Bioinformatics, 31(18), 2015, 3063–3065
doi: 10.1093/bioinformatics/btv299
Advance Access Publication Date: 13 May 2015
Applications Note

OXFORD

Systems biology

Fast parametric time warping of peak lists

Ron Wehrens[1,*], Tom G. Bloemberg[2,3] and Paul H.C. Eilers[1]

[1]Biometris, Wageningen UR, Wageningen, The Netherlands, [2]Educational Institute for Molecular Sciences and [3]Institute for Molecules and Materials, Radboud University, Nijmegen, The Netherlands

[*]To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on February 25, 2015; revised on April 28, 2015; accepted on May 4, 2015

Abstract

Summary: Alignment of peaks across samples is a difficult but unavoidable step in the data analysis for all analytical techniques containing a separation step like chromatography. Important application examples are the fields of metabolomics and proteomics. Parametric time warping (PTW) has already shown to be very useful in these fields because of the highly restricted form of the warping functions, avoiding overfitting. Here, we describe a new formulation of PTW, working on peak-picked features rather than on complete profiles. Not only does this allow for a much more smooth integration in existing pipelines, it also speeds up the (already among the fastest) algorithm by orders of magnitude. Using two publicly available datasets we show the potential of the new approach. The first set is a LC–DAD dataset of grape samples, and the second an LC–MS dataset of apple extracts.

Availability and implementation: Parametric time warping of peak lists is implemented in the ptw package, version 1.9.1 and onwards, available from Github (https://github.com/rwehrens/ptw) and CRAN (http://cran.r-project.org). The package also contains a vignette, providing more theoretical details and scripts to reproduce the results below.

Contact: ron.wehrens@wur.nl

1 Introduction

Hyphenated techniques, coupling a chromatographic separation step to a detector, are the norm in the analysis of complex biological samples, e.g. in proteomics and metabolomics. A common problem in all applications is that the separation, i.e. the exact time point at which a compound is measured by the detector, is not quite reproducible, hampering comparative analyses. When all measurements have been done in a single batch and peak shifts are expected to be small this is usually not a big problem, but when experimental difficulties arise during the batch, or comparisons with data from different batches or even different machines or labs need to be made, explicit retention time corrections are necessary.

Several procedures have been proposed in the literature, often indicated with the label 'Time Warping' (Bloemberg *et al.*, 2013). These transform ('warp') the time axis of a sample in such a way that the overlap with a reference signal is maximized. Parametric Time Warping (PTW, Eilers, 2004), for example, uses a polynomial transformation. This choice has several advantages: first, the allowed transformations are quite restricted, thus avoiding the common problem of overfitting (at least to some extent). Second, the use of such a polynomial warping function often describes processes like column aging quite well. Third, the explicit modeling allows for clever use of resources, like defining the warping using regularly injected quality control samples only, and interpolating to find the warping for the real experimental samples (Eilers, 2004).

Here, we describe a further development in the PTW algorithm: warping is no longer based on complete profiles, such as extracted ion chromatograms in liquid chromatography mass-spectrometry (LC–MS) applications, but rather on the peak-picked features. This has several important benefits. First of all, in the new approach the optimal warping is obtained by only considering the relevant, information-containing part of the signal. This increases sensitivity, and decreases the need for extensive preprocessing of the data (e.g. baseline correction). Second, the process is made faster, sometimes by orders of magnitude. Also the increase in calculation time with wider search windows for the weighted crosscorrelation

© The Author 2015. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com

Downloaded from https://academic.oup.com/bioinformatics/article/31/18/3063/240579 by guest on 21 August 2022

(WCC) optimization criterion (Bloemberg *et al.*, 2010) is virtually eliminated.

2 Data

We show the possibilities of **ptw** using two publicly available datasets. The first is on carotenoids in grape samples (Wehrens *et al.*, 2015), measured using diode-array detection coupled to liquid chromatography (LC–DAD). Multivariate curve resolution (MCR, de Juan and Tauler, 2006) was used to finally obtain peak lists, clustered in 14 groups according to spectral characteristics. The mean number of peaks per group varied from 4.4 to 14.6. Although these samples were analysed in a single batch, retention time differences are appreciable, due to the volatile nature of the solvent and the variable temperature conditions in the lab.

The second dataset consists of LC–MS measurements of 156 apple extracts, divided over seven different apple varieties. In this set, variation in retention times is appreciable, due to a leaking column which was only detected after the completion of all injections. A pooled sample was injected regularly as a quality control (QC). In total, 27 QC injections were measured. Since the apple varieties might have different metabolic compositions, leading to different peaks and peak intensities, it was decided to base the alignment on the warping functions for the QCs only, taking the first QC sample as the reference sample and calculating warping functions for the other 26 QCs. The alignment for the individual apple samples was then obtained by interpolation. This strategy is applicable in all cases where samples with potentially different compositions are compared (e.g. treatment–control experiments). In contrast to the grape data, the raw data for the apple LC–MS samples do not consist of continuous profiles over time, but rather of a sequence of mass spectra. It is possible to convert these into equispaced *m/z* bins, leading to a profile matrix in whose rows a profile-based alignment can be performed. A selection of bins containing relevant signals is usually advised (Bloemberg *et al.*, 2010). In virtually all cases, the data are further processed by software like XCMS (Smith *et al.*, 2006) to obtain peak lists for statistical analysis or metabolic network construction. Hence, doing retention time correction on the peaks does not constitute extra work.

Both datasets are publicly available from the Metabolights repository (http://www.ebi.ac.uk/metabolights) with identifiers MTBLS85 and MTBLS99, respectively.

3 Results

In all cases in this paper, quadratic warping functions are chosen. The grape DAD data are sufficiently small to be able to zoom in and assess the individual warping functions, yet big enough to present a challenge for the algorithm and to show the type of speed improvements that are possible. Indeed, warping the peaks rather than the profiles led to a decrease in computation time of almost an order of magnitude, whereas the results were identical for all practical purposes. Not only do the warping functions in both scenarios look very similar, also warping the peaks with the warping functions obtained from the continuous profiles leads to virtually the same values for the WCC optimization criterion: 0.164 against 0.170 (on a scale of 0–1).

Aligning the QC samples of the apple data using the new, peak-based PTW algorithm, leads to 26 warping functions, one for every non-reference QC. Interpolating the warping functions for the true apple samples on the basis of injection sequence leads to corrected peak positions for all peaks. Using the same warpings for the TICs, one can visualize the result more easily; this is shown in Figure 1. The top panel shows the TICs of the uncorrected apple samples, arranged per variety in order of injection (bottom to top). The effect of the leaking column is clearly visible: retention times tend to get much larger later in the injection sequence. In contrast, the corrected TICs shown in the bottom panel only show minor retention time variations, that are largely random in nature. These can typically be handled by the alignment methods available in the data processing software—the bigger shifts caused by, in this case, faulty equipment, usually cannot. In cases where a quadratic warping as used here is not sufficient to align the features, one can try higher-order warping functions.

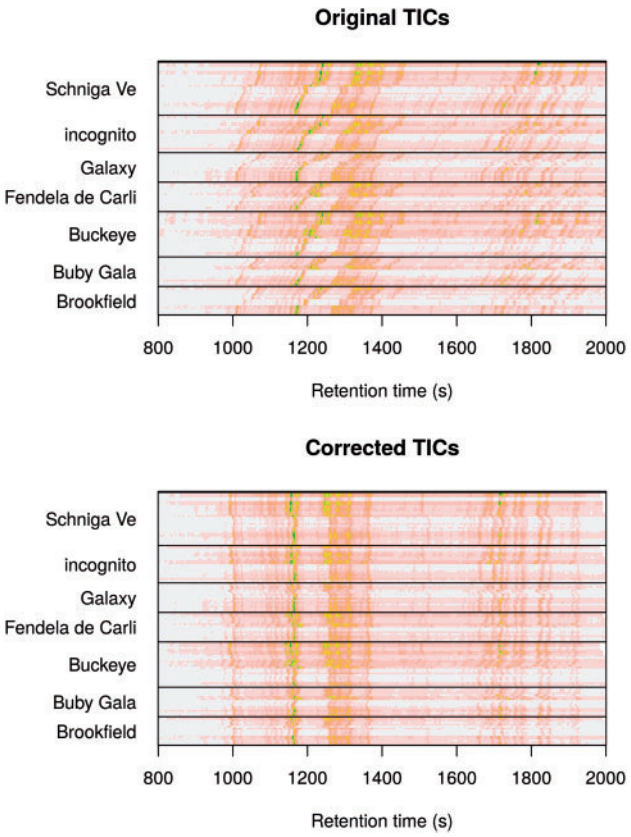

Fig. 1. Total ion chromatograms (TICs) of the apple samples, plotted class-wise in the original injection order (bottom to top). Only the part between 800 and 2000 seconds is shown

The strategy of obtaining warping coefficients using pooled QC samples also provides a structural solution to a common problem in metabolomics: when comparing two very different groups of samples, the number of common features may be too small, or too much concentrated in one part of the chromatograms, to obtain useable alignments (Bloemberg *et al.*, 2013). With technical replicates like the pooled QC samples used here, this problem is eliminated.

4 Conclusion

Time warping using peak lists, as described here, brings the alignment of large numbers of complex samples into the realm of the practically feasible. Instrument breakdown, for example, no longer necessarily forces samples to be remeasured, which is especially important when only small quantities of samples are available.

Downloaded from https://academic.oup.com/bioinformatics/article/31/18/3063/240579 by guest on 21 August 2022

Databases of standards, used for annotation and typically containing only information on peaks and not on raw data, can be used for alignment as well, thereby considerably increasing their usefulness.

Conflict of Interest: none declared.

Acknowledgement

Panagiotis Arapitsas is acknowledged for the apple data.

References

Bloemberg,T. *et al*. (2010) Improved parametric time warping for proteomics. *Chemom. Intell. Lab. Syst.*, **104**, 65–74.

Bloemberg,T. *et al*. (2013) Warping methods for spectroscopic and spectrometric signal alignment: a tutorial. *Anal. Chim. Acta*, **781**, 14–32.

de Juan,A. and Tauler,R. (2006) MCR from 2000: progress in concepts and applications. *Crit. Rev. Anal. Chem.*, **36**, 163–176.

Eilers,P. (2004) Parametric time warping. *Anal. Chem.*, **76**, 404–411.

Smith,C.A. *et al*. (2006) XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.*, **78**, 779–787.

Wehrens,R. *et al*. (2015) Metabolite profiling in LC–DAD using multivariate curve resolution: the alsace package for R. *Metabolomics*, **11**, 143–154.

Downloaded from https://academic.oup.com/bioinformatics/article/31/18/3063/240579 by guest on 21 August 2022