment of the target profile. For generality, it was decided that the method should align peaks by matching shapes rather than height/areas. The correlation coefficient gives a good measurement of covariations in data sets, and is thus the natural choice for calculations of similarity. The alignment quality function (ρ in this work) is termed the benefit function f, using the short notation $f(I) = \rho(I_{P'}, I_T)$, where I denotes an interval between two node positions. By defining the optimal combination of warpings as the one that gives the largest value of the summed correlation coefficients the problem is now directly solvable by dynamic programming.

The optimal combination of warpings defines the optimal set of node positions after warping $\boldsymbol{x}^*$. The problem is then described by

$$x_0 = 0 < x_1 < \cdots < x_{N-1} < x_N = L_T \tag{4}$$

$$u_i \in [\Delta - t;\ \Delta + t]; \quad i = 0, \ldots, N-1 \tag{5}$$

$$x_{i+1} = x_i + m + u_i; \quad i = 0, \ldots, N-1 \tag{6}$$

$$\begin{aligned} \boldsymbol{x}^* &= \arg\max_{\boldsymbol{x}} \left(\sum_{i=0}^{N-1} f([x_i; x_{i+1}]) \right) \\ &= \arg\max_{\boldsymbol{x}} \left(\sum_{i=0}^{N-1} \rho(P'[x_i; x_{i+1}], T[x_i; x_{i+1}]) \right) \\ &= \arg\max_{x} \left(\sum_{i=0}^{N-1} \frac{\mathrm{Cov}(P'[x_i; x_{i+1}], T[x_i; x_{i+1}])}{\sqrt{V(P'[x_i; x_{i+1}]) \cdot V(T[x_i; x_{i+1}])}} \right) \end{aligned} \tag{7}$$

where x_i and u_i are limited to the integer numbers in the constraint interval (Eqs. (4) and (5)). For multiple wavelength chromatographic profiles, ρ is calculated by treating data on different wavelengths as different observations of the same variable, and not as different variables in the same observation. This univariate way of calculating ρ can be illustrated by the procedure used for calculating the variance V:

$$V(P'[x_i; x_{i+1}]) = \frac{\sum_{j=x_i}^{x_{i+1}} \sum_{k=1}^{N_\lambda} \left(P'(j,k) - \overline{P'[x_i; x_{i+1}]} \right)}{(x_{i+1} - x_i) \cdot N_\lambda - 1} \tag{8}$$

where N_λ is the number of wavelengths and $\overline{P'[x_i; x_{i+1}]}$ is the mean value calculated over all wavelengths for all points in the time interval $[x_i; x_{i+1}]$.

2.2. Dynamic programming

Dynamic programming solves combinatorial optimisation problems by examining all possible combinations of the variables in a rational fashion: For each section i the optimal warping u_i is calculated for each possible position of the node x_i. The calculations are based on the previous section, which in the implementation used in this work is $i+1$. This variant of the method is called backward dynamic programming. Section by section, suboptimal combinations of warpings are discarded, and when all sections have been treated only the optimal combination remains. As all possible combinations are treated, dynamic programming will always yield the global optimum.

The algorithm is based on the use of two matrices, $\boldsymbol{F}$ and $\boldsymbol{U}$, that contain the benefit function values and

the corresponding control input, respectively. In this case the size of the matrices is $(N+1)\times(L_T+1)$.

For each node i there are only a certain number of positions, that can be reached after i sections with the given values of m and t. On row i of the $\boldsymbol{F}$ and $\boldsymbol{U}$ matrices, these positions are given the value of the cumulated benefit function (in matrix $\boldsymbol{F}$) or the warping of section i (matrix $\boldsymbol{U}$) that brought the node to that position. The cumulated value is calculated from the results from the previous section, here section $i+1$.

As an example, consider $L_T=40$, $m=10$, $\Delta=0$ and $t=5$. The structure of the matrices $\boldsymbol{F}$ and $\boldsymbol{U}$ will be as shown in Fig. 2.

The matrices have a row for each node, and a column for each possible node position. The first and the final nodes must be placed at the ends of the chromatographic matrix (Eq. (4)), but nodes 1 to 3 have intervals in which they may be placed. The interval is largest for node 2 (x_2), because it is the farthest from the ends of the profile, and therefore is least affected by the constraints. The interval I_i for node i is determined from Eqs. (4)–(6): The lower bound in Eq. (4) ($x_0=0$) yields

$$x_i \in I_1 = [i^*(m+\Delta-t);\ i^*(m+\Delta+t)];\quad i=0,\ \ldots,N \tag{9}$$

while the upper bound ($x_N=L_T$) yields

$$x_i \in I_2 = [L_T-(N-i)^*(m+\Delta+t);\ L_T-(N-i)^*(m+\Delta-t)];\quad i=0,\ \ldots,N \tag{10}$$

Thus

$$x_i \in I_i = I_1 \cap I_2;\quad i=0,\ \ldots,N \tag{11}$$

The value placed at $\boldsymbol{F}_{i,x}$ is the maximum possible cumulated value of the benefit function, if node i is to be placed at position x. In the example above, node 2 may be placed at position 29 in two ways: if node 3 was placed at position 35 ($u_3=-5$) then u_2 must be -4, but if node 3 was placed at position 34 ($u_3=-4$) then u_2 must be -5. The combination that gives the highest value of $f([x_3;\ x_4])+f([x_2;\ x_3])$ is the optimal combination for nodes 4 through 2 if node 2 is to be placed at position 29. The cumulated benefit function values are stored in $\boldsymbol{F}_{2,29}$, and the optimal u_2 is stored in $\boldsymbol{U}_{2,29}$.

Examining all possible positions for node 1, all legal values of u_1 will lead to a position that was examined for node 2. This means that the optimal combination of node placings for the rest of the nodes has already been determined, and the optimal cumulated value has been placed in row 2 of $\boldsymbol{F}$. In this way,

$$\boldsymbol{F}_{i,x} = \max\left(\boldsymbol{F}_{i+1,x+m+u_i} + \mathrm{f}([x;\ x+m+u_i])\right);\quad i=0,\ \ldots,N-1 \tag{12}$$

From this recursive definition of $\boldsymbol{F}_{i,x}$ it is seen, that

$$\begin{aligned}\boldsymbol{F}_{0,0} &= \max\left(\boldsymbol{F}_{1,m+u_0} + \mathrm{f}([0;\ m+u_0])\right)\\ &= \max\left(\boldsymbol{F}_{2,m+u_0+m+u_1} + \mathrm{f}([0;\ m+u_0]) + \mathrm{f}([m+u_0;\ m+u_0+m+u_1])\right)\\ &= \max\left(\mathrm{f}([0;\ m+u_0]) + \sum_{i=0}^{N-2}\mathrm{f}\left(\left[\sum_{j=0}^{i}(m+u_j);\ \sum_{j=0}^{i+1}(m+u_j)\right]\right)\right)\\ &= \max\left(\sum_{i=0}^{N-1}\mathrm{f}([x_i;\ x_{i+1}])\right)\end{aligned} \tag{13}$$

and comparing with Eq. (7) it is seen that $\boldsymbol{x}=\boldsymbol{x}^*$.

The optimal sequence of warpings may then be reconstructed by backtracking through the matrix $\boldsymbol{U}$, as $\boldsymbol{U}_{i,x}$ contains the optimal warping of section i. The algorithm is illustrated in Pascal-like pseudo code in Appendix A.

As it is seldom possible to divide chromatograms neatly into segments of the desired length, a special case must be made of the remainders $r_1=L_P-m^*N$ and $r_2=L_T-(m+\Delta)^*N$. In this implementation the remainders were taken as the final part of the chromatograms, and r_1 was simply warped to the length of r_2.

With the described formulation and solution pro-

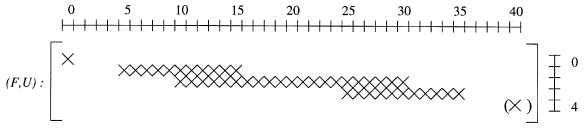

Fig. 2. Structure of the matrices $\boldsymbol{F}$ and $\boldsymbol{U}$ for $L_T=40$, $m=10$, $\Delta=0$ and $t=5$. The element (×) only appears in $\boldsymbol{F}$.

cedure, it is only necessary to give two parameters when using the method, as all other values are calculated from these. The parameters are the segment length m and the slack t. Whenever values of m and t are chosen, it should be borne in mind that it is the flexibility t/m that is important, and not the numeric value of t.

2.3. Alignment quality evaluation

It is a relatively simple task to evaluate the quality of alignment for single wavelength data by visual comparison. For multiple wavelength data, however, one or more traces must be chosen for visual comparison, thus making the evaluation more difficult and less reliable for complex chromatograms. An objective measure of alignment quality is therefore developed.

The correlation coefficient ρ is used as basis for the alignment quality measure, because of the complex nature of the chromatograms, and because the degree of covariation in the aligned chromatograms indicate the quality of peak matching, i.e., aligning. Using ρ, however, means that some considerations must be made.

Apart from the traditional definition of the correlation coefficient given in Eq. (7), ρ can also be expressed as an element-by-element multiplication of two datasets, after each dataset has been scaled to mean zero and variance one. For example, for two arbitrary data sets $\boldsymbol{a}$ and $\boldsymbol{b}$, ρ can be expressed as:

$$\rho(\boldsymbol{a}, \boldsymbol{b}) = \frac{\mathrm{Cov}(\boldsymbol{a}, \boldsymbol{b})}{\sqrt{\mathrm{V}(\boldsymbol{a})V(\boldsymbol{b})}} = \frac{1}{n}\sum_{i=0}^{n}\frac{a_i - \bar{a}}{s_a}\cdot\frac{b_i - \bar{b}}{s_b} \qquad (14)$$

where n is the number of observations in each data set, $\bar{a}$ and $\bar{b}$ are the means of the respective data sets, and s_a and s_b are the estimated standard deviations.

This means that when ρ is calculated over an area smaller than the width of a peak, the height of the peak has only little influence on ρ. When the area is so large, that both high and low peaks are present in the same interval, the lower peaks will be scaled to very small values because the higher peaks will have the largest influence on the variance. In this case, it will then be the alignment of the highest peaks that determine the value of the benefit function. The effect is, that for large areas ρ gives the alignment quality for the highest peaks. For small areas the general quality for both high and low peaks is given.

From these considerations it follows that calculating ρ for the aligned chromatographic profiles would only give a poor indication of the actual quality of the alignment. In order to obtain a good estimate of the alignment quality, the local variations must be given equal weight, and large-scale variations eliminated. This is done by applying a Wallis filter [14] (also known as statistical differencing) to the profiles prior to calculation of ρ. The Wallis filter generates a profile with specified local mean and variance using the formula

$$\mathrm{P}^*(t, \lambda) = \frac{A\sigma_d}{A\hat{\sigma} + \sigma_d}(\mathrm{P}(t, \lambda) - \hat{\mu}) + \alpha\mu_d + (1 - \alpha)\hat{\mu} \qquad (15)$$

where P is the profile being treated and P* is the resulting profile. $\hat{\sigma}$ is the local standard deviation, $\hat{\mu}$ is the local mean, σ_d and μ_d are the desired values, and A and α are weighting coefficients, indicating to which extent the local values should be transformed towards the desired values. A and α are chosen equal to 100 and 1, thus forcing the mean to μ_d and the standard deviation to nearly σ_d. The values of σ_d and μ_d are in this case of minor importance, as the correlation coefficient calculation will scale the data to mean 0 and variance 1. For consistency the values 0 and 1 were chosen for μ_d and σ_d, respectively.

Calculation of similarity is done by applying the Wallis filter to P′ and T, thus generating P′* and T*, and then calculating the correlation coefficient between P′* and T*. With the chosen values of A and α this is equivalent to calculating a local correlation coefficient for each data point and calculate the mean of these values. The combined operation using local area size l is termed the Wallis correlation coefficient ρ_l^{W}.

It follows from the discussion of locality that the size of the area in which $\hat{\sigma}$ and $\hat{\mu}$ are calculated determines the size of the features preserved by the filter. By varying the area size l the larger features or the local variations may be brought out by using a large or small value of l, respectively. This means that for large l, ρ_l^{W} indicates to what degree tall peaks have been aligned with tall peaks, while small values of l gives a measure of the similarity of local variations.

The discussion of locality is also relevant for choosing the COW section length *m*: For large values of *m*, the highest peaks are aligned, while the lower peaks may become poorly aligned. Small values of *m* will result in an equal quality of alignment for all peak heights.

In the following, the quality of correlation optimised warping will be examined, the influences of *m* and *t* will be evaluated, and guidelines for reasonable choices will be developed.

3. Experimental

Two kinds of experiments were conducted: firstly, two single wavelength chromatograms were constructed, in order to test the basic functioning of the algorithm. Secondly, chromatographic matrices from analyses of fungal cultures were used to test the algorithm on real world data, and to evaluate the influence of the values of *m* and *t*.

3.1. Construction of data sets

The two constructed data sets were created in Matlab™ version 5.0.0.4064 (The Mathworks, USA), and are shown in Fig. 3A. Sample intervals (i.e., data points or observations) is used as time axis unit, and mAU is used as absorbance axis units.

The chromatograms have different lengths, noise level and baseline drift. The number of peaks in the two data sets are different, as well as the position, width, height and shape of the individual peaks. The shorter profile will be referred to as P_1 and the longer as P_2.

All peaks are shaped like the normal distribution, except peak No. 1 in P_2 (at position 250), which has a triangular shape. The noise is normally distributed, with variance 0.2 in P_1 and with variance 1 in P_2. Both profiles have been added baseline drift with the shape of a sinus curve, the curve being shifted by 50 positions between P_1 and P_2.

COW was implemented in the C programming language on a HP 9000/755 workstation (Hewlett-Packard, CA, USA) and was applied with $(m, t) = (20, 3)$.

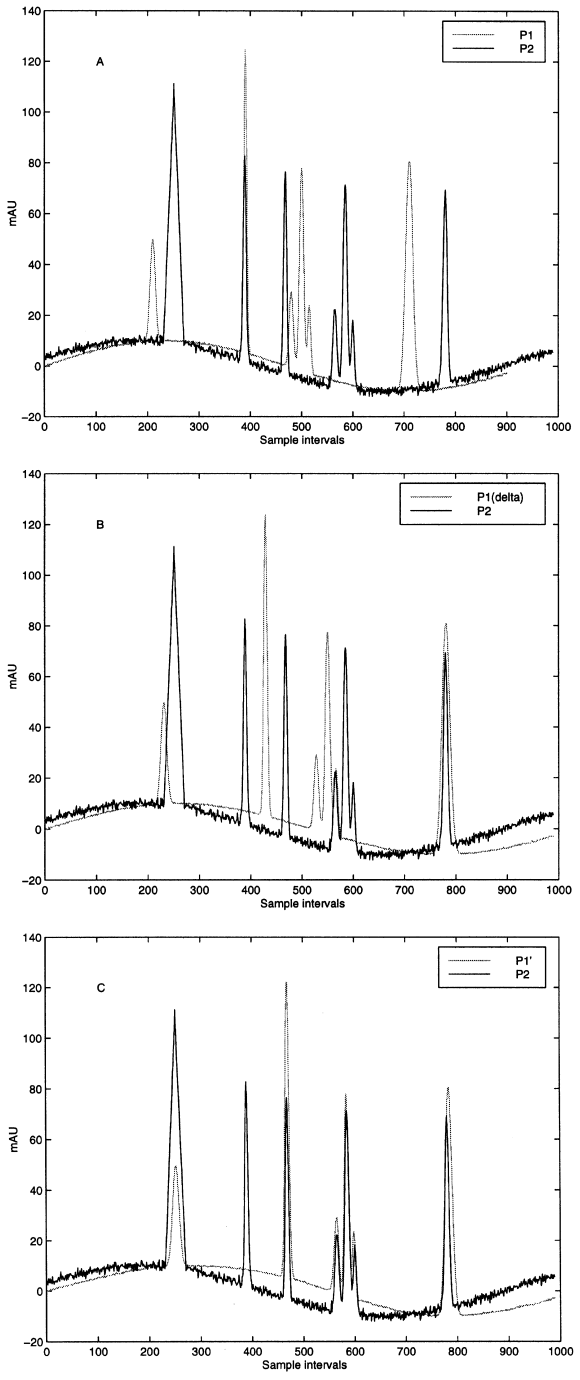

Fig. 3. Two constructed chromatographic profiles (P_1 and P_2), incorporating differences in peak height, width, retention time and shape, and differences in the number of peaks, the length of the profile and the baseline. (A) Prior to aligning, (B) after linear stretching of the entire profile, and (C) after aligning using COW.

3.2. Experimental data sets

Data from 16 of the HPLC–DAD analyses described in Ref. [13] were selected as test set for the COW algorithm.

The chromatograms represent analyses of the following isolates, all taken from the fungal collection at Department of Biotechnology, DTU, cultivated on Yeast Extract Sucrose (YES) agar: *P. cyclopium* (IBT 11415 and 15670), *P. aethiopicum* (IBT 5903 and 5753), *P. clavigerum* (IBT 5523 and 4899), *P. aurantiovirens* (IBT 12841 and 16769), *P. olsonii* (IBT 13065 and 14335), *P. vulpinum* (IBT 10606 and 11932), *P. sclerotigenum* (IBT 13826 and 13938), and *P. oxalicum* (IBT 10116 and 13309).

The data were collected at approx. 2 UV spectra/s from 200 nm to 600 nm with a bandwidth of 4 nm. This resulted in 100 data points in each UV spectrum and approximately 3300 in each chromatogram. The data were transferred from the HP Chemstation (version A4.02) software to a ASCII matrix containing approx. 3300 data points (time scale) in 100 columns (wavelength scale) by an in-laboratory written macro. The first column, representing the 202 ± 2 nm signal, was selected to examine the effect of aligning by COW.

Two series of COW runs were made: in the first series m was varied, while t was kept at a constant fraction of the value of m, while in the second series m was kept constant and t varied. The values used are shown in Table 1. This scheme was chosen in order to provide data for determining the optimal choice of m and t for the type of chromatographic data used in this study.

In a third series of runs COW was applied to the multiple wavelength data matrix, in order to examine the effect of including spectral information.

Table 1
The experiments conducted to examine the influence of m and t

First series		Second series	
m	t	m	t
5	1	20	1
20	4	20	2
100	20	20	4
500	100	20	6

The *P. cyclopium* isolate set was used for visual examination of the results from the different runs.

4. Results and discussion

4.1. Constructed data sets

In Fig. 3B it can be seen how the basic warping of each section from length $m = 20$ to length $m + \Delta = 22$ (the basic stretch) affects the peak positions. As described, the basic stretching is an integral part of COW, but it is shown here as a separate step in order to facilitate the interpretation of the result. Fig. 3C shows the result of the COW aligning of the two profiles. The peaks have been well aligned, and thus the method is seen not to be sensitive to the variations that are commonly found in chromatographic profiles. The method's basic ability to correct defects that are correctable by piecewise linear warping is hereby demonstrated.

Peak 2 in P_1 is aligned with peak 3 of P_2 and not peak 2, although it is positioned exactly in the middle between the two by the basic stretching (Fig. 3B). This is because peaks 3, 4, and 5 in P_1 can be matched very well with peaks 4, 5, and 6 in P_2, thus ''pulling'' peak 2 of P_1 towards peak 3 of P_2. In this way, characteristic patterns help give a better alignment of the parts of the profile, that are not as easily interpretable. This is very similar to the way humans would align chromatographic profiles: align sections that are obviously similar, and then use this information to judge the alignment of the rest of the profile.

The baseline variation does not influence the alignment, even though the variation patterns in the two profiles could be closely matched. Though the baseline variation pattern is very wide, spanning many sections and thus giving it a large weight in the benefit function, the slow variations mean that the difference in correlation between a good and a poor alignment is small, thus cancelling the increased weight.

4.2. Experimental data sets

In Fig. 4 the most ''peak rich'' section of the chromatographic trace at 202 nm is shown for the

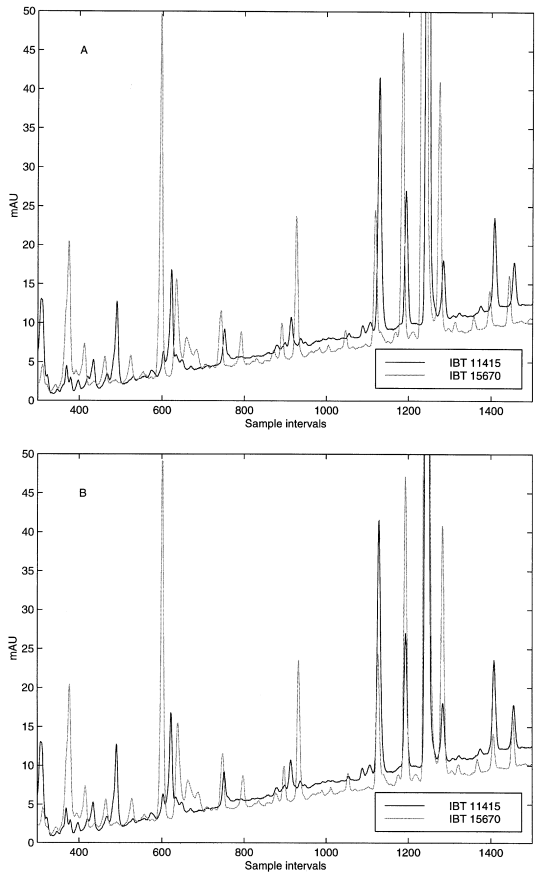

Fig. 4. Superimposed sections of the chromatographic trace at 202 nm from HPLC analysis of two isolates of *P. cyclopium*. (A) Prior to aligning and (B) after linear stretching of the entire profile.

two isolates in the *P. cyclopium* isolate set. It is seen, that the two traces are not well aligned originally, but after the basic stretching of the entire profile the alignment is much improved.

In the first series of COW runs, the effect of the choice of section length m is examined by varying m and keeping t at a constant fraction of m. All isolate sets were aligned, and the Wallis correlation coefficient calculated with the values of l indicated in Fig. 5. In this Figure the average over all isolate sets is shown for each value of m, for the unaligned profiles, and for the basic warping.

From the Figure it can be seen that all alignments are approximately equally good, when $l > m$. For values of l less than a given value of m the similarity score drops below the scores obtained for smaller values of m, i.e., local variations smaller than the given value of m are not matched as well as for smaller sections.

The conclusion must be that smaller values of m gives better alignment, since tall peaks are matched equally well by large and small m and small peaks are matched better using low values of m. This is in accordance with the theoretical discussion. It is also intuitively acceptable since smaller values of m makes the aligning procedure more flexible.

The data collection rate used in these analyses was approx. 2 scans/s resulting in 20–30 data point across the smallest peaks. Theory would then suggest that the optimal choice of m was 20 in order to align only the relevant features, because a smaller m would align features smaller than the smallest peaks, i.e., the noise. In Fig. 5 the gain in similarity from $m = 20$ to $m = 5$ is seen to be small, and if the smallest peaks are 20 sample intervals wide, the gain probably stems from overfitting, i.e., fitting not only to the data, but also to the noise.

The conclusion is that the optimal choice of m for the type of chromatographic profiles used in this work corresponds to the peak width in data points across the smallest peaks of interest.

In the second series of COW runs, m was kept constant while t was varied, in order to examine the effect of the choice of slack. The procedure was the same as in the first COW series, and the results are plotted in Fig. 6. The Figure shows that higher values of t gives better correlation, but also that the differences are not very large, especially when comparing to the gain in similarity score from unaligned profiles to basic stretch, and from basic stretch to $t = 1$. It appears that even a small amount of slack gives a large improvement, but raising the value of t only gives relatively small improvements.

These observations are in accordance with theory, which predicts that a larger slack gives larger flexibility, thus giving COW a larger degree of freedom to move and deform the features of the profiles to reach the closest match possible. Warping is a deformation of the profile, and using a large value of t will allow a strong deformation of the peaks, and thus make possible another form of overfitting: deformation of peaks to match features that they do not resemble originally.

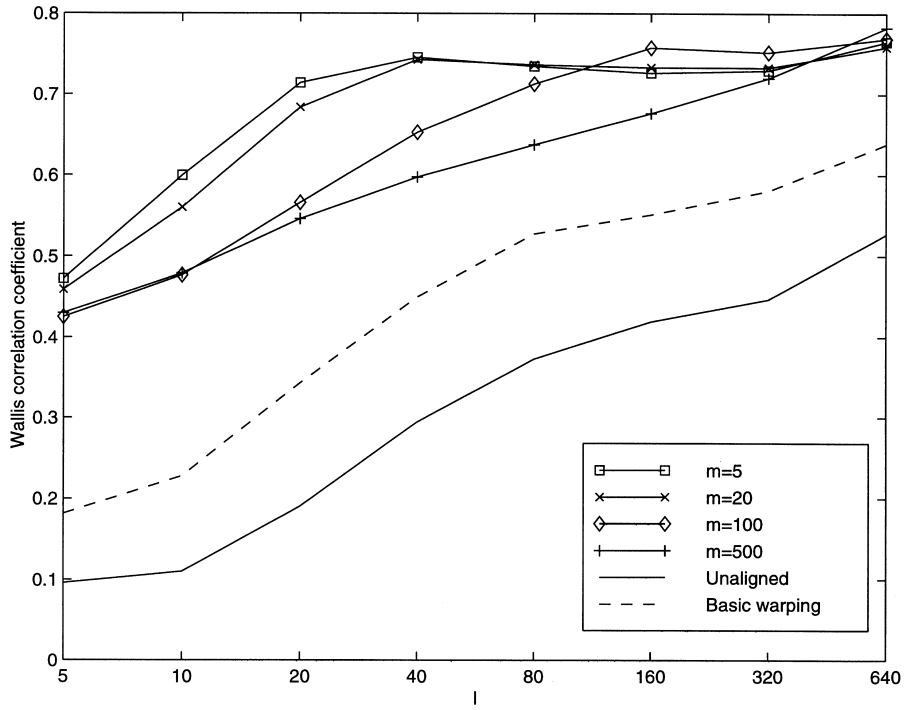

Fig. 5. Plot of the Wallis correlation coefficient ρ_l^{W} as a function of local area size l for the first series of experiments. The plotted values are the mean of the values obtained for the eight isolate sets by aligning using only the 202 nm trace. Left side of the figure indicates the quality of the fit in general, right hand side shows fit to largest peaks.

Fig. 7 shows aligned sections of the profiles from isolate set 1 for two of the runs: one with a low and one with a high value of t. It is seen that for $t=2$ the difference between the basic warping (Fig. 4B) and the aligned profile is small. Mostly, individual peaks have been moved a few sample intervals to produce a slightly better lineup of peaks that are already matched. For $t=6$ more dramatic changes are seen. The three overlapped peaks (1) are moved approx. thirty-five sample intervals and some peaks (2 and 3) have been visibly deformed to smaller (3) or larger (2) peak width.

$t=6$ is very likely a case of overfitting, but on the other hand the small changes observed for $t=2$ may be caused by the aligning being too ''stiff'' to produce a better alignment. The results and theory so far point to small values of t being optimal, but in order to draw definite conclusions spectral information must be included.

To examine the effect of including spectral information COW runs were made using the parameter combinations $(m, t)=(20, 2)$ and $(m, t)=(20, 6)$. The aligned 202 nm traces of the *P. cyclopium* isolate set are shown in Fig. 8. It is seen that the two alignments are very similar compared to the single wavelength case, the only noticeable difference in the peak matches being the ones at 850–950 sample intervals. Including spectral information is thus seen to stabilise the alignment.

A closer examination of the peak matches around 900 sample intervals is performed by plotting the normalised (mean zero and variance one) spectra at the top of the peaks at 899, 914 and 937 sample intervals in IBT 11415. These are the large peak and its two neighbours. Also plotted are the spectra at 890 and 926 sample intervals in IBT 15670 (the two large peaks). None of the eluting compounds show any absorption above 320 nm, and the plot is therefore limited to the interval 200–320 nm, better to show the relevant part of the spectra. Also, correction for baseline drift has been performed on each wavelength separately to bring out the spectra

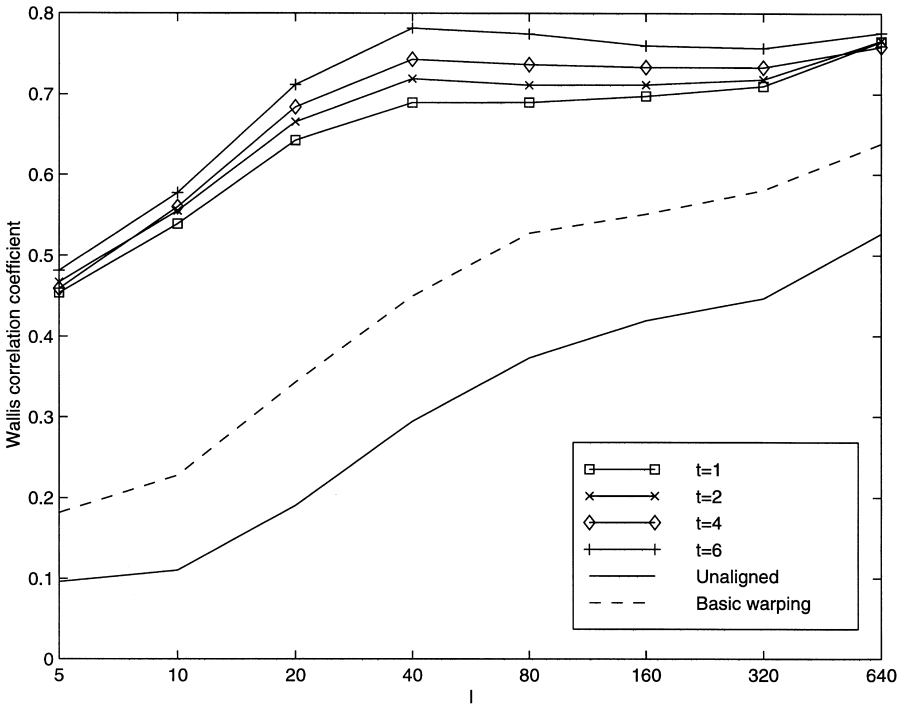

Fig. 6. Plot of ρ_l^W as a function of l for the second series of experiments. The plotted values are the mean of the values obtained for the eight isolate sets by aligning using only the 202 nm trace.

of the eluting substances (see below). The spectra are shown in Fig. 9 and it is obvious that neither spectrum of the large peaks in IBT 15670 match the spectrum of the large peak in IBT 11415, but that they match the small, neighbouring peaks' spectra very well. This is also reflected by a calculation of the correlation between the spectra, shown in Table 2: low scores are obtained for matching of the large peaks, but very high scores are obtained by matching the low peaks. It can thus be concluded that neither alignment in Fig. 8 is correct in the interval around data point 900.

In order to improve the alignment two preprocessing steps were introduced, and the benefit function modified slightly.

Firstly, the chromatographic matrices were trimmed to include only wavelengths 200–320 nm, thus extracting the relevant part of the spectra, and preventing the large amount of superfluous data from drowning out the relevant information.

Secondly, baseline correction was performed, to prevent baseline drift and differences in base signal level between the diodes from influencing the spectra. Baseline correction was performed on one wavelength at a time, by finding the minimum point in a 400 data point window for all possible window placements. Data points which were found as minimum more than twice were considered to be baseline points, and an estimated baseline for the current wavelength was created by linear interpolation between the detected baseline points. The resulting piecewise linear function was subtracted from the profile at the current wavelength, yielding a baseline corrected profile.

Finally, ρ^3 was used as the benefit function instead of ρ, thus favouring the highest correlation values relative to the lower ones. This makes COW prefer few, very good alignments to many, poorer alignments.

The results using the customised method are shown in Fig. 10, where the alignments are seen to be almost identical. It is noted that the peaks around 900 sample intervals (peaks 8–10 in Fig. 10B) are now correctly aligned, but also that the peaks around

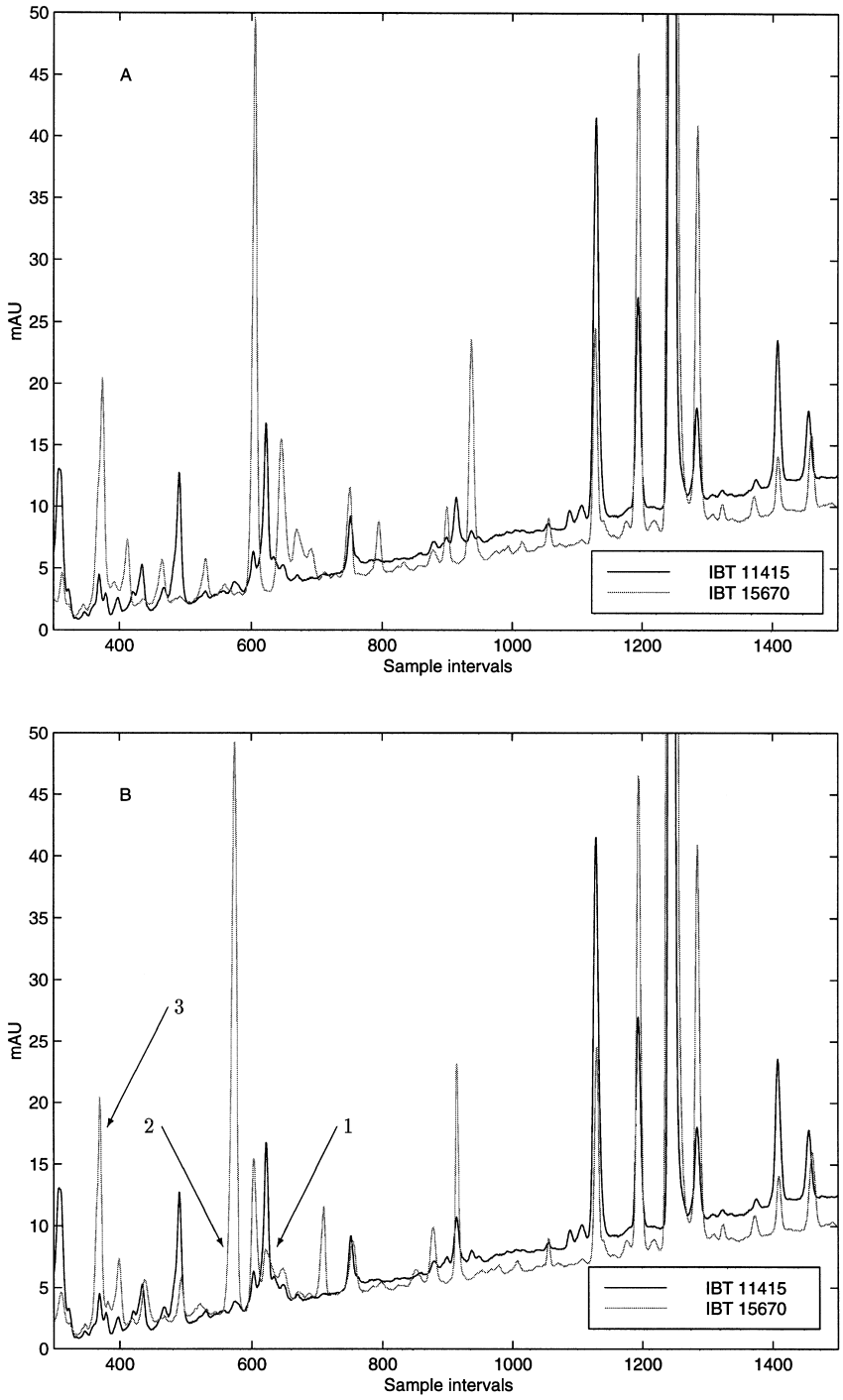

Fig. 7. Alignments for the *P. cyclopium* isolate set using only the 202 nm trace with A: $(m, t)=(20, 2)$ and B: $(m, t)=(20, 6)$.

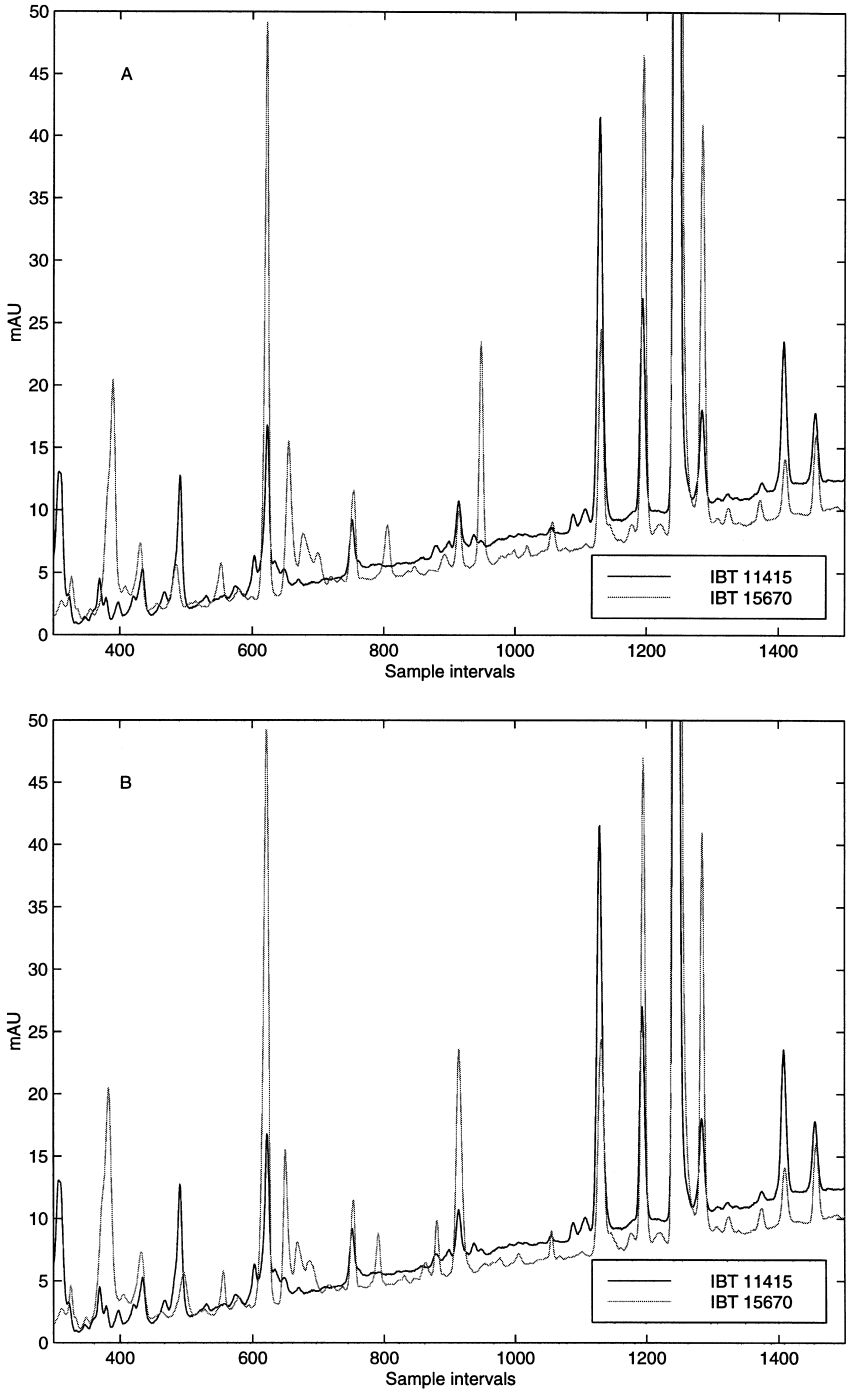

Fig. 8. 202 nm trace from the alignment of isolate set 1 using the entire chromatographic data matrix with A: (m, t)=(20, 2) B: (m, t)=(20, 6).

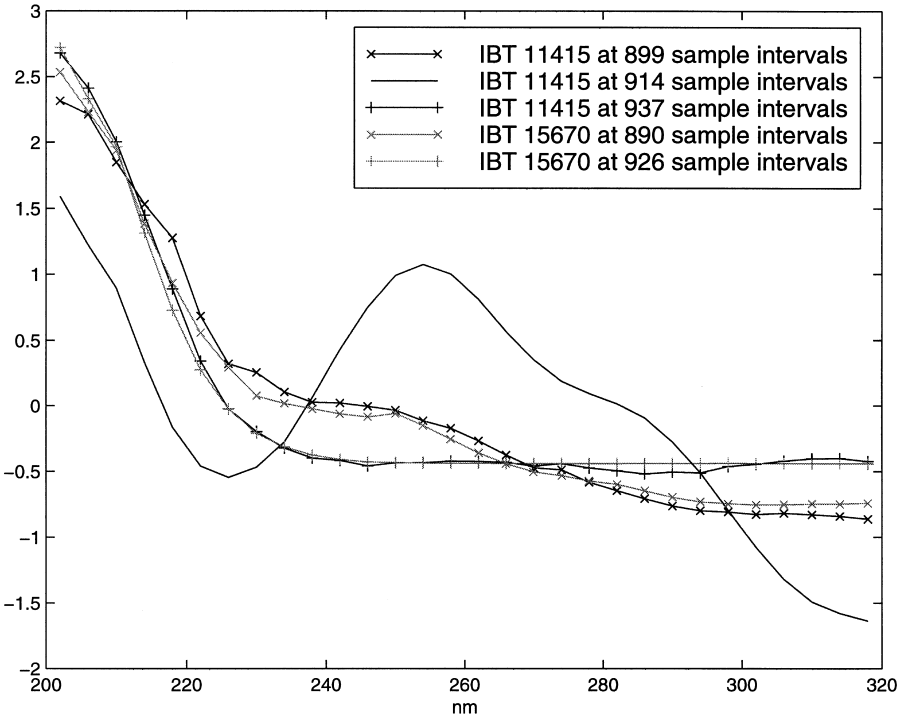

Fig. 9. Normalised peak top spectra from the chromatographic matrices in the *P. cyclopium* isolate set. The time axis positions given in the legend correspond to Fig. 4A.

600 sample intervals (peaks 1–6 in Fig. 10B) are aligned differently than with the basic COW algorithm. In order to compare the alignments with the ones obtained by using the basic COW algorithm, the spectra at the top of the peaks around 600 and 900 sample intervals in IBT 11415 (the target chromatogram) are compared with the spectra at the same time positions from the different alignments. The peaks are indicated and numbered in Fig. 10B, and the comparisons are done by calculating the correlation coefficient between the IBT 11415 spectrum and the spectra from the different alignments in the same way as described above for Table 2.

Table 2
Correlation coefficients for different combinations of the spectra in Fig. 9

	IBT 15670	
IBT 11415	Peak at 890	Peak at 926
Peak at 899	0.9933	–
Peak at 914	0.5741	0.4616
Peak at 937	–	0.9986

The correlations are listed in Table 3, along with the mean of the values and the mean of the cubed values. It is seen that the largest mean correlation is obtained with the customised COW, using $t=2$, while the customised COW using $t=6$ scores significantly lower. This is mainly due to the very low scores for peaks 4 and 9. However, correlations below ca. 0.8 usually means that the spectra are dissimilar, and once dissimilarity has been established, it is of lesser interest to know exactly how dissimilar they are. By cubing all correlations the lower values are transformed towards a lower, more uniform value, while the high values are transformed to relatively higher values, thus giving a better presentation of the relative similarities. The mean of the cubed values show the customised COW to yield significantly higher correlations than the basic COW, thus indicating that the alignments shown in Fig. 10 are the most correct.

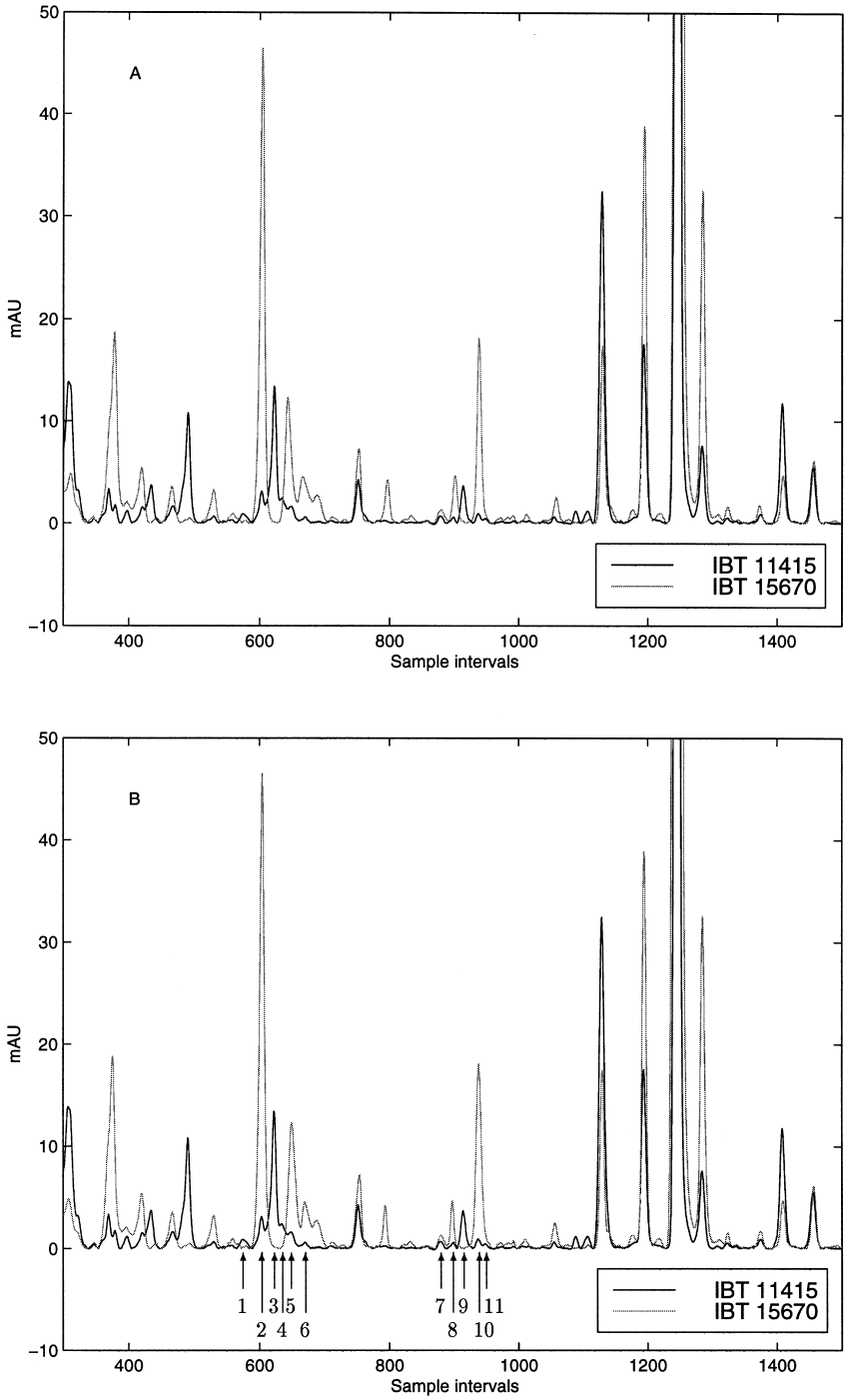

Fig. 10. 202 nm trace from the alignment of the *P. cyclopium* isolate set using the customised COW algorithm with A: (m, t)=(20, 2) B: (m, t)=(20, 6).

Table 3
Spectral correlation coefficients for the alignments in Figs. 8 and 10 at the time positions corresponding to the tops of the peaks in Fig. 10B

Peak No.	Position	Custom COW		Basic COW	
		$t=2$	$t=6$	$t=2$	$t=6$
1	574	0.7998	0.7998	0.7896	0.7819
2	603	0.9958	0.9958	0.6553	0.9921
3	622	−0.0988	0.1915	0.3725	0.3725
4	634	0.5057	−0.0741	0.1305	0.0940
5	648	0.8603	0.8718	0.8629	0.8721
6	670	0.9367	0.9361	0.9649	0.9368
7	879	0.8893	0.8872	0.4964	0.8292
8	899	0.9930	0.9930	0.9749	0.9098
9	914	0.3567	−0.2999	0.5778	0.4615
10	937	0.9985	0.9985	0.9745	0.1538
11	948	0.8306	0.8009	0.7842	0.6815
Mean		0.7334	0.6455	0.6894	0.6441
Mean of cubed values		0.5802	0.5589	0.4562	0.4304

The mean ρ_l^W using all eight isolate sets, calculated for alignments using the customised COW algorithm, is plotted in Fig. 11. The plots show values for $(m, t)=(20, 2)$, $(m, t)=(20, 6)$, unaligned chromatographic matrices and the basic warping. In all cases, baseline correction is performed, and ρ_l^W is calculated over wavelengths 200–320 nm. Comparing with Fig. 6, it is seen that the slope is now generally negative, and that the correlations are much higher. The slope is due to random effects, and

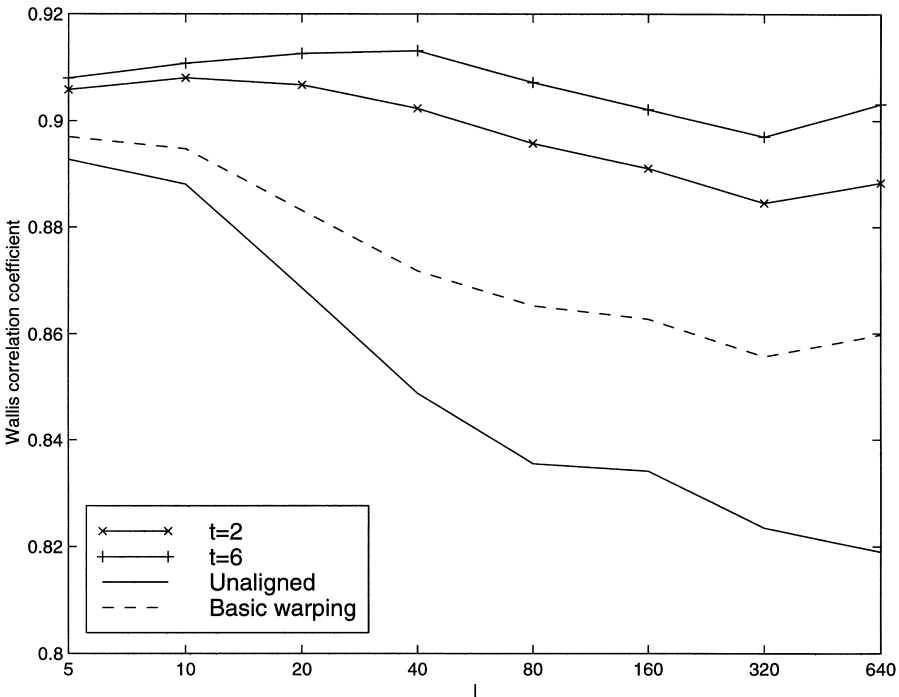

Fig. 11. Plot of ρ_l^W as a function of l for alignments using the customised COW algorithm. The plotted values are the mean of the values obtained for the eight isolate sets by aligning using wavelengths 200–320 nm.

it is unimportant whether it is positive or negative. The higher level of correlation is a reflection of the fact that emphasis is now put on spectral information, thus lessening the effect of a slightly misaligned peak, and the fact that even mismatched spectra can easily yield correlations above 0.7. The effect of the magnitude of t is similar to that found in Fig. 6: a large effect from applying COW, and a secondary, smaller effect from using a larger value of t.

The conclusions that can be drawn regarding the choice of t are thus: with a suitably customised COW algorithm, large values of t may be used without causing overfitting, because the spectral information ensures the proper alignment. Without spectral information, however, there is a real risk of overfitting, and thus in this case t should not be chosen too large. A lower limit for t can be found from the chosen value of m, and an estimate of the probable sizes of retention time shifts and the mean peak spacing.

For the data set used in this work the effect of raising the value of t was not large, and correct alignment was obtained using relatively small values of t. Considering that the amount of calculations, and thus calculation time, is a function of the square of t the proper parameter choice for the data set used here seems to be $m \approx 20$ and $t \approx 2$.

Computation time for aligning over all wavelengths in a 100×3349 chromatogram, using the basic COW algorithm with $m = 20$ and $t = 2$, was about 50 s on a HP 9000/755 workstation computer. Using the customised COW algorithm computation time was about 25 s. With $t = 6$ computation times were 5 min 20 s and 3 min 40 s, respectively.

5. Customisation of COW

A brief description of some of the possible customisations of COW are mentioned here, in order to inspire and illustrate the versatility of the method.

Firstly, preprocessing of the chromatographic matrices will improve the results, as demonstrated above. Apart from discarding superfluous data and correcting for baseline drift as in this work, all forms of data improvement techniques (such as transformations toward similar peak heights) are applicable.

The Wallis filter generates a local mean of 0 and local standard deviation of 1. Thus, a Wallis filtering of the chromatographic matrices prior to aligning would mean that ρ could be calculated directly from an element by element multiplication, without scaling. This would speed up the calculation process.

If the position of the starting or stopping point of the chromatographic profiles are unknown, and detecting them is impossible or impractical, a leading or trailing section may be added, containing only noise. If the estimated maximum difference in position of the endpoints is v then the section length s should be

$$s = m \frac{v}{t} \tag{16}$$

Secondly, changes may be made to different parts of COW itself: It has already been mentioned (Section 2.1) that the problem may be formulated with variable section lengths, and non-integer and non-uniformly spaced warping magnitudes. It is also possible to impose restrictions on the warpings u to insure that the aligned profile does not deviate from the basic warping by more than a given amount, or to limit strong compression or stretching by imposing restrictions on the length of runs of $u_i < \Delta$ or $u_i > \Delta$. A measure W of the degree of warping could be formulated as the deviation from the basic warping

$$W = \frac{1}{N} \sum_{i=1}^{N} x_i - i(m + \Delta) \tag{17}$$

or the deviation from the overall degree of stretch in the basic warping

$$W = \frac{1}{Nm} \sum_{i=1}^{N} (x_i - x_{i-1}) - (m + \Delta) \tag{18}$$

Other benefit functions may be used, e.g., the mean of the spectral correlations in the interval or a distance measure. In its basic form, COW only aligns according to shape. A possibility is to include peak height information by weighting the interval correlations according to the changes in height in the individual interval.

In this work linear warping was used, but it is possible to use non-linear warping as well, e.g., by using splines.

6. Conclusions

The proposed method for aligning chromatographic profiles gives a fast and accurate alignment without any kind of feature extraction. It performs well with profiles from complex mixtures, and is insensitive to noise and variations in baseline with time. Furthermore, it uses the full spectral information over the entire profile.

The method is automatic, and requires only two input parameters: section length m and flexibility t/m, where t is the allowed deformation of an individual section. m can be estimated from the peak width and the flexibility from retention time drift.

The general formulation of COW is universally applicable, but alignments are improved by customisation of the algorithm to exploit the characteristics of the data for individual problems.

COW alignment can be applied on all types of chromatographic data but also on data from e.g., capillary electrophoresis which can show considerable drift. In order to obtain sensible data it is important to record the data under similar conditions in order to make the differences correctable by piecewise linear warping.

COW alignment will allow creation of automated chromatographic database search procedures in a fashion as known from mass spectral databases, matching entire chromatograms.

7. List of symbols

P	Sample chromatogram
L_P	Pre-aligning length of chromatogram
m	Section length
N	Number of sections
P′	Sample chromatogram after aligning
T	Target chromatogram
L_T	Post-aligning length of chromatogram and length of target chromatogram
x	Position in data points on the time scale
$\boldsymbol{x}$	Vector set of node positions
t	The slack, i.e., the maximum warping
u	Warping in data points on the time scale
$\boldsymbol{u}$	Vector set of warpings
Δ	Difference in mean section length between P and T
ρ	Correlation coefficient (benefit function)
f	Same as ρ
Cov($\boldsymbol{a}$, $\boldsymbol{b}$)	Covariance between data sets $\boldsymbol{a}$ and $\boldsymbol{b}$
V($\boldsymbol{a}$)	Variance of data set $\boldsymbol{a}$
$\boldsymbol{F}$	Matrix containing optimal benefit function values
$\boldsymbol{U}$	Matrix containing optimal warping values
$\rho_l^{\,W}$	Wallis correlation coefficient, calculated for sections of length l

Acknowledgements

The authors would like to thank Angélique Grønborg Rasmussen for help with the description of dynamic programming theory.

Appendix A

Presented here is a pseudo code description of the COW algorithm. It shows how the sections, positions and warpings are gone through in three nested ''for'' loops, and how the $\boldsymbol{F}$ and $\boldsymbol{U}$ matrices are filled. Reconstruction of the optimal solution by use of the $\boldsymbol{U}$ matrix is also shown.

```
(*Perform dynamic programming*)
for i=0 to N
  for x=0 to L_T
    F(i,x)=-∞
  end
end
F(N,0)=0
for i=N-1 downto 0
  xstart=max(i*(m+Δ-t), L_T-(N-i)*(m+Δ+t))
  xend=min(i*(m+Δ+t), L_T-(N-i)*(m+Δ-t))
  for x=xstart to xend
    for u=Δ-t to Δ+t
      fsum=F(i+1, x+m+u)+f([x; x+m+u])
      if fsum>F(i, x) then
        F(i, x)=fsum
        U(i, x)=u
      end
    end
  end
```

```
end

(*reconstruct optimal solution*)
x(0)=0
for i=0 to N−1
  u(i)=U(i,x(i))
  x(i+1)=x(i)+m+u(i)
end
```

References

[1] D. Roberts, W. Bertsch, J. High Resolut. Chromatogr., Chromatogr. Commun. 10 (1987) 244.
[2] B.K. Logan, Anal. Chim. Acta 288 (1994) 111.
[3] T.O. Larsen, J.C. Frisvad, Mycol. Res. 99 (1995) 1167.
[4] L.S. Ramos, J. Chromatogr. Sci. 32 (1994) 219.
[5] R.L. White, P.D. Wentzell, M.A. Beasy, D.S. Clark, D.W. Grund, Anal. Chim. Acta 277 (1993) 333.
[6] H.C. Smit and E.J.v.d. Heuvel, Topics in Current Chemistry, Chemometrics and Species Identification, Vol. 147, Springer Verlag, Berlin, 1987, p. 63.
[7] M. Forina and L.C. Armanino, Topics in Current Chemistry, Chemometrics and Species Identification, Vol. 141, Springer Verlag, Berlin, 1987, p. 91.
[8] A.J. Round, M.I. Aguilar, M.T.W. Hearn, J. Chromatogr. A 661 (1994) 61.
[9] B. Grung, O.M. Kvalheim, Anal. Chim. Acta 304 (1995) 57.
[10] G. Malmquist, R. Danielsson, J. Chromatogr. A 687 (1994) 71.
[11] C.P. Wang, T.L. Isenhour, Anal. Chem. 59 (1987) 649.
[12] F.S. Hillier and G.J. Lieberman, Introduction to Mathematical Programming, McGraw-Hill, 1995, Ch. 10.
[13] J. Smedsgaard, J. Chromatogr. A 760 (1997) 264.
[14] R.H. Wallis, An Approach for the Space Variant Restoration and Enhancement of Images, Proceedings of the Symposium on Current Mathematical Problems in Image Scenes, Monterey, CA, 1976.