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 hypothesis testing using multivariate statistics. 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 (**?**). Patterns in complex chemical signatures can can yield information about kinship (**??**), genetic diversity (**??**), sexual maturation (**?**) or be used for species discrimination (**?**). 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 abundancies (**?**), while the additional implementation of mass-spectrometry (GC-MS) allows to identify specific substances (**?**).

However, before similarity patterns across samples can be analysed, it is essential to align compounds. 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 (**?**). Surprisingly, studies on mammalian or avian chemical communication often rely on manual alignment rather than (semi-)automated algorithms, 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 (**???**) and only some are easily accessible as web-based tools (**??**) or independent software (**?**).

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 that represent contaminations, which are identified due to their presence in negative control samples, henceforth called blanks. For an easy interpretation of the quality of an alignment we implemented several diagnostic plots that allow to access the aligned data visually. 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` 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. An easy workflow for the analysis of chemical data including `GCalignR` is shown in figure **??** and described below. The package vignette provides a detailed description of all functions

and their arguments and can be assessed via `browseVignettes('GCalignR')` once the package was installed.

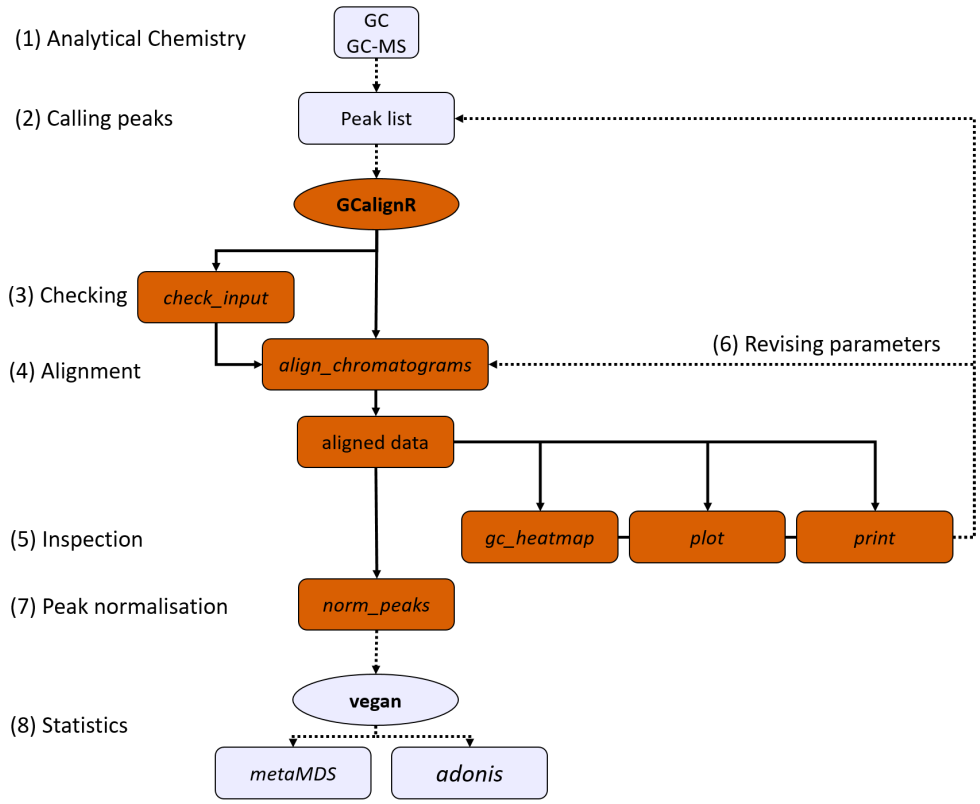


Figure 1: GCalignR workflow. In addition to the alignment of substances across samples, the package provides functions for checking and inspecting the data. The aligned data is ready to use for analyses in conjunction with other packages. Each function is explained within the text.

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 **?**. These data are available in a single text file, the standard input format of `GCalignR`, that is distributed with the package. The first two lines contain the names of all samples and variables respectively. From the third row onwards, data of all samples is included, whereby data frames are concatenated horizontally.

```
## Path to the dataset
fpath <- system.file("extdata", "peak_data.txt", package = "GCalignR")

# Open the file in an external editor
file.show(fpath)
```

Alignment of Gas-Chromatography peaks among samples

The alignment procedure is divided into five steps (figure **??**). 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 **??** 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.

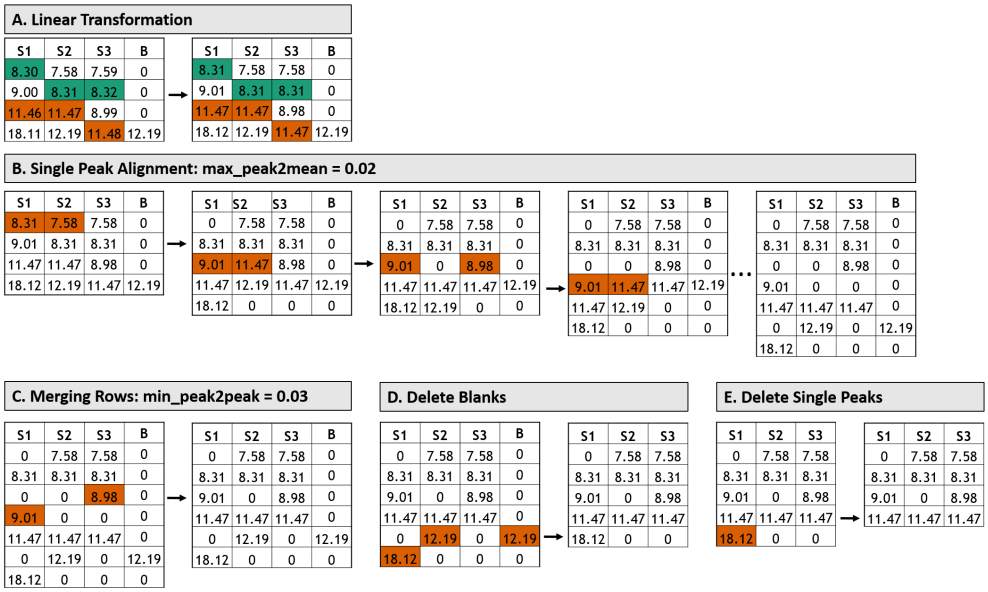


Figure 2: Overview of the algorithm performed by GCalignR. Rows of matrices correspond to substances, columns are samples. Zeros indicate absence of peaks and are ignored in calculations. **1**. Chromatograms are linearly shifted with respect to a reference (S2). **2**. From left to right the first four 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 that show a greater difference than specified. 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 removed 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 **??** 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 **??** 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 generally advised to include control samples. Within `align_chromatograms` those controls can be included in the data set in the same way as a normal sample. By specifying the name of one or more control samples with the `blanks = c("contr1", "contr2")`, all substances present in the control samples are removed from the dataset.

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

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 figure **??**, 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 those errors. This function ist executed internally prior to any alignment.

```
check_input(fpath)
```

```
#> All checks passed!
```

```
aligned_peak_data <- align_chromatograms(data = peak_data,
        rt_col_name = "time",
        max_diff_peak2mean = 0.02,
        min_diff_peak2peak = 0.08,
        max_linear_shift = 0.05,
        delete_single_peak = TRUE,
        blanks = c("C2","C3"),
        write_output = NULL) # change to generate text files
```

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:

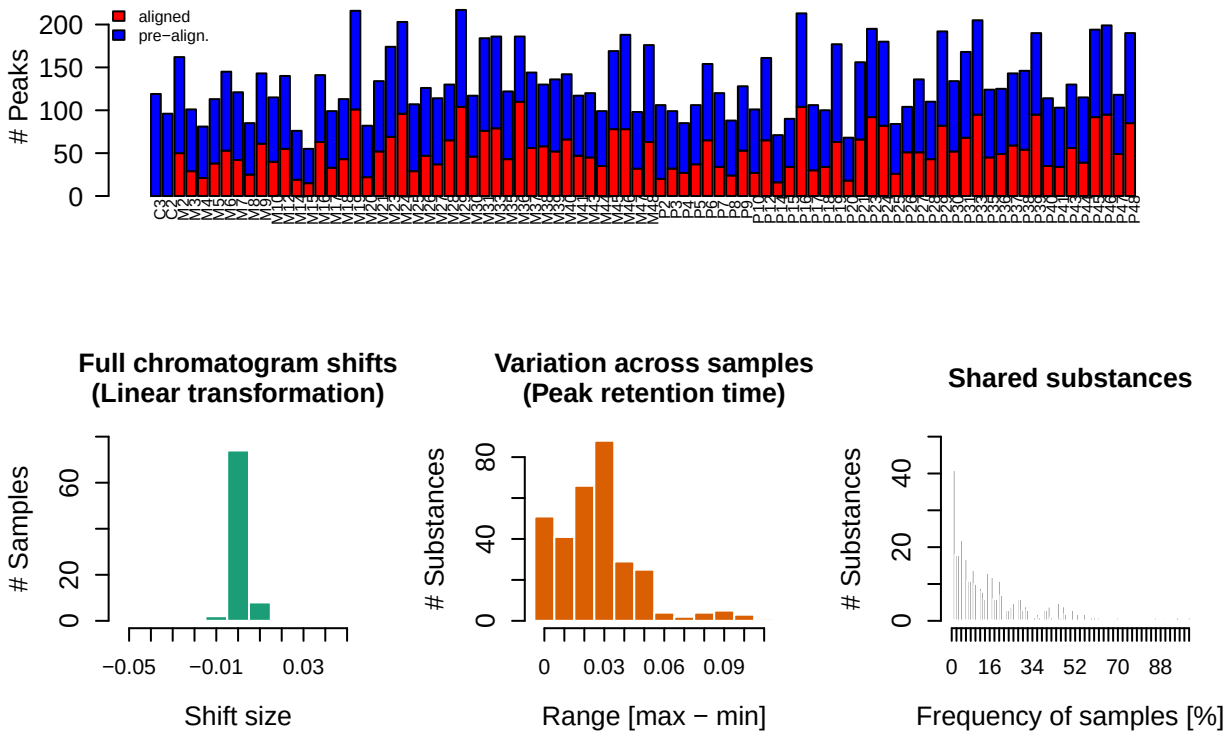
```
print(aligned_peak_data)
```

```
#> Summary of Peak Alignment running align_chromatograms
#> Input: peak_data
#> Start:  2017-02-01 18:04:11  Finished:  2017-02-01 18:41:11
#>
#> Call:
#>   GCalignR::align_chromatograms(data=peak_data, rt_col_name=time,
```

```
#>   max_linear_shift=0.05, blanks=(C2, C3), sep=\t, rt_cutoff_low=NULL,
#>   rt_cutoff_high=NULL, reference=NULL, max_diff_peak2mean=0.02,
#>   min_diff_peak2peak=0.08, delete_single_peak=FALSE)
#>
#> Summary of scored substances:
#>    total   blanks retained
#>      490      171      319
#>
#> In total 490 substances were identified among all samples. 171 substances were
#>   present in blanks. The corresponding peaks as well as the blanks were removed
#>   from the data set. 319 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(aligned_peak_data)' to retrieve heatmaps
#>   'plot(aligned_peak_data)' 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 four diagnostic plots that can be created alltogether as well as individually.

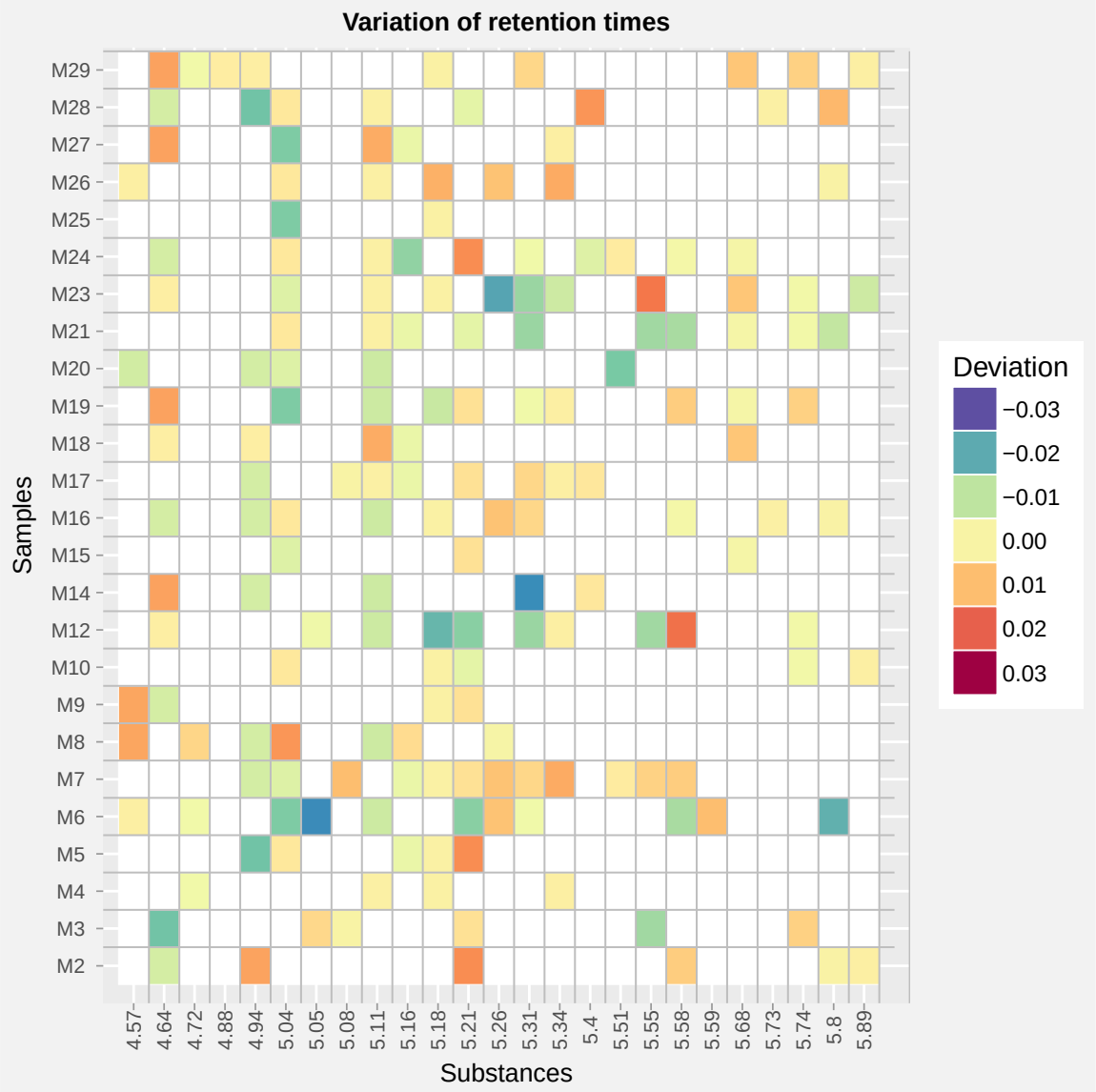
```
plot(aligned_peak_data)
```



The distribution of peak number before and after the alignment reveals a noticeable reduction of peaks in the aligned dataset. These changes can be explained by the removal of contamimations

(i.e. peaks present in blanks) and the removal of single peaks. Type `print(aligned_peak_data)` for details on both. The distribution of shifts sizes used for linear transformations shows a marginal linear trend across the chromatography run and will depend on the of samples relative to the reference while performing the chromatography. Besides the pure number of peaks it is of major interest to inspect the distribution of substances in the pool of samples and access the variation in retention times. This can be investigated simultaneously with a heatmap.

```
gc_heatmap(aligned_peak_data,type = "discrete", substance_subset = 1:25, samples_subset = 1:25)
```



The heatmap indicates the presence of a certain substance within a sample by a colour-filled box, whereas the absense is encoded by a white box. Furthermore a colour-gradient is used to indicate the deviation of each retention time from the mean value among all other samples as a measure of variation. The pattern shown here does not indicate any obvious issues with the aligned dataset. Hence, there is no need to adjust the aligning parameters further and we can move on to analyse the dataset.

Prior to analysing pattern within the aligned data we normalise the peak area to correct for difference in the toal concentration among samples and format the data to subsequent ordination approaches using **vegan**.

```
scent <- norm_peaks(data = aligned_peak_data,
                    rt_col_name = "time",
                    conc_col_name = "area",
                    out = "data.frame" )
```

Visualising patterns by ordination

vegan offers several methods for ordination approaches. We apply a non-metric-multidimensional scaling (NMDS) using a Bray-Curtis dissimilarity in order to investigate differences in chemical profiles between two colonies. The package contains a data frame `peak_factors` that is comprised of three factors ("colony","family" and "age") for all samples that denote the rownames. The visualisation is done using**ggplot2**.

\begin{Schunk}

```
library(vegan)
```

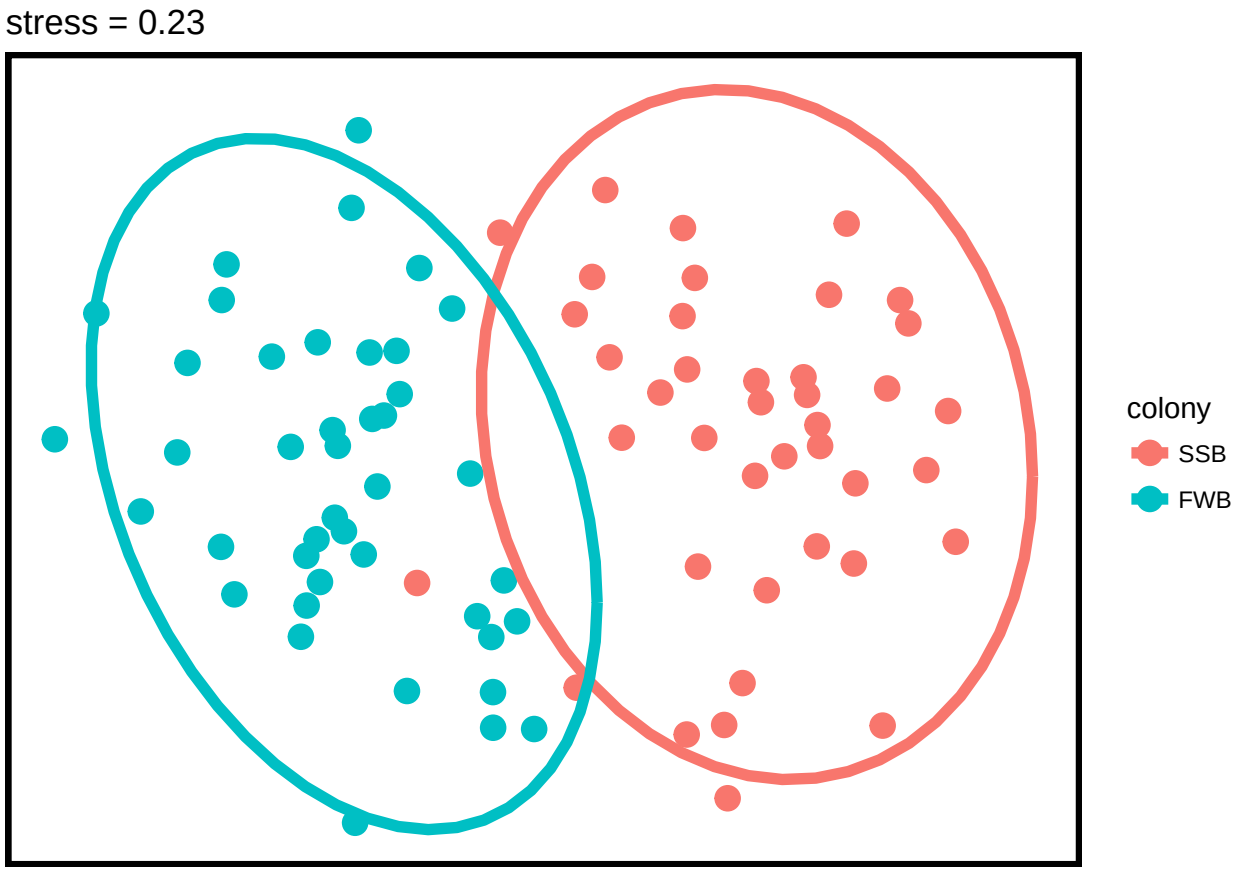
> Loading required package: permute

```
#> Loading required package: lattice
```

> This is vegan 2.4-2

\begin{Sinput} ## factors for each sample of the chemical dataset data("peak_factors") ## both datasets have the same rownames and can be sorted accordingly scent <- scent[match(row.names(peak_factors), ## standard log + 1 transformation scent <- log(scent + 1) ## NMDS using Bray-Curtis dissimilarities scent_nmds <- vegan::metaMDS(comm = scent,distance = "bray") ## extract the stress value stress <- scent_nmds[["stress"]] ## get the x and y coordinates scent_nmds <- as.data.frame(scent_nmds$points) ## add the factor of interest scent_nmds <- cbind(scent_nmds,colony = peak_factors[["colony"]]) \end{Sinput} \end{Schunk}

```
## ordiplot with ggplot2
library(ggplot2)
ggplot(data = scent_nmds,aes(MDS1,MDS2,color = colony)) +
    geom_point(size = 4) +
    stat_ellipse(size = 2) +
    labs(title = paste0("stress = ", round(stress,2)), x = "MDS1", y = "MDS2") +
    theme_void() +
    theme(panel.background = element_rect(colour = "black", size = 2,fill = NA))
```



The ordination plot shows a clear pattern that separates individuals by the breeding colony. **vegan** offers furthermore permutational test for multivariate analysis of variance ("permutational manova", (**?**)) that support the observed pattern.

```
## Testing for a location effect
vegan::adonis(scent ~ peak_factors[["colony"]],permutations = 9999)

#>
#> Call:
#> vegan::adonis(formula = scent ~ peak_factors[["colony"]], permutations = 9999)
#>
#> Permutation: free
#> Number of permutations: 9999
#>
#> Terms added sequentially (first to last)
#>
#>                          Df SumsOfSqs MeanSqs F.Model     R2 Pr(>F)
#> peak_factors[["colony"]]  1    2.5351 2.53514  11.492 0.1256  1e-04 ***
#> Residuals                80   17.6486 0.22061         0.8744
#> Total                    81   20.1837                 1.0000
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Testing for a dispersion effect
anova(vegan::betadisper(vegan::vegdist(scent,method = "bray"),peak_factors[["colony"]]))

#> Analysis of Variance Table
#>
#> Response: Distances
#>           Df   Sum Sq   Mean Sq F value Pr(>F)
#> Groups     1 0.000347 0.0003474   0.095 0.7587
#> Residuals 80 0.292452 0.0036557
```

Validation

We analysed the performance of `GcalignR` using datasets of three bumble bee species *Bombus bimaculatus*, *B. ephippiatus* and *B. flavifrons* where signatures have been obtained from cephalic labial gland secretions of 24, 20 and 11 individuals respectively. These data have been published as supplementary material by **?**. Moreover, for a subset of peaks, substances have been identified by GC-MS. We used all identified substances (*B. bimaculatus* = 32; *B. ephippiatus* = 42; *B. flavifrons* = 44) to determine error rates for our alignments. Hence, we defined assignments of a single peak as incorrect whenever the majority of other samples was assigned to another row in aligned matrices:

$$Error = \left[\frac{N_{missaligned}}{N_{total}} \right] \tag{3}$$

whereby N denotes the number of retention times. By systematically changing the two parameters `max_diff_peak2mean` and `min_diff_peak2peak` we explored 100 parameter combinations to demonstrate how parameter values affect the alignment accuracy.

Additionally, we show the effect of noise (i.e. bad quality chromatograms) on the error rate by addition or substraction of 0.02 or 0.01 minutes to a random subset of peaks per sample. Code and documentation are provided in a single PDF file written in Rmarkdown (S1) together with the data (S2).

Results

The parameter `min_diff_peak2peak` shows the strongest influence on the error rates based on all three bumblebee datsets (figure **??**). With values below 0.06 minutes the error

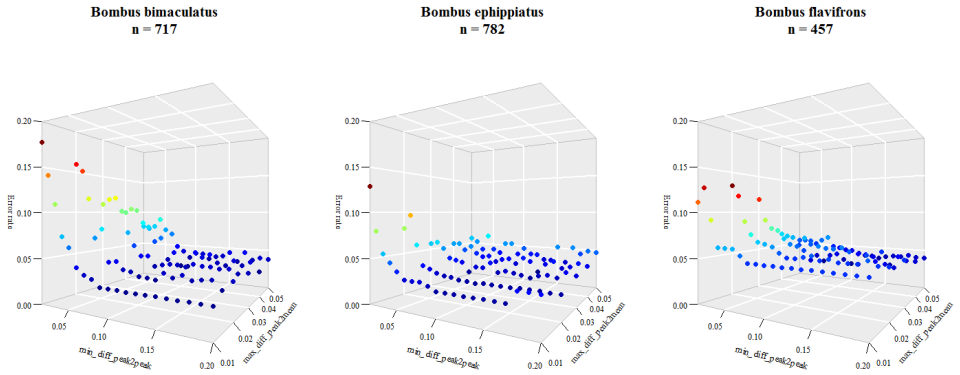


Figure 3: Effects of alignment parameters on the error rate of three datasets where the identity of a subset of peaks was confirmed by GC-MS [@Dellicour.2013]. Each point shows the error in aligning substances for a combination of max_diff_peak2mean and min_diff_peak2peak.

rate is modest for all datasets. The combination of `min_diff_peak2peak = 0.11` and `max_diff_peak2mean = 0.04` offers the lowest average error rate of 3.24 % (*B. bimaculatus* = 2.79%, *B. ephippiatus* = 3.20%, *B. flavifrons* = 3.72%).

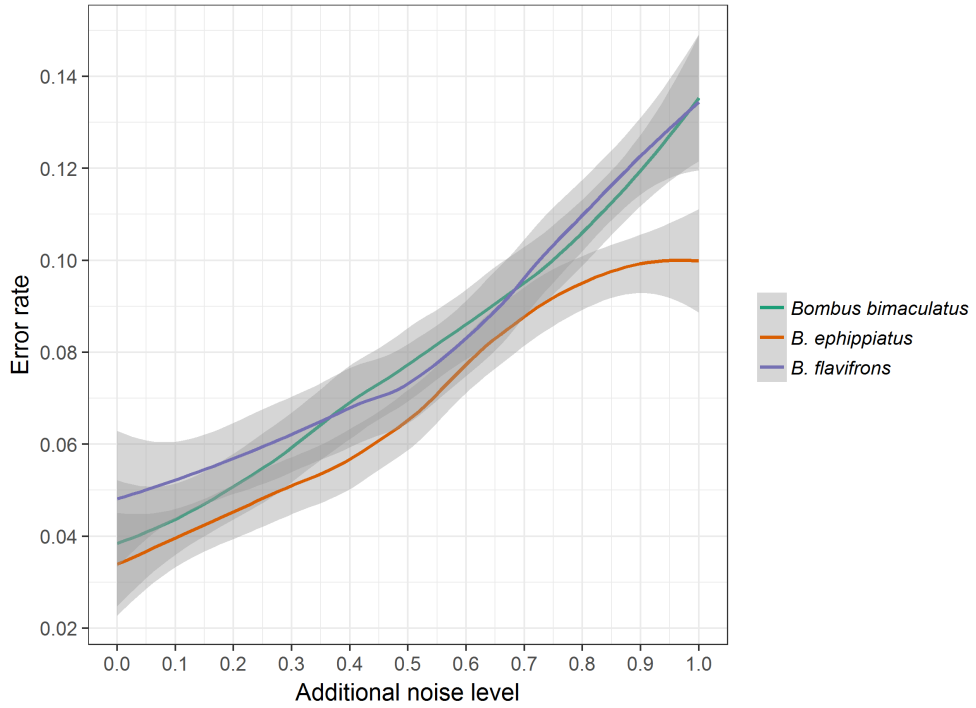


Figure 4: Additional noise in peak retention times increases the error rate substantially. Therefore, optimal alignments require clearly resolved peaks that need to be extracted prior to using **GCalignR**

Adding additional noise to the raw retention times increases the error rate in the aligned data substantially. Therefore, we emphasise the importance to check datasets prior to alignment.

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