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

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

2016-12-12

```
## Loading required package: permute
## Loading required package: lattice
## This is vegan 2.4-1
```

Testing GCalignR on empirical data

Data

We demonstrate the usage of `GcalignR` with two empirical data sets. A more detailed workthrough tutorial is given in the vignette.

Bumble bee cephalic secretions

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**.

```
# The data is comprised of 55 samples, distributed as follows:
bee_factors <- read.csv("data/d1/Table_S1_factors.csv",sep = ";")
row.names(bee_factors) <- bee_factors[["ID"]]
pandoc.table(summary(bee_factors$Species))
```

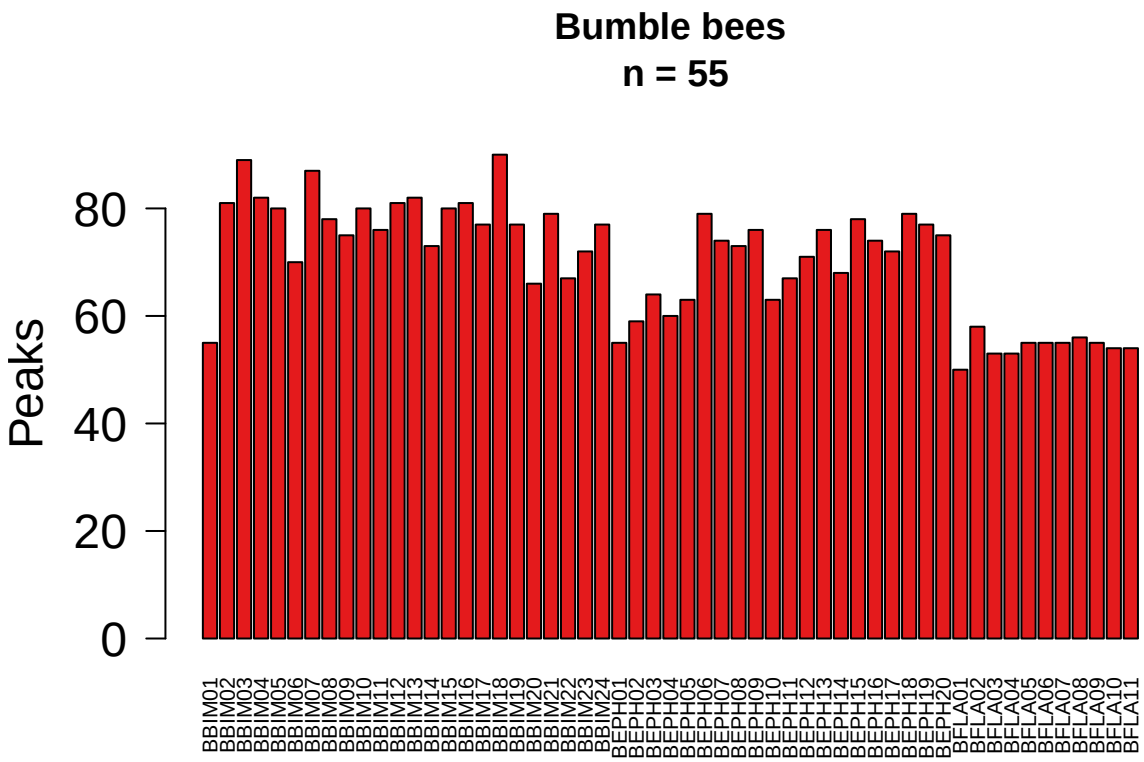
bimaculatus	ephippiatus	flavifrons
24	20	11

The chromatogram data was extracted from Table S1 of (Dellicour and Lecocq 2013) and can be downloaded here: {http://onlinelibrary.wiley.com/store/10.1002/jssc.201300388/asset/supinfo/jssc3437-sup-0001-TableS1.zip?v=1&s=57d5d58273d1d4207e70c72cecd5bba4b1fe95a1}{Supporting information}. Prior to excecuting any alignment, the conformity of the data with the requirements of `GCalignR`is tested "behind the scenes". Nonetheless, this can be invoked manually by calling the function `check_input`.

```
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
```

The output reveals that sample *BEPH06* is malformed. The last row contains both area and relative area but no retention time. It is unclear what this represents. The respective row is removed from the file we are going to use afterwards.

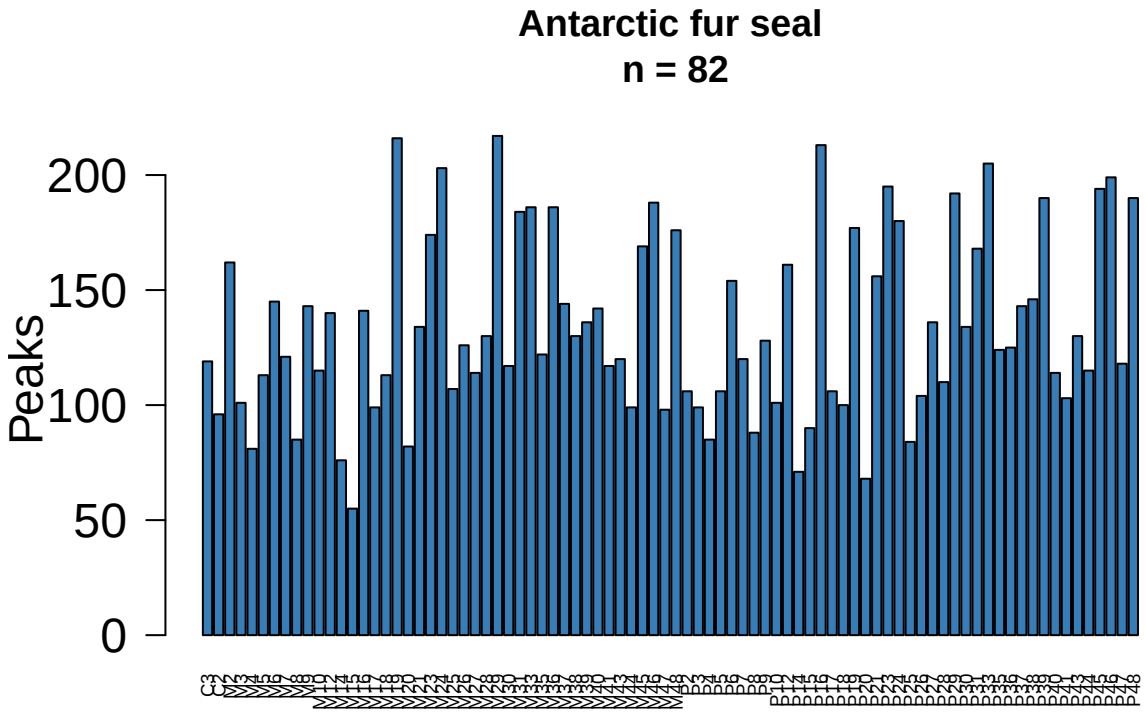
```
# By including "list_peaks = T" we can plot the peak distribution
# prior to alignment.
check_input(data = "data/d1/Table_S1_cleaned.txt", list_peaks = T,
            ylab = "Peaks", cex.names = 0.7, col = "#E41A1C",
            main = "Bumble bees\n n = 55")
#> All checks passed!
#> Ready for processing with align_chromatograms
```



Antarctic fur seal *Arctocephalus gazella* skin swabs

The second data set is comprised of skin swabs of 41 mother-pup pairs of Antarctic fur seals *Arctocephalus gazella* which among other things were shown to encode the membership to a breeding colony (Stoffel et al. 2015).

```
check_input(data = peak_data, list_peaks = T, cex.names = 0.6,
            ylab = "Peaks", col = "#377EB8", main = "Antarctic fur seal\n n = 82")
#> All checks passed!
#> Ready for processing with align_chromatograms
```



Alignment

- Aligning the bumblee bee data

```
bee_aligned <- align_chromatograms(data = "data/d1/Table_S1_cleaned.txt",
                    conc_col_name = "Area",
                    max_diff_peak2mean = 0.02,
                    min_diff_peak2peak = 0.05,
                    rt_col_name = "RT",
                    delete_single_peak = T,
                    iterations = 1)
save(bee_aligned, file = "data/d1/bee_aligned.RData")

# aligned data
load(file = "data/d1/bee_aligned.RData")
```

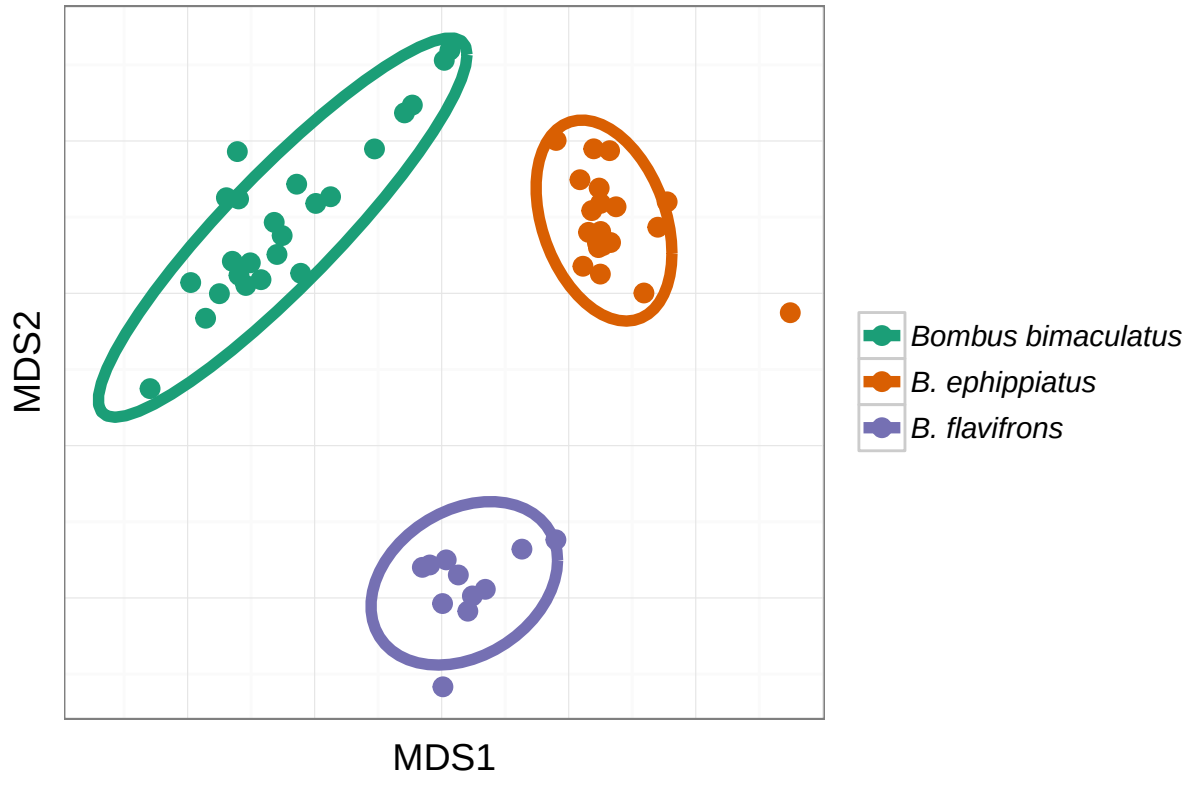
- Exploring species similarites with non-metric multidimensional scaling (NMDS)

```
# normalise abundancies within samples
bee_scent <- GCalignR::norm_peaks(bee_aligned,conc_col_name = "Area",rt_col_name = "RT",out = "data.fra
# Log + 1 Transformation
bee_scent <- log(bee_scent + 1)
bee_scent <- bee_scent[match(row.names(bee_factors),row.names(bee_scent)),]
# NMDS using bray-curtis in vegan
bee_scent_nmds <- vegan::metaMDS(comm = bee_scent,trymax = 9999)
# Get the coordinates
```

```
bee_scent_nmds <- as.data.frame(bee_scent_nmds$points)
bee_scent_nmds <- cbind(bee_scent_nmds,Species = bee_factors[["Species"]])
```

- Visualisation using ggplot2

```
ggplot2::ggplot(data = bee_scent_nmds,ggplot2::aes(MDS1,MDS2,color = Species)) +
ggplot2::geom_point(size = 3) +
ggplot2::stat_ellipse(size = 2) +
ggplot2::labs(title = "", x = "MDS1", y = "MDS2") +
ggplot2::theme_bw(base_size = 14) +
ggplot2::theme(axis.ticks = element_blank(), axis.text = element_blank()) +
    scale_colour_manual(values = RColorBrewer::brewer.pal(3,"Dark2"),
    name = "",
    breaks = c("bimaculatus", "ephippiatus", "flavifrons"),
    labels = c("Bombus bimaculatus","B. ephippiatus","B. flavifrons"),
    guide = guide_legend(label.theme = element_text(
    face = "italic", angle = 0, size = 11)))
```



* Multivariate statistics using `adonis` reveal a highly significant clustering by species

```
vegan::adonis(bee_scent ~ bee_factors$Species,permutations = 999)
#>
#> Call:
#> vegan::adonis(formula = bee_scent ~ bee_factors$Species, permutations = 999)
#>
#> Permutation: free
#> Number of permutations: 999
#>
```

```
#> Terms added sequentially (first to last)
#>
#>                     Df SumsOfSqs MeanSqs F.Model      R2 Pr(>F)
#> bee_factors$Species  2    7.0089  3.5045   32.73 0.55729  0.001 ***
#> Residuals           52    5.5678  0.1071         0.44271
#> Total               54   12.5768                 1.00000
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

R version and platform.

```
sessionInfo()
#> R version 3.3.2 (2016-10-31)
#> Platform: x86_64-w64-mingw32/x64 (64-bit)
#> Running under: Windows 10 x64 (build 14393)
#>
#> locale:
#> [1] LC_COLLATE=German_Germany.1252  LC_CTYPE=German_Germany.1252
#> [3] LC_MONETARY=German_Germany.1252 LC_NUMERIC=C
#> [5] LC_TIME=German_Germany.1252
#>
#> attached base packages:
#> [1] stats     graphics  grDevices utils     datasets  methods   base
#>
#> other attached packages:
#> [1] vegan_2.4-1         lattice_0.20-33     permute_0.9-0
#> [4] ggplot2_2.1.0       pander_0.6.0        bibtex_0.4.0
#> [7] knitcitations_1.0-2 GCalignR_0.0.9000
#>
#> loaded via a namespace (and not attached):
#>  [1] Rcpp_0.12.4         formatR_1.4         RColorBrewer_1.1-2
#>  [4] plyr_1.8.3          bitops_1.0-6        tools_3.3.2
#>  [7] digest_0.6.10       lubridate_1.6.0     evaluate_0.9
#> [10] tibble_1.1          gtable_0.2.0        nlme_3.1-127
#> [13] mgcv_1.8-12         Matrix_1.2-6        yaml_2.1.13
#> [16] parallel_3.3.2      RefManageR_0.13.1   stringr_1.1.0
#> [19] httr_1.2.1          knitr_1.14          cluster_2.0.4
#> [22] grid_3.3.2          R6_2.1.2            XML_3.98-1.5
#> [25] rmarkdown_1.1       RJSONIO_1.3-0       reshape2_1.4.2
#> [28] readr_1.0.0         magrittr_1.5        scales_0.4.0
#> [31] htmltools_0.3.5     MASS_7.3-45         assertthat_0.1
#> [34] colorspace_1.2-6    labeling_0.3        stringi_1.1.1
#> [37] RCurl_1.95-4.8      munsell_0.4.3
```

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.

Stoffel, Martin A., Barbara A. Caspers, Jaume Forcada, Athina Giannakara, Markus Baier, Luke Eberhart-Phillips, Caroline Müller, and Joseph I. Hoffman. 2015. "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.