Journal of Chromatography A, 961 (2002) 237–244

JOURNAL OF
CHROMATOGRAPHY A

www.elsevier.com/locate/chroma

Chromatographic alignment by warping and dynamic programming as a pre-processing tool for PARAFAC modelling of liquid chromatography–mass spectrometry data

Dan Bylund[a], Rolf Danielsson[a], Gunnar Malmquist[b], Karin E. Markides[a,*]

[a]*Department of Analytical Chemistry, Uppsala University, Box 531, 751 21 Uppsala, Sweden*

[b]*Amersham Biosciences, 751 84 Uppsala, Sweden*

Received 4 July 2001; received in revised form 6 May 2002; accepted 8 May 2002

Abstract

Solutes analysed with LC–MS are characterised by their retention times and mass spectra, and quantified by the intensities measured. This highly selective information can be extracted by multiway modelling. However, for full use and interpretability it is necessary that the assumptions made for the model are valid. For PARAFAC modelling, the assumption is a trilinear data structure. With LC–MS, several factors, e.g. non-linear detector response and ionisation suppression may introduce deviations from trilinearity. The single largest problem, however, is the retention time shifts not related to the true sample variations. In this paper, a time warping algorithm for alignment of LC–MS data in the chromatographic direction has been examined. Several refinements have been implemented and the features are demonstrated for both simulated and real data. With moderate time shifts present in the data, pre-processing with this algorithm yields approximately trilinear data for which reasonable models can be made. © 2002 Elsevier Science B.V. All rights reserved.

Keywords: PARAFAC modelling; Parallel factor analysis; Time warping algorithm

1. Introduction

Chromatographic systems coupled to sophisticated detection devices, e.g. liquid chromatography–mass spectrometry (LC–MS), generate large amounts of second-order data in chemical analysis. In order to process these data with traditional methods, the dimensionality of the data must be reduced. This can, for example, be achieved by generating total ion chromatograms or lists of integrated peak areas. An alternative approach, circumventing the inevitable loss of information in the dimension reduction step, is to directly model the raw data by the use of multiway analysis methods [1–5]. Thereby, the so-called second-order advantage [6] will be maintained, allowing for accurate quantitative analysis even in the presence of unknown interferents (as long as the sensitivity for the analyte is unaffected). A typical application of interest would be the detection of minor differences between complex samples, e.g. in peptide maps intended for quality control [7] or biomedical research.

A problem is that many of the chemometric techniques available for multiway modelling rely on trilinearity [8], a prerequisite seldom met due to variations in the chromatographic conditions affect-

*Corresponding author. Tel.: +46-18-471-3691; fax: +46-18-471-3692.

E-mail address: karin.markides@kemi.uu.se (K.E. Markides).

PII: S0021-9673(02)00588-5

ing both peak position and peak width. One way to tackle this problem is to pre-process the data by some kind of time alignment procedure. Several such methods have been proposed for one-dimensional chromatographic data [9–13] and there are also some methods for higher order data published, where the spectral information is used to guide the alignment procedure [14,15].

In 1998, Nielsen et al. introduced an algorithm for correlation optimised warping (COW) of chromatographic profiles [16]. By warping, a set of corresponding points (nodes) in the two profiles is selected to give maximal similarity between the intermediate sections. If the numbers of data points between two nodes are different for the two profiles, one of them is compressed or expanded by linear interpolation. The procedure reduces variations in time dependence of otherwise similar profiles, e.g. caused by column ageing or flow-rate variations. Thereby the effects from such chromatographic variations on multivariate or multiway models can be reduced, which greatly facilitates the interpretation. The correlation obtained with COW can also be directly used as a similarity measure for classification [17].

In this paper, we present a modification of COW, focusing on how to best utilise mass spectral information to align the chromatograms. The performance of the algorithm is evaluated for both real and simulated LC–MS data, and the effects on models obtained with principal component analysis (PCA) and parallel factor analysis (PARAFAC) [1] are shown.

2. Theory

2.1. The time warping algorithm

The notation from the original COW paper [16] will be used in this section, which mainly focuses on the modifications made to this procedure.

The first step is to select a target chromatogram T with size $s\times(L_T+1)$, where L_T is the number of sampling intervals and s is the number of single mass chromatograms. This target chromatogram should be as representative of the entire population as possible, for example the chromatogram acquired in the middle of the run sequence. Then T is divided into N time segments. These segments can be either of equal length m, giving a minimum need for user interaction, or of variable lengths m_n, which gives the possibility to focus on key features in the data. Since the segments are selected for T, it is more practical than in the original COW algorithm where the segments are chosen for the chromatograms to be aligned (P). The end points of the segments are referred to as nodes. Since the last position of a segment equals the first position of the next, there will be $N+1$ nodes in total.

In the next step, the chromatogram P $(s\times(L_P+1))$ is warped to match the segments in T. This can be seen as a combinatorial problem, where the optimal sequence of node positions in P is sought within a given candidate solution space. The size of this space depends on the value of the slack parameter t, which sets the limit for the change in the number of points in the segments of P. If $L_P \ll L_T$ it can be useful to make this limit asymmetric so that more stretching than contraction is allowed (and vice versa for $L_P \gg L_T$). A low setting of t speeds up the algorithm, but increases the risk of poor alignment if large time variations are present in the data. On the other hand, too high a setting may lead to mismatched peaks and/or deteriorated peak shapes.

The solution to the combinatorial problem can be found by, e.g. simplex optimization or genetic algorithms. However, the speed of current computers makes it possible to test the entire candidate solution space, thereby ensuring that the global optimum is found. In COW as well as in our modification, this is done in a systematic way by so-called dynamic programming. The results of the matching are stored in two matrices, **F** and **U**, both of size $(N+1)\times(L_P+1)$. Each row in **F** refers to a node, and within a row the accumulated value of the maximum benefit function (see below) are stored for the allowed time positions according to the constraints. The contribution from the current optimal segment corresponds to a certain position of the preceding node. The starting position of this segment is stored as the corresponding element in the matrix **U**. (In the original COW, **U** stores the optimal amount of warping.) Finally, the maximum value of **F** is localised and from the corresponding element in **U**, the node positions are backtracked to find the optimal node sequence from which the aligned chromatogram P′ ($s\times(L_T+1)$) can be derived.

An important difference compared to the original COW algorithm is the shape of the candidate solution space. In COW, the start and end points of P′ are fixed as being the start and end points of P, thereby giving the largest flexibility in the middle of the chromatogram. In our modification, there are no such fixed positions. Instead the user sets an uncertainty for the start position, and then the flexibility is allowed to increase with time. The lack of fixed positions means that the end portions of P can be discarded if they are not related to T. It is also possible to leave out the final part of T if it is not represented in P (this case gives a warning from the program). The negative effect of the modification of COW is increased computational time due to the enlarged candidate solution space, and a higher risk for mismatched peaks. However, the high specificity of mass spectral data makes the latter problem less significant.

Another difference is the choice of benefit function, i.e. the function that measures the similarity between the two data matrices. In the original COW algorithm, the correlation coefficient (or its cube value) is used as the benefit function, while the covariance is the default option for the benefit function in our modification of COW. The difference between correlation and covariance is whether the cross product is scaled by the variance in the segment. The scaling incorporated in the calculation of the correlation has the effect that matching segments with large peaks is not favoured compared to matching segments with smaller peaks. However, it also means that baseline variations may have a larger influence on the matching procedure. The preferred choice of benefit function is therefore dependent on the purpose of the analysis and the nature of the data. The two algorithms also differ in the centering procedure for second-order data. In the original COW algorithm, the matrices are centred with their total mean, while for the modified algorithm each row of the matrix, that is each mass channel, is centred to zero mean before the element-by-element cross products are summed. Subtracting the row mean rather than the total mean will reduce possible influence of background patterns on the covariance, and thereby reduce the need for data pre-processing. This is especially advantageous when working with LC–MS data, where the analyte signals are present for a few detection channels rather than over continuous spectra. Most of the detection channels will then only carry background and noise, and their contributions to the overall covariance are minimised when each channel is centred separately.

The MATLAB (The MathWorks Inc., Natick, MA) code for the modified COW algorithm is available from the authors on request.

2.2. The PARAFAC model

PARAFAC [1] can be considered as an extension of PCA into higher order data. For a three-way data array $\underline{\mathbf{X}}$ ($I\times J\times K$), the following trilinear model is fitted in the least squares sense

$$x_{ijk}=\sum_{f=1}^{F}a_{if}b_{jf}c_{kf}+e_{ijk} \tag{1}$$

Here F is the number of factors, e_{ijk} is an element in the residual array $\underline{\mathbf{E}}$ ($I\times J\times K$), and a_{if}, b_{jf}, and c_{kf} are elements of the loading matrices **A** ($I\times F$), **B** ($J\times F$), and **C** ($K\times F$), respectively. For LC–MS data, **A** will be related to the mass spectra, **B** to the elution profiles and **C** to the concentrations of the solutes, provided that the trilinear model is valid.

With the PARAFAC algorithm, different constraints can be applied for the elements of the loading matrices. For uncentred LC–MS data, non-negativity is a natural constraint for all elements of **A**, **B** and **C**.

3. Experimental

3.1. Real LC–MS data

The LC–MS system comprised a Rheos 2000 pump (Flux Instruments AG, Basle, Switzerland), a 1.0-μl external loop injector (Valco Instruments, Houston, TX), a PepMap C_{18} column (150×0.3 mm, 3 μm) from LC Packings (Amsterdam, The Netherlands) and an API100 mass spectrometer (PE Sciex, Concord, ON, Canada) operated in the positive ion scan mode. The pump was operated at 60 μl/min, with a flow splitting device before the

Table 1
Conditions for five LC–MS runs of a peptide standard mixture

Run #	Acquisition date	Comment
a	000303	
b	000303	Gradient slope slightly changed
c	000303	
d	000316	New mobile phase prepared
e	000317	One day storage at room temp.

injector, giving approximately 3 μl/min through the column. The solvents were A (95% water and 5% methanol) and B (20% water and 80% methanol), both acidified to 0.1% formic acid. Linear gradient elution was applied from 0 to 90% B within 40 min.

A peptide standard mixture (Sigma, St Louis, MO) was analysed five times under different conditions (Table 1) causing shifts in the retention and the relative intensity of the peaks. Portions of the data (matrices of size 1200×300) with seven of the peptides present was transferred to MATLAB (v.5.3) for subsequent analysis. One of the runs (d) was used as target with a constant segment size of $m = 20$ data points, corresponding to about two peak widths. The modified COW algorithm was then used to align all the data matrices with the start position within 10 data points and the slack parameter t set to 10.

PCA and PARAFAC models were made both for the raw data and for the aligned data. In order to perform PCA, it was necessary to reduce the dimensionality of the data. In this work, the base peak chromatogram (BPC, the maximum signal in each spectrum vs. time) was used as the one-dimensional representation of the LC–MS data. No scaling or centering was applied before calculating the models. In PARAFAC modelling, non-negativity constraints were used for all three modes and seven factors were calculated. PCA models were also made of the third mode loadings (the part of the PARAFAC model that is related to the concentrations of the solutes) with the purpose to further clarify the relationship between the samples.

3.2. Semi-simulated data

Following a factorial design (Table 3), changes in the chromatographic peak height, width and position were artificially introduced to the second peak in run (c) (data matrix size 1200×40) in order to compare different benefit functions with respect to their sensitivity for variations in the peak properties. The original data matrix (#1 in Table 3) was used as target and its similarity with the eight matrices was calculated. As the measure of similarity, four different benefit functions were tested (covariance and correlation with total or row mean centering).

4. Results and discussion

4.1. Real LC–MS data

Fig. 1 shows the BPCs of the five objects before and after alignment against run (d). Except for the retention time deviations present for the raw data, the

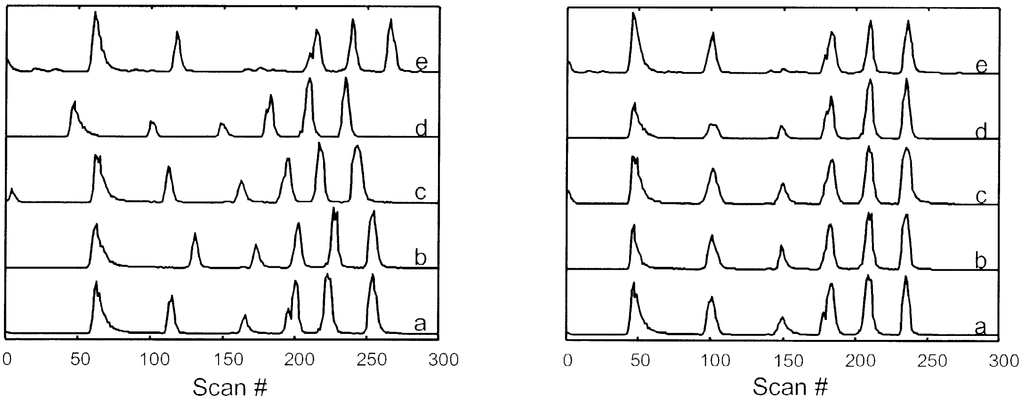

Fig. 1. BPCs of five LC–MS runs of a peptide standard mixture (cf. Table 1) before and after alignment with the modified COW algorithm.

most apparent difference between the objects is the almost total absence of the third peak in run (e). Possible explanations for this are deterioration or adsorption of this peptide due to storage in a plastic vial at room temperature (Table 1). The question was whether this sample-related variation could be discerned in a PARAFAC model or if it should be hidden by the variation caused by the chromatographic conditions.

One way to further visualise the degree of chromatographic alignment achieved is the contour plot of the covariance matrix of P and T. In Fig. 2 this is shown for run (a). The time shifts present are revealed as deviations between the spot centres and the main diagonal. The solution space searched and the optimal node sequence found for this run are also indicated in Fig. 2. After warping, the spots follow the main diagonal as seen in Fig. 3.

The effects of the alignment are also apparent for the results of the PCA models made for the BPCs of all runs. Table 2 gives the amount of variance explained by the principal components (PCs) of the data before and after warping. The time shifts present for the unaligned data make the PCA model less useful for characterisation of the sample variation, since the variance caused by the chromatography will have an influence on the principal components. The first two PCs, normally used for score plots, explained only 70% of the variance for the unaligned data. With alignment, PC1 and PC2 explained more than 98%, and in the score plot, run (e) was separated from the other runs, mainly in PC2. This could be expected from the shape of the aligned chromatograms (cf. Fig. 1). When PCA is carried out on a set of similar but non-normalised profiles, PC1 is dominated by total area variations, while varia-

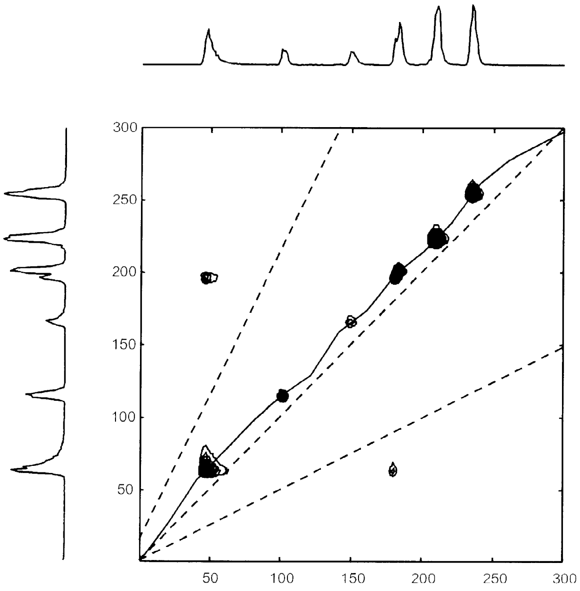

Fig. 2. The contour plot of the variance–covariance matrix for two LC–MS runs of a peptide standard mixture. The dotted lines indicate the diagonal for perfect alignment and the border limits of the candidate solution space, while the optimal time warping sequence found with the modified COW algorithm is indicated with solid lines. The BPCs of the two runs are also shown.

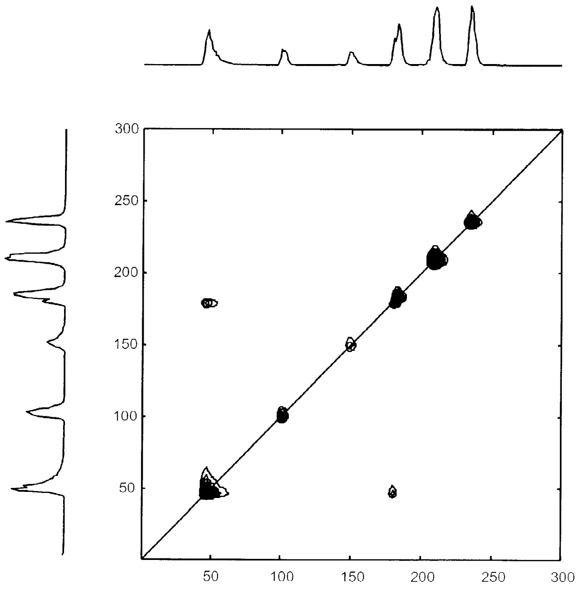

Fig. 3. The BPCs and the contour plot for the variance–covariance matrix of two LC–MS runs (cf. Fig. 2) after alignment with the modified COW algorithm. The main diagonal, corresponding to perfect retention time alignment, is indicated.

Table 2
Explained variances for PCA models of the five BPCs (cf. Fig. 1) before and after alignment with the modified COW algorithm

Component	Raw data (%)	Warped data (%)
PC1	48.4	96.6
PC2	21.6	1.8
PC3	15.3	0.7
PC4	9.2	0.6
PC5	5.5	0.3

tions in the area distribution between the peaks are manifest in the second and higher PCs. With large retention time shifts, however, the PCA model cannot be interpreted in such an intuitive way.

What is stated above for bilinear PCA modelling also holds for trilinear PARAFAC modelling. With unaligned data the first seven factors (corresponding to the number of peptides) only explained 59.5% of the total variance. The first mode loadings (Fig. 4), representing the mass spectra, showed overlaps (correlation) between some of the factors, and the third mode loadings, representing the peptide concentrations in the five samples, included several zeros. Because of the misalignment the same peptide could be modelled at different peak positions in different samples, i.e. a single peptide will influence more than one PARAFAC factor. In such a case, the corresponding first mode loadings are correlated (as actually found). Moreover the peak position separating versions of the same peptide will then appear only in selected samples, which could explain the third mode zeros. Finally, seven factors will not be sufficient to describe all systematic variation. As found, the lack of trilinearity for the unaligned data makes the PARAFAC model less appropriate in data analysis.

With properly aligned data, the seven-factor PARAFAC model could explain 96.7% of the total variance. Here the loadings in the first mode (Fig. 5a) and the second mode (Fig. 5b) gave good approximations of the mass spectra and elution profiles, respectively, of the seven peptides. Thus, the third mode loadings should relate to the concentrations of the peptides within the five runs. A two-component PCA model of this loading matrix was sufficient to explain 99% of the variance. The score plot (Fig. 6) separated run (e) from the others, much like the findings for the PCA model of the BPCs. The resolution obtained for peak #4 and peak #5 (Fig. 5b) exemplifies that the PARAFAC model

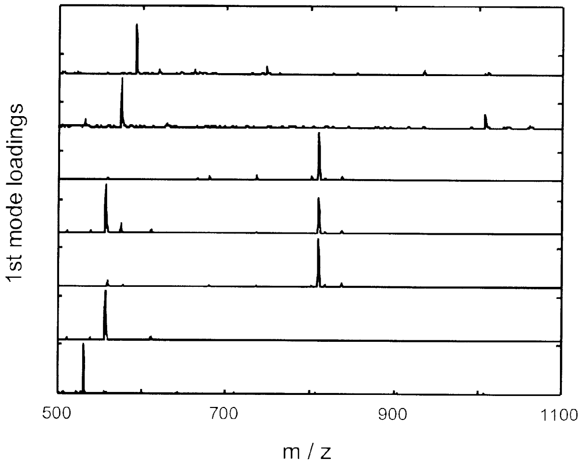

Fig. 4. First mode loadings (corresponding to mass spectra) of a PARAFAC model of unaligned LC–MS data for a peptide standard mixture.

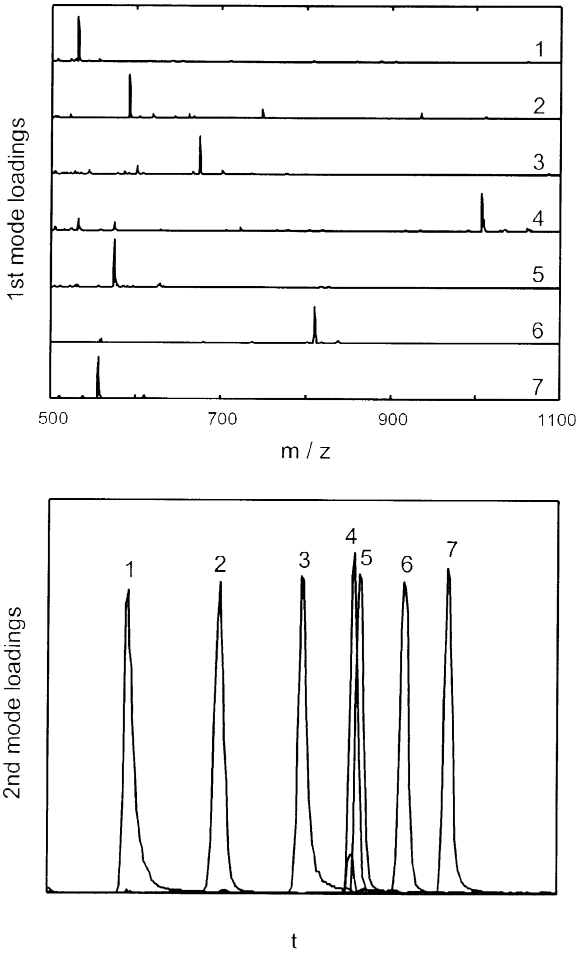

Fig. 5. First and second mode loadings of a PARAFAC model of aligned LC–MS data for a peptide standard mixture. The seven factors are related to the mass spectra and elution profiles of bradykinin (1), luteinizing releasing hormone (2), substance P (3), oxytocin (4), metionin enkephalin (5), bombesin (6) and leucin enkephalin (7).

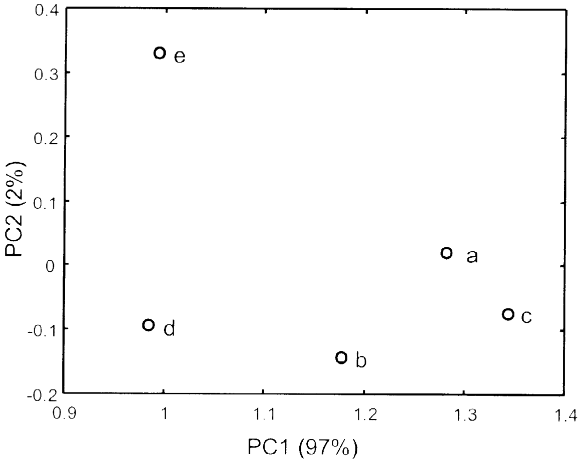

Fig. 6. PCA score plot of the third mode loadings (related to the concentrations) in a PARAFAC model of aligned LC–MS data for a peptide standard mixture. Run (e) is separated from the others mainly in PC2.

also constitutes an alternative to multivariate curve resolution and peak purity assessment tools [18–20], methods for which the rotational freedom may cause problems [21]. Due to the second-order advantage, such rotational ambiguity is not present for the PARAFAC model [1].

4.2. Semi-simulated data

The design of the simulations, with shifts in chromatographic peak position, height and width, is given in Table 3. The similarity between the target matrix (#1) and the others (#2–8) was calculated as correlation coefficient and covariance with total mean and row mean centering. The values that were found for these four benefit functions are included in Table 3 (where also the autocovariance and -correlation values for the target matrix are reported), while a summary of the interpretation of the results is given in Table 4.

Table 4
Main and interaction effects of peak property variations on different benefit functions (cf. Table 3)

Variable	Cov(rc)	Cov(tc)	Corr(rc)	Corr(tc)
$\lvert\Delta t_r\rvert$	−	−	−	−
I_{max}	+	+	0	0
σ	−	+	−	−
$\lvert\Delta t_r\rvert\times\sigma$	+	+	+	+

The sign (+ or −) indicates the influence of an increase in the corresponding variable.

A retention time shift (compare, e.g. #1 and #2) was found to decrease the values of all the benefit functions remarkably, with the largest effects obtained for row mean centering. This is of importance since even minor time shifts introduce extra factors in multivariate and multiway analysis.

Peak height variations (compare, e.g. #1 and #3) largely affects the covariance measures, while only minor effects were found for the correlation coefficients (how much will depend on the signal-to-background ratio and the choice of centering procedure). When using the covariance as the benefit function it is thus a larger risk present for alignment towards a large peak of a different compound instead of a smaller peak of the correct substance. The magnitude of this risk depends on the degree of spectral similarity and the amount of the erroneous compound, since the covariance of aligned chromatographic peaks is approximately proportional to the peak heights and the cosine of the spectral angle.

An increased peak width (compare, e.g. #1 and

Table 3
Evaluation of benefit functions on semi-simulated data with variations in chromatographic peak position (Δt_r), height (I_{max}) and width (σ)

Exp #	$\lvert\Delta t_r\rvert$/s	I_{max}/counts	σ/s	Cov(rc)×10^{-6}	Cov(tc)×10^{-6}	Corr(rc)	Corr(tc)
1	0	2800	7.5	27.940	43.556	1.0000	1.0000
2	8.5	2800	7.5	8.848	24.514	0.3172	0.5628
3	0	1400	7.5	13.971	21.784	1.0000	0.9998
4	8.5	1400	7.5	4.424	12.262	0.3172	0.5628
5	0	2800	15	24.268	54.464	0.8897	0.8949
6	8.5	2800	15	14.265	44.507	0.5248	0.7313
7	0	1400	15	12.134	27.236	0.8897	0.8949
8	8.5	1400	15	7.133	22.258	0.5248	0.7314

≠5) led to a slight decrease in all of the benefit functions studied, except for the covariance with total mean centering where a significant increase was noted. Therefore this benefit function may lead to peak broadening in the aligned chromatograms. This effect is restricted by the slack parameter *t*, and it will also be less pronounced when close eluting peaks are present. A significant interaction between the peak width and the peak position was also registered for all of the benefit functions studied (this can, for example, be verified by comparing the differences ≠1–≠5 and ≠2–≠6). This is due to the fact that a retention time difference has less impact when the peaks get broader, i.e. it is the resolution rather than the absolute retention time shift that is of importance.

To summarise, separate centering of all the detection channels gives the highest sensitivity for retention time alignment and should be the natural choice for LC–MS data. To select between correlation and covariance is more intricate. The covariance measure is sensitive to the peak height and will thus favour the alignment of large peaks. However, this also means that the covariance gives a larger risk for erroneous alignment when close eluting solutes with similar mass spectra are present.

5. Conclusions

It has been demonstrated that a proper time alignment is a necessity for successful PARAFAC modelling of LC–MS data, and that such alignment can be achieved using the modified COW algorithm presented. The result of the alignment is much dependent on the choice of benefit function, a fact also indicated by Pravdova et al. [22]. However, no general guidance for this choice can be given since both the properties of the raw data and the analytical purpose for the aligned data must be considered.

Acknowledgements

Rasmus Bro provided the PARAFAC algorithm used [23]. Financial support from the Swedish Natural Research Council project K-5104-706 and from the Swedish Foundation for Strategic Research is gratefully acknowledged.

References

[1] R. Bro, Chemometr. Intell. Lab. Syst. 38 (1997) 149.

[2] R. Bro, J. Chemometr. 10 (1996) 47.

[3] L.R. Tucker, Psychometrica 31 (1966) 279.

[4] E. Sanchez, B.R. Kowalski, Anal. Chem. 58 (1986) 496.

[5] A. de Juan, S.C. Rutan, R. Tauler, D.L. Massart, Chemometr. Intell. Lab. Syst. 40 (1998) 19.

[6] E. Sanchez, B.R. Kowalski, J. Chemometr. 2 (1990) 247.

[7] D. Bylund, J. Samskog, S.P. Jacobsson, K.E. Markides, J. Am. Soc. Mass Spectrom., submitted.

[8] A. de Juan, R. Tauler, J. Chemometr. 15 (2001) 749.

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

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

[11] T.L. Cecil, S.C. Rutan, Anal. Chem. 62 (1990) 1998.

[12] G. Malmquist, R. Danielsson, J. Chromatogr. A 687 (1994) 71.

[13] S.A. Cohen, M.W. Gorenstein, J.R. Henriksen, Genet. Eng. News 20 (2000) 32.

[14] B. Grung, O.M. Kvalheim, Anal. Chim. Acta 304 (1995) 57.

[15] B.J. Prazen, R.E. Synovec, B.R. Kowalski, Anal. Chem. 70 (1998) 218.

[16] N.-P.V. Nielsen, J.M. Carstensen, J. Smedsgaard, J. Chromatogr. A 805 (1998) 17.

[17] N.-P.V. Nielsen, J. Smedsgaard, J.C. Frisvad, Anal. Chem. 71 (1999) 727.

[18] F. Cuesta Sánchez, B. van den Bogaert, S.C. Rutan, D.L. Massart, Chemometr. Intell. Lab. Syst. 34 (1996) 139.

[19] A. Garrido Frenich, M. Martínez Galera, J.L. Martínez Vidal, D.L. Massart, J.R. Torres-Lapasió, K. De Braekeleer, J.-H. Wang, P.K. Hopke, Anal. Chim. Acta 411 (2000) 145.

[20] D. Bylund, R. Danielsson, K.E. Markides, J. Chromatogr. A 915 (2001) 43.

[21] R. Manne, Chemometr. Intell. Lab. Syst. 27 (1995) 89.

[22] V. Pravdova, B. Walczak, D.L. Massart, Anal. Chim. Acta 456 (2002) 77.

[23] C.A. Andersson, R. Bro, Chemometr. Intell. Lab. Syst. 52 (2000) 1.