GCalignR. An R package for aligning Gas-Chromatography data

by Meinolf Ottensmann, Martin A. Stoffel, Joseph I. Hoffman

Abstract This is just a placeholder for 150 words in the abstract This is just a placeholder for 150 words in the abstract This is just a placeholder for 150 words in the abstract This is just a placeholder for 150 words in the abstract This is just a placeholder for 150 words in the abstract This is just a placeholder for 150 words in the abstract This is just a placeholder for 150 words in the abstract This is just a placeholder for 150 words in the abstract This is just a placeholder for 150 words in the abstract This is just a placeholder for 150 words in the abstract This is just a placeholder for 150 words in the abstract This is just a placeholder for 150 words in the abstract This is just a placeholder for 150 words in the abstract This is just a placeholder for 150 words in the abstract

Introduction

Chemical cues are arguably the most common mode of communication among animals (**?**). By exploring broad patterns in complex chemical signatures, researchers make inferences on kinship (**??**), genetic diversity (**??**), sexual maturation (**?**) or species discrimination (**?**). One of the most common instruments to quantify the chemical composition of samples is gas-chromatography, a fast high-throughput method to detect individual chemicals and their abundances (**?**), while the additional implementation of mass-spectrometry (GC-MS) allows to identify specific substances (**?**).

However, before similarity patterns across can be analysed, it is essential to align compounds among samples, thereby accounting for drifts in the retention times of peaks caused by subtle, random and often unavoidable variations of the chromatography machine parameters (**?**). Many studies rely on manual alignment rather than (semi-)automated algorithms (citations necessary), but this approach bears three severe drawbacks: (1) In large scale studies this task becomes increasingly time consuming task and is impracticable. (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. (citations for the first two points necessary) Several alignment algorithms have been proposed to overcome these issues, but these focus nearly exclusively on GC-MS data (**???**) and only some a easily accessible as web-based tools (**??**) or independent software (**?**).

Here, we introduce **GCalignR**, a package that implements a simple and fast algorithm to align peaks from GC data and evaluate the resulting alignment using two data sets. **GCalignR** was specifically developed as a tool for pre-processing GC data from animal skin and preen glands prior to subsequent statistical analysis. In brief, 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 grouped iteratively together with homologous peaks of other samples and aligned within the same row in a retention time matrix . The quality of this grouping procedure can be adjusted to specific datasets through three parameters that are described in detail below. Among several optional processing steps, the package allows to remove peaks belonging to contaminations, which are identified due to their presence in control samples. For an easy interpretation of the quality of an alignment we implemented several ways to plot the outcome (You can change this to something more specific). Furthermore, we demonstrate a complete workflow from chemical raw data to multivariate analyses with the popular and widely used **vegan** (**?**) package. This allows the integration of the full analysis into **RMarkdown** documents (**?**) in order to meet the standards of reproducibility (**?**).

The Package

GCalignR consists of functions that allow the alignment of peaks from GC and GC-MS data based on retention times. The main aim of the package is to provide a simple tool that guides the user through the unbiased alignment of large data sets prior to hypothesis-testing of the multivariate data (**?**). We summarise the underlying algorithm and workflow (figure **??**) below and refer to the vignette that can be assessed via `browseVignettes('GCalignR')`.

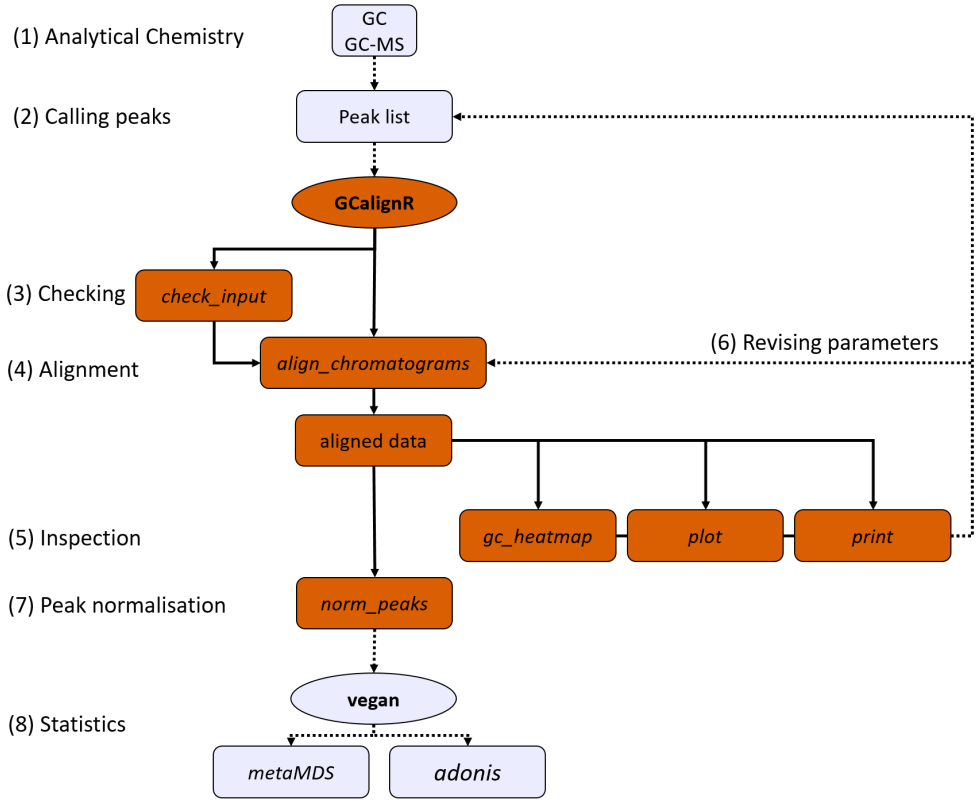

Figure 1: GCalignR workflow. In addition to the alignment of substances across samples, the package provides functions for checking and inspecting the data. The aligned data is ready to use for analyses in conjunction with other packages. Each function is explained within the text.

Example dataset

For demonstration purposes **GCalignR** includes data of chemical signatures that were obtained by sampling the skin of 82 Antarctic fur seals *Arctocephalus gazella*. It was previously shown that these signatures encode the membership to a breeding colony **?**. These data are available as a *list* with individual samples included as a *data.frame*. Two variables are available that represent the required retention time ("time") and concentration or peak abundance ("area") within a sample.

```
library(GCalignR)
# Seal scent data
data("peak_data")
# Data is organized in one list of data frames
str(peak_data[1:2])
```

```
#> List of 2
#>  $ C3:'data.frame':  217 obs. of  2 variables:
#>   ..$ time: num [1:217] 4.53 4.55 4.62 4.68 4.71 4.79 4.83 4.87 5.01 5.14 ...
#>   ..$ area: num [1:217] 3331224 1462381 4834211 7754401 1267617 ...
#>  $ C2:'data.frame':  217 obs. of  2 variables:
#>   ..$ time: num [1:217] 4.52 4.55 4.57 4.67 4.69 4.73 4.75 4.8 4.83 4.85 ...
#>   ..$ area: num [1:217] 2695110 5926253 10406833 6805905 1672849 ...
```

The package provides the function **check_input** to test the input file for typical formatting errors and incomplete data. We encourage to use unique names for samples that consist only of letters, numbers and underscores. If the data fails the test, indicative warnings are returned which guide in correcting the errors. Prior to the start of any alignment this function is used internally.

```
check_input(peak_data)
```

```
#> All checks passed!
#> Ready for processing with align_chromatograms
```

Aligning substances among samples

The alignment procedure is divided into five steps (figure **??**). All steps are executed by the main function *align_chromatograms* and will be explained in in the next sections.

Linear adjustments of chromatograms

At first, chromatograms are linearly shifted with respect to a reference sample to account for systematic shifts in retention times among homologous chemicals shared by samples. Therefore, same small linear adjustments are applied to the entire set of peaks in a chromatogram (figure **??** A), such that the number of peaks that are shared between the sample and the reference at a given threshold (i.e. 0.6 seconds) is maximised.The parameter *max_linear_shift* defines the maximum temporal range of linear shifts that are considered by the program.
Note: This method relies on the occurrence of substances that are shared among most substances to produce efficient adjustments. If those are absent, it is unlikely to find a suitable shift and chromatograms remain untransformed.

A reference may be selected automatically by searching for the sample with the highest average similarity to all other samples based on the number of shared peaks prior to alignment. Alternatively, a chromatogram may be included that contains peaks of an internal standard which peaks are *a-priori* known to occur in all samples. In this case, the sample is named `"reference"` and will be removed after the alignment was conducted.

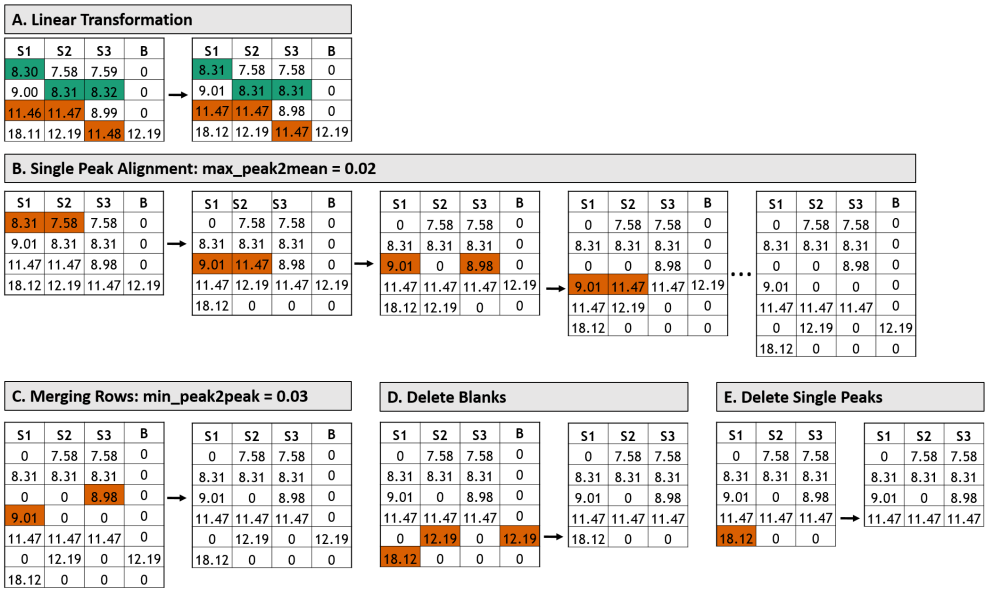

Figure 2: Overview of the algorithm performed by GCalignR. Rows of matrices correspond to substances, columns are samples. Zeros indicate absence of peaks and are ignored in calculations. **A**. Chromatograms are linearly shifted with respect to a reference (S2). StrongB. From left to right the first four steps from the input matrix to the final alignment are shown. Peaks are aligned row by row. Initially, always the second sample is compared to the first. Then the next sample is compared to all samples in previous columns until the last column is reached. Coloured cells represent conflicting retention times using $\texttt{max}_peak2mean = 0.02. C. After all peaks have been aligned, rows are merged depending on min_peak2peak, which defines the minimum difference that is expec$

Peak alignment

The core of the alignment procedure is based on clustering of individual peaks among samples. This is performed by examining retention times within single rows, where samples are compared consecutively with all previous samples starting with the second column (figure **??** B):

$$rt_m > \left(\frac{\sum_{i=1}^{m-1} rt_i}{m-1} \right) + max_peak2mean \quad (1)$$

If the examined peak is moved into the next row, whereas all previous samples are moved

$$rt_m < \left(\frac{\sum_{i=1}^{m-1} rt_i}{m-1} \right) - max_peak2mean \quad (2)$$

with rt = retention time; m = current column and $max_peak2mean defining the maximal deviation of the mean retention time. By considering the mean retention time among all previous samples the algorithm accounts for substance specific variations, such that less$

Merging

*Afterwards, rows with similar mean retention times are assessed for redundancy (figure **??** C), which applies whenever a merging does not cause any loss of any information (i.e. no sample exists that contains substances in both rows). The similarity threshold is given by min_peak2peak defining the minimal difference between peaks that is expected.*

Post processing

After aligning peaks the package offers several optional post processing steps that allow to cleanup the data.

Removing contaminations

Among other sources, residues of unwanted chemical substances in the gas chromtography column or within reagents used in the laboratory have the potential to contaminate chemical samples. To get rid of these substances

it is avdisable to include negative controls that have been treated in the very same way as the acutal samples but have not been used to sample and individual´s skin for instance. Within align$_{c}$*hromatogramsthosesamplescanbeincludedinthedatasetas*b

Removing single peaks

Sometimes substances occur that are only found within a single sample. For comparative approaches that calculate simmilarity mattrices these substances are not informative and can be removed from the data for reasons of simplicity. **GCalignR** allows to do so using the logical delete$_{s}$*ingle*$_{p}$*eak*.

Normalisation

Many multivariate analysis techniques,like those available in **vegan**,require a data frame of independent variables as input format. Moreover is generally advisable to normalise abundancies prior to statistical analysis to correct for variations in the total concentration of samples. This is utilised in **GCalignR** function normalise$_{p}$*eakswhichnormalisespeakabundancesbycal*

Workflow

Here,we demonstrate the typical workflow using our seal data. All alignment steps that have been described above are implemented within the function align_chromatograms. A list of all parameters and their description can be assessed from the documentation in the helpfile by typing ?align_chromatograms. As it is outlined in

```
seal_aligned <- align_chromatograms(data = peak_data,
                    conc_col_name = "area",
                    max_diff_peak2mean = 0.03,
                    min_diff_peak2peak = 0.05,
                    max_linear_shift = 0.05,
                    rt_col_name = "time",
                    delete_single_peak = TRUE,
                    blanks = c("C2","C3")) # negativ controls
```

```
#> All checks passed!
#> Ready for processing with align_chromatograms
#> Run GCalignR
#> Start: 12:25:58
#>
#> Data for 84 samples loaded.
#> A reference was not specified. Hence, 'P31' was selected on the basis of highest
#> average similarity to all samples (score = 37).
#> Start Linear Transformation with "P31" as a reference ... Done
#> Start Alignment of Peaks ...  This might take a while!
#> Iteration 1 out of 1  ...
#> Merged Redundant Peaks
#> Peak Alignment Done
#>
#> Blank Peaks deleted & Blanks removed
#>
#> Single Peaks deleted: 61 have been removed
#>
#> Alignment was successful!
#> Time: 12:43:48
```

Now,we can inspect the results by retrieving summaries of the alignment process. The printing method summarises the function call including defaults that have not been explicitly specified during the function call. We also get the relevant information to retrace every step in the alignment:

```
print(seal_aligned)
```

```
#>   Summary of Peak Alignment running align_chromatograms from package GCalignR
#>   Input: peak_data   Start:  2017-01-12 12:25:58     Finished:  2017-01-12 12:43:48
```

```
#>
#> Call:
#>   GCalignR::align_chromatograms(data=peak_data, conc_col_name=area,
#>   rt_col_name=time, max_linear_shift=0.05, max_diff_peak2mean=0.03,
#>   min_diff_peak2peak=0.05, blanks=(C2, C3), delete_single_peak=TRUE, sep=\t,
#>   rt_cutoff_low=NULL, rt_cutoff_high=NULL, reference=NULL, iterations=1)
#>
#> Summary of scored substances:
#>
#>     Peaks In_Blanks  Singular  Retained
#>       480       169        61       250
#>
#>   In total 480 substances were identified among all samples. NA substances were
#>   present in blanks. The corresponding peaks as well as the blanks were removed
#>   from the data set. 61 substances were present in just one single sample and were
#>   removed. 250 substances are retained after all filtering steps.
#>
#> Sample Overview  The following 84 Samples were aligned to the reference 'P31':
#>   M2, M3, M4, M5, M6, M7, M8, M9, M10, M12, M14, M15, M16, M17, M18, M19, M20,
#>   M21, M23, M24, M25, M26, M27, M28, M29, M30, M31, M33, M35, M36, M37, M38, M39,
#>   M40, M41, M43, M44, M45, M46, M47, M48, P2, P3, P4, P5, P6, P7, P8, P9, P10,
#>   P12, P14, P15, P16, P17, P18, P19, P20, P21, P23, P24, P25, P26, P27, P28, P29,
#>   P30, P31, P33, P35, P36, P37, P38, P39, P40, P41, P43, P44, P45, P46, P47, P48
#>
#> For further details:
#>   Type 'gc_heatmap(seal_aligned)' to retrieve a heatmap for the alignment accuracy
#>   Type 'plot(seal_aligned)' to retrieve further diagnostic plots
```

The quality of an alignment will depend on sensible parameters that facilitate the (i) correction of linear shifts that might fall in a larger range with increasing sample size and (ii) and the variability of retention times. Optimally,linear shifts do not exhaust the range given by max_linear_shift completely,which would in turn indicate that not all uncertainties haven been fully compensated for. This can be assessed by some diagnostic plots:

```
plot(seal_aligned)
```


0stoffel0hoffman$_f$iles/figure0latex/unnamed0chunk0501.png0stoffel0hoffman$_f$iles/figure0latex/unnamed0chunk0501.pdf0stoffel0hoffman$_f$iles/figure0latex/unnamed0chunk0501.jpg0stoffel0hoffman$_f$iles/figure0latex/unnamed0chunk0501.mps0stoffel0hoffman$_f$iles/figure0latex/unnamed0chunk0501.jpeg0stoffel0hoffman$_f$iles/figure0latex/unnamed0chunk0501.jbig20stoffel0hoffman$_f$iles/figure0latex/unnamed0chunk0501.jb20stoffel0hoffman$_f$iles/figure0latex/unnamed0chunk0501.PNG0stoffel0hoffman$_f$iles/figure0latex/unnamed0chunk0501.PDF0stoffel0hoffman$_f$iles/figure0latex/unnamed0chunk0501.JPG0stoffel0hoffman$_f$iles/figure0latex/unnamed0chunk0501.JPEG0stoffel0hoffman$_f$iles/figure0latex/unnamed0chunk0501.JBIG20stoffel0hoffman$_f$iles/figure0latex/unnamed0chunk0501.JB20stoffel0hoffman$_f$iles/figure0latex/unnamed0chunk0501.eps

```
gc_heatmap(seal_aligned,type = "continuous", substance_subset = 1:25, samples_subset = 1:25)
```


0stoffel0hoffman$_f$iles/figure0latex/unnamed0chunk0502.png0stoffel0hoffman$_f$iles/figure0latex/unnamed0chunk0502.pdf0stoffel0hoffman$_f$iles/figure0latex/unnamed0chunk0502.jpg0stoffel0hoffman$_f$iles/figure0latex/unnamed0chunk0502.mps0stoffel0hoffman$_f$iles/figure0latex/unnamed0chunk0502.jpeg0stoffel0hoffman$_f$iles/figure0latex/unnamed0chunk0502.jbig20stoffel0hoffman$_f$iles/figure0latex/unnamed0chunk0502.jb20stoffel0hoffman$_f$iles/figure0latex/unnamed0chunk0502.PNG0stoffel0hoffman$_f$iles/figure0latex/unnamed0chunk0502.PDF0stoffel0hoffman$_f$iles/figure0latex/unnamed0chunk0502.JPG0stoffel0hoffman$_f$iles/figure0latex/unnamed0chunk0502.JPEG0stoffel0hoffman$_f$iles/figure0latex/unnamed0chunk0502.JBIG20stoffel0hoffman$_f$iles/figure0latex/unnamed0chunk0502.JB20stoffel0hoffman$_f$iles/figure0latex/unnamed0chunk0502.eps

Meinolf Ottensmann
Department of Animal Behaviour
Bielefeld University
Morgenbreede 45
33615 Bielefeld
Meinolf.Ottensmann@web.de

Martin A. Stoffel
Department of Animal Behaviour
Bielefeld University
Morgenbreede 45
33615 Bielefeld
Martin.Adam.Stoffel@gmail.com

Joseph I. Hoffman
Department of Animal Behaviour
Bielefeld University
Morgenbreede 45
33615 Bielefeld
j_i_hoffman@hotmail.com