GCalignR. An R package for aligning Gas-Chromatography data

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

Abstract Chemical signals are among the most fundamental and oldest means of animal communication. The desire to unravel broader patterns of chemical communication in birds and mammals paved the way for two not entirely new techniques, gas-chromatography and mass-spectrometry, in the fields of ecology and evolution. Comparing chemical profiles or chromatograms across many individuals yields some major obstacles as even the newest GC machines have an inherent error when measuring the retention times of chemical substances. Here we present GCalignR, an R package for the alignment of chromatography peaks among samples prior to multivariate statistical analysis. GCalignR is specifically designed to be used by non-chemists by providing easy to use functions to check and align gas-chromatography data based on retention times. In addition, the package implements heatmaps and other plots to evaluate and potentially adjust the peak alignment. We hope that GCalignR will provide a tool that fits into a common biologist's workflow in R and that the package will facilitate the standardization and reproducibility of studies on chemical communication.

Introduction

Chemical cues are arguably the most common mode of communication among animals (Wyatt, 2014). Patterns in complex chemical signatures can can yield information about kinship (Krause et al., 2012; Stoffel et al., 2015), genetic diversity (Charpentier et al., 2010; Leclaire et al., 2012), sexual maturation (Caspers et al., 2011) or be used for species discrimination (de Meulemeester et al., 2011). One of the most common instruments to quantify the chemical composition of samples is gas-chromatography (GC), a fast high-throughput method to detect individual chemicals and their abundances (McNair and Miller, 2011), while the additional implementation of mass-spectrometry (GC-MS) allows to identify specific substances (Caspers et al., 2011).

However, before similarity patterns across can be analysed, it is essential to align compounds among samples. The alignment of samples has to account for drifts in the retention times of peaks which are caused by subtle, random and often unavoidable variations of the chromatography machine parameters (Pierce et al., 2005). Surprisingly, studies on mammalian or avian chemical communication often rely on manual alignment rather than (semi-)automated algorithms (citations necessary), but this approach bears three severe drawbacks: (1) For larger sample sizes, this task becomes extremely time consuming and inefficient (2) The researcher may bias the alignment due to subjective experience and expectations. (3) The data analytical 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 (Pierce et al., 2005; Robinson et al., 2007; Jiang et al., 2013) and only some are easily accessible as web-based tools (Hoffmann and Stoye, 2009; Wang et al., 2010) or independent software (Dellicour and Lecocq, 2013).

Here, we introduce `GCalignR`, an R package that implements a simple algorithm to align peaks purely from retention time data obtained by GC and provides sophisticated visualisations for the evaluation of the alignment quality. `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** (Oksanen et al., 2016) package. This allows the integration of the full analysis into **RMarkdown** documents (Allaire et al., 2016) in order to meet the standards of reproducibility (Peng, 2011).

The Package

GCalignR contains functions to align peaks from GC and GC-MS data based on retention times and evaluate the respective alignments. The main aim of the package is to provide a simple tool that guides the user through the alignment of large data sets prior to the statistical analysis of multivariate chemical data (Anderson, 2001) (why a citation here?). An easy workflow for the analysis of chemical data including GCalignR is shown below (figure 1). The package vignette provides a detailed description of all functions and their argumentsand can be assessed via `browseVignettes('GCalignR')`. However, the

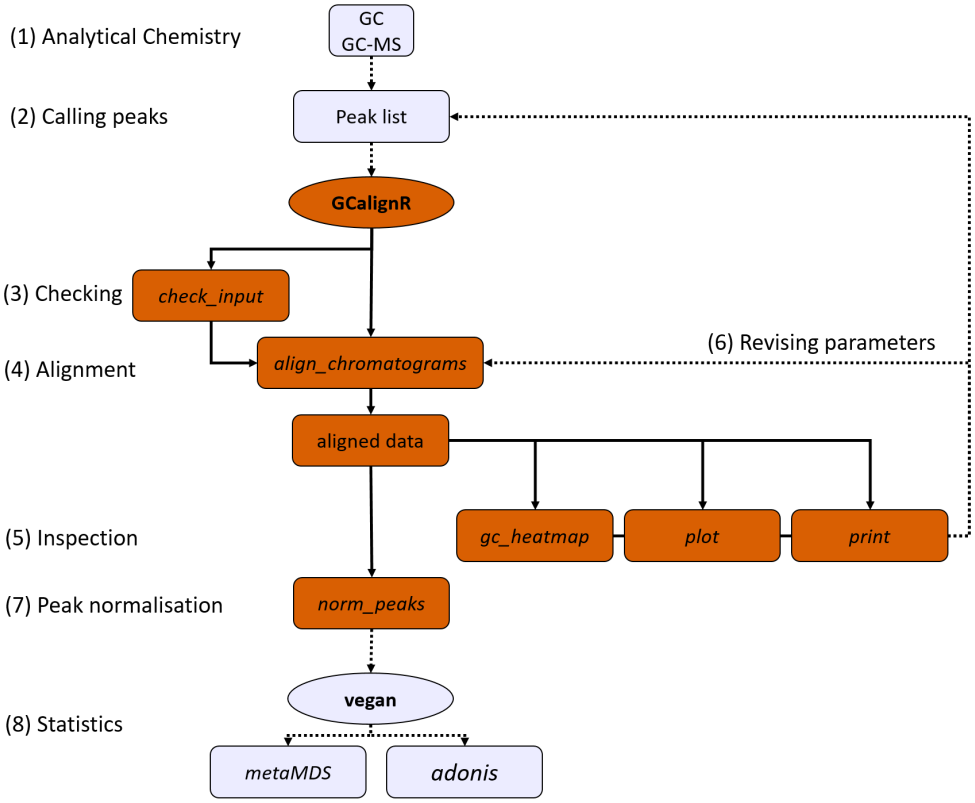


Figure 1: GCalignR workflow. In addition to the alignment of substances across samples, the package provides functions for checking and inspecting the data. The aligned data is ready to use for analyses in conjunction with other packages. Each function is explained within the text.

Example dataset For demonstration purposes GCalignR includes data of skin chemicals from 82 Antarctic fur seals *Arctocephalus gazella*. It was previously shown that these signatures encode the membership to a breeding colony Stoffel et al. (2015). These data are available as a `list` of `data.frame`'s. Each `data.frame` contains the peak data of one individual, whereby one variable contains the retention time ("time") and the other variable the concentration or peak abundance ("area") within a sample. **we should use a txt file as an example here**

```
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
```

Alignment of Gas-Chromatography peaks among samples

The alignment procedure is divided into five steps (figure 2). All steps are executed within the main function `align_chromatograms` and will be explained in in the next sections.

(1) Linear adjustments of chromatograms At first, all peaks within a chromatogram are shifted with respect to a reference chromatogram to account for systematic shifts in retention times among homologous chemicals shared by samples (figure 2 A). This is done for all samples in relation to the reference sample such that the number shared peaks 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 is 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 can 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.

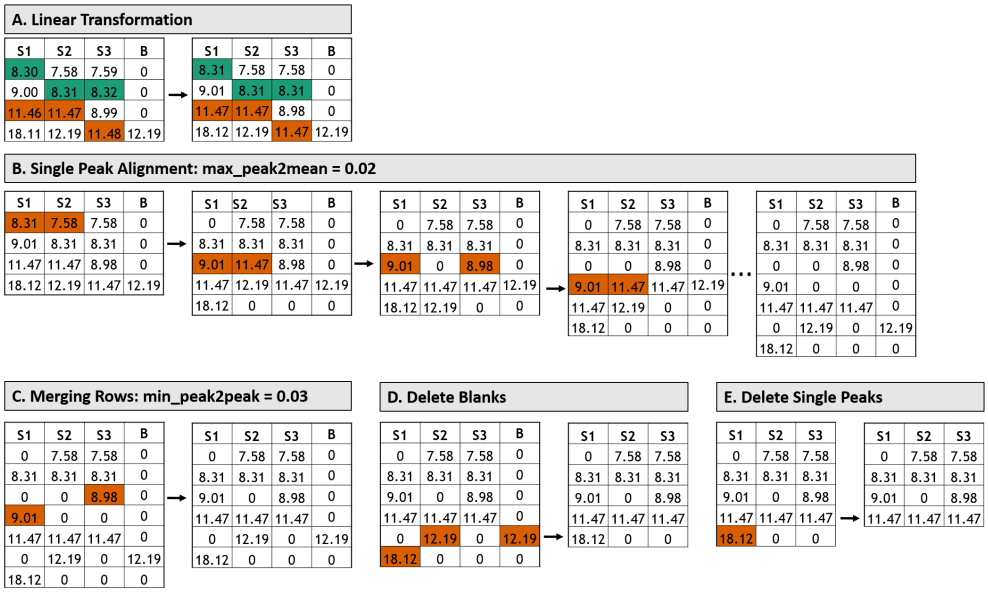


Figure 2: Overview of the algorithm performed by GCalignR. Rows of matrices correspond to substances, columns are samples. Zeros indicate absence of peaks and are ignored in calculations. **1**. Chromatograms are linearly shifted with respect to a reference (S2). **2**. From left to right the first four 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. **3**. Coloured cells represent conflicting retention times using a minumum `max_diff_peak2mean = 0.02`, which defines the minimum difference expected between two distinct substances. If merging does not result in the loss of any data, rows are merged. **4**. If specified, all peaks found in one or more blanks (negative controls) are removes as well as the blank itself. Unique peaks present in only a single individual are not of interest for similarity analyses and can be removed as well.

(2) Peak alignment The core of the alignment procedure is based on clustering of individual peaks across 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 2 B):

$$rt_m > \left(\frac{\sum_{i=1}^{m-1} rt_i}{m-1} \right) + max\_peak2mean \tag{1}$$

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

$$rt_m < \left(\frac{\sum_{i=1}^{m-1} rt_i}{m-1} \right) - max\_peak2mean \tag{2}$$

with rt = retention time; m = current column and *max_diff_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 variable retention times are treated more stringent than chemicals exhibiting higher variability. Once the last retention time of a row was evaluated the whole procedure is repeated with the next row until the end of the retention time matrix was reached.

(3) Merging Sometimes, a single substance has been split up into two different rows. However, the emerging pattern is very clear, as part of the samples will have the substance in a given row, but no substance in the adjacent row and vice versa for another part of the samples. Knowing this pattern, rows will be merged when this does not cause any loss of any information (i.e. no sample exists that contains substances in both rows).(figure 2 C). Again, the user can change the threshold for the minimal difference in the retention time between two mergeable peaks with *min_diff_peak2peak*.

(4) 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 advised to include negative controls that have been treated in the very same way as the samples but have not been used to actually sample something. Within `align_chromatograms` those controls can be included in the data set as *blanks*. Blanks are treated as normal samples during all alignment steps and are used to identify contaminants afterwards, which in turn can be removed from the data as well as the negative controls itself.

Removing single peaks Sometimes, substances occur purely in a single sample. For comparative approaches that calculate similarity matrices these substances are often not informative and can be removed from the data. `GCalignR` allows to do so by setting the `delete_single_peak` argument to `TRUE`.

Normalisation Many multivariate analysis techniques, like those available in **vegan**, require a `data,frame` of independent variables as input format. Moreover it is generally advisable to normalise substance abundances prior to statistical analysis to correct for variations in the total concentration of samples. This can be done in `GCalignR` with the function `normalise_peaks` which calculate relative abundances within each sample.

Workflow

** maybe include check_input **

Here, we demonstrate a typical workflow in `GCalignR` 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: 21:39:28
#>
#> 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: 62 have been removed
#>
#> Alignment was successful!
#> Time: 21:56:00
```

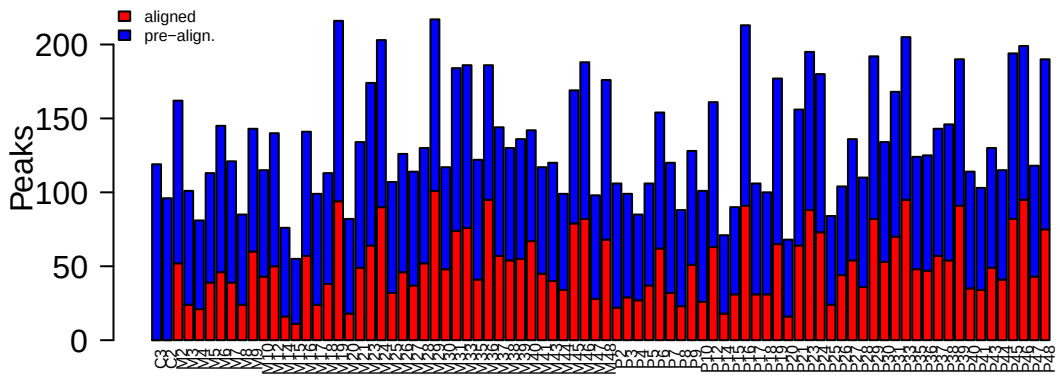
Now, we can inspect the results by retrieving summaries of the alignment process. The printing method summarises the function call including defaults that have not been explicitly specified during the function call. We also get the relevant information to retrace every step in the alignment:

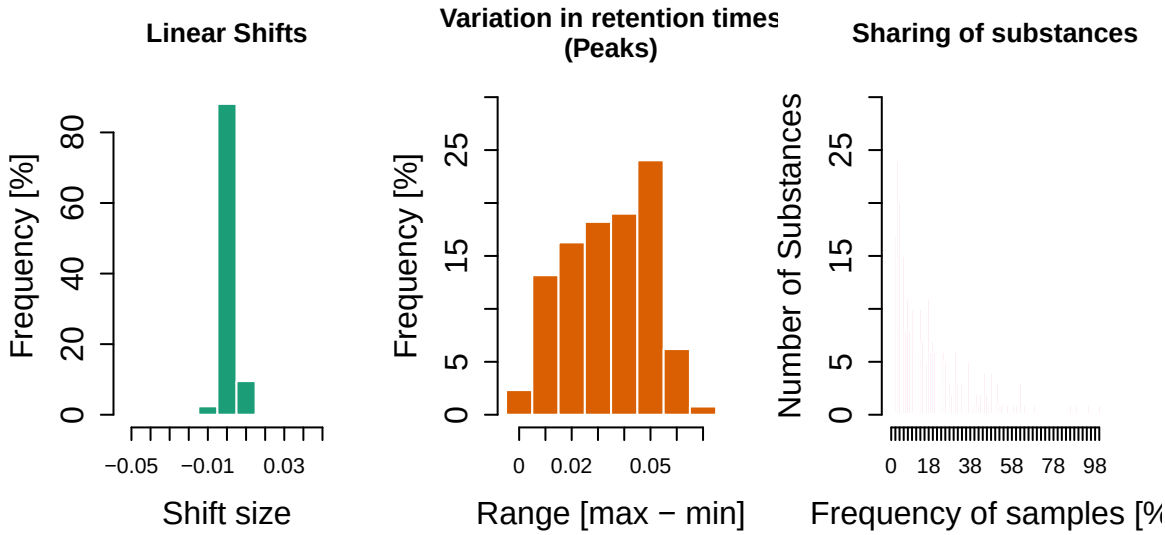
```
print(seal_aligned)
```

```
#>   Summary of Peak Alignment running align_chromatograms from package GCalignR
#>   Input: peak_data   Start:  2017-01-12 21:39:28     Finished:  2017-01-12 21:56:00
#>
#> 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
#>       485       165        62       258
#>
#>   In total 485 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. 62 substances were present in just one single sample and were
#>   removed. 258 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
```

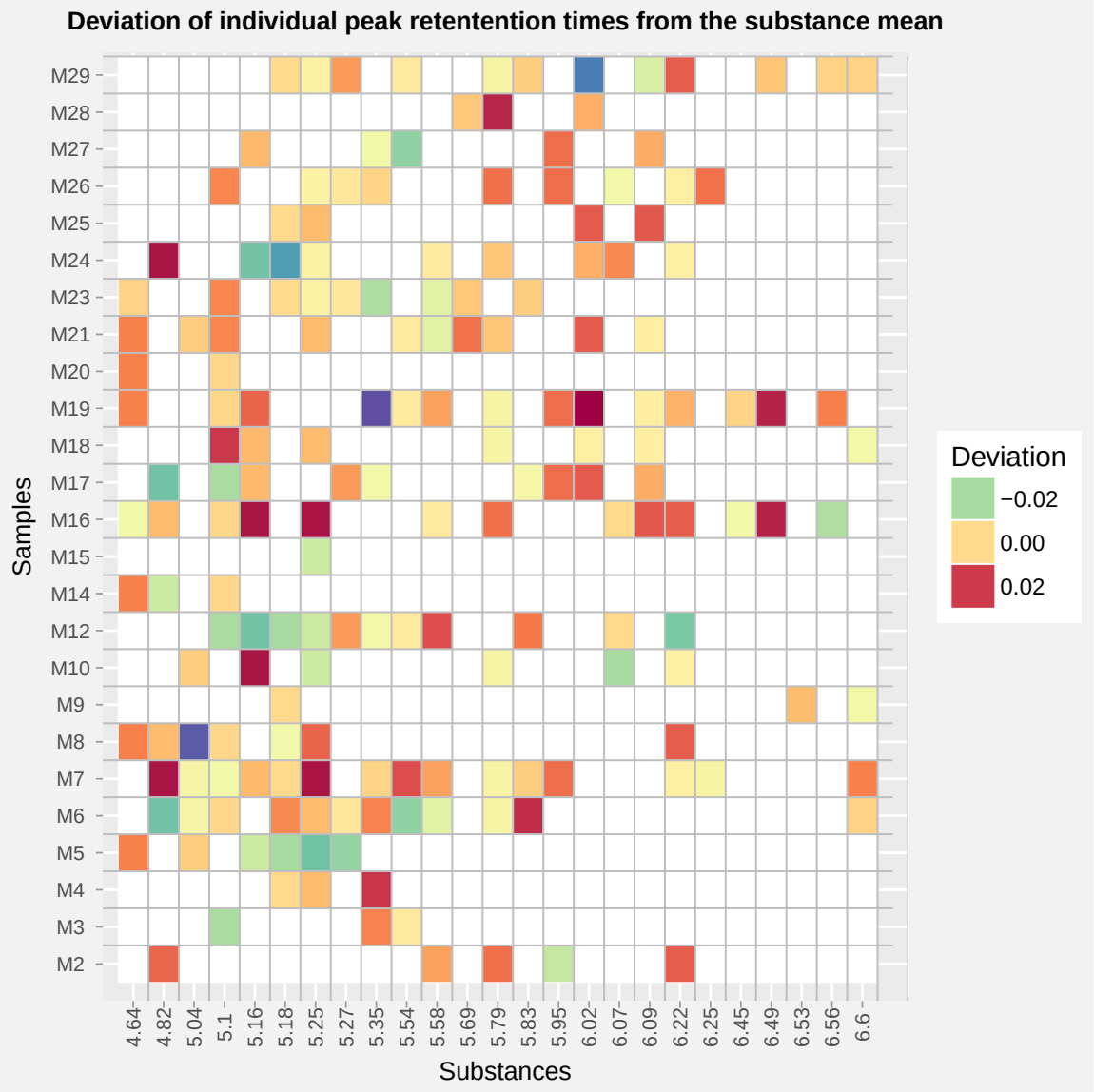
The quality of an alignment will depend on sensible parameters that facilitate the (i) correction of linear shifts that might fall in a larger range with increasing sample size and (ii) and the variability of retention times. Optimally, linear shifts do not exhaust the range given by `max_linear_shift` 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)
```






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



Bibliography

J. J. Allaire, J. Cheng, Y. Xie, J. McPherson, W. Chang, J. Allen, H. Wickham, A. Atkins, and R. Hyndman. rmarkdown: Dynamic documents for r, 2016. URL https://CRAN.R-project.org/package=rmarkdown. [p1]

M. J. Anderson. A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, 26(1):32–46, 2001. ISSN 14429985. doi: 10.1111/j.1442-9993.2001.01070.pp.x. [p2]

B. A. Caspers, F. C. Schroeder, S. Franke, and C. C. Voigt. Scents of adolescence: the maturation of the olfactory phenotype in a free-ranging mammal. *PloS one*, 6(6):e21162, 2011. [p1]

M. J. Charpentier, J. C. Crawford, M. Boulet, and C. M. Drea. Message 'scent': Lemurs detect the genetic relatedness and quality of conspecifics via olfactory cues. *Animal Behaviour*, 80(1):101–108, 2010. ISSN 00033472. doi: 10.1016/j.anbehav.2010.04.005. [p1]

T. de Meulemeester, P. Gerbaux, M. Boulvin, A. Coppée, and P. Rasmont. A simplified protocol for bumble bee species identification by cephalic secretion analysis. *Insectes Sociaux*, 58(2):227–236, 2011. ISSN 0020-1812. doi: 10.1007/s00040-011-0146-1. [p1]

S. Dellicour and T. Lecocq. 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):3206–3209, 2013. ISSN 1615-9306. doi: 10.1002/jssc.201300388. URL http://onlinelibrary.wiley.com/doi/10.1002/jssc.201300388/full. [p1]

N. Hoffmann and J. Stoye. Chroma: signal-based retention time alignment for chromatography-mass spectrometry data. *Bioinformatics (Oxford, England)*, 25(16):2080–2081, 2009. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp343. [p1]

W. Jiang, Z.-M. Zhang, Y. Yun, D.-J. Zhan, Y.-B. Zheng, Y.-Z. Liang, Z. Y. Yang, and L. Yu. Comparisons of five algorithms for chromatogram alignment. *Chromatographia*, 76(17-18):1067–1078, 2013. ISSN 0009-5893. doi: 10.1007/s10337-013-2513-8. [p1]

E. T. Krause, O. Krüger, P. Kohlmeier, and B. A. Caspers. Olfactory kin recognition in a songbird. *Biology letters*, 8(3):327–329, 2012. ISSN 1744-9561. [p1]

S. Leclaire, T. Merkling, C. Raynaud, H. Mulard, J.-M. Bessière, É. Lhuillier, S. A. Hatch, and É. Danchin. 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):1185–1193, 2012. [p1]

H. M. McNair and J. M. Miller. *Basic gas chromatography*. John Wiley & Sons, 2011. ISBN 1118211200. [p1]

J. Oksanen, F. G. Blanchet, M. Friendly, R. Kindt, P. Legendre, D. McGlinn, P. R. Minchin, R. B. O'Hara, G. L. Simpson, P. Solymos, M. H. H. Stevens, E. Szoecs, and H. Wagner. vegan: Community ecology package, 2016. URL https://CRAN.R-project.org/package=vegan. [p1]

R. D. Peng. Reproducible research in computational science. *Science (New York, N.Y.)*, 334(6060): 1226–1227, 2011. ISSN 0036-8075. doi: 10.1126/science.1213847. [p1]

K. M. Pierce, J. L. Hope, K. J. Johnson, B. W. Wright, and R. E. Synovec. 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):101–110, 2005. [p1]

M. D. Robinson, D. P. de Souza, W. W. Keen, E. C. Saunders, M. J. McConville, T. P. Speed, and V. A. Likić. A dynamic programming approach for the alignment of signal peaks in multiple gas chromatography-mass spectrometry experiments. *BMC bioinformatics*, 8(1):419, 2007. ISSN 1471-2105. [p1]

M. A. Stoffel, B. A. Caspers, J. Forcada, A. Giannakara, M. Baier, L. Eberhart-Phillips, C. Müller, and J. I. Hoffman. 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, 2015. [p1, 2]

S.-Y. Wang, T.-J. Ho, C.-H. Kuo, and Y. J. Tseng. Chromaligner: a web server for chromatogram alignment. *Bioinformatics (Oxford, England)*, 26(18):2338–2339, 2010. ISSN 1367-4803. doi: 10.1093/bioinformatics/btq337. [p1]

T. D. Wyatt. *Pheromones and animal behavior: chemical signals and signatures*. Cambridge University Press, 2014. ISBN 1107647495. [p1]

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