Contents lists available at ScienceDirect

Journal of Chromatography A

journal homepage: www.elsevier.com/locate/chroma

No-alignment-strategies for exploring a set of two-way data tables obtained from capillary electrophoresis–mass spectrometry

M. Daszykowski[a], R. Danielsson[b], B. Walczak[a,*]

[a] Department of Chemometrics, Institute of Chemistry, Silesian University, 9 Szkolna Street, 40-006 Katowice, Poland
[b] Department of Physical and Analytical Chemistry, Analytical Chemistry, Uppsala University, P.O. Box 599, SE-751 24 Uppsala, Sweden

ARTICLE INFO

Article history:
Received 25 December 2007
Received in revised form 8 March 2008
Accepted 11 March 2008
Available online 15 March 2008

Keywords:
Comparing data tables
Hyphenated techniques
Two-dimensional fingerprints
Alignment
Warping
Chemometrics

ABSTRACT

Hyphenated techniques such as capillary electrophoresis–mass spectrometry (CE–MS) or high-performance liquid chromatography with diode array detection (HPLC–DAD), etc., are known to produce a huge amount of data since each sample is characterized by a two-way data table. In this paper different ways of obtaining sample-related information from a set of such tables are discussed. Working with original data requires alignment techniques due to time shifts caused by unavoidable variations in separation conditions. Other pre-processing techniques have been suggested to facilitate comparison among samples without prior peak alignment, for example, 'binning' and/or 'blurring' the data along the time dimension. All these techniques, however, require optimization of some parameters, and in this paper an alternative parameter-free method is proposed. The individual data tables ($\mathbf{X}$) are represented as Gram matrices ($\mathbf{XX}^T$), where the summation is taken over the time dimension. Hence the possible variations in time scale are eliminated, while the time information is at least partly preserved by the correlation structure between the detection channels. For comparison among samples, a similarity matrix is constructed and explored by principal component analysis and hierarchical clustering. The Gram matrix approach was tested and compared to some other methods using 'binned' and 'blurred' data for a data set with CE–MS runs on urine samples. In addition to data exploration by principal component analysis and hierarchical clustering, a discriminant partial least squares model was constructed to discriminate between the samples that were taken with and without the prior intake of a drug. The result showed that the proposed method is at least as good as the others with respect to cluster identification and class prediction. A distinct advantage is that there is no need for parameter optimization, while a potential drawback is the large size of the Gram matrices for data with high mass resolution.

1. Introduction

Nowadays, hyphenated chromatographic techniques, such as high-performance liquid chromatography–mass spectrometry (LC–MS) are frequently used in proteomic and metabolomic studies. They provide as output a data table of each individual chromatographic run in which a sample is analyzed. Therefore, the data collected in the analysis of several samples can be viewed as a three-way array, e.g. *retention time* × *multiple detector responses* × *samples*. A major advantage of the hyphenated techniques over the chromatographic methods equipped with monochannel detectors is the possibility to overcome problems with co-elution and to verify the purity of chromatographic peaks [1]. The hyphenated chromatographic techniques are often the methods of choice in order to obtain fingerprints of complex mixtures like biofluids (e.g. urine and serum samples), environmental samples, peptides, food samples, etc., where the goal is to find differences among samples and to identify components responsible for these differences. Moreover, the obtained data are very complex and their chemometric exploration is still an ongoing challenge [2].

Usually the collected three-way data are matricized (unfolded), for example, like *samples* ×(*retention time* × *multiple detector responses*) and further explored with unsupervised chemometric techniques designed to process two-way data. The applications of principal component analysis (PCA) [3] and different clustering techniques [4] seem to be dominant in the literature. With PCA, data visualization and a study of similarities among samples are possible by projecting samples onto selected pairs of principal components that describe the majority of the data variance. Additionally, PCA helps to analyze relationships among the explanatory variables and their contributions to individual principal components. In addition to PCA, the hierarchical clustering approaches

* Corresponding author. Tel.: +48 32 3592115; fax: +48 32 2599978.
E-mail address: beata@us.edu.pl (B. Walczak).

0021-9673/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.chroma.2008.03.027

are often a preferred exploratory tool due to their good visualization properties. Their purpose is to form groups of similar samples or variables. The degree of similarity is evaluated using distance measures or correlation coefficients.

Processing unfolded data seems to be a simple and straightforward strategy to deal with the three-way data arrays obtained with LC–MS, etc. There are, however, some pitfalls when applying PCA and similar techniques on chromatographic data. Unstable chromatographic conditions may cause peak shifts along the time axis, and the retention times of individual samples, put one over another in the unfolded matrix, do not correspond to each other. Such undesired instrumental variations could mask the chemical variability among the samples being studied, and the applied chemometric techniques might lead to improper conclusions.

In general, two strategies of instrumental signal pre-processing are considered to enable proper multivariate data analysis of chromatographic data. In the first strategy, the chromatographic signals are represented as a table of peak intensities or peak areas for selected mixture components. The peak table is then subjected to multivariate data exploration and/or modeling. A major difficulty when using this strategy is the requirement of standards for identification and quantification of peaks. The other strategy does not include detection of peaks; the chromatographic signals are seen as fingerprints that eventually need to be further analyzed using chemometric methods. This data pre-processing strategy involves the use of alignment techniques to adjust the peak shifts for a set of samples [5]. The time axis for each sample is warped in such a way that the overall correlation with a chosen target sample is maximized [6]. It should be emphasized that the performance of alignment techniques strongly depends on the choice of, for example, the correlation measure and warping parameters. Most warping methods found in the literature utilize the correlation with respect to the chromatographic direction only, based on total ion or base peak chromatograms. With complex samples, comprising several hundreds or even thousands of peaks, the mass spectral direction should be included in order to avoid misalignment. Although such methods have been presented, it could be questioned whether it is feasible to perfectly align the two-dimensional fingerprints obtained from complex biological samples. The correlation optimized warping and related methods assume a rather high degree of similarity (or correlation) between samples. This condition can be hard to fulfill when, for example, different individuals are studied.

The obstacle caused by the time shifts to PCA scoring and sample clustering is their detrimental effect on the correlation or measure of similarity between two two-dimensional fingerprints. In this context, it should be noted that the PCA scores can be obtained merely from a table of correlations between samples, an approach sometimes referred to as ‘kernel PCA’. If the time shifts cannot be fully eliminated, one may look for other ways to reduce their influence. It is well known, for example, that time binning (putting data points within a certain time interval together) often renders less time shift effects. Ultimately, time binning results in a single mass spectrum, summed over the total period of time. Also other time operations, aiming at reducing or eliminating the time shift effects have been reported in the literature. In PARAFAC2 [7] the two-dimensional data table (fingerprint) is represented by the covariance matrix in a way that the time dimension was eliminated, and time binning was combined with time blurring resulting in a less shift sensitive ‘fuzzy’ correlation measure [8]. The crucial point with the non-alignment alternatives is whether or not the time representation is essential for the sample characterization. With high enough mass resolution one would expect unique masses for all ions in the sample, and the time position would be needless for proper sample characterization. In such a case the total spectra should be preferred to the unfolded two-dimensional data tables, where the inevitable time shifts may negatively affect the pair-wise comparisons. With overlapping masses, however, the retention time information is needed to distinguish the response from different ions and time binning is only acceptable to a certain degree. The purpose of additional time blurring with a moving average filter [8] is to allow for peak shifts across the time bins without loosing the contribution to the correlation measure. With the mass covariance data representation, here denoted as the Gram matrix, the simultaneous appearance of different masses along the time dimension is preserved. Hence multiple mass peaks where not all masses overlap will result in separable information in spite of the time elimination. It should be noted that the Gram matrices, like the summed spectra, are independent of the time order; the data points may be randomly reordered with respect to time without affecting the result. As a consequence the time shifts are supposed to have little, if any, influence on the sample correlations or similarities based on the Gram data representation. On the other hand, the partial loss of time information may reduce the possibility to achieve relevant clustering results.

In this paper, different non-alignment approaches for comparison of two-dimensional data tables (fingerprints) by a similarity matrix are reviewed and brought into a common computational framework. They use different representations of the original data, including time binning and blurring, Gram matrices as well as mass spectra. Of special interest is the strategy proposed here that represents the two-way data tables as Gram matrices, which enables a comparison of samples without prior peak alignment and with no parameters to be optimized. The different strategies of data exploration will be illustrated with a real chromatographic data set obtained by CE–MS from urine samples with and without the prior intake of paracetamol [9]. In addition to the exploratory analysis, discriminant partial least squares models are constructed from the similarity matrices. The ability to discriminate between the samples with respect to drug intake reflects how well the data representation preserves the relevant information.

2. Theory

2.1. Towards comparing a set of tables (samples) representing chromatographic runs

A major difficulty in comparing a set of data tables obtained from, e.g. LC–MS, CE–MS, etc., arises due to irreproducible retention times observed for different chromatographic runs. Therefore, a synchronization of the time axes for all samples is usually considered necessary prior to multivariate chemometric data analysis. Different alignment techniques can be employed [10] for this purpose.

Another approach for handling retention time shifts in chromatographic signals without prior peak alignment has been proposed by Danielsson et al. [8]. With this method, a ‘fuzzy’ version of the correlation or covariance matrix based on a ‘blurred’ version of the original data matrix $\mathbf{X}$ is obtained and used in the multivariate data analysis. The ‘blurred’ $\mathbf{X}$, denoted as $\mathbf{X}^*$, is derived by averaging the values of the matrix elements within a moving window of a specified size. The data ‘blurring’ can be done either along the time or m/z direction or in both directions. More details about the method, as well as its applications, can be found in Refs. [8,9,11].

It should be noted that matrix representation of the chromatographic data may result in ‘binning’, i.e. the summation of several data points into the same matrix cell. By the use of wider time bins the effects of time shifts will be reduced, although even small time shifts may affect the distribution between adjacent time bins.