Supplementary for "GCalignR: An R package for aligning Gas-Chromatography data"

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

2016-12-01

Published alignment tools for gas-chromatography

There is a small number of alignment procedures that are available to handle gas-chromatography data *without* the aid of additional information about mass spectra which require the use of GC-MS as analytical pathway.

Table 1: Peer-reviewed alignment tools for gas-chromatography data. Note: Unlisted are programs using mass-spectra of GC-MS runs

Program	Availability	Platform	Visualisation	Limitations	Year	Source
GCALIGNER 1.0	freeware	Java	None	Last sample remains unaligned	2013	{Dellicour 2013 #8}

Testing GCalignR with published data

Dellicour and Lecocq (2013) present data for three North America bumble bee species *Bombus bimaculatus*, *B. ephippiatus* and *B. flavifrons*. Samples represent cephalic labial gland secretions and are supposed to show species specific patterns. Hence, this in an ideal data set to test both (i) the alignment efficiency of **GCalignR** and (ii) the functionality to explore similarity patterns by multidimensional scaling within one pipeline in **R**.

```
library(GCalignR)
check_input(data = "data/d1/Table_S1_raw.txt")
#> Warning: BEPH06 violate(s) the requirements.
#> Warning: Every sample needs to have the same number of values for each
#> variable!
```

Not all checks have been passed. Read warning messages and change data accordingly

Sample *BEPH06* is malformed. The last substance has not retention time and needs to be exluded from the data set. This is an error in the supporting information of the paper.

```
check_input(data = "data/d1/Table_S1_cleaned.txt")
```

All checks passed! Ready for processing with align_chromatograms

```
aligned <- align_chromatograms(data = "data/d1/Table_S1_raw.txt",
                    conc_col_name = "Area",
                    max_diff_peak2mean = 0.03,
                    min_diff_peak2peak = 0.1,
                    rt_col_name = "RT",
                    delete_single_peak = F,
                    merge_rare_peaks = F)
#> Warning: BEPH06 violate(s) the requirements.
```

```
#> Warning: Every sample needs to have the same number of values for each
#> variable!
```

Not all checks have been passed. Read warning messages and change data accordinglyRun GCalignR Start: 16:38:40

GC-data for 55 samples loaded

A reference was not specified. Hence 'BBIM03' was selected on the basis of highest average similarity to all samples (score = 18)

Start Linear Transformation with "BBIM03" as a reference ... Done

Start Alignment of Peaks ... This might take a while!

Do you know how well birds can see? The Northern Hawk Owl (Surnia ulula) can detect primarily by sight a vole to eat up to a half a mile away.

Iteration 1 out of 1 ...

Merged Redundant Peaks

Peak Alignment Done

Alignment was Successful! Time: 16:41:11

```
save(aligned, file = "data/d1/Table_S1_aligned.RData")
```

```
# aligned data
load(file = "data/d1/Table_S1_aligned.RData")
# factors
factors <- read.csv("data/d1/Table_S1_factors.csv",sep = ";")
row.names(factors) <- factors[["ID"]]
```

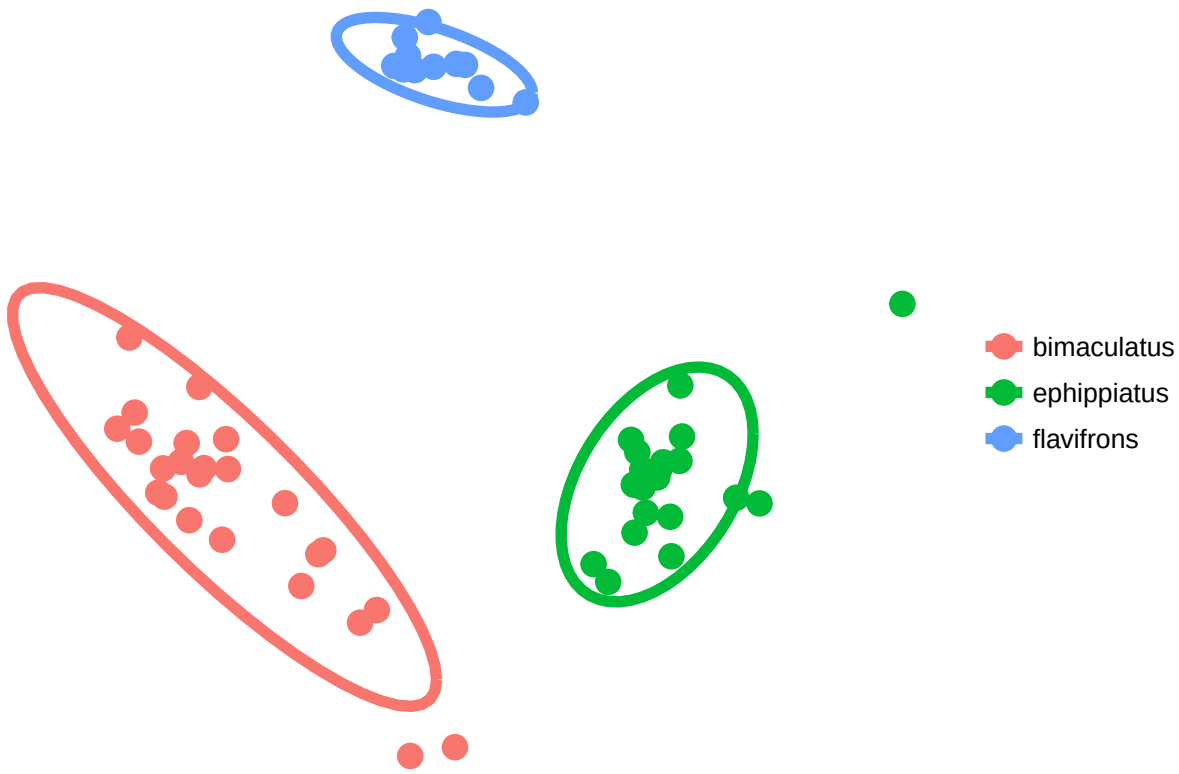
Dellicour and Lecocq (2013) validated the alignment of their tool by GC-MS. For *B. flavifrons* they report a low error rate of 0.3 %. Hence, this data set is a good source to explore acceptable variation among retention times that are mapped to the same substance.

```
t2 <- read.csv("data/d1/Table_S2_modified.txt", skip = 1, sep = "\t", header = FALSE)
# get the retention times
t2 <- t2[3:61,seq(1,31,3)]
t2 <- apply(t2, MARGIN = 2, as.numeric)
t2 <- data.frame(rt = rowMeans(t2, na.rm = T), var = apply(t2, 1, var, na.rm = T), range = apply(t2,1,
```

NMDS

```
scent <- norm_peaks(aligned,conc_col_name = "Area",rt_col_name = "RT",out = "data.frame")
scent <- log(scent + 1)
library(vegan)
#> Loading required package: permute
#> Loading required package: lattice
#> This is vegan 2.4-1
scent <- scent[match(row.names(factors),row.names(scent)),]
scent_nmds <- vegan::metaMDS(comm = scent)
#> Run 0 stress 0.1556261
#> Run 1 stress 0.1427999
#> ... New best solution
#> ... Procrustes: rmse 0.05323072  max resid 0.2428411
#> Run 2 stress 0.1429259
#> ... Procrustes: rmse 0.01092181  max resid 0.07636847
```

```
#> Run 3 stress 0.1427999
#> ... Procrustes: rmse 2.834907e-05  max resid 0.0001252045
#> ... Similar to previous best
#> Run 4 stress 0.1556259
#> Run 5 stress 0.2688805
#> Run 6 stress 0.1556259
#> Run 7 stress 0.1519242
#> Run 8 stress 0.2483444
#> Run 9 stress 0.2479965
#> Run 10 stress 0.232269
#> Run 11 stress 0.2653986
#> Run 12 stress 0.2667564
#> Run 13 stress 0.1556266
#> Run 14 stress 0.1447737
#> Run 15 stress 0.1519241
#> Run 16 stress 0.2349557
#> Run 17 stress 0.1556259
#> Run 18 stress 0.1445487
#> Run 19 stress 0.1523269
#> Run 20 stress 0.2339989
#> *** Solution reached
scent_nmds <- as.data.frame(scent_nmds$points)
scent_nmds <- cbind(scent_nmds,Species = factors[["Species"]])
ggplot2::ggplot(data = scent_nmds,ggplot2::aes(MDS1,MDS2,color = Species)) +
    ggplot2::geom_point(size = 4) + ggplot2::stat_ellipse(size = 2) + ggplot2::labs(title = "", x = "MDS
```



References

Dellicour, Simon, and Thomas Lecocq. 2013. "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): 3206–9. doi:10.1002/jssc.201300388.