Package 'GCalignR'

August 26, 2020

Title Simple Peak Alignment for Gas-Chromatography Data

Version 1.0.3

Date 2020-08-20

Encoding UTF-8

Description Aligns peak based on peak retention times and matches homologous peaks across samples. The underlying alignment procedure comprises three sequential steps. (1) Full alignment of samples by linear transformation of retention times to maximise similarity among homologous peaks (2) Partial alignment of peaks within a user-defined retention time window to cluster homologous peaks (3) Merging rows that are likely representing homologous substances (i.e. no sample shows peaks in both rows and the rows have similar retention time means). The algorithm is described in detail in Ottensmann et al., 2018 <doi:10.1371/journal.pone.0198311>.

Depends R (>= 3.2.5)

Imports ggplot2 (>= 2.2.1), graphics, stats, readr, reshape2, stringr, utils, pbapply, tibble

License GPL (>= 2) | file LICENSE

Language en-GB

LazyData true

RoxygenNote 7.1.1

Suggests knitr, pander, rmarkdown, testthat, vegan (>= 2.4.2)

VignetteBuilder knitr

BugReports https://github.com/mottensmann/GCalignR/issues

URL https://github.com/mottensmann/GCalignR

NeedsCompilation no

Author Meinolf Ottensmann [aut, cre] (<https://orcid.org/0000-0003-3112-6928>), Martin Stoffel [aut], Hazel J. Nichols [aut], Joseph I. Hoffman [aut]

Maintainer Meinolf Ottensmann <meinolf.ottensmann@web.de>

Repository CRAN

Date/Publication 2020-08-26 08:00:07 UTC

R topics documented:

`aligned_peak_data` *Aligned Gas-Chromatography data*

Description

This is in example of an aligned gas-chromatography dataset processed with `align_chromatograms`. The raw data is accessible within this package as **peak_data.RData** and is comprised of 41 Mother-Pup pairs of Antarctic Fur Seals (*Arctocephalus gazella*) sampled from two different colonies at Bird Island, South Georgia. In addition two blanks are included.

Format

Object of class "GCalign" including three lists. List **"aligned"** includes data.frames for all variables present in the raw data ("time" and "area"). The list **"heatmap_input"** holds data frames with retention times of the input data, linearly adjusted retention times as well as the final output, were peaks are aligned among samples. This file is primarily used in `gc_heatmap`. The list **"Logfile"** summarises the alignment process and the data structure before, during and after running `align_chromatograms`. For a convenient overview use `print.GCalign`.

Source

http://www.pnas.org/content/suppl/2015/08/05/1506076112.DCSupplemental/pnas.1506076112.sd02.xlsx

References

Stoffel, M.A.; Caspers, B.A.; Forcada, J.; Giannakara, A.; Baier, M.; Eberhart-Phillips, L.; Mueller, C.; Hoffman, J.I. (2015): Chemical fingerprints encode mother-offspring similarity, colony membership, relatedness, and genetic quality in fur seals. In: Proceedings of the National Academy of Sciences of the United States of America 112 (36), S. E5005-12. DOI: 10.1073/pnas.1506076112.

`align_chromatograms` *Aligning peaks based on retention times*

Description

This is the core function of `GCalignR` to align peak data. The input data is a peak list. Read through the documentation below and take a look at the vignettes for a thorough introduction. Three parameters `max_linear_shift`, `max_diff_peak2mean` and `min_diff_peak2peak` are required as well as the column name of the peak retention time variable `rt_col_name`. Arguments are described among optional parameters below.

Usage

```
align_chromatograms(
  data,
  sep = "\t",
  rt_col_name = NULL,
  write_output = NULL,
  rt_cutoff_low = NULL,
  rt_cutoff_high = NULL,
  reference = NULL,
  max_linear_shift = 0.02,
  max_diff_peak2mean = 0.02,
  min_diff_peak2peak = 0.08,
  blanks = NULL,
  delete_single_peak = FALSE,
  remove_empty = FALSE,
  permute = TRUE,
  ...
)
```

Arguments

`data`	Dataset containing peaks that need to be aligned and matched. For every peak a arbitrary number of numerical variables can be included (e.g. peak height, peak area) in addition to the mandatory retention time. The standard format is a tab-delimited text file according to the following layout: (1) The first row contains sample names, the (2) second row column names of the corresponding peak lists. Starting with the third row, peak lists are included for every sample that needs to be incorporated in the dataset. Here, a peak list contains data for individual peaks in rows, whereas columns specify variables in the order given in

the second row of the text file. Peak lists of individual samples are concatenated horizontally and need to be of the same width (i.e. the same number of columns in consistent order). Alternatively, the input may be a list of data frames. Each data frame contains the peak data for a single individual. Variables (i.e.columns) are named consistently across data frames. The names of elements in the list are used as sample identifiers. Cells may be filled with numeric or integer values but no factors or characters are allowed. NA and 0 may be used to indicate empty rows.

`sep` The field separator character. The default is tab separated (`sep = '\t'`). See the "sep" argument in `read.table` for details.

`rt_col_name` A character giving the name of the column containing the retention times. The decimal separator needs to be a point.

`write_output` A character vector of variable names. For each variable a text file containing the aligned dataset is written to the working directory. Vector elements need to correspond to column names of data.

`rt_cutoff_low` A numeric value giving a retention time threshold. Peaks with retention time below the threshold are removed in a preprocessing step.

`rt_cutoff_high` A numeric value giving a retention time threshold. Peaks with retention time above the threshold are removed in a preprocessing step.

`reference` A character giving the name of sample from the dataset. By default, a sample is automatically selected from the dataset using the function `choose_optimal_reference`. The reference is used for the full alignment of peak lists by linear transformation.

`max_linear_shift`
Numeric value giving the window size considered in the full alignment. Usually, the amplitude of linear drift is small in typical GC-FID datasets. Therefore, the default value of 0.05 minutes is adequate for most datasets. Increase this value if the drift amplitude is larger.

`max_diff_peak2mean`
Numeric value defining the allowed deviation of the retention time of a given peak from the mean of the corresponding row (i.e. scored substance). This parameter reflects the retention time range in which peaks across samples are still matched as homologous peaks (i.e. substance). Peaks with retention times exceeding the threshold are sorted into a different row.

`min_diff_peak2peak`
Numeric value defining the expected minimum difference in retention times among homologous peaks (i.e. substance). Rows that differ less in the mean retention time, are therefore merged if every sample contains either one or none of the respective compounds. This parameter is a major determinant in the classification of distinct peaks. Therefore careful consideration is required to adjust this setting to your needs (e.g. the resolution of your gas-chromatography pipeline). Large values may cause to merge truly different substances with similar retention times, if those are not simultaneously occurring within at least one individual, which might occur by chance for small sample sizes. By default set to 0.2 minutes.

`blanks` Character vector of names of negative controls. Substances found in any of the blanks will be removed from the aligned dataset, before the blanks are deleted from the aligned data as well. This is an optional filtering step.

`delete_single_peak` Boolean, determining whether substances that occur in just one sample are removed or not.

`remove_empty` Boolean, allows to remove samples which lack any peak after the alignment finished. By default FALSE

`permute` Boolean, by default a random permutation of samples is conducted prior for each row-wise alignment step. Setting this parameter to FALSE causes alignment of the dataset as it is.

order of samples is constantly randomised during the alignment. Allows to prevent this behaviour for maximal repeatability if needed.

`...` optional arguments passed to methods, see `barplot`.

Details

This function aligns and matches homologous peaks across samples using a three-step algorithm based on user-defined parameters that are explained in the next section. In brief: **(1)** A full alignment of peak retention times is conducted to correct for systematic linear drift of retention times among homologous peaks from run to run. Thereby a coarse alignment is achieved that maximises the similarity of retention times across homologous peaks prior to a **(2)** partial alignment and matching of peaks. This and the next step in the alignment is based on a retention time matrix that contains all samples as columns and peak retention times in rows. The goal is to match homologous peaks within the same row that represents a chemical substance. Here, peaks are recognised as homologous when the retention time matches within a user-defined error span (see `max_diff_peak2mean`) that approximates the expected retention time uncertainty. Here, the position of every peak in the matrix is evaluated in comparison to similar peaks across the complete dataset and homologous peaks are gradually grouped together row by row. After all peaks were matched, a **(3)** adjacent rows of similar retention time are scanned to detect redundancies. A pair of rows is identified as redundant and merged if mean retention times are closer than specified by `min_diff_peak2peak` and single samples only contain peak in one but not both rows. Therefore, this step allows to match peaks that are associated with higher variance than allowed during the previous step. Several optional processing steps are available, ranging from the removal of peaks representing contaminations (requires to include blanks as a control) to the removal of uninformative peaks that are present in just one sample (so called singletons).

Value

Returns an object of class "GCalign" that is a a list containing several objects that are listed below. Note, that the objects "heatmap_input" and "Logfile" are best inspected by calling the provided functions `gc_heatmap` and `print`.

`aligned` Aligned Gas Chromatography peak data subdivided into individual data frames for every variable. Samples are represented by columns, rows specify homologous peaks. The first column of every data frame is comprised of the mean retention time of the respective peak (i.e. row). Retention times of samples resemble the values of the raw data. Internally, linear adjustments are considered where appropriate

heatmap_input	Used internally to create heatmaps of the aligned data
Logfile	A protocol of the alignment process.
input_list	Input data in form of a list of data frames.
aligned_list	Aligned data in form of a list of data frames.
input_matrix	List of matrices. Each matrix contains the input data for a variable

Author(s)

Martin Stoffel (martin.adam.stoffel@gmail.com) & Meinolf Ottensmann (meinolf.ottensmann@web.de)

Examples

```
## Load example dataset
data("peak_data")
## Subset for faster processing
peak_data <- peak_data[1:3]
peak_data <- lapply(peak_data, function(x) x[1:50,])
## align data with default settings
out <- align_chromatograms(peak_data, rt_col_name = "time")
```

---

as.data.frame.GCalign *Output aligned data in form of a data frame for each variable*

Description

Based on an object of class "GCalign" that was created using `align_chromatograms`, a list of data frames for each variable in the dataset is returned. Within data frames rows represent substances and columns are variables (i.e. substances).

Usage

```
## S3 method for class 'GCalign'
as.data.frame(x, row.names = NULL, optional = FALSE, ...)
```

Arguments

x	An object of class "GCalign". See `align_chromatograms` for details.
row.names	`NULL` or a character vector giving the row names for the data frame. Missing values are not allowed.
optional	logical. If `TRUE`, setting row names and converting column names (to syntactic names: see `make.names`) is optional. Note that all of R's **base** package `as.data.frame()` methods use `optional` only for column names treatment, basically with the meaning of `data.frame(*,check.names = !optional)`. See also the `make.names` argument of the `matrix` method.
...	additional arguments to be passed to or from methods.

Author(s)

Meinolf Ottensmann (meinolf.ottensmann@web.de) & Martin Stoffel (martin.adam.stoffel@gmail.com)

Examples

```
data("aligned_peak_data")
out <- as.data.frame(x = aligned_peak_data)
```

`check_input` *Check input prior to processing in GCalignR*

Description

Checks input files for common formatting problems.

Usage

```
check_input(data, plot = FALSE, sep = "\t", message = TRUE, ...)
```

Arguments

`data`	Dataset containing peaks that need to be aligned and matched. For every peak a arbitrary number of numerical variables can be included (e.g. peak height, peak area) in addition to the mandatory retention time. The standard format is a tab-delimited text file according to the following layout: (1) The first row contains sample names, the (2) second row column names of the corresponding peak lists. Starting with the third row, peak lists are included for every sample that needs to be incorporated in the dataset. Here, a peak list contains data for individual peaks in rows, whereas columns specify variables in the order given in the second row of the text file. Peak lists of individual samples are concatenated horizontally and need to be of the same width (i.e. the same number of columns in consistent order). Alternatively, the input may be a list of data frames. Each data frame contains the peak data for a single individual. Variables (i.e.columns) are named consistently across data frames. The names of elements in the list are used as sample identifiers. Cells may be filled with numeric or integer values but no factors or characters are allowed. NA and 0 may be used to indicate empty rows.
`plot`	Boolean specifying if the distribution of peak numbers is plotted.
`sep`	The field separator character. The default is tab separated (`sep = '\t'`). See the "sep" argument in `read.table` for details.
`message`	Boolean determining if passing all checks is indicated by a message.
`...`	optional arguments passed to methods, see `barplot`.

Details

Sample names should contain just letters, numbers and underscores and no whitespaces. Each sample has to contain the same number of columns, one of which is the retention time and the others are arbitrary variables in consistent order across samples. Retention times are expected to be numeric, i.e. they are only allowed to contain numbers from 0-9 and "." as the only decimal character. Have a look at the vignettes for examples.

Author(s)

Martin Stoffel (martin.adam.stoffel@gmail.com) & Meinolf Ottensmann (meinolf.ottensmann@web.de)

Examples

```
## gc-data
data("peak_data")
## Checks format
check_input(peak_data)
## Includes a barplot of peak numbers in the raw data
check_input(peak_data, plot = TRUE)
```

`choose_optimal_reference` *Select the optimal reference for full alignments of peak lists*

Description

Full alignments of peak lists require the specification of a fixed reference to which all other samples are aligned to. This function provides an simple algorithm to find the most suitable sample among a dataset. The so defined reference can be used for full alignments using `linear_transformation`. The functions is evoked internally by `align_chromatograms` if no reference was specified by the user.

Usage

```
choose_optimal_reference(data = NULL, rt_col_name = NULL, sep = "\t")
```

Arguments

`data`	Dataset containing peaks that need to be aligned and matched. For every peak a arbitrary number of numerical variables can be included (e.g. peak height, peak area) in addition to the mandatory retention time. The standard format is a tab-delimited text file according to the following layout: (1) The first row contains sample names, the (2) second row column names of the corresponding peak lists. Starting with the third row, peak lists are included for every sample that needs to be incorporated in the dataset. Here, a peak list contains data for individual peaks in rows, whereas columns specify variables in the order given in

the second row of the text file. Peak lists of individual samples are concatenated horizontally and need to be of the same width (i.e. the same number of columns in consistent order). Alternatively, the input may be a list of data frames. Each data frame contains the peak data for a single individual. Variables (i.e.columns) are named consistently across data frames. The names of elements in the list are used as sample identifiers. Cells may be filled with numeric or integer values but no factors or characters are allowed. NA and 0 may be used to indicate empty rows.

`rt_col_name` A character giving the name of the column containing the retention times. The decimal separator needs to be a point.

`sep` The field separator character. The default is tab separated (`sep = '\t'`). See the "sep" argument in `read.table` for details.

Details

Every sample is considered in determining the optimal reference in comparison to all other samples by estimating the similarity to all other samples. For a reference-sample pair, the deviation in retention times between all reference peaks and the always nearest peak in the sample is summed and divided by the number of reference peaks. The calculated value is a similarity score that converges to zero the more similar reference and sample are. For every potential reference, the median score of all pair-wise comparisons is used as a similarity proxy. The optimal sample is then defined by the minimum value among these scores. This functions is used internally in `align_chromatograms` to select a reference if non was specified by the user.

Value

A list with following elements

`sample` Name of the sample with the highest average similarity to all other samples

`score` Median number of shared peaks with other samples

Author(s)

Martin Stoffel (martin.adam.stoffel@gmail.com) & Meinolf Ottensmann (meinolf.ottensmann@web.de)

Examples

```
## 1.) input is a list
## using a list of samples
data("peak_data")
## subset for faster processing
peak_data <- peak_data[1:3]
choose_optimal_reference(peak_data, rt_col_name = "time")
```

`draw_chromatogram` *Visualise peak lists as a pseudo-chromatogram*

Description

Creates a graphical representation of one or multiple peak lists in the form of a pseudo- chromatogram. Peaks are represented by gaussian distributions centered at the peak retention time. The peak height is arbitrary and does not reflect any measured peak intensity.

Usage

```
draw_chromatogram(
  data = NULL,
  rt_col_name = NULL,
  width = 0.1,
  step = NULL,
  sep = "\t",
  breaks = NULL,
  rt_limits = NULL,
  samples = NULL,
  show_num = FALSE,
  show_rt = FALSE,
  plot = TRUE,
  shape = c("gaussian", "stick"),
  legend.position = "bottom"
)
```

Arguments

`data`	The input data can be either a GCalignR input file or an GCalign object. See `align_chromatograms` for details on both.
`rt_col_name`	A character giving the name of the column containing the retention times. The decimal separator needs to be a point.
`width`	Numeric value giving the standard deviation of gaussian peaks. Decrease this value to separate overlapping peaks within samples. Default is 0.01.
`step`	character allowing to visualise different steps of the alignment when a GCalign object is used. By default the aligned data is shown.
`sep`	The field separator character. The default is tab separated (`sep = '\t'`). See the "sep" argument in `read.table` for details.
`breaks`	A numeric vector giving the breakpoints between ticks on the x axis.
`rt_limits`	A numeric vector of length two giving min and max values or retention times to plot.
`samples`	A character vector of sample names to draw chromatograms of a subset.
`show_num`	Boolean indicating whether sample numbers are drawn on top of each peak.

show_rt	Boolean indicating whether peak retention times are drawn on top of each peak.
plot	Boolean indicating if the plot is printed.
shape	A character determining the shape of peaks. Peaks are approximated as "gaussian" by default. Alternatively, peaks can be visualised as "sticks".
legend.position	See theme for options of legend positions.

Details

Peaks from the are depicted as gaussian distributions. If the data is an "GCalign" object that was processed with align_chromatograms, chromatograms can be drawn for the dataset prior to alignment (**"input"**), after correcting linear drift (**"shifted"**) or after the complete alignment was conducted (**"aligned"**). In the latter case, retention times refer to the mean retention time of a homologous peaks scored among samples and do not reflect any between-sample variation anymore. Depending on the range of retention times and the distance among substances the peak width can be adjusted to enable a better visual separation of peaks by changing the value of parameter width. Note, homologous peaks (= exactly matching retention time) will overlap completely and only the last sample plotted will be visible. Hence, the number of samples can be printed on top of each peak. The function returns a list containing the ggplot object along with the internally used data frame to allow for maximum control in adapting the plot (see examples section in this document).

Value

A list containing the data frame created for plotting and the ggplot object. See ggplot.

Author(s)

Meinolf Ottensmann (meinolf.ottensmann@web.de) & Martin Stoffel (martin.adam.stoffel@gmail.com)

Examples

```
## load data
path <- (system.file("extdata", "simulated_peak_data.txt", package = "GCalignR"))
## run with defaults
x <- draw_chromatogram(data = path, rt_col_name = "rt")
## Customise and split samples in panels
x <- draw_chromatogram(data = path, rt_col_name = "rt", samples = c("A2","A4"),
 plot = FALSE, show_num = FALSE)
x[["ggplot"]] + ggplot2::facet_wrap(~ sample, nrow = 2)
## plot without numbers
x <- draw_chromatogram(data = path, show_num = FALSE, rt_col_name = "rt")
```

find_peaks *Detect local maxima in time series*

Description

Detects peaks in a vector and calculates the peak height. This function is only appropriate for symmetric gaussian peaks and does not take into account any baseline correction as it required in 'real word' data. Therefore, it does not substitute sophisticated peak detection and integration tools and is only used for illustration purposes in our vignettes.

Usage

```
find_peaks(df)
```

Arguments

df	A data frame containing x and y coordinates.

Value

A data frame containing x and y coordinates of peaks.

Author(s)

Meinolf Ottensmann (meinolf.ottensmann@web.de) & Martin Stoffel (martin.adam.stoffel@gmail.com)

Examples

```
## create df
df <- data.frame(x = 1:1000, y = dnorm(1:1000,300,20))
## plot
with(df, plot(x,y))
## detect peak
find_peaks(df)
```

GCalignR *GCalignR: A Package to Align Gas Chromatography Peaks Based on Retention Times*

Description

GCalignR contains the functions listed below. Follow the links to access the documentation of each function.

align_chromatograms executes all alignment steps.

as.data.frame.GCalign exports aligned data to data frames.

check_input tests the input data for formatting issues.

draw_chromatogram visualises peak lists in form of a chromatogram.

find_peaks detects and calculates peak heights in chromatograms. Not intended to be used for peak integration in empirical data. Used for illustration purposes only.

gc_heatmap visualises aligned datasets using heatmaps that can be customised.

norm_peaks allows to compute the relative abundance of peaks with samples.

peak_interspace gives a histogram of the distance between peaks within samples over the whole dataset.

read_peak_list reads the content of a text file and converts it to a list.

remove_blanks removes peaks resembling contaminations from aligned datasets.

remove_singletons removes peaks that are unique for one individual sample.

simple_chroma creates simple chromatograms for testing and illustration purposes.

Details

More details on the package are found in the vignettes that can be accessed via browseVignettes("GCalignR").

---

gc_heatmap *Visualises peak alignments in form of a heatmap*

Description

The goal of aligning peaks is to match homologous peaks that are thought to represent homologous substances in the same row across samples, although peaks have slightly different retention times across samples. This function makes it possible to evaluate the alignment quickly by inspecting the (i) distribution of peaks across samples, (ii) the variation for each homologous peak (column) as well as (iii) patterns that might hint at splitting peaks across rows. The mean retention time per homologous peak is here defined as the "true" retention time and deviations of individual peaks can be seen by a large deviation in the retention time to the mean. Subsetting of the retention time range (i.e. selecting peaks by the mean retention time) and samples (by name or by position) allow to quickly inspect regions of interest. Two types of heatmaps are available, a binary heatmap allows to determine if the retention time of single samples deviates by more than the user defined threshold from the mean. Optionally, a discrete heatmap allows to check deviations quantitatively. Large deviation can have multiple reasons. The most likely explanation is given by the fact that adjacent rows were merged as specified by the value min_diff_peak2peak in align_chromatograms. Here clear cases, in which peaks of multiple samples have been grouped in either of two or more rows can be merged and cause relatively high variation in peak retention times.

Usage

```
gc_heatmap(
  object = NULL,
  algorithm_step = c("aligned", "shifted", "input"),
  substance_subset = NULL,
  legend_type = c("legend", "colourbar"),
  samples_subset = NULL,
  type = c("binary", "discrete"),
  threshold = NULL,
  label_size = NULL,
  show_legend = TRUE,
  main_title = NULL,
  label = c("y", "xy", "x", "none")
)
```

Arguments

object	Object of class "GCalign", the output of a call to align_chromatograms.
algorithm_step	Character indicating which step of the algorithm is plotted. Either "input", "shifted" or "aligned" specifying the raw, linearly shifted or aligned data respectively. Default is the heatmap for the aligned dataset.
substance_subset	A vector of integers containing indices of substances in ascending order of retention times to plot.
legend_type	A character specifying how to present deviations of retention times from the mean. Either in form of discrete steps or on a gradient scale using 'legend' or 'colourbar' respectively. Changes are only possible when type = "discrete"
samples_subset	A vector indicating which samples are plotted on the heatmap by giving either indices or names of samples.
type	A character specifying whether a deviations of retention times are shown 'binary' (i.e. in comparison to the threshold value) or on a 'discrete' scale with respect to the mean retention time.
threshold	A numeric value denoting the threshold above which the deviation of individual peak retention times from the mean retention time of the respective substance are highlighted in heatmaps. By default, the value of parameter max_diff_peak2mean (see align_chromatograms) that was used in aligning the data is used.
label_size	An integer determining the size of labels on y and x axis. By default a fitting label_size is calculate (beta!) to compromise between readability and messiness due to a potentially large number of substances and samples.
show_legend	Boolean determining whether a legend is included or not.
main_title	Character giving the title of the heatmap. If not specified, titles are generated automatically.
label	Character determining if labels are shown on axes. Depending on the number of peaks and/or samples, labels are difficult to read. Use subsets instead. Possible option are "xy", "x", "y" or "none"

Value

object of class "ggplot"

Author(s)

Martin Stoffel (martin.adam.stoffel@gmail.com) & Meinolf Ottensmann (meinolf.ottensmann@web.de)

Examples

```
## aligned gc-dataset
data("aligned_peak_data")
## Default settings: The final output is plotted
gc_heatmap(aligned_peak_data, algorithm_step = "aligned")

## Plot the input data
gc_heatmap(aligned_peak_data,algorithm_step = "input")

## Plot a subset of the first 50 scored substances
gc_heatmap(aligned_peak_data,algorithm_step="aligned",substance_subset = 1:50)

## Plot specific samples, apply a stricter threshold
gc_heatmap(aligned_peak_data,samples_subset = c("M2","P7","M13","P13"),threshold = 0.02)
```

---

linear_transformation *Full Alignment of Peak Lists by linear retention time correction.*

Description

Shifts all peaks within samples to maximise the similarity to a reference sample. For optimal results, a sufficient number of shared peaks are required to find a optimal solution. A reference needs to be specified, for instance using `choose_optimal_reference`. Linear shifts are evaluated within a user-defined window in discrete steps. The highest similarity score defines the shift that will be applied. If more than a single shift step yields to the same similarity score, the smallest absolute value wins in order to avoid overcompensation. The functions is envoked internally by `align_chromatograms`.

Usage

```
linear_transformation(
  gc_peak_list,
  reference,
  max_linear_shift = 0.05,
  step_size = 0.01,
  rt_col_name,
  Logbook = NULL
)
```

Arguments

gc_peak_list	List of data.frames. Each data.frame contains GC-data (e.g. retention time, peak area, peak height) of one sample. Variables are stored in columns. Rows represent distinct peaks. Retention time is a required variable.
reference	A character giving the name of a sample included in the dataset. All samples are aligned to the reference.
max_linear_shift	Numeric value giving the window size considered in the full alignment. Usually, the amplitude of linear drift is small in typical GC-FID datasets. Therefore, the default value of 0.05 minutes is adequate for most datasets. Increase this value if the drift amplitude is larger.
step_size	Integer giving the step size in which linear shifts are evaluated between `max_linear_shift` and `-max_linear_shift`.
rt_col_name	A character giving the name of the column containing the retention times. The decimal separator needs to be a point.
Logbook	A list. If present, a summary of the applied linear shifts in full alignments of peak lists is appended to the list. If not specified, a list will be created automatically.

Details

A similarity score is calculated as the sum of deviations in retention times between all reference peaks and the closest peak in the sample. The principle idea is that the appropriate linear transformation will reduce the deviation in retention time between homologous peaks, whereas all other peaks should deviate randomly. Among all considered shifts, the minimum deviation score is selected for subsequent full alignment by shifting all peaks of the sample by the same value.

Value

A list containing two items.

chroma_aligned	List containing the transformed data
Logbook	Logbook, record of the applied shifts

Author(s)

Martin Stoffel (martin.adam.stoffel@gmail.com) & Meinolf Ottensmann (meinolf.ottensmann@web.de)

Examples

```
dat <- peak_data[1:10]
dat <- lapply(dat, function(x) x[1:50,])
x <- linear_transformation(gc_peak_list = dat, reference = "C2", rt_col_name = "time")
```

merge_redundant_rows *Merge redundant rows*

Description

Sometimes, redundant rows (i.e. groups of resembling a homologous peak) remain in an aligned dataset. This is the case when two or more adjacent rows exhibit a difference in the mean retention time that is greater than `min_diff_peak2peak`, the parameter that determines a threshold below that redundancy is checked within `align_chromatograms`. Therefore, this function allows to raise the threshold for a post processing step that groups the homologous peaks together without the need of repeating a potentially time-consuming alignment with adjusted parameters.

Usage

```
merge_redundant_rows(data, min_diff_peak2peak = NULL)
```

Arguments

`data`	An object of class "GCalign". See `align_chromatograms` for details.
`min_diff_peak2peak`	A numerical giving a threshold in minutes below which rows of similar retention time are checked for redundancy.

Details

Based on the value of parameter `threshold`, possibly redundant rows are identified by comparing mean retention times. Next, rows are checked for redundancy. When one or more samples contain peaks in a pair of compared rows, no redundancy is existent and the pair is skipped.

Value

a list of two items

`GCalign`	input data with updated input to `gc_heatmap`
`peak_list`	a list of data frames containing the updated dataset

Author(s)

Meinolf Ottensmann (meinolf.ottensmann@web.de) & Martin Stoffel (martin.adam.stoffel@gmail.com)

Examples

```
## Load example dataset
data("peak_data")
## Subset for faster processing
peak_data <- peak_data[1:3]
peak_data <- lapply(peak_data, function(x) x[1:50,])
## align data whith strict parameters
```

```
out <- align_chromatograms(peak_data, rt_col_name = "time",
max_diff_peak2mean = 0.01, min_diff_peak2peak = 0.02)
## relax threshold to merge redundant rows
out2 <- merge_redundant_rows(data = out, min_diff_peak2peak = 0.05)
```

`norm_peaks` *Normalisation of peak abundancies*

Description

Calculates the relative abundance of a peak by normalising an intensity measure with regard to the cumulative abundance of all peaks that are present within an individual sample. The desired measure of peak abundance needs to be included in a column of the input dataset aligned with `align_chromatograms`.

Usage

```
norm_peaks(
  data,
  conc_col_name = NULL,
  rt_col_name = NULL,
  out = c("data.frame", "list")
)
```

Arguments

`data`	Object of class GCalign created with `align_chromatograms` or a list of data frames that contain peak list of individual samples.
`conc_col_name`	Character giving the name of a column in `data` containing a variable describing the abundance of peaks (e.g. peak area or peak height).
`rt_col_name`	A character giving the name of the column containing the retention times. The decimal separator needs to be a point.
`out`	character defining the format of the returned data. Either "List" or "data.frame".

Value

Depending on `out` either a list of data frame or a single data frame were rows represent samples and columns relative peak abundances. Abundances are given as percentages.

@author Martin Stoffel (martin.adam.stoffel@gmail.com) & Meinolf Ottensmann (meinolf.ottensmann@web.de)

Examples

```
## aligned gc-dataset
data("aligned_peak_data")
## returns normalised peak area
norm_peaks(data = aligned_peak_data, conc_col_name = "area", rt_col_name = "time")
```

peak_data *Gas-chromatography data for Antarctic Fur Seals (*Arctocephalus gazella*)*

Description

This is an example of a typical gas-chromatography output file, listing a number of peaks with their respective retention times and abundance measures. Peaks were detected using Xcalibur 2.0.5 (Thermo Fisher Scientific). The data consists of 41 mother-pup pairs of two different colonies from Bird Island, South Georgia. In addition two blanks (i.e. negative controls) are included.

Format

A `list` of `data.frame`'s. Each `data.frame` contains gas-chromatography peak data of a single sample. The variables within each `data.frame` are: "time" (peak retention time) and "area" (integral of the peak curve). Each list element i.e. each `data.frame`is named with the unique sample ID.

Source

http://www.pnas.org/content/suppl/2015/08/05/1506076112.DCSupplemental/pnas.1506076112.sd02.xlsx

References

Stoffel, M.A.; Caspers, B.A.; Forcada, J.; Giannakara, A.; Baier, M.; Eberhart-Phillips, L.; Mueller, C.; Hoffman, J.I. (2015): Chemical fingerprints encode mother-offspring similarity, colony membership, relatedness, and genetic quality in fur seals. In: Proceedings of the National Academy of Sciences of the United States of America 112 (36), S. E5005-12. DOI: 10.1073/pnas.1506076112.

peak_factors *Grouping factors corresponding to gas-chromatography data of Antarctic Fur Seals (*Arctocephalus gazella*)*

Description

List of factors corresponding to samples in `peak_data`

Format

A data frame where columns represent factors, rows are samples.

Source

http://www.pnas.org/content/suppl/2015/08/05/1506076112.DCSupplemental/pnas.1506076112.sd02.xlsx

References

Stoffel, M.A.; Caspers, B.A.; Forcada, J.; Giannakara, A.; Baier, M.; Eberhart-Phillips, L.; Mueller, C.; Hoffman, J.I. (2015): Chemical fingerprints encode mother-offspring similarity, colony membership, relatedness, and genetic quality in fur seals. In: Proceedings of the National Academy of Sciences of the United States of America 112 (36), S. E5005-12. DOI: 10.1073/pnas.1506076112.

`peak_interspace` *Estimate the observed space between peaks within chromatograms*

Description

The parameter `min_diff_peak2peak` is a major determinant in the alignment of a dataset with `align_chromatograms`. This function helps to infer a suitable value based on the input data. The underlying assumption here is that distinct peaks within a separated by a larger gap than homologous peaks across samples. Tightly spaced peaks within a sample will appear on the left side of the plotted distribution and can indicate the presence of split peaks in the data.

Usage

```
peak_interspace(
  data,
  rt_col_name = NULL,
  sep = "\t",
  quantiles = NULL,
  quantile_range = c(0, 1),
  by_sample = FALSE
)
```

Arguments

`data`	Dataset containing peaks that need to be aligned and matched. For every peak a arbitrary number of numerical variables can be included (e.g. peak height, peak area) in addition to the mandatory retention time. The standard format is a tab-delimited text file according to the following layout: (1) The first row contains sample names, the (2) second row column names of the corresponding peak lists. Starting with the third row, peak lists are included for every sample that needs to be incorporated in the dataset. Here, a peak list contains data for individual peaks in rows, whereas columns specify variables in the order given in the second row of the text file. Peak lists of individual samples are concatenated horizontally and need to be of the same width (i.e. the same number of columns in consistent order). Alternatively, the input may be a list of data frames. Each data frame contains the peak data for a single individual. Variables (i.e.columns) are named consistently across data frames. The names of elements in the list are used as sample identifiers. Cells may be filled with numeric or integer values but no factors or characters are allowed. NA and 0 may be used to indicate empty rows.

rt_col_name	A character giving the name of the column containing the retention times. The decimal separator needs to be a point.
sep	The field separator character. The default is tab separated (sep = '\t'). See the "sep" argument in read.table for details.
quantiles	A numeric vector. Specified quantiles are calculated from the distribution.
quantile_range	A numeric vector of length two that allows to subset an arbitrary interquartile range.
by_sample	A logical that allows to calculate peak interspaces individually for each sample. By default all samples are combined to give the global distribution of next-peak differences in retention times. When by_sample = TRUE, a series of plots (one for each sample) is created and a keystroke is required to proceed.

Value

List containing summary statistics of the peak interspace distribution

Author(s)

Martin Stoffel (martin.adam.stoffel@gmail.com) & Meinolf Ottensmann (meinolf.ottensmann@web.de)

Examples

```
## plotting with defaults
peak_interspace(data = peak_data, rt_col_name = "time")
## plotting up to the 0.95 quantile
peak_interspace(data = peak_data,rt_col_name = "time",quantile_range = c(0,0.95))
## return the 0.1 quantile
peak_interspace(data = peak_data,rt_col_name = "time", quantiles = 0.1)
```

---

plot.GCalign *Plot diagnostics for an GCalign Object*

Description

Visualises the aligned data based on four diagnostic plots. One plot shows the distribution of peak numbers per sample in the raw data and after alignment. A second plot gives the distribution of linear shifts that were applied in order to conduct a full alignment of samples with respect to reference. A third sample gives a distribution of the variation in retention times of homologous peaks. The fourth plot shows a frequency distribution of peaks shared among samples.

Usage

```
## S3 method for class 'GCalign'
plot(
  x,
  which_plot = c("all", "shifts", "variation", "peak_numbers", "peaks_shared"),
  ...
)
```

Arguments

x	Object of class GCalign, created with align_chromatograms
which_plot	A character defining which plot is created. Options are "shifts", "variation", "peak_numbers" and "peaks_shared". By default all four are created.
...	Optional arguments passed on to methods. See plot, hist and barplot. Note that optional arguments are not passed on when plotting all figures.

Value

Depending on the selected plot a data frame containing the data source of the respective plot is returned. If all plots are created, no output will be returned.

Author(s)

Martin Stoffel (martin.adam.stoffel@gmail.com) & Meinolf Ottensmann (meinolf.ottensmann@web.de)

Examples

```
## GCalign object
data("aligned_peak_data")

## All plots are shown by default
plot(aligned_peak_data)

## Distribution of peak numbers
plot(aligned_peak_data, which_plot = "peak_numbers")

## variation of retention times
plot(aligned_peak_data, which_plot = "variation")
```

---

print.GCalign *Summarising Peak Alignments with GCalignR*

Description

print method for class "GCalign"

Usage

```
## S3 method for class 'GCalign'
print(x, write_text_file = FALSE, ...)
```

Arguments

x	Object of class GCalign, created with align_chromatograms
write_text_file	A boolean allowing to write a text file.
...	Optional arguments passed on to methods are currently not supported.

Author(s)

Martin Stoffel (martin.adam.stoffel@gmail.com) & Meinolf Ottensmann (meinolf.ottensmann@web.de)

Examples

```
## GCalign object
data("aligned_peak_data")
## print summary
print(aligned_peak_data)
```

`read_empower2` *Import data from single EMPOWER2 HPLC files*

Description

reads output files of the EMPOWER 2 SOFTWARE (Waters). Input files must contain data of single samples deposited within the same directory.

Usage

```
read_empower2(
  path = NULL,
  pattern = ".txt",
  sep = "\t",
  skip = 2,
  id = "SampleName"
)
```

Arguments

`path`	path to a folder containing input files
`pattern`	pattern used to select files. By default ".txt"
`sep`	The field separator character. The default is tab separated (`sep = '\t'`). See the "sep" argument in `read.table` for details.
`skip`	rows to skip before reading data
`id`	column containing sample name

Value

a list of data frames (each corresponding to a sample)

read_peak_list *Read content of a text file and convert it to a list*

Description

Reads the content of text file that is formatted as described in `align_chromatograms` and converts it to a list.

Usage

```
read_peak_list(data, sep = "\t", rt_col_name, check = T)
```

Arguments

`data`	A text file containing a peak list. See `align_chromatograms` for details.
`sep`	The field separator character. The default is tab separated (`sep = '\t'`). See the "sep" argument in `read.table` for details.
`rt_col_name`	A character giving the name of the column containing the retention times. The decimal separator needs to be a point.
`check`	logical

Value

A list of data frames containing peak data for every sample of `data`.

Author(s)

Meinolf Ottensmann (meinolf.ottensmann@web.de) & Martin Stoffel (martin.adam.stoffel@gmail.com)

Examples

```
path <- system.file("extdata", "simulated_peak_data.txt", package = "GCalignR")
x <- read_peak_list(data = path, rt_col_name = "rt")
```

remove_blanks *Remove peaks present in negative control samples*

Description

Removes peaks that are present in blanks (i.e. negative control samples) to eliminate contaminations in the aligned data. Afterwards, blanks are deleted itself. This function is only applicable when blanks were not discarded during a previous alignment using `align_chromatograms`.

Usage

```
remove_blanks(data, blanks)
```

Arguments

data	An object of class "GCalign". See align_chromatograms for details. Alternatively, a list of data frames. Whereby each data frame contains the peak list for an individual sample.
blanks	Character vector of names of negative controls. Substances found in any of the blanks will be removed from the aligned dataset, before the blanks are deleted from the aligned data as well. This is an optional filtering step.

Value

a list of data frames for each individual.

Author(s)

Meinolf Ottensmann (meinolf.ottensmann@web.de) & Martin Stoffel (martin.adam.stoffel@gmail.com)

Examples

```
data("peak_data")
## subset for faster processing
data <- lapply(peak_data[1:5], function(x) x[20:35,])
x <- align_chromatograms(data, rt_col_name = "time")
out <- remove_blanks(data = x, blanks = c("C2","C3"))
## number of deleted peaks
nrow(x[["aligned_list"]][["M2"]]) - nrow(out[["M2"]])
```

---

remove_singletons *Remove singletons*

Description

Identifies and removes singletons (i.e. peaks that are unique for one sample) from the aligned dataset.

Usage

```
remove_singletons(data)
```

Arguments

data	An object of class "GCalign". See align_chromatograms for details. Alternatively, a list of data frames. Whereby each data frame contains the peak list for an individual sample.

Value

a list of data frames for each individual.

Author(s)

Meinolf Ottensmann (meinolf.ottensmann@web.de) & Martin Stoffel (martin.adam.stoffel@gmail.com)

Examples

```
data("peak_data")
## subset for faster processing
data <- lapply(peak_data[1:5], function(x) x[20:35,])
x <- align_chromatograms(data, rt_col_name = "time")
out <- remove_singletons(data = x)
```

---

simple_chroma *Simulate simple chromatograms*

Description

Creates chromatograms with user defined peaks for illustrative purposes. Linear drift is applied in sample order if more than one sample is created. See parameters of the function.

Usage

```
simple_chroma(
  peaks = c(10, 13, 25, 37, 50),
  N = 1,
  min = 0,
  max = 30,
  Names = NULL,
  sd = NULL
)
```

Arguments

peaks	A numeric vector giving the retention times on which gaussian distribution, defining peaks, are centered. If more than one sample is generated N > 1, peaks defines a population of peaks, from which samples are generated.
N	An integer giving the number of chromatograms to create. By default N = 1.
min	A numeric giving the minimum retention time.
max	A numeric giving the maximum retention time.
Names	A character vector giving sample names. If not specified, names are generated automatically.
sd	A numeric vector of the same length as peaks giving the standard deviation of each peak. Only supported if N = 1.

Value

A data frame containing x and y coordinates and corresponding sample names.

Author(s)

Meinolf Ottensmann (meinolf.ottensmann@web.de) & Martin Stoffel (martin.adam.stoffel@gmail.com)

Examples

```
## create a chromatogram
x <- simple_chroma(peaks = c(5,10,15), N = 1, min = 0, max = 30, Names = "MyChroma")
## plot chromatogram
with(x, plot(x,y, xlab = "time", ylab = "intensity"))
```

Index