Enrichment Analysis

priesgo

August 9, 2016

Using Object Oriented (OO) programming

Using OO allows achieving **high cohesion** and **low coupling**, which facilitates all the software development cycle from development to maintenance, including testing. Following the principle **"Don't repeat yourself"** (https://en.wikipedia.org/wiki/Don't_repeat_yourself) is also enhanced with a good OO design. There are 3 OO models in R, we will be using S4 as it is the model most used in Bioconductor. Furthermore, we want to foster **code reuse** from existing Bioconductor S4 classes.

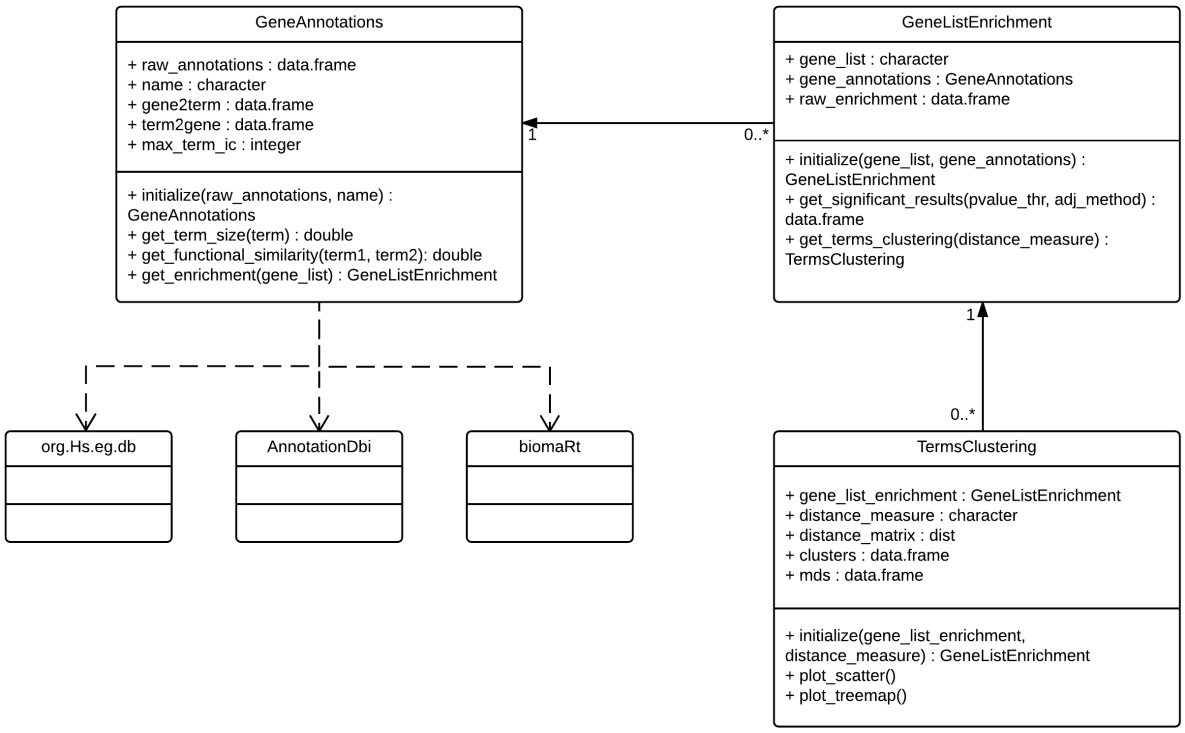

Figure 1: Enrichment class model

Annotations

This model allows us to support multiple annotations for enrichment. So far the supported annotations based on AnnotationDbi data source and some custom resources are:

- Gene Ontology (from org.Hs.eg.db)
- KEGG pathways (from org.Hs.eg.db)
- OMIM diseases (from org.Hs.eg.db)
- HPO phenotypes (from http://compbio.charite.de/jenkins/job/hpo.annotations.monthly/lastStableBuild/artifact/annotation/ALL_SOURCES_TYPICAL_FEATURES_phenotype_to_genes.txt)

These annotations can be loaded as follows:

```
goa <- TCGAome::load_goa()
kegg <- TCGAome::load_kegg()
omim <- TCGAome::load_omim()
hpo <- TCGAome::load_hpo()
```

The object created contains the following information:

```
kegg <- TCGAome::load_kegg()
str(kegg, list.len = 5, vec.len = 5)
```

```
## Formal class 'GeneAnnotations' [package "TCGAome"] with 6 slots
##   ..@ raw_annotations        :'data.frame':  16313 obs. of  2 variables:
##   .. ..$ Gene: chr [1:16313] "A2M" "NAT1" "NAT1" "NAT1" "NAT2" ...
##   .. ..$ Term: chr [1:16313] "04610" "00232" "00983" "01100" "00232" ...
##   ..@ name                   : chr "KEGG-Human"
##   ..@ gene2term              :'data.frame':  5869 obs. of  2 variables:
##   .. ..$ Gene: chr [1:5869] "A2M" "A4GALT" "AAAS" "AACS" "AADAT" ...
##   .. ..$ Term:List of 5869
##   .. .. ..$ 0001: chr "04610"
##   .. .. ..$ 0002: chr [1:2] "00603" "01100"
##   .. .. ..$ 0003: chr "03013"
##   .. .. ..$ 0004: chr "00650"
##   .. .. ..$ 0005: chr [1:4] "00300" "00310" "00380" "01100"
##   .. .. .. [list output truncated]
##   ..@ term2gene              :'data.frame':  229 obs. of  2 variables:
##   .. ..$ Term: chr [1:229] "00010" "00020" "00030" "00040" "00051" ...
##   .. ..$ Gene:List of 229
##   .. .. ..$ 001: chr [1:65] "ADH1A" "ADH1B" "ADH1C" "ADH4" "ADH5" ...
##   .. .. ..$ 002: chr [1:30] "ACLY" "ACO1" "ACO2" "CS" "DLAT" ...
##   .. .. ..$ 003: chr [1:27] "ALDOA" "ALDOB" "ALDOC" "FBP1" "G6PD" ...
##   .. .. ..$ 004: chr [1:32] "ALDH2" "ALDH1B1" "ALDH3A2" "AKR1B1" "GUSB" ...
##   .. .. ..$ 005: chr [1:36] "ALDOA" "ALDOB" "ALDOC" "AKR1B1" "FBP1" ...
##   .. .. .. [list output truncated]
##   ..@ max_term_freq          : int 1130
##   .. [list output truncated]
```

Metrics

For any term within the **TCGAome::GeneAnnotations** object we can obtain its **relative frequency** of annotation, which might be useful to compute the **Information Content** within a term.

```
random_kegg_term = kegg@term2gene$Term[runif(1, max = length(kegg@term2gene$Term))]
random_kegg_term
```

```
## [1] "00785"
```

```
TCGAome::get_term_freq(kegg, random_kegg_term)
```

```
## [1] 0.002654867
```

For any two terms within the GeneAnnotations object we can obtain its **functional similarity** based on the binary distances implemented in TCGAome.

```
random_kegg_term1 = kegg@term2gene$Term[runif(1, max = length(kegg@term2gene$Term))]
random_kegg_term2 = kegg@term2gene$Term[runif(1, max = length(kegg@term2gene$Term))]
random_kegg_term1
```

```
## [1] "05020"
```

```
random_kegg_term2
```

```
## [1] "04722"
```

```
TCGAome::get_functional_similarity(kegg, random_kegg_term1, random_kegg_term2, method = "UI")
```

```
## [1] 0.03184713
```

The supported binary distances are: UI, cosine, Bray-Curtis and binary.

This model should be extended to support **semantic similarity** measures based on the ontology structure. This is a special case for some of these annotation resources which are also backed by an ontology, like GO and HPO. On the previous implementation of TCGAome we used **GoSemSim** for the semantic similarity within GO terms. We need to study if this can be easily extended to other ontologies.

Extend annotation support

It is relatively simple to extend the support to additional annotations, we just need to provide the association between genes and terms in a tall-skinny data frame with columns "Gene" and "Term".

```
uniKeys <- AnnotationDbi::keys(org.Hs.eg.db::org.Hs.eg.db, keytype="SYMBOL")
cols <- c("PATH")
kegg_raw <- AnnotationDbi::select(org.Hs.eg.db::org.Hs.eg.db, keys=uniKeys, columns=cols, keytype="SYMB
kegg_raw <- kegg_raw[, c(1, 2)]
colnames(kegg_raw) <- c("Gene", "Term")
kegg <- new("GeneAnnotations", raw_annotations = kegg_raw, name="KEGG-Human")
```

Enrichment

The previous TCGAome version used the package **topGO** for computing the enrichment of GO terms. This package is limited to GO. The computation employed was a Fisher's test, we were not making use of the advanced functionality in topGO. Thus, in order to gain flexibility the enrichment computation was reimplemented inside the class **TCGAome::GeneListEnrichment**.

We can compute enrichment for a given list of genes based on a preloaded annotation:

```
gene_list <- c("ZNF638", "HNRNPU", "PPIAL4G", "RAPH1", "USP7", "SUMO1P3",
          "TMEM189.UBE2V1", "ZNF837", "LPCAT4", "ZFPL1", "STAT3", "XRCC1",
          "STMN1", "PGR", "RB1", "KDR", "YBX1", "YAP1", "FOXO3", "SYK", "RAB17",
          "TTC8", "SLC22A5", "C3orf18", "ANKRA2", "LBR", "B3GNT5", "ANP32E",
          "JOSD1", "ZNF695", "ESR1", "INPP4B", "PDK1", "TSC2", "AR", "HSPA1A",
```

```
          "CDH3", "SMAD4", "CASP7", "GMPS", "NDC80", "EZH2", "MELK", "CDC45",
          "CRY2", "KLHDC1", "MEIS3P1", "FBXL5", "EHD2", "CCNB1", "GSK3A",
          "DVL3", "NFKB1", "COL6A1", "CCND1", "BAK1")
kegg_enrichment <- TCGAome::get_enrichment(kegg, gene_list = gene_list)
str(kegg_enrichment, list.len = 5, vec.len = 5)
```

```
## Formal class 'GeneListEnrichment' [package "TCGAome"] with 3 slots
##   ..@ gene_list       : chr [1:56] "ZNF638" "HNRNPU" "PPIAL4G" "RAPH1" "USP7" ...
##   ..@ gene_annotations:Formal class 'GeneAnnotations' [package "TCGAome"] with 6 slots
##   .. .. ..@ raw_annotations        :'data.frame':    16313 obs. of  2 variables:
##   .. .. .. ..$ Gene: chr [1:16313] "A2M" "NAT1" "NAT1" "NAT1" "NAT2" ...
##   .. .. .. ..$ Term: chr [1:16313] "04610" "00232" "00983" "01100" "00232" ...
##   .. .. ..@ name                   : chr "KEGG-Human"
##   .. .. ..@ gene2term              :'data.frame':    5869 obs. of  2 variables:
##   .. .. .. ..$ Gene: chr [1:5869] "A2M" "A4GALT" "AAAS" "AACS" "AADAT" ...
##   .. .. .. ..$ Term:List of 5869
##   .. .. .. .. ..$ 0001: chr "04610"
##   .. .. .. .. ..$ 0002: chr [1:2] "00603" "01100"
##   .. .. .. .. ..$ 0003: chr "03013"
##   .. .. .. .. ..$ 0004: chr "00650"
##   .. .. .. .. ..$ 0005: chr [1:4] "00300" "00310" "00380" "01100"
##   .. .. .. .. .. [list output truncated]
##   .. .. ..@ term2gene              :'data.frame':    229 obs. of  2 variables:
##   .. .. .. ..$ Term: chr [1:229] "00010" "00020" "00030" "00040" "00051" ...
##   .. .. .. ..$ Gene:List of 229
##   .. .. .. .. ..$ 001: chr [1:65] "ADH1A" "ADH1B" "ADH1C" "ADH4" "ADH5" ...
##   .. .. .. .. ..$ 002: chr [1:30] "ACLY" "ACO1" "ACO2" "CS" "DLAT" ...
##   .. .. .. .. ..$ 003: chr [1:27] "ALDOA" "ALDOB" "ALDOC" "FBP1" "G6PD" ...
##   .. .. .. .. ..$ 004: chr [1:32] "ALDH2" "ALDH1B1" "ALDH3A2" "AKR1B1" "GUSB" ...
##   .. .. .. .. ..$ 005: chr [1:36] "ALDOA" "ALDOB" "ALDOC" "AKR1B1" "FBP1" ...
##   .. .. .. .. .. [list output truncated]
##   .. .. ..@ max_term_freq          : int 1130
##   .. .. .. [list output truncated]
##   ..@ raw_enrichment  :'data.frame': 229 obs. of  3 variables:
##   .. ..$ Term  : chr [1:229] "00010" "00020" "00030" "00040" "00051" ...
##   .. ..$ pvalue: num [1:229] 1 1 1 1 1 1 1 1 1 1 1 ...
##   .. ..$ freq  : num [1:229] 0.0575 0.0265 0.0239 0.0283 0.0319 0.0239 ...
```

Or alternatively:

```
kegg_enrichment <- new("TCGAome::GeneListEnrichment", gene_annotations = kegg, gene_list = gene_list)
```

And extract significant results:

```
TCGAome::get_significant_results(kegg_enrichment, significance_thr = 0.05, adj_method = "none")
```

```
##      Term       pvalue       freq   adj_pvalue
## 112 04110 0.0061590702 0.10973451 0.0061590702
## 114 04115 0.0263433137 0.06017699 0.0263433137
## 200 05145 0.0360194893 0.11681416 0.0360194893
## 204 05200 0.0334067677 0.28849558 0.0334067677
## 207 05212 0.0004880615 0.06194690 0.0004880615
```

```
## 210 05215 0.0098109166 0.07876106 0.0098109166
## 215 05220 0.0048909621 0.06460177 0.0048909621
## 216 05221 0.0165441885 0.05044248 0.0165441885
## 217 05222 0.0464121410 0.07522124 0.0464121410
## 218 05223 0.0143072153 0.04778761 0.0143072153
```

To be continued....