ELSEVIER

Journal of Chromatography A, 687 (1994) 71–88

JOURNAL OF CHROMATOGRAPHY A

Alignment of chromatographic profiles for principal component analysis: a prerequisite for fingerprinting methods☆

Gunnar Malmquist[1], Rolf Danielsson*

Institute of Chemistry, Department of Analytical Chemistry, Uppsala University, P.O. Box 531, S-751 21 Uppsala, Sweden

First received 29 October 1993; revised manuscript received 4 July 1994

Abstract

In fingerprinting methods, small differences between chromatograms with rather complex appearance have to be detected. Pattern recognition methods based on principal component analysis (PCA) could be a useful tool, but for chromatographic profiles as input data a severe problem is the great impact of chromatographic variations compared with true variations in sample composition. The problem has been analysed in terms of parameter variations for exponentially modified Gaussian peaks, and a procedure has been developed to align a sample chromatogram towards a target chromatogram in order to compensate for (i) small shifts in retention time (not due to different sample components), (ii) common variations in peak area (not due to sample composition) and (iii) variations in level and slope of the baseline. The effects of the alignment procedure on the PCA is demonstrated for a set of chromatographic profiles intended for peptide mapping.

1. Introduction

Several important analytical techniques used for biological samples rely on a comparison of chromatographic profiles. Examples of application areas are food and beverage analysis [1], DNA fingerprinting [2], pyrolysis–GC [3] and peptide mapping [4,5]. Sometimes the original sample is analysed for its components, whereas in other instances the sample is fragmented before the analysis. The analysis is performed by a separation of the components or fragments, e.g., by electrophoresis or chromatography, and the resulting profile is used as a pattern or fingerprint for the sample. In many instances, the evaluation consists of a direct visual comparison with a reference sample, in order to detect profile differences.

An important characteristic of these fingerprinting methods, is that the significant information may be contained in the presence or absence of certain components or fragments. Small differences between samples that overall are similar can, for instance, be detected by the altered retention time for a certain fragment containing the modified site. This differentiates these methods from techniques in which the determination of certain components in the samples provide the information. Another aspect is that it is seldom necessary, or even possible, to identify and quantify all peaks in the profile. The

☆ Parts of this material were previously presented at *Analysis of Peptides, Stockholm, 2–4 June 1993*, and *Analysdagarna, Lund, 14–18 June 1993*.

* Corresponding author.

[1] Present address: Pharmacia Biotech, R&D, S-751 82 Uppsala, Sweden.

0021-9673/94/$07.00
SSDI 0021-9673(94)00726-8

overall appearance of the fingerprint is instead used for discrimination purposes, making these methods qualitative rather than quantitative.

Peptide mapping is a fingerprinting method frequently used for quality control in the biotechnological production of recombinant DNA-derived proteins [6]. It is necessary to establish that the amino acid sequence is the same for different production batches. The protein is fragmented, by chemical cleavage or enzymatic digestion, and the resulting peptide fragments are separated, usually by reversed-phase liquid chromatography (RPLC). This technique will be used as an illustrative example in the present paper, while a more detailed discussion of the experimental aspects will be given in the accompanying paper [7].

The chromatographic separation is associated with several sources of variation that may have a large impact on the overall pattern. Variations in the mobile phase composition, gradient reproducibility, temperature variations and column variability lead to shifts in the chromatographic pattern, making the evaluation more difficult. If several chromatograms are to be compared, these matters should be considered. On the other hand, if two sequential chromatograms, run with the same mobile phase preparation and on the same column, are compared, these chromatographic variations might not be very severe.

The result of a manual comparison of chromatographic profiles may depend on the individuals performing the comparison. Multivariate methods that can cope with the variations in the digestion of reference samples make the evaluation less subjective. The idea is to gather a whole set of reference chromatograms that represents the normal variations in the chromatographic profile caused by the experimental conditions. Principal component analysis (PCA) [8] can be used to identify the main variation sources and to highlight the peaks where the variations are reflected. New test samples can subsequently be classified by multivariate classification methods.

The multivariate data analysis is conducted on a data set where the objects, the reference chromatograms in this case, are described by a number of variables. Some of the different approaches for conversion of chromatographic traces to variable values that have been suggested will be briefly discussed here.

The most intuitive approach is to represent the chromatograms by integration reports, i.e., the retention time and peak area for all detected peaks. The peak areas for all, or selected, peaks may constitute the variable values, provided a correct assignment of the peaks can be made between the chromatograms. An important prerequisite for multivariate analysis is that the variations should be expressed as different levels of the variables, not as shifts between variables [9]. Incorrect peak assignments imply that the quantity of a certain solute will be contained in different variables between the objects, thus reducing the quality of the data set.

The peak assignment is usually based entirely on retention time matching, unless a specific detector, e.g., a mass spectrometer, can identify the peaks. The assignment process is simplified if the retention times of the peaks are synchronized between the chromatograms. Several methods for peak synchronization, where the retention times of the detected peaks are adjusted using reference peaks that can be identified in all chromatograms, have been developed for multivariate analysis in fingerprinting contexts. These methods are conceptually related to the use of retention indices in GC. The retention time matching is only qualitative, as the retention times are used for identification purposes only. Crawford and Hellmuth [10] used two internal standard peaks to make minor linear adjustments of the retention times. More elaborate methods using multiple reference peaks have also been presented, e.g., by Mayfield and Bertsch [11], Pino et al. [12], and Parrish et al. [13].

Multivariate analysis of chromatograms represented by peak areas for a number of peaks is of limited value in peptide mapping, where the purpose is to detect modifications of the amino acid sequence. All peaks correspond to potential modification sites, and a single amino acid substitution may lead to a large change in retention.

This not only complicates the peak assignment, but implies that no relevant assignment can be made for the modified fragment as it actually constitutes a new peak. No variable will be present in the data set to contain its peak area.

The variable values can instead be defined by window summation, where the chromatogram is divided into a number of consecutive retention time segments. Each segment corresponds to one variable in the data set, and the variable value is calculated as the sum of the total signal within the window. This approach has been used, for instance, by Headley and Hardy [1] for detection of contaminants in whiskey based on GC profiles. Recently, Armanino et al. [14] presented a similar method for extraction of information regarding air pollution from GC profiles. In the latter instance, the window summation was preceded by retention alignment. A serious drawback of the window summation approach is the inherent loss of resolution. The effect of the summation is essentially a decrease in the sampling frequency, where each variable value corresponds to an averaged signal within the window. This means that variations in the size of a small peak may not be detected if it occurs in the same window as a large peak.

Fingerprinting methods in general, and peptide mapping in particular, put strong demands on the resolution and peak capacity, thus enhancing the previously mentioned drawback of window summation. A more adequate approach would be to use the entire chromatographic profile, i.e., the digitalized detector signal, directly. The data set will then have one variable for each data point in the chromatogram. Such large data sets were previously considered impractical, forcing variable reductions prior to the multivariate analysis, but the advent of powerful personal computers allows the direct analysis of large data sets. The present paper deals with some special aspects of chromatographic profiles in multivariate analysis. A preprocessing procedure is presented that facilitates the characterization of a set of reference chromatograms by PCA. An application of this procedure concerning multivariate classification of peptide mapping chromatograms is discussed in the accompanying paper [7].

2. Chromatographic profiles in PCA

It is important to realize that PCA is sensitive to all variations in the data set, i.e., not only those emanating from differences between the samples but also to variations caused by the chromatographic process. The latter aspect will be treated in some detail, illustrating the influence of chromatographic variations in PCA.

A commonly used model for chromatographic peaks is the exponentially modified Gaussian function, EMG [15]. With this function the response $y(t)$ is characterized by four parameters:

$$y(t) = \frac{A}{\tau} \exp\left[\frac{1}{2}\left(\frac{\sigma}{\tau}\right)^2 - \left(\frac{t - t_r}{\tau}\right)\right] \int_{-\infty}^{z\sqrt{2}} \exp(-x^2)\,dx \quad (1)$$

where A is the peak area, t_r and σ are the retention time and width (standard deviation) of a unit area Gaussian peak and τ is the time constant of the modifying exponential decay (tailing). The upper limit of integration is given by $z = (t - t_r)/\sigma - \sigma/\tau$.

2.1. Linear approximation for peak difference

Small shifts in the EMG parameters give rise to variations in the resulting peaks. In order to compare two similar peaks we can look at the difference Δy, which can be approximated by

$$\Delta y = \frac{\partial y}{\partial A} \cdot \Delta A + \frac{\partial y}{\partial t_r} \cdot \Delta t_r + \frac{\partial y}{\partial \sigma} \cdot \Delta \sigma + \frac{\partial y}{\partial \tau} \cdot \Delta \tau \quad (2)$$

i.e., a linear combination of the parameter shifts with the corresponding partial derivatives as coefficients. These partial derivatives with respect to the parameters are derived in the Appendix. Although the EMG function is rather complex, the results can be related to the origi-

nal peak in a simple way. Except for scaling constants they are

$$\frac{\partial y}{\partial A} = y(t) \tag{3}$$

(i.e., the EMG function itself);

$$\frac{\partial y}{\partial t_r} = \frac{dy}{dt} \tag{4}$$

(i.e., the time derivative);

$$\frac{\partial y}{\partial \sigma} = \frac{d^2y}{dt^2} \tag{5}$$

(i.e., the second time derivative);

$$\frac{\partial y}{\partial \tau} = \frac{dy}{dt} * e^{-t/\tau} \tag{6}$$

(i.e., the time derivative, exponentially modified by convolution once more).

The first two partial derivatives can easily be perceived. The area parameter is just a scaling constant, and a change in retention time appears as a shift along the time axis.

The linear approximation holds for each point in time, i.e., the response difference $\Delta y(t)$ can be regarded as a linear combination of the partial derivatives (as functions of time) with the parameter shifts as coefficients. It is valid only for small deviations in the parameters, except for the scaling parameter A, where a linear relation holds for all values. In Fig. 1 the validity of the approximation is demonstrated for certain values of parameter deviation. These values are chosen to represent a practical limit for the linear approximation. The area deviation is merely chosen to give a realistic picture of all parameter variations.

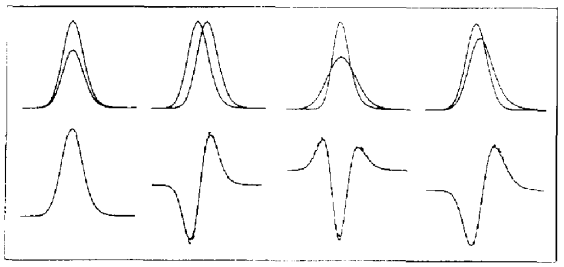

Fig. 1. Top: EMG peaks with deviations in A, t_r, σ, and τ, respectively (from left to right). Bottom: the resulting differences (solid lines) and the linear approximations using partial derivatives (dashed lines). The derivatives were obtained for A = 1 (arbitrary units), t_r = 40%, σ = 8% and τ = 5% compared with the length of the time scale. The parameter deviations were ΔA = ±0.2, Δt_r = ±4%, $\Delta\sigma$ = ±2.5% and $\Delta\tau$ = ±2.5%.

2.2. PCA of the EMG parameter shifts

To characterize the variations for a set of similar chromatographic peaks (e.g., a set of single-peak chromatograms), we can apply PCA [8]. With this technique, the chromatograms are described as linear combinations of deviations $p_1(t)$, $p_2(t)$, ... from the mean chromatogram $y_{mean}(t)$. For chromatogram i one obtains

$$y_i(t) = y_{mean}(t) + s_{i1}p_1(t) + s_{i2}p_2(t) + \cdots \tag{7}$$

Although written as time functions, the chromatogram $y_i(t)$ and the deviations $p_1(t)$, $p_2(t)$, ... are represented as sampled values at t_j, for $j = 1, 2, 3, \ldots$ In the context of PCA, the sampled values of the chromatograms are assigned to separate variables, one for each point in time. The results of PCA are the coefficients $s_{i1}, s_{i2}, \ldots$, one set for each chromatogram, and the sampled values $p_1(t_j)$, $p_2(t_j)$, ... The latter are referred to as loadings (loading vectors), deviation patterns that are common for the whole set of chromatograms. The loadings are constructed one by one, starting with $p_1(t_j)$, in such a way that the current approximations are as close to $y_i(t)$ as possible in the sense of least squares (all chromatograms included). The individual chromatograms are then represented by the coefficients $s_{i1}, s_{i2}, \ldots$, which are called the scores for chromatogram i.

Using PCA, the variations around the mean chromatogram are thus described by a number of components with contributions $s_{i1}p_1(t), s_{i2}p_2(t)$, etc. The first component accounts for as much variation as possible in the original set of chromatograms, the second plays the same role for the remaining residuals, and so on. This is reflected in a decreasing series of measures for "explained variance", the sum of which approaches 100% when the number of components

equals the number of independent sources of variation. Owing to the effect of noise and non-linearities, the optimum number of components must be determined by some more elaborate procedure, usually by cross-validation [16].

2.3. PCA for single peak variations

If the variations arise from shifts in only one of the EMG parameters, we will obtain one main component from the PCA. The loadings for this component, i.e., the shape of the variation, then corresponds to the partial derivative for that parameter (since the deviation from the average curve can be approximated by Eq. 2 for each curve). To verify this, a series of five EMG peaks were generated for each parameter (A, t_r, σ and τ), with parameter values evenly distributed within the same range as before (cf., Fig. 1). Except for a normalizing constant, the loadings obtained with PCA were almost identical with the partial derivatives (cf., Fig. 1), and the scores were related to the parameter deviations in a linear way. This is demonstrated by the correlation coefficients listed in Table 1, where also the amount of explained variance is given. The deviation from unity (100%) indicates the influence of non-linear relationships for the parameter in question.

When there are variations in more than one of the parameters, there will be additional principal components. In Fig. 2 a design is shown for deviations in area and also in retention time,

Table 1
Results from PCA of chromatograms with variations in one parameter

Parameter	Correlation coefficient		Explained variance (%)
	a[a]	b[b]	
ΔA	1.0000	1.0000	100
Δt_r	0.9993	0.9999	97.4
$\Delta\sigma$	0.9965	0.9946	98.4
$\Delta\tau$	0.9978	0.9974	98.4

[a] Correlation between the PC1 loadings and the partial derivative.
[b] Correlation between the PC1 scores and the parameter values.

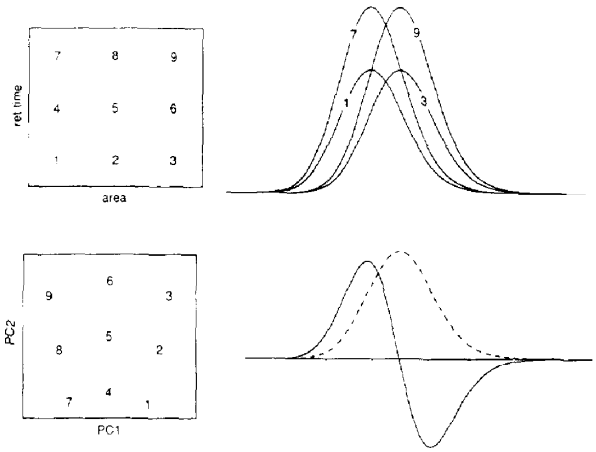

Fig. 2. Top: design of variations in A and t_r and the EMG peaks for the corner points, ($\pm\Delta A$, $\pm\Delta t_r$). Bottom: score plot for PC1 and PC2 (left) and the loadings for PC1 (solid line) and PC2 (dashed line).

together with the resulting curves corresponding to the corners of the square design. Two principal components are obtained, which together explain 98.6% of the variance. The PCA results are also depicted in Fig. 2. The two-dimensional score plot reflects the design pattern, with the deviations in retention time along the first principal component (PC1) and those in area along the second (PC2). In the two loading plots, the peak derivative shape connected with Δt_r dominates PC1 and the peak shape for ΔA dominates PC2. Thus the shifts in retention time had the greatest influence on the peaks, and the first component alone could explain 64.4% of the variations. The distortion of the design pattern shows the influence of non-linearities and interactions between the parameters.

A similar design for deviations in σ and τ, and the resulting curves for the corner points, are shown in Fig. 3 together with the results of PCA. Now the design seems to be tilted in the score plot, and the two loading plots are actually linear combinations of the two partial derivatives. Thus PCA does not separate the true sources of variations, the principal components are constructed as linearly independent combinations (orthogonal loading vectors). In the preceding case the two partial derivatives were actually

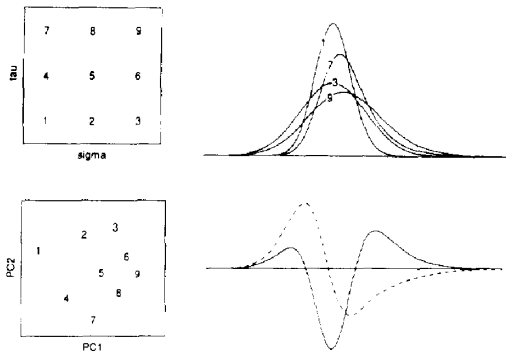

Fig. 3. Top: design of variations in σ and τ and the EMG peaks for the corner points, ($\pm\Delta\sigma$, $\pm\Delta\tau$). Bottom: score plot for PC1 and PC2 (left) and the loadings for PC1 (solid line) and PC2 (dashed line).

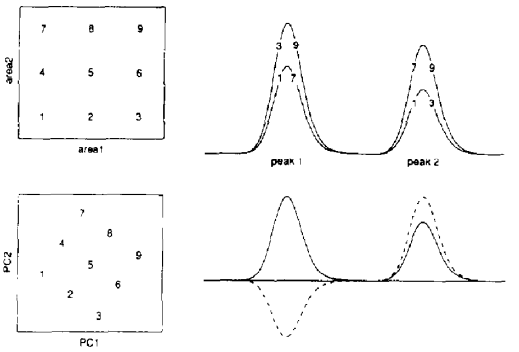

Fig. 4. Top: design of area variations for the two peaks (A_1 = 1, ΔA_1 = ±0.2, A_2 = 0.8, ΔA_2 = ±0.2) and the EMG peaks for the corner points. Bottom: score plot for PC1 and PC2 (left) and the loadings for PC1 (solid line) and PC2 (dashed line).

almost linearly independent (orthogonal functions), which explains the successful separation. For the deviations in σ and τ the first component, PC1, accounts for 74.7% of the variations, and the second, PC2, for 22.6%. Again, non-linearities and interactions are responsible for the distortion of the regular design pattern.

It should be noted that the main features of the parameter design patterns were revealed by PCA without any evaluation of the parameter values from the curves. Such values must be obtained from the statistical moments, with much problems related to the baseline, or by non-linear regression, which for the EMG function is not a trivial task.

2.4. PCA for multiple peak variations

With more than one peak in the chromatogram, there are more parameters with possible deviations to be accounted for. A simple example will show some features of PCA in the case of multiple peaks. In Fig. 4 the design for area deviations in two peaks is shown, together with the curves corresponding to the corners. The results from PCA, i.e., the scores and the loadings, are also shown in Fig. 4. Again, the design pattern is tilted in the score plot and the diagonal direction of PC1 maximizes the variations.

When the area for two peaks varies independently, the loading plots for PC1 and PC2 will reflect the sum and difference, respectively, of the single peak variation pattern. In a real case, however, the area variations are often coupled. Variations in the injected amount of sample have the same influence on both peaks, corresponding to one principal component with loadings in accordance with the sum (cf., PC1 in Fig. 4). On the other hand, another common situation is variations in the area distribution between peaks. When this is the case, the loadings for the corresponding principal component will be the difference (cf., PC2 in Fig. 4).

3. Alignment of chromatographic profiles

When PCA is applied to a set of chromatographic profiles with multiple peaks, several types of variation sources may be encountered. Usually we are looking for variations between samples, and the variations induced by the chromatographic procedure are then a severe complication. The maximum variation principle in PCA implies mixing of these sources within the principal components, and the variations of interest may be difficult to discern. In general the main source of variation is small shifts in retention time, systematic for all peaks as well as random for individual peaks, caused by, for example, variations in flow-rate, mobile phase composition, or gradient slope. Furthermore, the

overall variations in the chromatographic signal due to the injected amount of sample, detector sensitivity, etc., are reflected in the first principal components. The true sample variations may be small in comparison, and hence difficult to distinguish in the results of PCA. To detect differences between the samples that are independent of the chromatographic variations, these variations have to be reduced. Otherwise, the retention shifts and differences in the injected amount may conceal significant information regarding the samples.

The unwanted variations can be minimized by increased reproducibility in the chromatographic procedure, but multivariate analysis with chromatographic profiles as input data is sensitive to even minute variations. This requires a post-chromatographic alignment of the profiles, much in the same spirit as the retention time matching methods described in the Introduction. However, it must be stressed that chromatographic profiles as input data requires a quantitative alignment of the time scale in contrast to the more qualitative matching of peaks in the former instance.

As early as 1979, Reiner et al. [17] suggested a method to compensate for retention shifts. Each data point in the profile was adjusted towards a reference chromatogram according to a "time-warping" function. The aligned chromatograms were used for visual comparisons only, without any use of multivariate methods. Another solution has recently been proposed by Andersson and Hämäläinen [18], commercially available as the software ChromPro [19]. The entire profile is adjusted towards a selected target chromatogram, using two parameters corresponding to linear displacement of the profile and compression/expansion of the time scale, respectively. The parameter values are determined by non-linear regression using a simplex procedure, maximizing the correlation within two selected retention windows. The aligned chromatogram is calculated by linear interpolation between the two windows, preferably situated in the start and the end of the chromatogram. This approach is useful for chromatograms with relatively broad peaks. In some instances, the chromatographic profile has to be divided into several segments, each aligned individually [18]. A similar procedure has been used by Wathelet and Marlier [20] for alignment of migration distances in electrophoresis. In the work of Armanino et al. [14], the chromatograms are synchronized using multiple peaks prior to the window summation. Unfortunately, no details were given about the synchronizing algorithm.

The high peak capacity and separation power necessary for peptide mapping fingerprints often require segmented gradients in order to obtain adequate resolution within a reasonable analysis time [21]. Linear expansion or compression of the time scale will not be sufficient in these cases. The large number of peaks and the sensitivity of the retention towards the mobile phase composition further emphasize that the necessary alignment function may be non-linear. This requires a more elaborate method with individual alignment of peaks throughout the whole chromatogram. A method for such alignment for NMR profiles has been used by Vogels et al. [22]. In this instance, individual groups of lines in the two spectra are matched and brought to the same resonance position. In chromatography, however, it is seldom possible to match all peaks individually. There is a need for some interpolation of the alignment between the selected peaks.

Normalization to constant area is a common procedure used to compensate for the different amounts of injected sample. This method implies so-called closure of the data set, i.e., if one peak increases the size of other peaks must decrease [23]. This may lead to artificial correlations in the data set and thus degrade the quality of the data. In the proposed method, a selective normalization is made, by considering only selected peaks in the calculation of the normalization factor.

In this work, a combined procedure was developed to reduce the chromatographic variations. The idea is to align the sample chromatograms towards a target chromatogram in order to compensate for (i) small shifts in retention time (not due to different sample components), (ii) common variations in peak area (not due to

sample composition) and (iii) variations in level and slope of the baseline.

The alignment procedure consists of several steps, which will be illustrated with an example concerning peptide mapping. From a set of chromatograms obtained as described under Experimental, one representative chromatogram was selected as the target chromatogram. The sample chromatogram was arbitrarily chosen from the others. Each chromatogram is represented by 4900 data points, the peak width corresponding to about 30 points. Later the effects on PCA for the total set, when all chromatograms are aligned to the selected target, will be discussed.

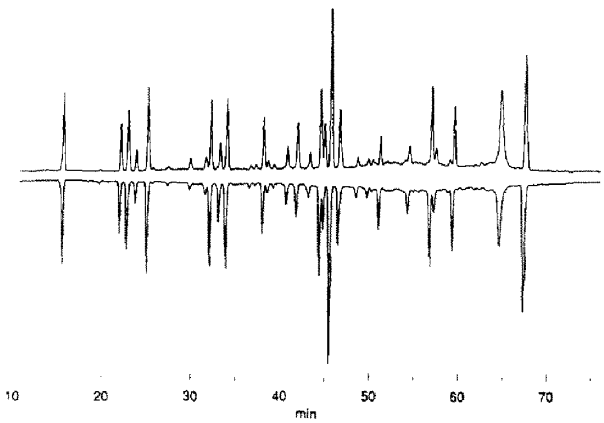

Fig. 6. Mirror plot of (top) the sample and (bottom) the target chromatograms.

3.1. Comparative chromatogram plot

To facilitate the alignment procedure a descriptive plot of the chromatograms is utilized. The sample chromatogram (the one to adjust) is plotted versus the target chromatogram in a comparative chromatogram plot (CCP). The features of this plot is demonstrated by a simplified example. Two simulated chromatograms (Fig. 5, left), with peaks differing in selected ways, illustrate the corresponding effects on the CCP (Fig. 5, right). Especially small retention shifts, the main obstacle when applying PCA to chromatograms, are clearly revealed as "loops". Although well indicated by a curved line, deviations in band broadening (σ) are not accounted for in the alignment procedure to be described.

For our real example the two chromatograms are shown in the traditional "mirror plot" (Fig. 6) and also the comparative chromatogram plot (Fig. 7). The dominant feature in the CCP is the wide loops, indicating that the retention time

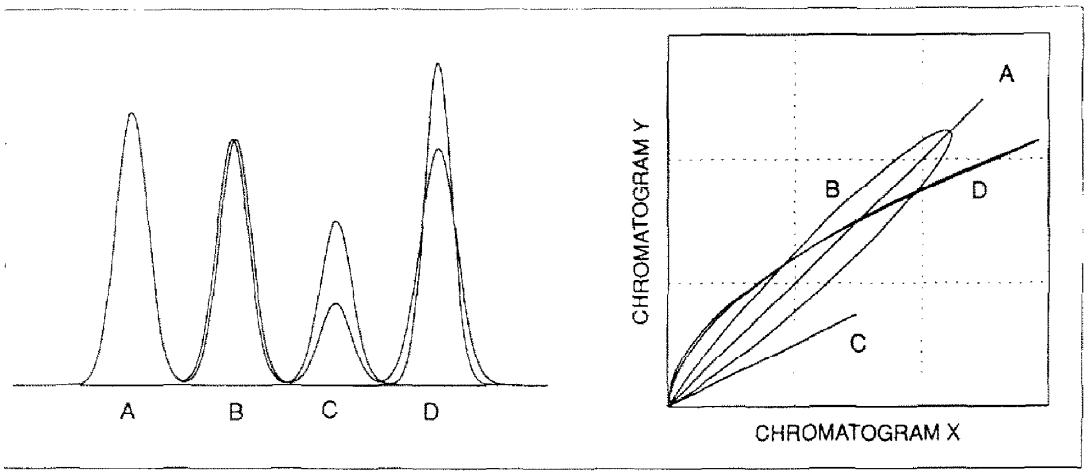

Fig. 5. Left: two chromatograms with four EMG peaks: A = no parameter shifts; B = shift in t_r; C = shift in A; D = shift in σ. Right: the CCP of the two chromatograms.

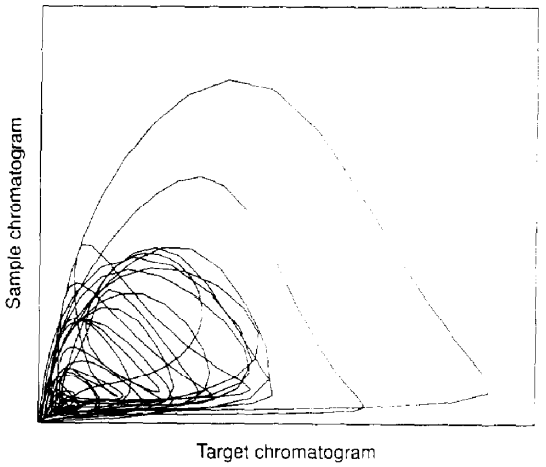

Fig. 7. The CCP for the untreated sample chromatogram.

scale is shifted between the two chromatograms (cf. peak B in Fig. 5).

3.2. Retention alignment

The first part of the alignment procedure is to align the retention for peaks that are assumed to correspond to the same sample component. For the two chromatograms $y_{sample}(t)$ and $y_{target}(t)$ the cross-correlation function (ccf) is defined as

$$ccf(\Delta t) = \int y_{sample}(t \quad \Delta t) y_{target}(t)\, dt \qquad (8)$$

which for equidistant data points is evaluated as the sum

$$ccf(\Delta t) = \sum y_{sample}(t_i - \Delta t) y_{target}(t_i) \qquad (9)$$

This function is calculated for discrete values $\Delta t = 0, \pm 1, \pm 2, \ldots$, where the integers refer to the time displacement expressed as number of sampling intervals.

In the first step, the cross-correlation function is calculated for all data points in the two chromatograms, and a maximum is obtained for some Δt. This integer time shift value is used as an overall time shift for a coarse precorrection of the test chromatogram by renumbering the data points. This corrects for any large shift of the chromatogram along the time axis.

In the second step, a limited number of peaks, those with the largest peak heights, are selected from the target chromatogram. If the number is not too high, the selected peaks correspond to main components and should appear in the sample chromatogram also, as we are dealing with sets of apparently similar chromatograms. For each of these peaks, a section of the target chromatogram is taken around the peak centre. The length of this section is typically a few peak widths, and related to the largest time shift allowed. The cross-correlation function for each section and the corresponding section of the sample chromatogram is calculated, and from its maximum the time correction for each selected peak is obtained. The position of the ccf maximum indicates how many data points (i.e., sampling intervals) the section of the sample chromatogram must be shifted to match that of the target chromatogram as close as possible.

The individual time shifts, which are valid at the centre of the selected peaks, are used to construct a time displacement function $\delta t(t)$ for the sample chromatogram. The time shift for all points between these fix points is calculated by linear interpolation as shown in Fig. 8, where the fix points are indicated by circles. By visual inspection of the time displacement function, any mis-matches for the selected peaks are detected, e.g., as obvious outliers in the curve, and can be removed.

For the sample chromatogram a corrected time scale, $t' = t + \delta t(t)$, is now applied. The sample chromatogram is evaluated at the points

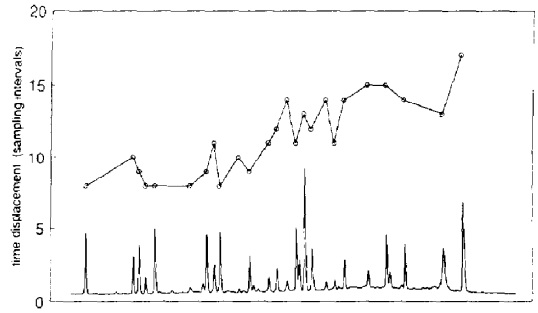

Fig. 8. The time displacement function for the coarse time alignment. The target chromatogram is shown for identification of selected peaks.

in time where the target chromatogram is sampled

$$y'_{\text{sample}}(t) = y_{\text{sample}}(t') = y_{\text{sample}}[t + \delta t(t)] \qquad (10)$$

When the corrected time value lies between two adjacent points in the sample chromatogram, $y'_{\text{sample}}(t)$ is calculated from the adjacent points by linear interpolation. The CCP for the sample chromatogram at this stage is shown in Fig. 9.

In the third step, a larger number of peaks in the target chromatogram are involved. For all peaks identified as local maxima, the cross-correlation function with the same section of the pre-aligned sample chromatogram is calculated. This time a smaller number of shifts (sampling intervals) are allowed, corresponding to about half the peak width. The maximum value of the *ccf* for each peak is used together with the peak height in the target chromatogram to sort the peaks, with highest preference for large peaks that correlates well with the corresponding portion of the sample chromatogram.

So far the time shifts have been obtained as integer values of the sampling interval. In the first two steps the sample chromatogram was actually recalculated (coarse alignment of selected main peaks). The result of the third step was a list of integer time shift values for a larger number of peaks, but so far no adjustment was made.

Because even small variations in retention times have a great influence on the principal components, a fourth fine-tuning step is desirable. A certain number of peaks are taken from the sorted list, and for each one the *ccf* is calculated for five points around the maximum found in the previous step. To these values a polynomial is fitted, and if the maximum of the polynomial lies within the five points, the peak is a candidate for matching. Again a piece-wise linear time displacement function $\delta' t(t)$ is constructed from the positions of the *ccf* maxima (now non-integer values), the validity of which should be checked by visual inspection as before.

The final alignment of the sample chromatogram according to

$$y''_{\text{sample}}(t) = y'_{\text{sample}}[t + \delta' t(t)] \qquad (11)$$

is obtained in the same way as before. The comparative chromatogram plot after time alignment is displayed in Fig. 10, which should be compared with that for the untreated data (Fig. 7).

3.3. Response corrections

In Fig. 10, the CCP so far, a dashed line indicates the trace for identical peaks, i.e., with

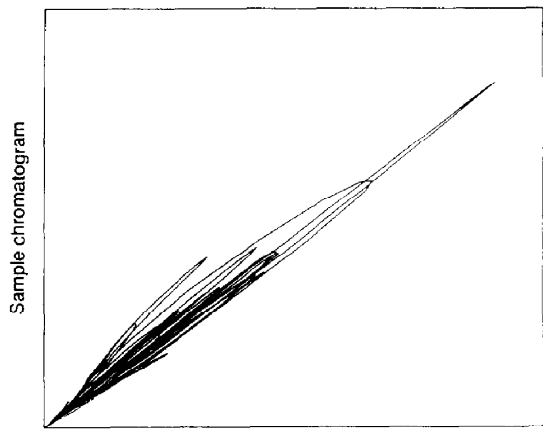

Fig. 9. The CCP after the coarse time alignment.

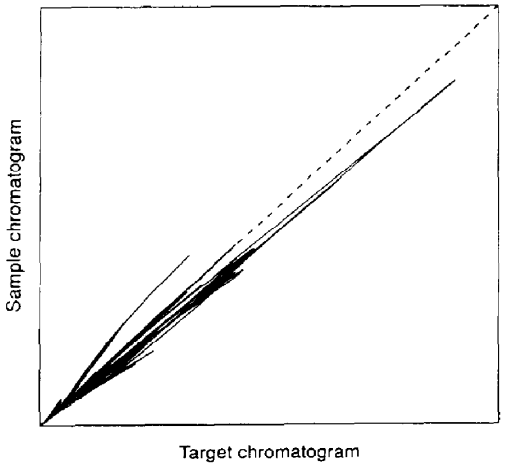

Fig. 10. The CCP after the final time alignment. The dashed line is the trace for identical peaks.

the same chromatographic response at all sampling events. It is seen that the majority of points fall on one side of this line. Obviously there is a need for a common correction factor for all peaks, but some smaller peaks do show deviations from the main trace. These true differences in peak area must be considered when the normalization is performed. The proposed method applies a least-squares fit, which also takes into account possible differences in the baseline characteristics (level and slope). The linear regression model is

$$y''_{\text{sample}}(t) = a + bt + cy_{\text{target}}(t) \quad (12)$$

The regression is obtained iteratively, initially using all data points. Then all points too far from the ideal trace are excluded from the regression, according to the condition

$$|y''_{\text{sample}}(t) - a - bt - cy_{\text{target}}(t)| < k + ly_{\text{target}}(t) \quad (13)$$

The constants k and l define a wedge-shaped strip around the ideal diagonal trace, and points outside this strip are rejected. The regression is repeated without these points, leading to new values for a, b and c and a new test for exclusion. Finally, the set of excluded points is constant, which ends the regression procedure.

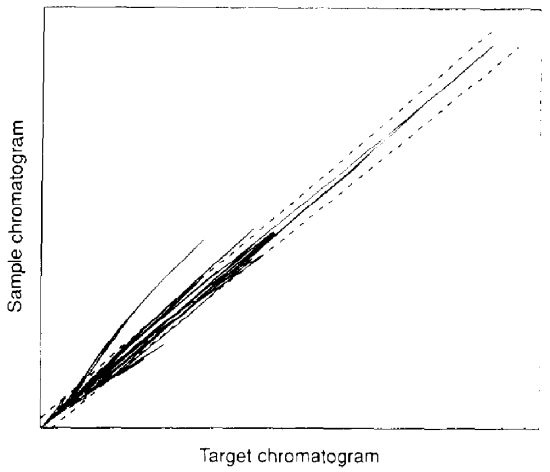

Fig. 11. The CCP after time alignment and response correction. The dashed lines indicate conditions for data points used in regression.

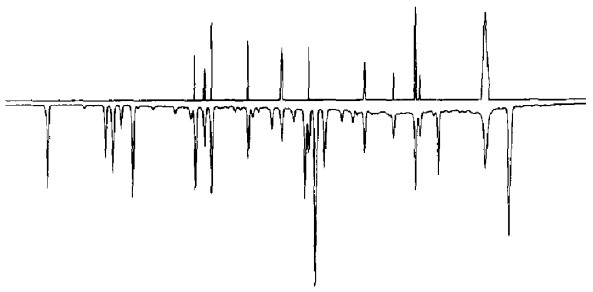

Fig. 12. Peaks with true area differences in the sample chromatogram (top) compared to the target chromatogram (bottom).

The final correction of the sample chromatogram is then

$$y'''_{\text{sample}}(t) = [y''_{\text{sample}}(t) - a - bt]/c \quad (14)$$

and the resulting comparative chromatogram plot is shown in Fig. 11, where also the lines of rejection are indicated. The segments that fall outside these lines are regarded as peaks with a true area difference, and a chromatogram where only such peaks are shown is a tool for pinpointing the sample differences (Fig. 12). A similar method for response correction, multiplicative scattering correction (MSC), has been used to pretreat near-infrared reflectance spectra prior to multivariate calibration [24].

4. Experimental

4.1. Tryptic digests

The tryptic digests of equine cytochrome *c* (Sigma, St. Louis, MO, USA) were prepared according to the procedure described by Renlund et al. [25], with the exception that the concentration of trypsin (Sigma) was decreased to 0.2 μg/μl [26]. The procedure was also scaled up five times, by increasing the volumes in all steps. Four replicated preparations of the reagents for denaturation and cysteine reduction, desalting buffer and the trypsin solution were used.

4.2. Chromatographic procedure

The tryptic digests were injected by a CMA 200/240 refrigerated (4°C) autosampler (CMA Microdialysis, Stockholm, Sweden), and separated on a SuperPac Pep-S C_2/C_{18} (5 μm, 100 Å) column (250 × 4 mm I.D.) using a precolumn (10 × 4 mm I.D.) packed with the same material. The separations were performed with a Model 2249 gradient pump with detection at 215 nm by a Model 2141 variable-wavelength monitor. The chromatographic system was controlled by HPLCmanager software, also used to store the chromatograms prior to the multivariate analysis. All chromatographic columns and instruments were from Pharmacia Biotech (Uppsala, Sweden), except where indicated.

The separations were performed by gradient elution (flow-rate 1 ml/min). The mobile phases were consistently prepared by weighing instead of volumetric measurements. The aqueous phase (A) consisted of 50 m*M* phosphate buffer (pH 2.5), prepared by mixing fixed amounts of stock solutions of phosphoric acid and sodium dihydrogenphosphate (both from Merck, Darmstadt, Germany). The organic phase (B) consisted of acetonitrile–A (80:20). The acetonitrile was of gradient grade (Merck). The mobile phases were degassed by sparging with helium for 5 min (A) and 10 min (B). The samples (125 μl) were eluted with a linear gradient from 0 to 60% B in 96 min, corresponding to a gradient slope of 0.5% acetonitrile/ml. All calculations were implemented in the programming environment ASYST (Macmillan Software, New York, USA).

5. Results and discussion

To test the performance of the proposed alignment and selective normalization procedure, 27 replicated digests of equine cytochrome *c* were prepared as described under Experimental. The digestions were divided into two sets, consisting of fourteen and thirteen samples, respectively. The protein samples in each set were simultaneously digested in the same thermostated digestion block. Within each set, digestions were made with two separate preparations of all reagents, e.g., denaturing agent, cysteine reducing agent and trypsin solution. After the complete digestion, the digests were stored at 4°C and chromatographed twice, each replicate performed with freshly prepared mobile phases. The time between the digestion and the last injection were in all instances less than 4 days (this storage time can be accepted, as shown by Dougherty et al. [27]). The training set for PCA consists of 54 chromatograms, as described in Table 2.

The training set was pretreated by the procedure presented above. A suitable target chromatogram can be chosen by arbitrarily selecting one of the training set chromatograms with intermediate retention for the majority of the peaks. Another approach is to perform PCA on the raw, non-aligned, chromatograms and select the chromatogram that correlates best with the first principal component [20]. For this data set, chromatogram No. 27 (see Table 2) was selected, as it fulfilled both these criteria.

The proposed alignment and selective normalization procedure was tested by performing PCA on the data set in different phases of the procedure. The three versions of the data set correspond to (i) raw data, (ii) retention aligned chromatograms and (iii) the final data set where the selective normalization and baseline correction is included. It is important to realize that the pretreatment of the chromatograms has reduced the total amount of variation by removing the retention shifts and differences in the injected amount. The total sum of squares is calculated by

$$SS_{\text{tot}} = \sum_{i=1}^{N} \sum_{j=1}^{P} [y_i(t_j) - \bar{y}(t_j)]^2 \qquad (15)$$

where N and P are the number of objects and data points (variables), respectively. The average chromatogram is denoted by $\bar{y}(t_j)$. Comparison of the total sum of squares between the three phases of the pretreatment (see Table 3) shows that the variations have been reduced to about 1.6% of the initial amount.

Table 2
Design of the training set

Digest	Amount of protein (mg)	Chromatogram	Digestion set	Reagent preparation	Mobile phase preparation
1	0.62	1/15	1	a	A/C
2	0.56	2/16	1	b	A/C
3	0.64	3/17	1	a	A/C
4	0.48	4/18	1	b	A/C
5	0.65	5/19	1	a	A/C
6	0.53	6/20	1	b	A/C
7	0.48	7/21	1	a	A/C
8	0.55	8/22	1	b	B/D
9	0.47	9/23	1	a	B/D
10	0.53	10/24	1	b	B/D
11	0.45	11/25	1	a	B/D
12	0.59	12/26	1	b	B/D
13	0.60	13/27[a]	1	a	B/D
14	0.55	14/28	1	b	B/D
15	0.47	29/42	2	c	E/G
16	0.51	30/43	2	d	E/G
17	0.45	31/44	2	c	E/G
18	0.49	32/45	2	d	E/G
19	0.44	33/46	2	c	E/G
20	0.61	34/47	2	d	E/G
21	0.52	35/48	2	d	F/H
22	0.61	36/49	2	c	F/H
23	0.55	37/50	2	d	F/H
24	0.57	38/51	2	c	F/H
25	0.43	39/52	2	d	F/H
26	0.53	40/53	2	c	F/H
27	0.51	41/54	2	d	F/H

Each digest was chromatographed twice with replicated mobile phase preparations.
[a] Indicates the selected target chromatogram.

Table 3
Results from PCA on the data set in different phases of the pretreatment

Data set	SS_{tot} [a]	Rank[b]	Explained variance (% of SS_{tot})						
			PC1	PC2	PC3	PC4	PC5	PC6	Total[c]
Raw data	1186	6	41.2	24.1	16.0	7.6	3.1	2.0	94.0
Aligned data	64.7	4	84.8	9.9	1.6	1.4			97.7
Final data	19.2	5	67.8	10.9	6.8	5.5	2.4		93.4

[a] Total sum of squares in the data set, calculated by Eq. 15.
[b] Number of significant principal components according to cross-validation.
[c] Cumulative explained variance with the significant components.

5.1. Characterization by principal component analysis

The data set consists of 54 objects (chromatograms), each described by 4900 variables (sampling interval 0.8 s). Before the PCA, the (54 × 4900)-dimensional matrix was mean centred, i.e., for each individual variable the mean over all chromatograms was subtracted from each chromatogram.

The number of significant principal components is not very important if the main purpose of the PCA is to characterize the data set, looking for patterns and groupings between the chromatograms. The principal components are calculated so that the amount of variance that is explained by the individual components decreases for each additional component (see Table 3). This means that the main information regarding the variations in the data set is found in the first few components. In this case it is sufficient to calculate an arbitrarily chosen number of principal components and then examine the score plots for the first components. Nevertheless, cross-validation [16] can be used to determine the number of components corresponding to the model with the best predictive ability.

To assess the influence of the chromatographic variations in the data set, PCA was first performed on the raw, non-aligned chromatograms (see Table 3). The loading plot for the first principal component, explaining about 41% of the initial variation, is shown in Fig. 13. From the complicated pattern, and the many regions resembling the first time derivative, it is obvious that the retention shifts influence the data set to a very high degree. The score plot of the two first principal components for the raw data (Fig. 14) shows a strong clustering according to the mobile phase preparations. The replicated chromatograms of the same digest are in most instances very far from each other. The pattern can be interpreted as a general trend, going from the first chromatograms in the upper right part of the score plot via the middle left part to the last chromatograms in the lower right part. This indicates that the underlying factor might be related to time and not to random deviations in the mobile phase composition. One possible explanation is a gradual degradation of the performance of the chromatographic column, e.g., by the acidic mobile phase. Such prominent variations caused by the chromatographic process will certainly obscure the interesting, sample-dependent, variations in the data set. The use of a multivariate classification method for detection of abnormal samples will probably not be successful with non-aligned chromatograms.

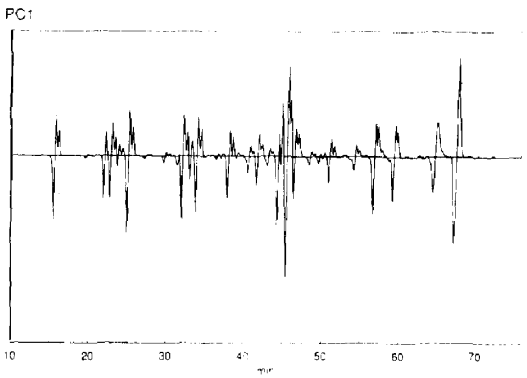

Fig. 13. PC1 loadings with the raw data set.

In the next phase, PCA was performed on the retention aligned chromatograms, prior to the selective normalization and baseline adjustment (see Table 3). The loadings for PC1, explaining about 84% (data not shown), are all positive and the overall pattern is very similar to PC0, i.e., the average chromatogram. The similarity can be quantified by calculation of the correlation between the variable averages and the loadings for PC1. The good correlation, $r = 0.96$, indicates that the dominant source of variation is related to the injected amount of sample. The correlation between the scores in PC1 and the initial amount of protein (see Table 2), is lower, $r = 0.88$. This suggests that the injected amount is influenced also by other factors, e.g., the recovery in the sample pretreatment or the digestion efficiency. The differences in the injected amount will prevent the characterization of variations between samples.

Finally, after retention alignment, selective normalization and baseline adjustment, the data set was characterized by PCA (see Table 3). If

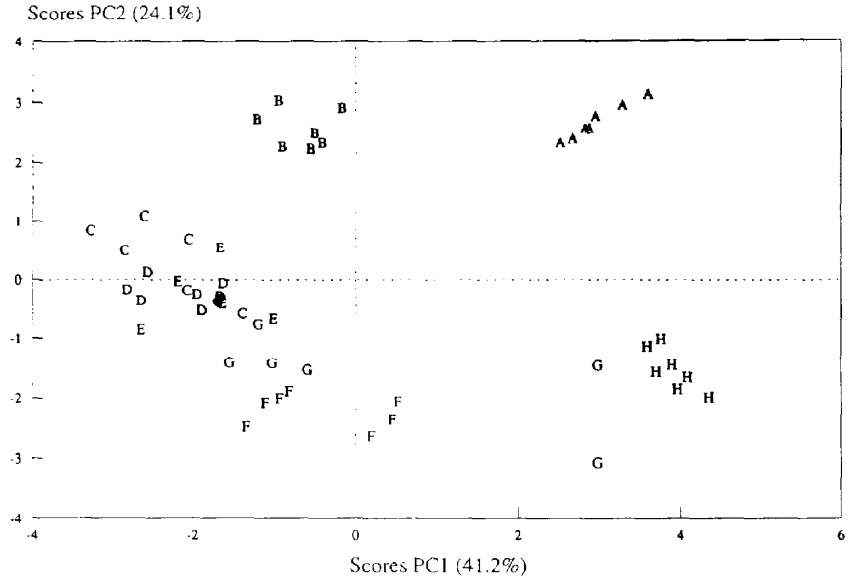

Fig. 14. Score plot of the two first principal components calculated for the raw data. The letters A–H designate the mobile phase preparations (see Table 2).

the scores for the two first principal components (together explaining about 79% of all variations in the final data set) are plotted against each other (see Fig. 15), an interesting pattern is revealed. The objects (chromatograms) are separated according to the preparation of the reagents, with the preparation labelled "a" (see Table 2), situated in the upper part of the score plot, i.e. with high scores in the second component. In the lower part of the score plot, a less

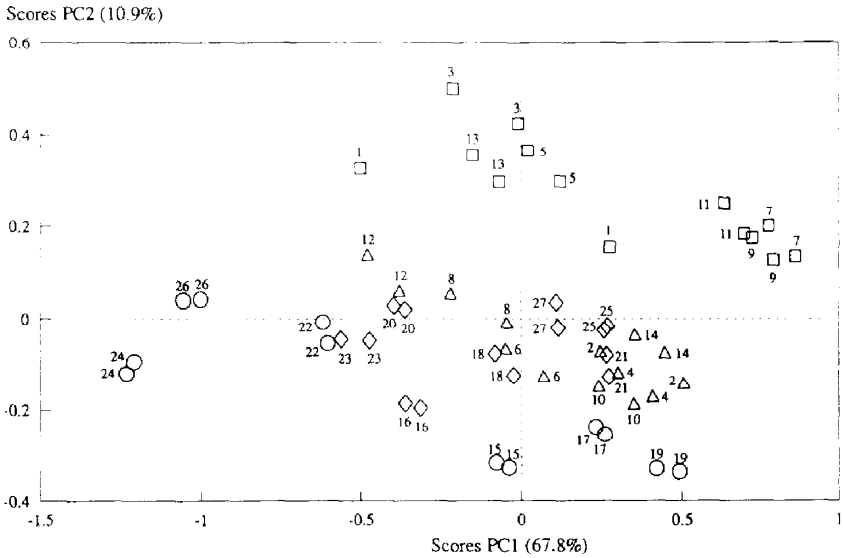

Fig. 15. Score plot for the two first principal components calculated for the final data set. The symbols denote the replicated reagent preparations (see Table 2): □ = a; △ = b; ○ = c; ◇ = d. The numbers indicate the individual digests (see Table 2).

pronounced separation is observed for the preparation labelled "c". The two preparations "b" and "d" are not separated and situated in the middle of the score plot.

The different digests (see Table 2) are also indicated in Fig. 15. The two replicated chromatograms of the same digest are generally close to each other in the score plot, showing that the chromatographic variations are to a large extent removed from the data set. One exception can be seen for digest 1, where the two chromatograms are fairly far apart in the score plot. This deviation can be mainly attributed to the observed large difference in the peak, probably corresponding to undigested cytochrome *c*, as discussed below.

No systematic information that can be easily interpreted is found in the scores for the higher principal components (data not shown), but these components are nevertheless important to characterize the data set fully.

The loading plot for the first principal component, explaining about 68% of all variations, is shown in Fig. 16. The component is dominated by negative loadings for the broad peak eluted at approximately 64 min (cf., Fig. 6). This peak is believed to be connected with the undigested cytochrome *c*, which means that the most important variation in the final data set is the degree of digestion. It is also natural that this variation is preserved after the selective normalization procedure, as the amount of undigested protein is negatively correlated with the amount of digested protein, i.e., the majority of peaks. The first component is not correlated with the average chromatogram (by loadings) or the initial amount of protein (by scores).

The second component, responsible for the separation between reagent preparations, has a more complicated pattern (Fig. 17). Some parts of the loading plot, e.g., the first peak at 16 min and the last peak at 68 min, show an anomalous pattern. This indicates that there are remaining variations in peak shape.

There are some types of chromatographic variations that cannot be removed by the current method. Unfortunately, there is to our knowledge no other procedure that would be successful in the following cases either. In the presence of overlapping peaks, it is not possible to correct for variations in the overlap between the peaks. This could probably be achieved by fitting an appropriate peak shape model, e.g., Gaussian or exponentially modified Gaussian, to each peak in the chromatogram. These fitted peaks could then be aligned and the resulting chromatogram calculated. This is not an attractive solution, however, owing to the large number of peaks present in most fingerprint chromatograms. The proposed strategy does not compensate for variations in peak width and tailing. Such variations could be caused by column degradation, which is a serious problem in all fingerprinting methods, regardless of the type of evaluation and interpretation that is used [12].

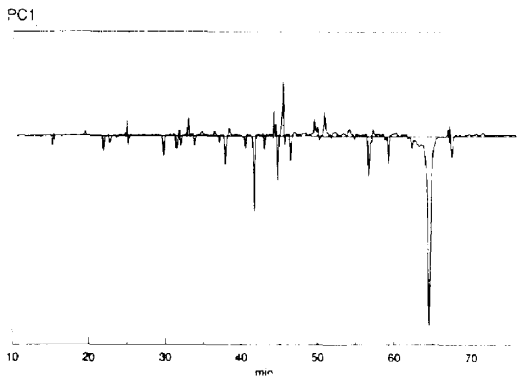

Fig. 16. PC1 loadings for the final data set.

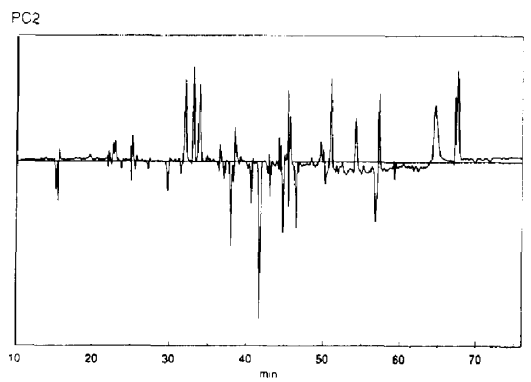

Fig. 17. PC2 loadings for the final data set.

The proposed alignment procedure is nevertheless capable of reducing the chromatographic variations in the current data set to an acceptable level. This is indicated in the score plot (Fig. 15) by the generally small deviations for the replicated chromatograms of the same digest. The characterization by PCA could describe differences between the digests that were unrelated to variations in peak overlap and column degradation.

6. Conclusions

By using the proposed retention alignment procedure and the selective normalization of peak heights, it is possible to perform principal component analysis on complex chromatographic data sets. When there are many peaks in the chromatograms, and there is a possibility that peaks can disappear and new peaks appear, it is beneficial to use the whole chromatographic profile for the analysis. The proposed method can be made fully automatic and allows the processing of numerous chromatographic profiles in a data set.

For the peptide mapping data set used in this study it was possible to detect differences between the digests that would have been obscured by the chromatographic variations if proper alignment had not been performed.

The proposed pretreatment method is also applicable to other situations where complex chromatographic data sets are treated by multivariate data analysis, e.g., pyrolysis–GC [3]. In the accompanying paper [7], multivariate classification of tryptic digests is suggested as an objective evaluation method for peptide mapping.

Acknowledgements

We are grateful to our former colleague Niklas Lundell, now at Pharmacia Bioscience Center, Stockholm, Sweden, for inspiring ideas during the initiation of this project and for fruitful discussions during the later stages.

Appendix

Despite the complex appearance of the EMG function in the time domain (Eq. 1), the calculations can readily be performed in the frequency domain. The Fourier transform of the EMG function is

$$\tilde{y}(\omega) = A \exp[-(\omega\sigma')^2] \exp(-j\omega t'_r)/(1 + j\omega\tau') \quad \text{(A1)}$$

The primed versions of the time related parameters t_r, σ and τ are scaled with the constant $2\pi/T$, where T is the duration of the time function.

The equation can be interpreted according to three frequency-dependent factors:

(i) $\exp[-(\omega\sigma')^2]$, a Gaussian peak of width σ with unit area around $t = 0$;

(ii) $\exp(-j\omega t'_r)$, shifting this peak along the time axis by t_r;

(iii) $1/(1 + j\omega\tau')$, convolution of the shifted peak with an exponential decay with time constant τ (tailing).

Finally this unit area peak is multiplied with the area parameter A.

The transform can easily be calculated for $\omega = 0, 1, 2, \ldots, N$, and the time function is then obtained for $2N$ equally spaced points in time by the inverse fast Fourier transform (IFFT). All simulated chromatograms with EMG peaks in this work were obtained in this way.

Moreover, the separation of the parameters in the four factors facilitates the calculation of the partial derivatives:

$$\frac{\partial\tilde{y}}{\partial A} = (1/A)\tilde{y}(\omega) \quad \text{(A2)}$$

$$\frac{\partial\tilde{y}}{\partial t'_r} = (-j\omega)\tilde{y}(\omega) \quad \text{(A3)}$$

$$\frac{\partial\tilde{y}}{\partial\sigma'} = -2\sigma'(\omega^2)\tilde{y}(\omega) = 2\sigma'(j\omega)(j\omega)\tilde{y}(\omega) \quad \text{(A4)}$$

$$\frac{\partial\tilde{y}}{\partial\tau'} = [(-j\omega)/(1 + j\omega\tau')]\tilde{y}(\omega) \quad \text{(A5)}$$

By including the scaling constant $2\pi/T$, the partial derivatives with respect to the original, unprimed, parameters are obtained. They can be

transformed to the time domain by IFFT. However, the shape of the partial derivatives in the time domain can be predicted without numerical calculations. Apart from constants, i.e., factors not containing ω, the derivatives are the original EMG function, possibly multiplied with the factors $j\omega$ and $1/(1+j\omega\tau)$. Multiplication with these factors in the frequency domain corresponds to the time operations d/dt (time differentiation) and convolution with $\exp(-t/\tau)$, respectively. Thus partial differentiation corresponds to the following shape modifications of the EMG time function:

$\partial/\partial A$ unmodified;

$\partial/\partial t_r$ time differentiation;

$\partial/\partial\sigma$ time differentiation twice;

$\partial/\partial\tau$ time differentiation and exponential convolution.

References

[1] L.M. Headley and J.K. Hardy, *J. Food Sci.*, 57 (1992) 980.

[2] J.S.C. Smith and O.S. Smith, *Adv. Agron.*, 47 (1992) 85.

[3] J.A. Pino, J.E. McMurry, P.C. Jurs, B.K. Lavine and A.M. Harper, *Anal. Chem.*, 57 (1985) 295.

[4] F.E. Regnier, *LC·GC*, 5 (1987) 392.

[5] F.E. Regnier, *LC·GC*, 5 (1987) 472.

[6] W.S. Hancock, *LC·GC*, 5 No. 4 (1992) 30.

[7] G. Malmquist, *J. Chromatogr.* 687 (1994) 89.

[8] S. Wold, K. Esbensen and P. Geladi, *Chemometr. Intell. Lab. Syst.*, 2 (1987) 37.

[9] S. Wold, C. Albano, W.J. Dunn, III, U. Edlund, K. Esbensen, P. Geladi, S. Hellberg, E. Johansson, W. Lindberg and M. Sjöström, in B.R. Kowalski (Editor), *Chemometrics: Mathematics and Statistics in Chemistry*, Reidel, Dordrecht, 1984, p. 17.

[10] N.R. Crawford and W.W. Hellmuth, *Fuel*, 69 (1990) 443.

[11] H.T. Mayfield and W. Bertsch, *Comput. Appl. Lab.*, 1 (1983) 130.

[12] J.A. Pino, J.E. McMurry, P.C. Jurs, B.K. Lavine and A.M. Harper, *Anal. Chem.*, 57 (1985) 295.

[13] M.E. Parrish, B.W. Good, F.S. Hsu, F.W. Hatch, D.M. Ennis, D.R. Douglas, J.H. Shelton and D.C. Watson, *Anal. Chem.*, 53 (1981) 826.

[14] C. Armanino, M. Forina, L. Bonfanti and M. Maspero, *Anal. Chim. Acta*, 284 (1993) 73.

[15] J.P. Foley and M.S. Jeansonne, *J. Chromatogr. Sci.*, 29 (1991) 258.

[16] S. Wold, *Technometrics*, 20 (1978) 397.

[17] E. Reiner, L.E. Abbey, T.F. Moran, P. Papamichalis and R.W. Schafer, *Biomed. Mass Spectrom.*, 6 (1979) 491.

[18] R. Andersson and M.D. Hämäläinen, *Chemometr. Intell. Lab. Syst.*, 22 (1994) 49.

[19] *ChromPro*, BioTriMark, Björkkulla, Funbo, Uppsala, Sweden.

[20] B. Wathelet and M. Marlier, *Chemometr. Intell. Lab. Syst.*, 4 (1988) 327.

[21] R.C. Chloupek, W.S. Hancock and L.R. Snyder, *J. Chromatogr.*, 594 (1992) 65.

[22] J.T.W.E. Vogels, A.C. Tas, F. van den Berg and J. van der Greef, *Chemometr. Intell. Lab. Syst.*, 21 (1993) 249.

[23] E. Johansson, S. Wold and K. Sjödin, *Anal. Chem.*, 56 (1984) 1685.

[24] P. Geladi, D. MacDougall and H. Martens, *Appl. Spectrosc.*, 39 (1985) 491.

[25] S. Renlund, I.-M. Klintrot, M. Nunn, J.L. Schrimsher, C. Wernstedt and U. Hellman, *J. Chromatogr.*, 512 (1990) 325.

[26] S. Renlund, personal communication, 1992.

[27] J.J. Dougherty, Jr., L.M. Snyder, R.L. Sinclair and R.H. Robins, *Anal. Biochem.*, 190 (1990) 7.