An introduction to FSim

Qiang Hu, Song Liu

May 7, 2013

Contents

1 Introduction

FSim is an R package to search functionally similar genes for objective gene, a set of functional keywords or a biological pathway. The function compare the functional relation between genes based on the Gene Ontology annotation. A new algorithm is proposed to analyze the relation between genes based on the GO annotation of genes. Our package is able to search the most functionally similar genes by comparing the GO terms between genes.

```
> library(FSim)
```

```
groupGOTerms:         GOBPTerm, GOMFTerm, GOCCTerm environments built.
```

2 Similar score calculation

2.1 Comparison between two genes

```
> calSim(g1="9", g2="10", ontology="ALL", an.go=an.Hs.egGO)
```

```
     value        ASE    z.value
 0.81818146  0.06827444 11.98371591
```

2.2 Comparison between gene and GO terms

```
> terms <- names(get("10", org.Hs.egGO))
> calSim(g1="9", tids=terms, an.go=an.Hs.egGO)
```

```
     value        ASE    z.value
 0.81818146  0.06827444 11.98371591
```

3 Search example

3.1 Search by gene

The function SearchGene can be used to search functionally similar genes for an objective gene. For example, we can use the function to find the most funcionally related genes for an objective gene ”NAT2” in the GO annotation database.

```
> SearchGene(symbol="NAT2", an.go=an.Hs.egGO, targets="ALL", n=10)
```

```
        Symbol sharedTerms      value        ASE   z.value
9         NAT1           4 0.81818146 0.06827444 11.983716
126      ADH1C           3 0.44740139 0.10899085  4.104945
119391   GSTO2           3 0.43045407 0.10557113  4.077384
125      ADH1B           3 0.43796235 0.10875409  4.027089
124      ADH1A           3 0.37303624 0.10602482  3.518386
10380    BPNT1           3 0.34734390 0.10439168  3.327314
1312      COMT           4 0.21696033 0.08272130  2.622787
2950     GSTP1           4 0.10838889 0.06248277  1.734700
11069  RAPGEF4           3 0.11917485 0.07958140  1.497521
100        ADA           3 0.05626705 0.05342171  1.053262
```

By specifying n is 10, the function return 10 functionally related genes ordered by Z values. The search database can be set by the targets option, which can be ”ALL” to search in all GO annotated genes and abso can be a set of customized genes to specify search range. For example, the function between gene ”9” and a gene set can be compared d by the option ”targets”.

```
> SearchGene(gene="9", targets=c("10", "100", "124"), an.go=an.Hs.egGO)
```

```
    Symbol sharedTerms      value        ASE    z.value
10    NAT2           4 0.81818146 0.06827444 11.9837159
124  ADH1A           3 0.42759849 0.10843732  3.9432779
100    ADA           2 0.04894472 0.05421093  0.9028571
```

3.2 Search by GO terms

A group of GO terms can aslo be used to search functionally related genes. For exmaple,

```
> t1 <- names(get("9", org.Hs.egGO))
> t1
```

```
[1] "GO:0006805" "GO:0044281" "GO:0005829" "GO:0004060"

> SearchGene(terms=t1, an.go=an.Hs.egGO, n=5)

       Symbol sharedTerms     value        ASE   z.value
9        NAT1           4 1.0000000 0.00000000       Inf
10       NAT2           4 0.8181815 0.06827444 11.983716
119391  GSTO2           3 0.5000095 0.10658958  4.690979
124     ADH1A           3 0.4275985 0.10843732  3.943278
10380   BPNT1           3 0.3942334 0.10711596  3.680436
```

3.3 Search by keywords

The function can also be used to analyze the functionally similar genes with a group of biological keywords. For example, we try to search for genes related with function ”chromatin remodeling” and ”histone binding”.

```
> t2 <- SearchTerm(fun=c("chromatin remodeling", "histone binding"))
> t2

        GOID Ontology                                           Term
1 GO:0006338       BP                           chromatin remodeling
2 GO:0031055       BP             chromatin remodeling at centromere
4 GO:0043044       BP             ATP-dependent chromatin remodeling
5 GO:0043156       BP chromatin remodeling in response to cation stress
6 GO:0016585       CC                   chromatin remodeling complex
3 GO:0031011       CC                                  Ino80 complex
7 GO:0031493       MF                     nucleosomal histone binding
8 GO:0042393       MF                                histone binding
```

Then we select part of the returned GO terms to search function related genes.

```
> SearchGene(terms=t2$GOID[c(1,6,8)], an.go=an.Hs.egGO, n=5)

       Symbol sharedTerms     value        ASE  z.value
10361    NPM2           2 0.6027630 0.07355932 8.194244
5928    RBBP4           2 0.4361119 0.07215941 6.043729
54617   INO80           2 0.2930305 0.06404111 4.575663
373861  HILS1           2 0.3547751 0.08244450 4.303199
51773    RSF1           2 0.3158120 0.07814011 4.041612
```

3.4 Search by gene set

In order to search functionally similar genes for a gene set, we need summary the objective gene set to a group of GO terms first. GO over-represent analysis can be used to discover major biological functions for a gene set. The `ovreGO` integrated functions from *topGO* can be used to find over-represented GO terms for a gene set. For example, we have a gene set from KEGG database.

```
> library(KEGG.db)
> geneset <- get("hsa00232", KEGGPATHID2EXTID)
> geneset
```

```
[1] "10"   "1544" "1548" "1549" "1553" "7498" "9"
```

The function `ovreGO` can be used to find major represent terms. All human genes from KEGG database are used as backgroud.

```
> paths <- as.list(KEGGPATHID2EXTID)
> paths <- paths[grep("^hsa", names(paths))]
> allgenes <- unique(unlist(paths))
> BPterms <- ovreGO(genes=geneset, allgenes=allgenes, ontology="BP", nterm=10)
```

```
Building most specific GOs .....        ( 6732 GO terms found. )

Build GO DAG topology ..........        ( 9963 GO terms and 21580 relations. )

Annotating nodes ...............        ( 5305 genes annotated to the GO terms. )

                         -- Parent-Child Algorithm --

                 the algorithm is scoring 268 nontrivial nodes
                 parameters:
                         test statistic:  fisher : joinFun = union

         Level 12:        3 nodes to be scored.

         Level 11:        6 nodes to be scored.

         Level 10:        12 nodes to be scored.

         Level 9:        18 nodes to be scored.

         Level 8:        22 nodes to be scored.

         Level 7:        28 nodes to be scored.

         Level 6:        41 nodes to be scored.

         Level 5:        50 nodes to be scored.

         Level 4:        49 nodes to be scored.

         Level 3:        27 nodes to be scored.

         Level 2:        11 nodes to be scored.
```

```
> BPterms
```

```
        GO.ID                                        Term Annotated Significant
1  GO:0006805                xenobiotic metabolic process       123           5
2  GO:0009410             response to xenobiotic stimulus       124           5
```

```
3  GO:0071466    cellular response to xenobiotic stimulus       123           5
4  GO:0042737                      drug catabolic process         9           2
5  GO:0019748                 secondary metabolic process        22           2
6  GO:0009403                  toxin biosynthetic process         1           1
7  GO:0017144                     drug metabolic process         22           2
8  GO:0042738             exogenous drug catabolic process        7           2
9  GO:1900746 regulation of vascular endothelial growt...         1           1
10 GO:0006725 cellular aromatic compound metabolic pro...       189           3
   Expected result1    Score   wScore
1      0.14 2.3e-07 6.640261 5.039745
2      0.14 8.9e-06 5.050291 2.555339
3      0.14 0.00011 3.977005 2.515349
4      0.01 0.00020 3.704480 2.646057
5      0.02 0.00043 3.362149 1.287642
6      0.00 0.00050 3.303628 1.957705
7      0.02 0.00051 3.294099 1.765308
8      0.01 0.00066 3.177825 2.723850
9      0.00 0.00114 2.942504 2.675004
10     0.21 0.00240 2.619535 1.307371
```

The results from ovreGO show the most over-represented terms ordered by p values. Top 10 GO terms are used to stand for the the major biological functions of the gene set.

```
> SearchGene(terms=BPterms$GO.ID, targets="ALL", an.go=an.Hs.egGO, n=5)

     Symbol sharedTerms     value        ASE  z.value
1544 CYP1A2           5 0.2668618 0.08916362 2.992944
1548 CYP2A6           3 0.2918221 0.10571045 2.760579
1555 CYP2B6           3 0.2983153 0.11184116 2.667312
1559 CYP2C9           4 0.2325906 0.10332004 2.251167
1576 CYP3A4           4 0.2062567 0.09761784 2.112900
```

The returned genes are all from ”CYP” gene family because most of the interested gene set are also from this family.

4 Evaluation

The functionally related genes should get higer similar scores as our method proposed. Here we use a set of genes from KEGG pathway to simply evaluate our method. First, we use GO over-represented algorithm to find the major functions of the gene set. Then the over-represented GO terms are used to calculate the similar scores with the gene set and other randomly selected genes using our method.

```
> MFterms <- ovreGO(genes=geneset, allgenes=allgenes, ontology="MF", nterm=10)

Building most specific GOs .....        ( 2721 GO terms found. )

Build GO DAG topology .........         ( 3182 GO terms and 3779 relations. )
```

```
Annotating nodes ..............         ( 5581 genes annotated to the GO terms. )

                         -- Parent-Child Algorithm --

                 the algorithm is scoring 58 nontrivial nodes
                 parameters:
                         test statistic:  fisher : joinFun = union

         Level 8:        3 nodes to be scored.

         Level 7:        3 nodes to be scored.

         Level 6:        9 nodes to be scored.

         Level 5:        14 nodes to be scored.

         Level 4:        13 nodes to be scored.

         Level 3:        11 nodes to be scored.

         Level 2:        4 nodes to be scored.
> CCterms <- ovreGO(genes=geneset, allgenes=allgenes, ontology="CC", nterm=10)
Building most specific GOs .....        ( 888 GO terms found. )

Build GO DAG topology ..........        ( 1069 GO terms and 2041 relations. )

Annotating nodes ..............         ( 5659 genes annotated to the GO terms. )

                         -- Parent-Child Algorithm --

                 the algorithm is scoring 27 nontrivial nodes
                 parameters:
                         test statistic:  fisher : joinFun = union

         Level 10:        1 nodes to be scored.

         Level 9:        2 nodes to be scored.

         Level 8:        3 nodes to be scored.

         Level 7:        3 nodes to be scored.

         Level 6:        2 nodes to be scored.

         Level 5:        3 nodes to be scored.

         Level 4:        4 nodes to be scored.
```

```
          Level 3:        4 nodes to be scored.

          Level 2:        4 nodes to be scored.
```

```
> allterms <- c(BPterms$GO.ID, MFterms$GO.ID, CCterms$GO.ID)
```

The gene set from the pathway is a group of genes related in biological process, molecular function and cellular component, so GO terms from the three ontologies are used. The `allterms` are the over-represented GO terms to do the comparisons.

```
> score1 <- SearchGene(terms=allterms, targets=geneset, an.go=an.Hs.egGO, ontology="ALL", term2
> score1

      Symbol sharedTerms      value        ASE  z.value
1548  CYP2A6           5 0.53532411 0.08800991 6.082543
1553 CYP2A13           3 0.55718496 0.09715194 5.735191
1549  CYP2A7           2 0.52492639 0.10563474 4.969259
1544  CYP1A2          11 0.40508670 0.08183676 4.949936
9       NAT1           1 0.24921834 0.14525798 1.715695
10      NAT2           1 0.20762385 0.13625675 1.523769
7498     XDH           2 0.09217004 0.07167409 1.285960
```

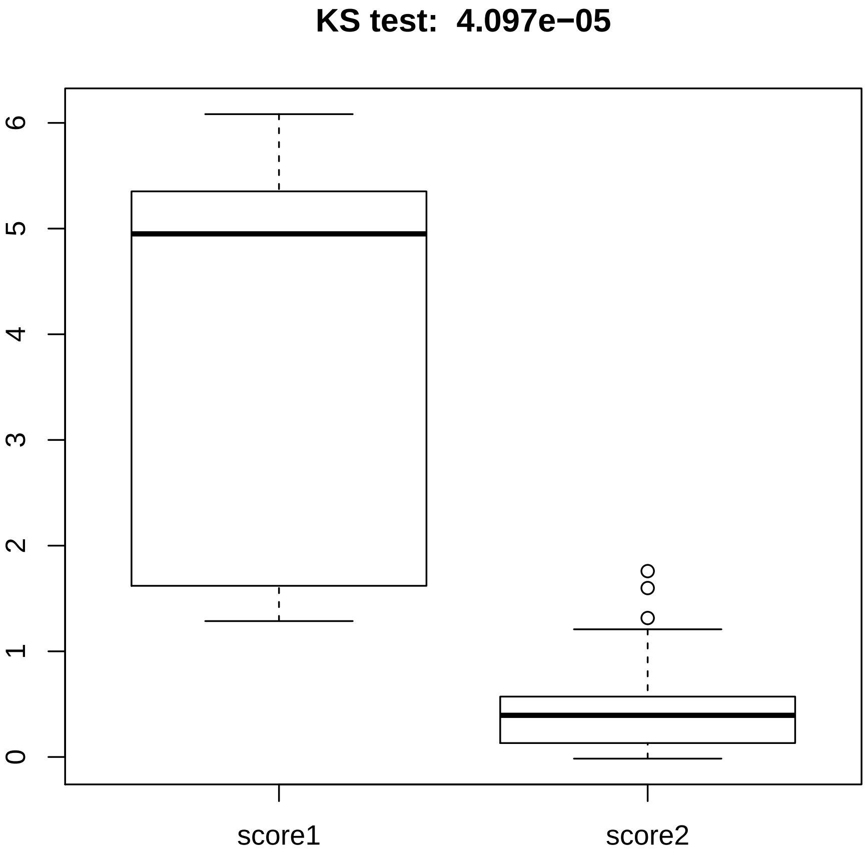
Then we randomly select 100 genes as control group to calculate the similar scores.

```
> set.seed(1)
> ctlgene <- sample(setdiff(allgenes, geneset), 50)
> score2 <- SearchGene(terms=allterms, targets=ctlgene, an.go=an.Hs.egGO, ontology="ALL", term2
> head(score2)

       Symbol sharedTerms      value        ASE  z.value
126129  CPT1C           1 0.19789514 0.11250666 1.758964
23225  NUP210           1 0.18225567 0.11400423 1.598675
3040     HBA2           0 0.14380074 0.10932575 1.315342
8540     AGPS           0 0.14535307 0.12028357 1.208420
1080     CFTR           0 0.09669977 0.08363717 1.156182
125965 COX6B2           0 0.15568241 0.13504771 1.152796
```

```
> pv <- suppressWarnings(ks.test(score1$z.value, score2$z.value))
> pv

        Two-sample Kolmogorov-Smirnov test

data:  score1$z.value and score2$z.value
D = 0.9388, p-value = 4.097e-05
alternative hypothesis: two-sided
```

The `ks.test` shows the two scores are significantly different. The scores from the `geneset` are significantly higer than the randomly selected genes. A boxplot can be used to show the detailed distribution of the two groups of scores.

```
> boxplot(score1$z.value, score2$z.value, names=c("score1", "score2"), main=paste("KS test: ",
```

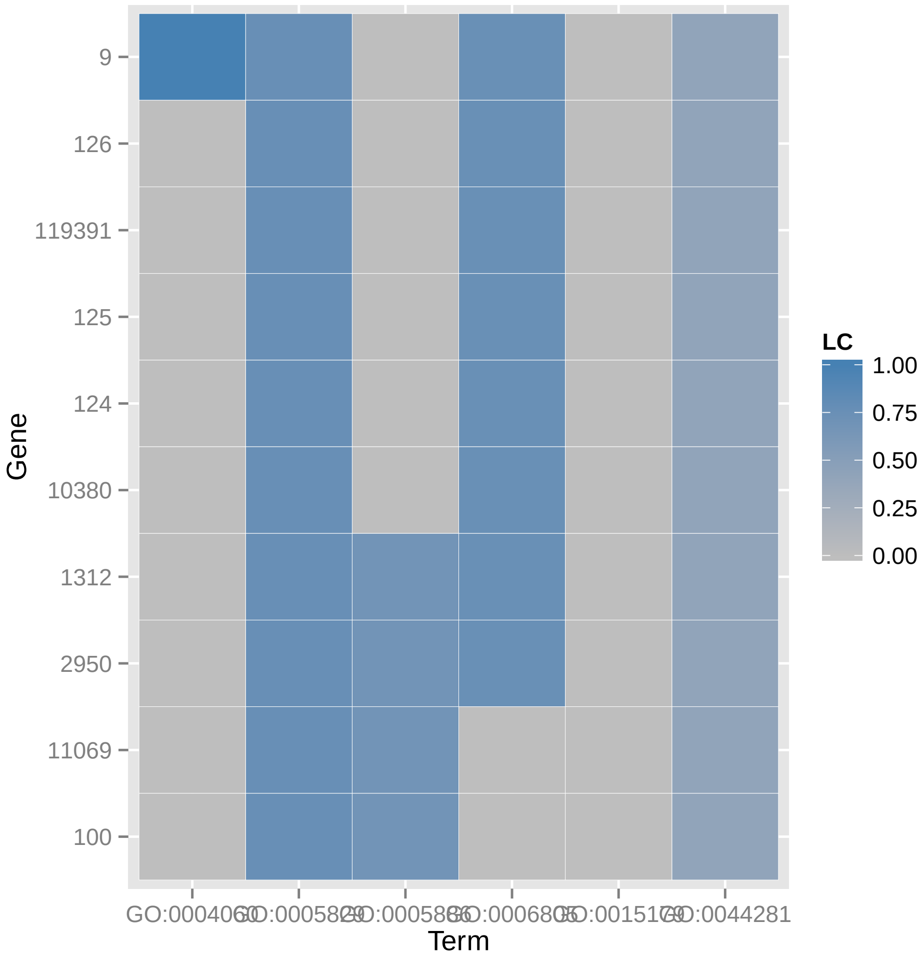


5 Visualization

5.1 Heatmap

Basically, functionally related genes share part of GO terms. The search results only show the number of shared terms. The details can be plotted with heatmap.

```
> res1 <- SearchGene(symbol="NAT2", an.go=an.Hs.egGO, targets="ALL", n=10, plot=TRUE)
```



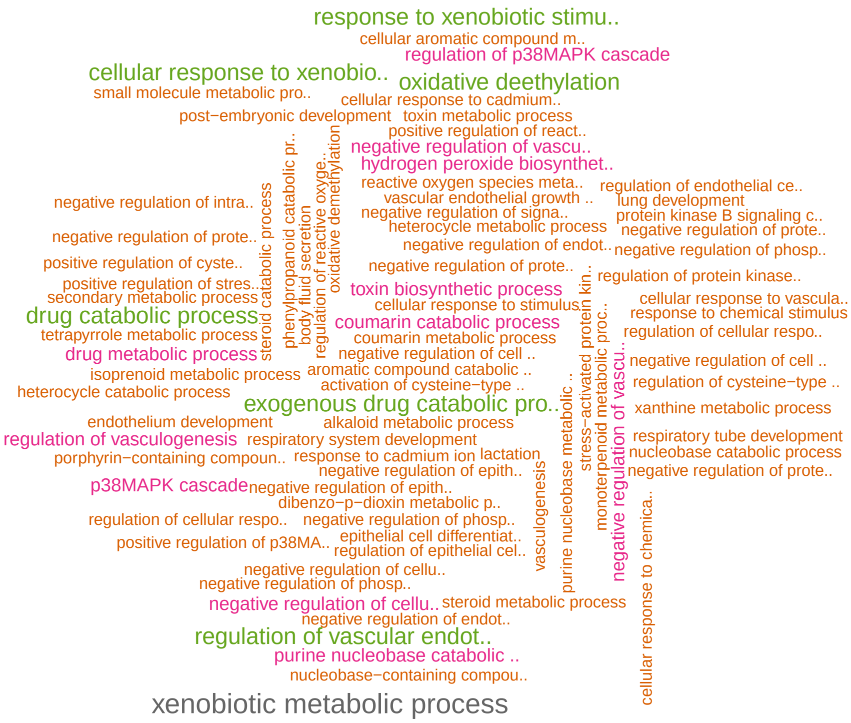
5.2 Wordcloud

The GO over-represent analysis return the major biological functions of a gene set. The results can also be visualized by word cloud plot. The function `ovreGO` can also be used to plot the wordcloud with the option ”plot=TRUE”.

```
> res2 <- ovreGO(genes=geneset, allgenes=allgenes, ontology="BP", plot=TRUE, scale=c(1,0.5))
```

```
Building most specific GOs .....        ( 6732 GO terms found. )

Build GO DAG topology ..........        ( 9963 GO terms and 21580 relations. )

Annotating nodes ...............        ( 5305 genes annotated to the GO terms. )

                        -- Parent-Child Algorithm --

                 the algorithm is scoring 268 nontrivial nodes
                 parameters:
                         test statistic:  fisher : joinFun = union
```

```
Level 12:        3 nodes to be scored.

Level 11:        6 nodes to be scored.

Level 10:        12 nodes to be scored.

Level 9:        18 nodes to be scored.

Level 8:        22 nodes to be scored.

Level 7:        28 nodes to be scored.

Level 6:        41 nodes to be scored.

Level 5:        50 nodes to be scored.

Level 4:        49 nodes to be scored.

Level 3:        27 nodes to be scored.

Level 2:        11 nodes to be scored.
```



6 Session information

```
> sessionInfo()

R version 2.15.2 (2012-10-26)
Platform: x86_64-unknown-linux-gnu (64-bit)

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=C
 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=C                 LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] grid      stats     graphics  grDevices utils     datasets  methods
[8] base
```

```
other attached packages:
 [1] KEGG.db_2.8.0       FSim_0.1.1          reshape2_1.2.2
 [4] ggplot2_0.9.3.1     wordcloud_2.4       RColorBrewer_1.0-5
 [7] Rcpp_0.10.3         vcd_1.2-13          colorspace_1.2-2
[10] MASS_7.3-22         topGO_2.10.0        SparseM_0.97
[13] graph_1.36.2        org.Hs.eg.db_2.8.0  GO.db_2.8.0
[16] RSQLite_0.11.3      DBI_0.2-5           AnnotationDbi_1.20.7
[19] Biobase_2.18.0      BiocGenerics_0.4.0

loaded via a namespace (and not attached):
 [1] IRanges_1.16.6  dichromat_2.0-0 digest_0.6.3    gtable_0.1.2
 [5] labeling_0.1    lattice_0.20-15 munsell_0.4     parallel_2.15.2
 [9] plyr_1.8        proto_0.3-10    scales_0.2.3    slam_0.1-28
[13] stats4_2.15.2   stringr_0.6.2   tools_2.15.2
```