GCalignR: An R package for aligning Gas-Chromatography data

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

2016-11-25

```
## Skipping install of 'knitcitations' from a github remote, the SHA1 (8bc14b57) has not changed since
##   Use `force = TRUE` to force installation
```

1 Abstract

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

1.1 Introduction

Importance in Zoology Increasing use of R Reproducibility and Bias Metabolomics approaches such as Gas-Chromatography (GC) and Gas Chromatography-Mass Spectrometry (GC-MS) are increasingly used by biologists to unravel the chemical basis of animal behaviour, such as olfactory communication or navigation (citations). For the detection of broader patterns in chemical samples most 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 the retention times of peaks vary across samples due to subtle, random and often unavoidable variation of the GC-MS machine parameters (Pierce et al. 2005). 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.

Despite the existence of automated alignment algorithms (e.g. Smith et al. 2006; Stein 1999; Robinson et al. 2007) (also gcaligner here), most researchers in the relatively young fields of mammalian and avian chemical communication align chromatograms manually (Charpentier, Boulet, and Drea 2008; Setchell et al. 2010; Caspers et al. 2011; Leclaire et al. 2012; Theis et al. 2013) or identify all compounds prior to analysis (Whittaker et al. 2010) rather than using (semi-)automated alignment software. This approach yields three main drawbacks: (1) For larger sample sizes, the method is time intensive and can take up to several weeks. (2) Humans are prone to detect patterns in noise (some citation) which is why the researcher may bias the alignment due to subjective experience and expectations. (3) The data analytic pipeline from the raw GC data to the results of the statistical analysis is not reproducible.

Here, we introduce `GCalignR`, a package developed in R, which provides a simple means of aligning peaks from Gas Chromatography data and evaluate the quality of the alignment. `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 alignment functions can easily be embedded in reproducible research tools such as `Rmarkdown` (citation?) and piped as input into further statistical packages such as `vegan`.

2 The package

2.0.0.1 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

2.1 Data preprocessing

The statistical analysis of GC or GC-MS data is usually based on the detection of signal peaks within the chromatograms, which is can be done by proprietary software or free programs such as AMDIS (Stein 1999). The peak data of a chromatogramm usually contain the retention time of a given peak plus additional information such as the area under the peak or its height which are used in the subsequent analysis. `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.

3 Example dataset

3.0.0.1 explanation of example dataset

4 GCalignR workflow

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

5 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
```

6 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")
```

7 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_siz
```

8 Algorithm

9 Evaluation with empirical data and simulations

9.1 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.

9.2 Data accessibility

References

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). Public Library of Science: e21162.

Charpentier, Marie JE, MarylENe Boulet, and Christine M Drea. 2008. “Smelling Right: The Scent of Male Lemurs Advertises Genetic Quality and Relatedness.” *Molecular Ecology* 17 (14). Wiley Online Library: 3225–33.

Leclaire, Sarah, Thomas Merkling, Christine Raynaud, Hervé Mulard, Jean-Marie Bessière, Émeline Lhuillier, Scott A Hatch, and Étienne Danchin. 2012. “Semiochemical Compounds of Preen Secretion Reflect Genetic Make-up in a Seabird Species.” *Proceedings of the Royal Society of London B: Biological Sciences* 279 (1731). The Royal Society: 1185–93.

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). Elsevier: 101–10.

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). BioMed Central Ltd: 419.

Setchell, Joanna M, Stefano Vaglio, Kristin M Abbott, Jacopo Moggi-Cecchi, Francesca Boscaro, Giuseppe Pieraccini, and Leslie A Knapp. 2010. “Odour Signals Major Histocompatibility Complex Genotype in an Old World Monkey.” *Proceedings of the Royal Society of London B: Biological Sciences*. The Royal Society, rspb20100571.

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). ACS Publications: 779–87.

Stein, Stephen E. 1999. "An Integrated Method for Spectrum Extraction and Compound Identification from Gas Chromatography/Mass Spectrometry Data." *Journal of the American Society for Mass Spectrometry* 10 (8). Elsevier: 770–81.

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). National Acad Sciences: E5005–E5012.

Theis, Kevin R, Arvind Venkataraman, Jacquelyn A Dycus, Keith D Koonter, Emily N Schmitt-Matzen, Aaron P Wagner, Kay E Holekamp, and Thomas M Schmidt. 2013. "Symbiotic Bacteria Appear to Mediate Hyena Social Odors." *Proceedings of the National Academy of Sciences* 110 (49). National Acad Sciences: 19832–7.

Whittaker, Danielle J, Helena A Soini, Jonathan W Atwell, Craig Hollars, Milos V Novotny, and Ellen D Ketterson. 2010. "Songbird Chemosignals: Volatile Compounds in Preen Gland Secretions Vary Among Individuals, Sexes, and Populations." *Behavioral Ecology* 21 (3). ISBE: 608–14.