Journal of Chromatography A, 1256 (2012) 150–159

Contents lists available at SciVerse ScienceDirect

Journal of Chromatography A

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

New supervised alignment method as a preprocessing tool for chromatographic data in metabolomic studies

Wiktoria Struck, Paweł Wiczling, Małgorzata Waszczuk-Jankowska, Roman Kaliszan, Michał Jan Markuszewski*

Department of Biopharmaceutics and Pharmacodynamics, Medical University of Gdańsk, Al. Gen. Hallera 107, 80-416 Gdańsk, Poland

ARTICLE INFO

Article history:
Received 19 March 2012
Received in revised form 19 June 2012
Accepted 26 July 2012
Available online 2 August 2012

Keywords:
Metabolomics
HPLC
Retention time shift
Data alignment
Warping
Preprocessing methods

ABSTRACT

The purpose of this work was to develop a new aligning algorithm called supervised alignment and to compare its performance with the correlation optimized warping. The supervised alignment is based on a “supervised” selection of a few common peaks presented on each chromatogram. The selected peaks are aligned based on a difference in the retention time of the selected analytes in the sample and the reference chromatogram. The retention times of the fragments between known peaks are subsequently linearly interpolated. The performance of the proposed algorithm has been tested on a series of simulated and experimental chromatograms. The simulated chromatograms comprised analytes with a systematic or random retention time shifts. The experimental chromatographic (RP-HPLC) data have been obtained during the analysis of nucleosides from 208 urine samples and consists of both the systematic and random displacements. All the data sets have been aligned using the correlation optimized warping and the supervised alignment. The time required to complete the alignment, the overall complexity of both algorithms, and its performance measured by the average correlation coefficients are compared to assess performance of tested methods. In the case of systematic shifts, both methods lead to the successful alignment. However, for random shifts, the correlation optimized warping in comparison to the supervised alignment requires more time (few hours *versus* few minutes) and the quality of the alignment described as correlation coefficient of the newly aligned matrix is worse 0.8593 *versus* 0.9629. For the experimental dataset supervised alignment successfully aligns 208 samples using 10 prior identified peaks. The knowledge about retention times of few analytes' in the data sets is necessary to perform the supervised alignment for both systematic and random shifts. The supervised alignment method is faster, more effective and simpler preprocessing method than the correlation optimized warping method and can be applied to the chromatographic and electrophoretic data sets.

1. Introduction

Now, in the era of evolving bioinformatics methods, there is a clear trend toward the analysis of the entire chromatographic data matrix, rather than the selected peaks detected in the chromatograms. This broad approach does not require choosing the individual analytes for their integration and subsequent analysis, therefore, does not cause data loss. Such a holistic approach overcomes the problem with enormity of the data in the areas like metabolomics, which refers, *inter alia*, to the analysis of metabolic profiles, metabolic fingerprinting, as well as examines the interactions between levels of not necessarily identified metabolites. By analyzing the entire chromatographic data matrix instead of concentrations or peak areas of selected analytes, more relevant information can be extracted about the analyzed sample using appropriate classification and prediction methods. However prior to such chemometric analyses, it is necessary to align retention time shifts that occur either globally or in small sections of the chromatograms. The peaks are shifted because of the unavoidable changes of the experimental conditions caused by the minor changes in the mobile phase composition, stationary phase properties or by the impact of sample matrix (particularly in case of biological sample matrix such as urine or serum). Two types of peak shifts can be distinguished in a real set of chromatograms. In the first one, called systematic, the difference between retention times of the corresponding analytes on the two consecutive chromatograms *versus* the retention time is a continuous function. It is a very common situation and might be a consequence of column ageing, changes in chromatographic conditions, minor changes in the mobile phase composition, *etc.* Contrary, for the random displacement the difference between retention times of corresponding analytes is a random variable, so it affects each peak

* Corresponding author. Tel.: +48 58 349 3260; fax: +48 58 349 3262.
E-mail address: markusz@gumed.edu.pl (M.J. Markuszewski).

0021-9673/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.chroma.2012.07.084

in a different way. It might be a consequence of some unexplained variability, *i.e.* caused by interaction between analytes. Very often both types of displacements are present on a chromatogram.

Numerous methods have been applied to align retention time shifts. They are based on different mathematical transformation, namely correlation optimized warping [1–8], dynamic time warping [4–8], parametric time warping [4,7,9], semi-parametric time warping [4], peak alignment with genetic algorithm [10], local warping [11], automated alignment [12,13], fuzzy warping [14] as well as alignment using differential evolution (DE) [15]. The aim of above mentioned methods is to align all shifted signals so that different chromatograms appear at the same retention time. There are also aligning methods that are suitable for mass spectrometry data [16,17]. Recently a tool for chromatogram alignment has been developed that is freely accessible for the users [18]. Among different warping methods the correlation optimized warping seems to be the most often used. This method was first time reported by Nielsen et al. [19] and concerns the alignment of two chromatographic profiles by piecewise linear stretching and compression among the time axis of one of the profiles. Since 1998 it has been implemented in case of both chromatographic [1–6] and electrophoretic peak shifts [7,8]. The quality of the alignment is calculated as correlation coefficient among the newly aligned chromatograms. Moreover, the alignment is performed without any preliminary information about the peaks' correspondence. However, the main disadvantage is that total procedure is based on choosing two parameters the so called segment length and slack parameter. That is very time consuming and does not guarantee the best alignment for the signals that contain a large set of data points (such as thousands of data points). Moreover, this method does not give any satisfactory results when it is adapted to the random peak's shifts [4,12].

In this paper we would like to propose a new, fast and simple alignment method that can be adapted for both systematic and random data shifts and is not limited to the data matrix size. To achieve the chosen goal, firstly two artificial data sets have been created. First one contains systematic retention time shifts and the second one is composed of random retention time displacements. For both simulated data sets, the correlation optimized warping as well as the newly developed method, called supervised alignment (SA), have been performed and obtained results have been compared. The proposed algorithm has been also applied to the real chromatographic data set from the analysis of nucleosides and the other metabolites derived from urine samples.

2. Theory and implementation

2.1. Correlation optimized warping

The correlation optimized warping (COW) is a well known and popular alignment method that is usually applied to align peak shifts from the chromatographic data. This method is based on a piecewise linear stretching and compression among the time axis relative to the reference chromatogram. COW was previously successfully applied and described by us for alignment of electrophoretic data wherein systematic shifts were observed [7,8]. For the random shifts this method seems not to be the best choice [4,12]. However to prove this we applied correlation optimized warping method for simulated data set of both systematic and random peak shifts. The highest average correlation coefficient among all samples within the matrix served to find the reference chromatogram. Subsequently the peak displacement has been performed by choosing, usually by trials and errors, two parameters, the so called segment length and slack parameter. In short, the segment is a range of points among time axis wherein the alignment is performed. Furthermore, a slack parameter is the number of points about the value that can move the peak (flexibility). In other words, slack parameter is a maximum range of warping in a segment length. An important advantage of the method is that in order to compensate peak shifts there is no information on peak identity required. However, this method is very time-consuming especially when a large data set has to be aligned (over thousand points).

2.2. Supervised alignment

Our goal was to create a method that in a relatively short time will enable the alignment of the shifted peaks based on the known retention times of the few common analytes in all analyzed chromatograms. We called our method as the supervised alignment (SA) because the shift of peaks is based on the shifts of peaks corresponding to the same analyte and which are pointed by the user (supervised by the user). Similarly to the correlation optimized warping, the SA procedure consists of selecting the reference chromatogram. The selection is based on the calculation of the correlation coefficient calculated for each pair of the samples in the matrix. The sample, which has the highest average correlation coefficient among samples matrix, is chosen as the reference chromatogram, T. Here end the similarities with COW. After selecting the target chromatogram, peaks that are common on each chromatogram from the analyzed matrix and reference chromatogram are selected. Therefore, identification of peaks present in all chromatograms is required. This step seems to be the main drawback of our method, however it could be easily automated, which would make this method very fast. It is also not required to confirm the identity of all common peaks existing in the chromatograms. Only some of them occurring at the beginning, middle and end of the time vector are crucial for the whole procedure. Depending on the nature of the data it is necessary to select minimum 2–3 out of the total number of peaks presented in the chromatogram. In the next step the retention times of each peak are determined (*i.e.* the times of maximal absorbance of each selected peak) for the sample and target chromatograms. In the final step, the linear interpolation is used to map the whole range of times from the reference chromatogram to the target chromatogram based on the retention times of the selected peaks as it is illustrated in Fig. 1. We decided to preserve the heights of peaks, as height is a more useful measure of concentration for the subsequent chemometric analysis. The Matlab code is given in Appendix 1.

Undoubtedly, the most time consuming phase of analysis is to properly choose the peaks that exist in each chromatogram, because it requires inspection of each single chromatogram from, in fact, a very large data set. However, after this step the rest of the analysis is performed automatically and lasts from few seconds to few minutes depending on the computer computational abilities. It has also to be underlined that the time required for the identification of those several peaks, is incomparably shorter than the time needed to optimize the key parameters of the correlation optimized warping (segment length and slack parameter). Finally, to check potential of supervised alignment method to align randomly and systematically shifted peaks the method was applied to the simulated data sets. We also decided to examine the number of the identified peaks necessary to correctly align both data sets.

3. Materials and methods

All simulations and calculations have been conducted using Matlab 9.1 (The MathWorks, Inc., USA). We used the COW algorithm that was developed at the Department of Analytical Chemistry and Pharmaceutical Technology, Vrije University, Brussels, Belgium [7,8]. To calculate quality of the alignment two parameters have been compared, namely root mean square error (RMSE) and the