An example for using of goSorensensen R-Package

Abstract

This document provides an example for the manipulation of `goSorensen` R Package, using real feature lists obtained from external files

Starting from a fast introduction about the installation of package and description of the data, we explain how to: i) perform the equivalence test from contingency tables of joint enrichment or directly from features lists (either using a normal asymptotic or a bootstrap approximation), ii) collect specific fields of the test results like the p-value, the upper limit of the confidence interval or standard errors iii) obtain another statistics related to the Sorensen-Dice dissimilarity, iv) compute new results for updated input values (i.e. confidence level, irrelevance limit, used distribution, etc) and v) perform all pairwise computes

This document is based on the vignette "*An introduction to package goSorensen*" available on the `goSorensen` R-Package

Installation

goSorensen package has to be installed with a working R version `(R >= 4.2.0)`. Installation could take a few minutes on a regular desktop or laptop. Package can be installed from Bioconductor or `devtools` package, then it needs to be loaded using `library(goSorensen)`

To install from Bioconductor (recommended):

```
## Only if BiocManager is not previosly installed:
install.packages("BiocManager")

## otherwise, directly:
BiocManager::install("goSorensen")
```

To install from Github

```
devtools::install_github("pablof1988/goSorensen", build_vignettes = TRUE)
```

Data.

We use the external data file "c5.go.bp.v2022.1.Hs.entrez.gmt" obtained from https://www.gsea-msigdb.org/gsea/msigdb/human/collections.jsp#C5 section C5: ontology gene sets, BP: subset of GO.

First step is import the external file with the list of genes:

```
library(qusage)
dataBPGO <- read.gmt("c5.go.bp.v2022.1.Hs.entrez.gmt")
```

This data set contain 7763 lists of genes. So, only the six largest (more than 1850 gene identifiers) are taken into account:

```
filtBPGO <- NULL
for(i in 1: length(dataBPGO)){
  filtBPGO[i] <- length(dataBPGO[[i]])
}
BPGOGeneList <- dataBPGO[filtBPGO >= 1850]
```

Name of the selected gene lists to be compared are:

```
names(BPGOGeneList)
```

```
## [1] "GOBP_APOPTOTIC_PROCESS"
## [2] "GOBP_IMMUNE_RESPONSE"
## [3] "GOBP_TISSUE_DEVELOPMENT"
## [4] "GOBP_POSITIVE_REGULATION_OF_MACROMOLECULE_BIOSYNTHETIC_PROCESS"
## [5] "GOBP_PHOSPHORYLATION"
## [6] "GOBP_CELLULAR_RESPONSE_TO_STRESS"
```

Performing the equivalence test

Building 2 x 2 contingency table of mutual enrichment

It is possible to build the enrichment contingency table between gene lists (for example "GOBP_APOPTOTIC_PROCESS" and "GOBP_PHOSPHORYLATION") for an specific Ontology and GO level (for example Ontology BP and GO level 5

```
library(goSorensen)
data("humanEntrezIDs") # necessary to recognize features in ENTREZ id's
enrichTab <- buildEnrichTable(BPGOGeneList$GOBP_APOPTOTIC_PROCESS,
                              BPGOGeneList$GOBP_PHOSPHORYLATION,
                              geneUniverse = humanEntrezIDs,
                              orgPackg = "org.Hs.eg.db",
                              onto = "BP", GOLevel = 5,
                              listNames = c("APOPTOTIC_PROCESS",
                                            "PHOSPHORYLATION"))
enrichTab
```

```
##                               Enriched in PHOSPHORYLATION
## Enriched in APOPTOTIC_PROCESS TRUE FALSE
##                          TRUE  1306   311
##                          FALSE  188  8343
```

Equivalence test

Equivalence test from contingency table

Equivalence test from contingency table for an equivalence (or irrelevance) limit $d_0 = 0.2857$ and a significance level $\alpha = 0.05$

```
contin_testResult <- equivTestSorensen(enrichTab, d0 = 0.2857,
                                       conf.level = 0.95)
contin_testResult
```

```
##
##  Normal asymptotic test for 2x2 contingency tables based on the
##  Sorensen-Dice dissimilarity
##
## data:  enrichTab
## (d - d0) / se = -17.674, p-value < 2.2e-16
## alternative hypothesis: true equivalence limit d0 is less than 0.2857
## 95 percent confidence interval:
##  0.0000000 0.1720596
## sample estimates:
## Sorensen dissimilarity
##              0.1603986
## attr(,"se")
## standard error
##    0.007089421
```

This process is using by default the normal approximation to the sample distribution of the $(\widehat{d} - d)/\widehat{se}$ statistic. Alternatively, it is possible to estimate this distribution by means of bootstrap:

```
boot_testResult <- equivTestSorensen(enrichTab, d0 = 0.2857,
                                       conf.level = 0.95, boot = T)
boot_testResult
```

```
##
##  Bootstrap test for 2x2 contingency tables based on the Sorensen-Dice
##  dissimilarity (10000 bootstrap replicates)
##
## data:  enrichTab
## (d - d0) / se = -17.674, p-value = 9.999e-05
## alternative hypothesis: true equivalence limit d0 is less than 0.2857
## 95 percent confidence interval:
##  0.0000000 0.1723683
## sample estimates:
## Sorensen dissimilarity
##              0.1603986
## attr(,"se")
## standard error
##    0.007089421
```

For low frequencies in the contingency table, bootstrap is a more conservative but preferable approach, with better type I error control.

Equivalence test from feature lists.

Obtain the test directly from gene lists is also possible:

```
lists_testResult <- equivTestSorensen(BPGOGeneList$GOBP_APOPTOTIC_PROCESS,
                                      BPGOGeneList$GOBP_PHOSPHORYLATION,
                                      d0 = 0.2857,
                                      geneUniverse = humanEntrezIDs,
                                      orgPackg = "org.Hs.eg.db",
                                      onto = "BP", GOLevel = 5,
                                      listNames = c("APOPTOTIC_PROCESS",
                                                    "PHOSPHORYLATION"))
lists_testResult
```

```
##
##  Normal asymptotic test for 2x2 contingency tables based on the
##  Sorensen-Dice dissimilarity
##
## data:  tab
## (d - d0) / se = -17.674, p-value < 2.2e-16
## alternative hypothesis: true equivalence limit d0 is less than 0.2857
## 95 percent confidence interval:
##  0.0000000 0.1720596
## sample estimates:
## Sorensen dissimilarity
##              0.1603986
## attr(,"se")
## standard error
##    0.007089421
```

Accesing to specific fields

To access specific fields from the test result:

```
# Contingency table from equivalence test using normal approximation:
getTable(lists_testResult)
```

```
##                              Enriched in PHOSPHORYLATION
## Enriched in APOPTOTIC_PROCESS TRUE FALSE
##                         TRUE  1306   311
##                         FALSE  188  8343
```

```
# Sorensen-Dice dissimilarity from equivalence test using normal approximation:
getDissimilarity(lists_testResult)
```

```
## Sorensen dissimilarity
##              0.1603986
## attr(,"se")
## standard error
##    0.007089421
```

```
# p value from equivalence test using normal approximation:
getPvalue(lists_testResult)
```

```
##      p-value
## 3.300726e-70
```

```
# p value from equivalence test using bootstrap approximation:
getPvalue(boot_testResult)
```

```
##   p-value
## 9.999e-05
```

and the same for other specific fields: standard error `getSE`, upper bound equivalence limit `getUpper` and efective number of bootstrap and resamples `getNboot` (Only available for bootstrap tests)

Other statistics related to the Sorensen-Dice dissimilarity

Sometimes, it would be interesting not to perform the full equivalence test but to compute other statistics related to the Sorensen-Dice dissimilarity. This computes would be done from contingency tables or directly from feature lists:

- The dissimilarity:

```
dSorensen(enrichTab)
```

```
## [1] 0.1603986
```

```
dSorensen(BPGOGeneList$GOBP_APOPTOTIC_PROCESS,
          BPGOGeneList$GOBP_PHOSPHORYLATION,
          geneUniverse = humanEntrezIDs,
          orgPackg = "org.Hs.eg.db",
          onto = "BP", GOLevel = 5,
          listNames = c("APOPTOTIC_PROCESS", "PHOSPHORYLATION"))
```

```
## [1] 0.1603986
```

- The Upper limit of the confidence interval for the true distance:

```
duppSorensen(enrichTab)
```

```
## [1] 0.1720596
```

```
duppSorensen(enrichTab, boot = T)
```

```
## [1] 0.1721678
## attr(,"eff.nboot")
## [1] 10000
```

```
duppSorensen(BPGOGeneList$GOBP_APOPTOTIC_PROCESS,
          BPGOGeneList$GOBP_PHOSPHORYLATION,
          geneUniverse = humanEntrezIDs,
          orgPackg = "org.Hs.eg.db",
          onto = "BP", GOLevel = 5,
          listNames = c("APOPTOTIC_PROCESS", "PHOSPHORYLATION"))
```

```
## [1] 0.1720596
```

- the same for `seSorensen`

Updating the results:

When some inputs (i.e. confidence level, irrelevance limit, used distribution, etc) have to be updated, it is no necessary to make computes again, only updating is enough

```
upgrade(lists_testResult, d0 = 0.175, conf.level = 0.99, boot = T)
```

```
##
##  Bootstrap test for 2x2 contingency tables based on the Sorensen-Dice
##  dissimilarity (10000 bootstrap replicates)
##
## data:  tab
## (d - d0) / se = -2.0596, p-value = 0.023
## alternative hypothesis: true equivalence limit d0 is less than 0.175
## 99 percent confidence interval:
##  0.0000000 0.1773906
## sample estimates:
## Sorensen dissimilarity
##              0.1603986
## attr(,"se")
## standard error
##    0.007089421
```

All pairwise tests (or other computations)

For objects of class list, all these functions (`equivTestSorensen`, `dSorensen`, `seSorensen`, `duppSorensen`) assume a list of character objects containing gene identifiers and all pairwise computations are performed.

For example, to obtain the matrix of all pairwise Sorensen-Dice dissimilarities:

```
dSorensen(BPGOGeneList, onto = "BP", GOLevel = 5,
          geneUniverse = humanEntrezIDs, orgPackg = "org.Hs.eg.db")
```

```
##                                                                GOBP_APOPTOTIC_PROCESS
## GOBP_APOPTOTIC_PROCESS                                                      0.0000000
## GOBP_IMMUNE_RESPONSE                                                        0.4530854
## GOBP_TISSUE_DEVELOPMENT                                                     0.2137150
## GOBP_POSITIVE_REGULATION_OF_MACROMOLECULE_BIOSYNTHETIC_PROCESS              0.2197840
## GOBP_PHOSPHORYLATION                                                        0.1603986
## GOBP_CELLULAR_RESPONSE_TO_STRESS                                            0.2081289
##                                                                GOBP_IMMUNE_RESPONSE
## GOBP_APOPTOTIC_PROCESS                                                    0.4530854
## GOBP_IMMUNE_RESPONSE                                                      0.0000000
## GOBP_TISSUE_DEVELOPMENT                                                   0.5720974
## GOBP_POSITIVE_REGULATION_OF_MACROMOLECULE_BIOSYNTHETIC_PROCESS            0.5147279
## GOBP_PHOSPHORYLATION                                                      0.4400357
```

```
## GOBP_CELLULAR_RESPONSE_TO_STRESS                                          0.4662638
##                                                                  GOBP_TISSUE_DEVELOPMENT
## GOBP_APOPTOTIC_PROCESS                                                         0.2137150
## GOBP_IMMUNE_RESPONSE                                                           0.5720974
## GOBP_TISSUE_DEVELOPMENT                                                        0.0000000
## GOBP_POSITIVE_REGULATION_OF_MACROMOLECULE_BIOSYNTHETIC_PROCESS                 0.2184778
## GOBP_PHOSPHORYLATION                                                           0.2544255
## GOBP_CELLULAR_RESPONSE_TO_STRESS                                               0.3246951
##                                                                  GOBP_POSITIVE_REGULATION_OF_MACROMOLE
## GOBP_APOPTOTIC_PROCESS
## GOBP_IMMUNE_RESPONSE
## GOBP_TISSUE_DEVELOPMENT
## GOBP_POSITIVE_REGULATION_OF_MACROMOLECULE_BIOSYNTHETIC_PROCESS
## GOBP_PHOSPHORYLATION
## GOBP_CELLULAR_RESPONSE_TO_STRESS
##                                                                  GOBP_PHOSPHORYLATION
## GOBP_APOPTOTIC_PROCESS                                                         0.1603986
## GOBP_IMMUNE_RESPONSE                                                           0.4400357
## GOBP_TISSUE_DEVELOPMENT                                                        0.2544255
## GOBP_POSITIVE_REGULATION_OF_MACROMOLECULE_BIOSYNTHETIC_PROCESS                 0.2743814
## GOBP_PHOSPHORYLATION                                                           0.0000000
## GOBP_CELLULAR_RESPONSE_TO_STRESS                                               0.2500915
##                                                                  GOBP_CELLULAR_RESPONSE_TO_STRESS
## GOBP_APOPTOTIC_PROCESS                                                                  0.2081289
## GOBP_IMMUNE_RESPONSE                                                                    0.4662638
## GOBP_TISSUE_DEVELOPMENT                                                                 0.3246951
## GOBP_POSITIVE_REGULATION_OF_MACROMOLECULE_BIOSYNTHETIC_PROCESS                          0.2629466
## GOBP_PHOSPHORYLATION                                                                    0.2500915
## GOBP_CELLULAR_RESPONSE_TO_STRESS                                                        0.0000000
```

Similarly, the following code performs all pairwise tests for all Ontologies and GO levels. In this case the Holm-Bonferroni criteria to avoid the inflation of p values for multiple tests is used::

```
allBPGO <- equivTestSorensen(BPGOGeneList, geneUniverse = humanEntrezIDs,
                             orgPackg = "org.Hs.eg.db")
```

Visualization of these computes is not good due to the huge number of results obtained, for this reason they were saved in the available .rda file using `save(allBPGO, file = "allBPGO.rda")`. If user wishes to see these results, he can easily load them using `load("allBPGO.rda")`. In addition, an excel file is also provided with a summary of these results.

Remember that, it is possible to access to specific fields of interest for specifics Ontologies and GO levels:

```
load("allBPGO.rda")
allBPGO$CC$`level 7`$GOBP_CELLULAR_RESPONSE_TO_STRESS$GOBP_APOPTOTIC_PROCESS
```

```
##
##  Normal asymptotic test for 2x2 contingency tables based on the
##  Sorensen-Dice dissimilarity
##
## data:  tab
## (d - d0) / se = 0.55598, p-value = 0.7109
## alternative hypothesis: true equivalence limit d0 is less than 0.4444444
```

```
## 95 percent confidence interval:
##  0.0000000 0.5818909
## sample estimates:
## Sorensen dissimilarity
##              0.4791667
## attr(,"se")
## standard error
##     0.06245192
```

```
getPvalue(allBPGO$CC$`level 7`$GOBP_CELLULAR_RESPONSE_TO_STRESS$GOBP_APOPTOTIC_PROCESS)
```

```
##   p-value
## 0.7108889
```

Session information

All software and respective versions used to produce this document are listed below.

```
sessionInfo()
```

```
## R version 4.2.2 (2022-10-31 ucrt)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 22621)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=Spanish_Ecuador.utf8  LC_CTYPE=Spanish_Ecuador.utf8
## [3] LC_MONETARY=Spanish_Ecuador.utf8 LC_NUMERIC=C
## [5] LC_TIME=Spanish_Ecuador.utf8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] goSorensen_1.0.0 qusage_2.32.0    limma_3.54.0
##
## loaded via a namespace (and not attached):
##   [1] ggtree_3.6.2           fgsea_1.24.0           colorspace_2.0-3
##   [4] gson_0.0.9             estimability_1.4.1     qvalue_2.30.0
##   [7] XVector_0.38.0         aplot_0.1.8            rstudioapi_0.14
##  [10] farver_2.1.1           graphlayouts_0.8.3     ggrepel_0.9.2
##  [13] bit64_4.0.5            scatterpie_0.1.8       AnnotationDbi_1.60.0
##  [16] fansi_1.0.3            mvtnorm_1.1-3          codetools_0.2-18
##  [19] splines_4.2.2          cachem_1.0.6           GOSemSim_2.24.0
##  [22] knitr_1.41             polyclip_1.10-4        jsonlite_1.8.3
##  [25] GO.db_3.16.0           png_0.1-7              ggforce_0.4.1
##  [28] compiler_4.2.2         httr_1.4.4             emmeans_1.8.3
##  [31] lazyeval_0.2.2         assertthat_0.2.1       Matrix_1.5-3
##  [34] fastmap_1.1.0          cli_3.4.1              tweenr_2.0.2
##  [37] htmltools_0.5.3        tools_4.2.2            igraph_1.3.5
```

```
##  [40] gtable_0.3.1           glue_1.6.2             GenomeInfoDbData_1.2.9
##  [43] reshape2_1.4.4         dplyr_1.0.10           fastmatch_1.1-3
##  [46] Rcpp_1.0.9             enrichplot_1.18.1      Biobase_2.58.0
##  [49] vctrs_0.5.1            Biostrings_2.66.0      ape_5.6-2
##  [52] nlme_3.1-160           ggraph_2.1.0           xfun_0.35
##  [55] stringr_1.4.1          CompQuadForm_1.4.3     lifecycle_1.0.3
##  [58] clusterProfiler_4.6.0  DOSE_3.24.2            org.Hs.eg.db_3.16.0
##  [61] zlibbioc_1.44.0        MASS_7.3-58.1          scales_1.2.1
##  [64] tidygraph_1.2.2        parallel_4.2.2         RColorBrewer_1.1-3
##  [67] yaml_2.3.6             memoise_2.0.1          gridExtra_2.3
##  [70] ggplot2_3.4.0          downloader_0.4         ggfun_0.0.9
##  [73] HDO.db_0.99.1          yulab.utils_0.0.5      stringi_1.7.8
##  [76] RSQLite_2.2.18         S4Vectors_0.36.0       tidytree_0.4.1
##  [79] BiocGenerics_0.44.0    BiocParallel_1.32.1    GenomeInfoDb_1.34.3
##  [82] rlang_1.0.6            pkgconfig_2.0.3        bitops_1.0-7
##  [85] evaluate_0.18          lattice_0.20-45        purrr_0.3.5
##  [88] treeio_1.22.0          patchwork_1.1.2        shadowtext_0.1.2
##  [91] cowplot_1.1.1          bit_4.0.5              tidyselect_1.2.0
##  [94] plyr_1.8.8             magrittr_2.0.3         R6_2.5.1
##  [97] IRanges_2.32.0         fftw_1.0-7             generics_0.1.3
## [100] DBI_1.1.3              pillar_1.8.1           withr_2.5.0
## [103] KEGGREST_1.38.0        RCurl_1.98-1.9         tibble_3.1.8
## [106] crayon_1.5.2           goProfiles_1.60.0      utf8_1.2.2
## [109] rmarkdown_2.18         viridis_0.6.2          grid_4.2.2
## [112] data.table_1.14.6      blob_1.2.3             digest_0.6.30
## [115] xtable_1.8-4           tidyr_1.2.1            gridGraphics_0.5-1
## [118] stats4_4.2.2           munsell_0.5.0          viridisLite_0.4.1
## [121] ggplotify_0.1.0
```

References

Flores, P., Salicrú, M., Sánchez-Pla, A. et al. An equivalence test between features lists, based on the Sorensen-Dice index and the joint frequencies of GO term enrichment. BMC Bioinformatics 23, 207 (2022). https://doi.org/10.1186/s12859-022-04739-2