Journal of Chromatography A, 1217 (2010) 1375–1385

Contents lists available at ScienceDirect

Journal of Chromatography A

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

Probability of failure of the watershed algorithm for peak detection in comprehensive two-dimensional chromatography

Gabriel Vivó-Truyols [a,*], Hans-Gerd Janssen [a,b]

[a] Analytical-Chemistry Group, van't Hoff Institute for Molecular Sciences, University of Amsterdam, Nieuwe Achtergracht 166, 1018 WV Amsterdam, The Netherlands
[b] Unilever Research and Development Vlaardingen, Advanced Measurement and Data Modelling, P.O. Box 114, 3130 AC Vlaardingen, The Netherlands

ARTICLE INFO

Article history:
Received 29 September 2009
Received in revised form
15 December 2009
Accepted 22 December 2009
Available online 4 January 2010

Keywords:
Two-dimensional chromatography
Watershed algorithm
Peak detection

ABSTRACT

The watershed algorithm is the most common method used for peak detection and integration in two-dimensional chromatography. However, the retention time variability in the second dimension may render the algorithm to fail. A study calculating the probabilities of failure of the watershed algorithm was performed. The main objective was to calculate the maximum second-dimension retention time variability, $\Delta^2 t_{R,crit}$, above which the algorithm fails. Several models to calculate $\Delta^2 t_{R,crit}$ were developed and evaluated: (a) exact model; (b) simplified model and (c) simple-modified model. Model (c) gave the best performance and allowed to deduce an analytical expression for the probability of failure of the watershed algorithm as a function of experimental $\Delta^2 t_R$, modulation time and peak width in the first and second dimensions. It could be demonstrated that the probability of failure of the watershed algorithm under normal conditions in GC × GC is around 15–20%. Small changes of $\Delta^2 t_R$, modulation time and/or peak width in the first and second dimension could induce subtle changes in the probability of failure of the watershed algorithm. Theoretical equations were verified with experimental results from a diesel sample injected in GC × GC and were found to be in good agreement with the experiments.

1. Introduction

The growing complexity of the data generated by modern liquid chromatography (LC) and gas chromatography (GC) systems requires the development of new data analysis algorithms. The algorithms to be applied depend on the application, and range from base-line treatment to chromatogram alignment methods. In most of the applications, peak detection (and peak integration) is one of the key steps in the analysis process. Peak detection might be troublesome when complex chromatograms are being analysed, with peak numbers easily exceeding the thousands.

In one-dimensional chromatography with single-channel detection, peak-detection methods are almost fully developed. They are based on detecting a raise of the signal coming from the detector and applying the condition of unimodality (i.e., the signal should have only one maximum). Two main families of peak-detection methods have been developed [1]: those that make use of derivatives, and those that make use of matched filters. When a multi-channel detector is used (e.g., MS), new possibilities for peak detection are possible. Different algorithms have been developed in order to make use of the relational information provided by the existence of more than one detection channel. In particular, the advent of the "-omics" disciplines has stimulated the development of a significant quantity of statistical tools, including novel methods for peak detection in chromatography. For a review, see [2–4].

Peak-detection methods in comprehensive two-dimensional chromatography are less advanced. This is mainly due to the fact that these techniques are not completely mature yet. Adapting the peak-detection algorithms developed for hyphenated techniques to two-dimensional chromatography is not straightforward for two reasons. First, in two-dimensional chromatography, the condition of unimodality holds for both dimensions (a chromatographic peak has only a single retention time in both the first and the second dimension). This condition is normally not met in multi-channel detection. Second, a modulation cycle in comprehensive two-dimensional chromatography is normally several orders of magnitude longer than the detector's sampling rate. This makes a chromatogram in two dimensions to appear undersampled in the first dimension as opposed to the highly sampled chromatogram obtained with multi-channel detection. One should note that this second condition does not apply when the two-dimensional separation is performed in space (such as in 2D-PAGE electrophoresis, or two-dimensional thin layer chromatography). Opposed to separations in space, LC × LC or GC × GC are two-dimensional chromatographic methods in time. In these methods, a peak is analysed only a limited number of times by a (fast) second dimension during its elution in the (slow) first dimension, hence the low sampling rate in the first dimension. This article

* Corresponding author. Tel.: +31 205256531.
E-mail address: g.vivotruyols@uva.nl (G. Vivó-Truyols).

0021-9673/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.chroma.2009.12.063

is devoted only to time-driven two-dimensional separations (i.e., GC × GC or LC × LC), so it is not applicable to spatially separated chromatograms (e.g., 2D-PAGE).

So far only a limited number of peak-detection methods for (time-driven) two-dimensional chromatography has been described in the literature [5–7]. Only two main families of methods are available, those based on the watershed algorithm [8], and those based on an extension of the one-dimensional peak-detection algorithms [9,10]. The main difference between the two families of methods relies in the fact that watershed-algorithm based methods make use of the true two-dimensional image generated in two-dimensional chromatography, whereas the extended one-dimensional algorithms are based on the analysis of the one-dimensional raw signal arising from the detector. The watershed algorithm was originally developed to delimitate single catchment areas of geographic zones [11] (see Section 2.2.1 for a detailed explanation), and has been adapted to peak detection in two-dimensional chromatography by Reichenbach et al. [8]. Methods of the second family are normally based on a two-step procedure. In a first step, one-dimensional peak-detection algorithms are applied to the raw, one-dimensional signal. In a second step, the previously detected peaks are then "merged" after it has been decided that they belong to the same modulated compound.

As it will be demonstrated in this paper, one of the main drawbacks of the watershed algorithm is its intolerance to second-dimension retention time variability. This intolerance may bring the algorithm to fail, splitting a two-dimensional peak into two peaks (or two catchment areas), when there is only a single two-dimensional peak. Unfortunately, second-dimension retention time variability is unavoidable, and hence so is failure of the watershed algorithm. In this article, a study is performed to predict in which situations the watershed algorithm will fail. A model for time-driven two-dimensional chromatographic peaks is developed. The model (applicable to both LC × LC and GC × GC) is used to calculate which combination of values for second-dimension retention time variability, first- and second-dimension peak width, modulation time and peak phase are not tolerated. An experimental study is performed in GC × GC to compare data calculated using the model (and its approximations) with experiments.

2. Theory

2.1. Peak model for two-dimensional chromatography

Let us suppose a two-dimensional chromatographic peak with known first- and second-dimension retention times (^1t_R and ^2t_R) and known first- and second-dimension peak widths ($^1\sigma$ and $^2\sigma$). The raw signal from the two-dimensional chromatograph (prior to any manipulation, including "folding" the data into a two-dimensional data table) is represented in Fig. 1. This signal can be modelled as a sum of sub-peaks:

$$y(t) = \sum_{i=-\infty}^{\infty} a_i y_i = \sum_{i=-\infty}^{\infty} a_i \frac{1}{{}^2\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{t-t_i}{{}^2\sigma}\right)^2\right] \quad (1)$$

where the subindex $i = -\infty, \ldots, -2, -1, 0, 1, 2, \ldots, \infty$ corresponds to the sub-peaks resulting from the modulated fractions of the first-dimension peak injected into the second dimension, a_i is the relative abundance of the ith modulated peak (see Eq. (2)), y_i represents the equation for the ith sub-peak, t_i is the retention time where the peak is represented (see Eq. (3)), and $^2\sigma$ is the peak width (measured as the standard deviation) of the sub-peak in the second dimension. Note that this model has two underlying assumptions. First, it assumes a constant value of $^2\sigma$ for the different sub-peaks. Second, it assumes a Gaussian, symmetric peak model. Both assumptions are not strictly true in practice, but these assumptions are not significant for our computations. As the quantity of material injected into the second dimension corresponds to the fraction of the first-dimension peak contained between $t_i - m/2$ and $t_i + m/2$ (i.e., one modulation period), a_i can be defined as:

$$a_i = \frac{A}{{}^1\sigma\sqrt{2\pi}} \int_{t_i-(m/2)}^{t_i+(m/2)} \exp\left[-\frac{1}{2}\left(\frac{t}{{}^1\sigma}\right)^2\right] dt \quad (2)$$

where m is the modulation time and A is a factor expressing the total abundance of the compound. Note that the expression inside the integral corresponds to an unmodulated peak arisen in the first dimension, and the integral limits correspond to the fraction of this peak contained between $t_i - m/2$ and $t_i + m/2$. This is a condition for the two-dimensional chromatography to be comprehensive. In practice, it may be possible to inject in the second dimension only part of the sample eluted from the first. In this case, as long as this split of the first-dimension eluent is constant along the elution, Eq. (2) is still valid (only parameter A has to be corrected). For simplicity, the first-dimension peak is centred around $t = 0$, but in practice

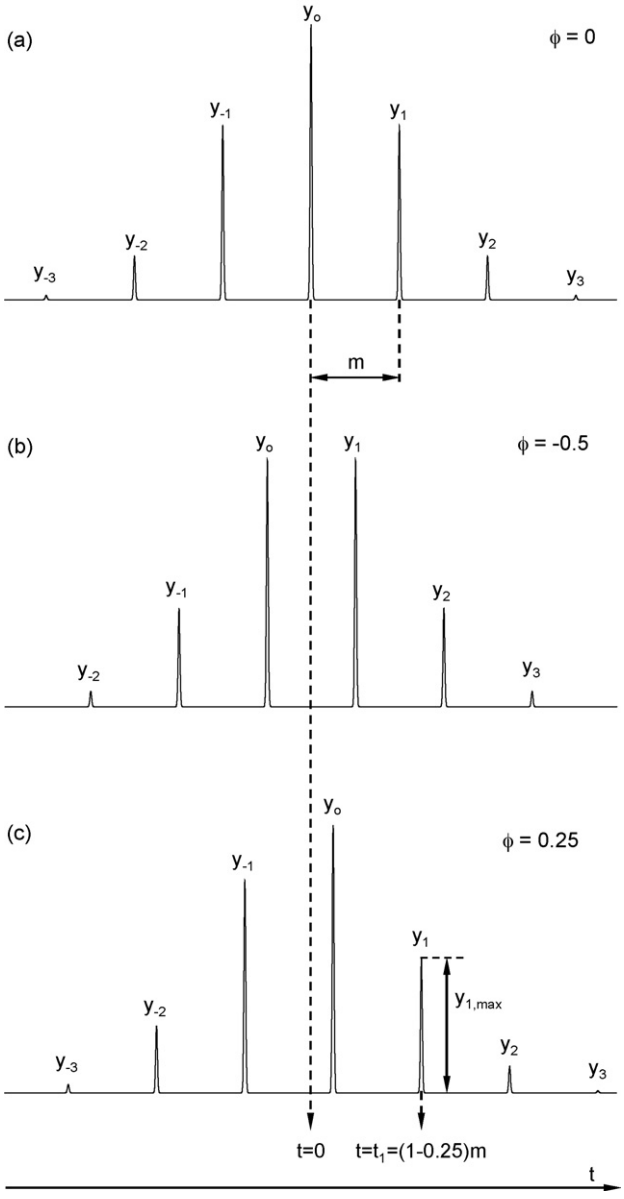

Fig. 1. Schematic representation of a modulated peak following Eqs. (1)–(4), with different values of ϕ ((a) $\phi = 0$; (b) $\phi = -0.5$; (c) $\phi = 0.25$). Only sub-peaks $i = -2$, -1, 0, 1, and 2 are represented. The value of the modulation time (m) is overlaid.