GCalignR: An R package for aligning Gas-Chromatography data

Meinolf Ottensmann, Martin A. Stoffel, Barbara Caspers, Joseph I. Hoffman

2017-01-09

1 Abstract

Key-words: GC-MS,gas-chromatography, chemical communication, olfactory communication, alignment

1.1 Introduction

Chemical cues are arguably the most common mode of communication in animals (Wyatt 2014), where the role is evident in kin (Porter and Moore 1981; Krause et al. 2012; Gilad et al. 2016) and mate recognition (Martin, Helanterä, and Drijfhout 2008), mate-choice (Penn 2002) and signalling of genetic quality (Charpentier et al. 2010; Stoffel et al. 2015). The investigation of chemical signatures requires the use of metabolomic approaches in order to characterise and compare the involved chemicals. Gas-chromatography is a widespread analytical method to unravel the compostition of samples with high efficiency (McNair and Miller 2011). For the detection of broader patterns in chemical samples researchers use an untargeted approach and analyse the whole spectrum of sampled chemicals rather than targeting specific compounds. However, chromatography data across multiple samples are not directly comparable as they need to be aligned first. Furthermore, the retention times of peaks vary across samples due to subtle, random and often unavoidable variation of the chromatograph machine parameters (Pierce et al. 2005 and references within). For studies that seek to identify chemical patterns across samples it becomes essential to account for these retention time drifts by using an appropriate alignment method. A number of automated tools is available for two-dimensional chromatograms (LC-MS, GC-MS) which offer mass spectra in addition to retention times (Pierce et al. 2005; C. A. Smith et al. 2006; Robinson et al. 2007; Luedemann et al. 2008; Koh et al. 2010; Jiang et al. 2013; Zhang et al. 2012; Niu et al. 2014; R. Smith, Ventura, and Prince 2015), while there is to our knowledge only one application proposed for the needs of one-dimensional chromatograms (Dellicour and Lecocq 2013). As a consequence, many researchers rely on a manual alignment of their data (Charpentier et al. 2010; Caspers et al. 2011; Harris, Davies, and Nicol 2012); (Citations are critical. Because nothing is mentioned on alignment within papers, cf. http://bit.ly/2gYJAZw). This approach bears three main drawbacks: (1) In large scale studies the task becomes a difficult and time consuming task. (2) Humans are prone to detect patterns in noise which is why the researcher may bias the alignment due to subjective experience and expectations. (3) The data analytic pipeline from the raw gas-chromatography data to the results of the statistical analysis is not reproducible. Here, we introduce `GCalignR`, a package developed in the language `R` (R Core Team 2016), which provides a simple and fast algorithm to align peaks from GC data and evaluate the alignment of empirical data.`GCalignR` was specifically developed and tested as a preprocessing tool prior to the statistical analysis of chemical samples from animal skin and preen glands (see Stoffel et al. (2015) for an application of the underlying algorithm). The main focus of the package is set on the alignment of homologous retention times and the inspection of the resulting data. The algorithm consists of two main steps: (1) Systematic shifts of chromatograms are corrected by applying appropriate linear shifts to whole chromatograms based on a single reference. (2) Retention times of individual peaks are step-wise grouped together with homologous peaks of other samples within one row. The outcome of this grouping procedures can be altered by specifying three parameters that are described below. Among several optional processing steps, the package allows to remove peaks belonging to contaminations, which are identified due to there presence in control samples. Furthermore we demonstrate the easy integration of the R-packge `vegan`(Oksanen et al. 2016) into a solid workflow for multivariate analyses of chemical data,

which can be fully integrated into in `Rmarkdown` documents (Allaire et al. 2016) to fullfil good standards of reproducibility (Peng 2011).

1.2 Material and Methods

1.2.0.1 maybe a flowdiagram (package DiagramR) to illustrate the complete workflow

1. GC / GC-MS analysis
2. Peak detection software
3. GCalignR workflow
4. statistical analysis

1.3 Input data

Peaks have to be extracted from raw chromatogram files using proprietary or freew software prior to alignment with GCalignR. The standard working format of `GCalignR`is text file containing peak retention times and a arbitrary number of further variables (see Supporting information A). Additionally the use of `lists` in R is supported. `GCalignR` aligns peaks via their retention times (and not their mass-spectra, which may not be available, e.g. when using gas-chromatography coupled to a flame ionization detector (FID)) to align the peaks across individuals for subsequent chemometric analysis and pattern detection .The simple assumption is that peaks with similar retention times represent the same substances. However, it is highly recommended to verify this assumption by comparing also the mass-spectra (if available) of the substances of interest. The final data is returned either in form of a list within R or saved as text files (.txt)

2 Example dataset

2.0.0.1 explanation of example dataset

3 GCalignR workflow

- GCalignR steps: Checking the input, aligning chromatograms, evaluating alignment
- adjust parameters, align again, evaluate again (if first alignment wasn´t satisfactory)

4 Input

- Quickly describe input formats
- Check input and what it checks

```
check_input(data = peak_data,show_peaks = T, col= "red") # If show_peaks = T, a histogram of peaks is p
```

5 Aligning peaks

- describe main features of the main function

```
peak_data_aligned <- align_chromatograms(data = gc_peak_data, # input data
    conc_col_name = "area", # peak abundance variable
    rt_col_name = "time", # retention time
```

```
    rt_cutoff_low = 5, # cut peaks with retention times below 5 Minutes
    rt_cutoff_high = 45, # cut peaks with retention times above 45 Minutes
    reference = "M3", # name of reference
    max_linear_shift = 0.05, # maximum linear shift of chromatograms
    max_diff_peak2mean = 0.03, # maximum distance of a peak to the mean
    min_diff_peak2peak = 0.03, # maximum distance between the mean of two peaks
    blanks = NULL, # no blanks. Specify blanks by names (e.g. c("blank1", "blank2"))
    delete_single_peak = TRUE, # delete peaks that are present in just one sample
    write_output = NULL) # add c("time","area") to write data frames to .txt file

data("aligned_peak_data")
```

6 Evaluating the quality of the alignment

```
library(ggplot2)
library(gridExtra)

gc_heatmap(aligned_peak_data,threshold = 0.01, samples_subset = 1:20, substance_subset = 1:30, label_si
```

7 Algorithm

8 Evaluation with empirical data and simulations

8.1 Availability

The latest version of `GCalignR` can be downloaded from GitHub.

```
install.packages("devtools")
devtools::install_github("mastoffel/GCalignR")
```

We welcome any contributions or feedback on the package.

8.2 Data accessibility

References

Allaire, J. J., Joe Cheng, Yihui Xie, Jonathan McPherson, Winston Chang, Jeff Allen, Hadley Wickham, Aron Atkins, and Rob Hyndman. 2016. "Rmarkdown: Dynamic Documents for R." https://CRAN.R-project.org/package=rmarkdown.

Caspers, Barbara A., Frank C. Schroeder, Stephan Franke, and Christian C. Voigt. 2011. "Scents of Adolescence: The Maturation of the Olfactory Phenotype in a Free-Ranging Mammal." *PloS One* 6 (6): e21162.

Charpentier, Marie J.E., Jeremy Chase Crawford, Marylène Boulet, and Christine M. Drea. 2010. "Message 'Scent': Lemurs Detect the Genetic Relatedness and Quality of Conspecifics via Olfactory Cues." *Animal Behaviour* 80 (1): 101–8. doi:10.1016/j.anbehav.2010.04.005.

Dellicour, Simon, and Thomas Lecocq. 2013. "GCALIGNER 1.0: An Alignment Program to Compute a Multiple Sample Comparison Data Matrix from Large Eco-Chemical Datasets Obtained by Gc." *Journal of*

Separation Science 36 (19): 3206–9. doi:10.1002/jssc.201300388.

Gilad, Oranit, Ronald R. Swaisgood, Megan A. Owen, and Xiaoping Zhou. 2016. "Giant Pandas Use Odor Cues to Discriminate Kin from Nonkin." *Current Zoology* 62 (4): 333–36. doi:10.1093/cz/zow025.

Harris, Rachel L., Noel W. Davies, and Stewart C. Nicol. 2012. "Chemical Composition of Odorous Secretions in the Tasmanian Short-Beaked Echidna (Tachyglossus Aculeatus Setosus)." *Chemical Senses* 37 (9): 819–36. doi:10.1093/chemse/bjs066.

Jiang, Wei, Zhi-Min Zhang, YongHuan Yun, De-Jian Zhan, Yi-Bao Zheng, Yi-Zeng Liang, Zhen Yu Yang, and Ling Yu. 2013. "Comparisons of Five Algorithms for Chromatogram Alignment." *Chromatographia* 76 (17-18): 1067–78. doi:10.1007/s10337-013-2513-8.

Koh, Yueting, Kishore Kumar Pasikanti, Chun Wei Yap, and Eric Chun Yong Chan. 2010. "Comparative Evaluation of Software for Retention Time Alignment of Gas Chromatography/Time-of-Flight Mass Spectrometry-Based Metabonomic Data." *Journal of Chromatography. A* 1217 (52): 8308–16. doi:10.1016/j.chroma.2010.10.101.

Krause, E. Tobias, Oliver Krüger, Philip Kohlmeier, and Barbara A. Caspers. 2012. "Olfactory Kin Recognition in a Songbird." *Biology Letters* 8 (3): 327–29.

Luedemann, Alexander, Katrin Strassburg, Alexander Erban, and Joachim Kopka. 2008. "TagFinder for the Quantitative Analysis of Gas Chromatography–mass Spectrometry (Gc-Ms)-Based Metabolite Profiling Experiments." *Bioinformatics (Oxford, England)* 24 (5): 732–37. doi:10.1093/bioinformatics/btn023.

Martin, Stephen J., Heikki Helanterä, and Falko P. Drijfhout. 2008. "Evolution of Species-Specific Cuticular Hydrocarbon Patterns in Formica Ants." *Biological Journal of the Linnean Society* 95 (1): 131–40. doi:10.1111/j.1095-8312.2008.01038.x.

McNair, Harold M., and James M. Miller. 2011. *Basic Gas Chromatography*. John Wiley & Sons.

Niu, Weihuan, Elisa Knight, Qingyou Xia, and Brian D. McGarvey. 2014. "Comparative Evaluation of Eight Software Programs for Alignment of Gas Chromatography-Mass Spectrometry Chromatograms in Metabolomics Experiments." *Journal of Chromatography. A* 1374: 199–206. doi:10.1016/j.chroma.2014.11.005.

Oksanen, Jari, F. Guillaume Blanchet, Michael Friendly, Roeland Kindt, Pierre Legendre, Dan McGlinn, Peter R. Minchin, et al. 2016. "Vegan: Community Ecology Package." https://CRAN.R-project.org/package=vegan.

Peng, Roger D. 2011. "Reproducible Research in Computational Science." *Science (New York, N.Y.)* 334 (6060): 1226–7. doi:10.1126/science.1213847.

Penn, Dustin J. 2002. "The Scent of Genetic Compatibility: Sexual Selection and the Major Histocompatibility Complex." *Ethology* 108 (1): 1–21. doi:10.1046/j.1439-0310.2002.00768.x.

Pierce, Karisa M., Janiece L. Hope, Kevin J. Johnson, Bob W. Wright, and Robert E. Synovec. 2005. "Classification of Gasoline Data Obtained by Gas Chromatography Using a Piecewise Alignment Algorithm Combined with Feature Selection and Principal Component Analysis." *Journal of Chromatography A* 1096 (1): 101–10.

Porter, R., and J. Moore. 1981. "Human Kin Recognition by Olfactory Cues." *Physiology & Behavior* 27 (3): 493–95. doi:10.1016/0031-9384(81)90337-1.

R Core Team. 2016. "R: A Language and Environment for Statistical Computing." Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Robinson, Mark D., David P. de Souza, Woon W. Keen, Eleanor C. Saunders, Malcolm J. McConville, Terence P. Speed, and Vladimir A. Likić. 2007. "A Dynamic Programming Approach for the Alignment of Signal Peaks in Multiple Gas Chromatography-Mass Spectrometry Experiments." *BMC Bioinformatics* 8 (1): 419.

Smith, Colin A., Elizabeth J. Want, Grace O'Maille, Ruben Abagyan, and Gary Siuzdak. 2006. "XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching,

and Identification." *Analytical Chemistry* 78 (3): 779–87.

Smith, Rob, Dan Ventura, and John T. Prince. 2015. "LC-Ms Alignment in Theory and Practice: A Comprehensive Algorithmic Review." *Briefings in Bioinformatics* 16 (1): 104–17. doi:10.1093/bib/bbt080.

Stoffel, Martin A., Barbara A. Caspers, Jaume Forcada, Athina Giannakara, Markus Baier, Luke Eberhart-Phillips, Caroline Müller, and Joseph I. Hoffman. 2015. "Chemical Fingerprints Encode Mother–offspring Similarity, Colony Membership, Relatedness, and Genetic Quality in Fur Seals." *Proceedings of the National Academy of Sciences* 112 (36): E5005–E5012.

Wyatt, Tristram D. 2014. *Pheromones and Animal Behavior: Chemical Signals and Signatures*. Cambridge University Press.

Zhang, Zhi-Min, Yi-Zeng Liang, Hong-Mei Lu, Bin-Bin Tan, Xiao-Na Xu, and Miguel Ferro. 2012. "Multiscale Peak Alignment for Chromatographic Datasets." *Journal of Chromatography. A* 1223: 93–106. doi:10.1016/j.chroma.2011.12.047.