# Manual of GO-function

Jing Wang, Zheng Guo

April 24, 2017

## 1 Introduction

The GO-function package is an enrichment analysis tool for Gene Ontology (GO) [1]. According to several explicit rules, it is designed for treating the redundancy resulting from the GO structure or multiple annotation genes. Different from current redundancy treatment tools [2, 3, 4, 5] simply based on some numerical considerations, GO-function can find terms which are both statistically interpretable and biologically meaningful.

In this manual, we use one gene expression profile [6] for colorectal cancer to show the input and output of the GO-function package. The gene expression profile consists of 32 pairs of colorectal adenomas tissue and adjacent normal mucosa. The raw data were preprocessed using Robust Multi-array Average (RMA) [7]. The SOURCE database [8] (April, 2009) was used for the translation of ProbeIDs to GeneIDs. 9201 differentially expressed (DE) genes are selected using the Significance Analysis of Microarrays (SAM) method [9] with an FDR of 1%. These 9201 DE genes are analysed by GO-function as an example.

## 2 Environment and Input

### 2.1 Environment

The R version is at least 2.11.1, which can be downloaded in our web. Before applying GO-function to analyse data, the users should install the following packages: Biobase (>=2.8.0), graph (>=1.26.0), Rgraphviz (>=1.26.0), GO.db (>=2.4.1), AnnotationDbi (>=1.10.2), SparseM (>=0.85) and a gene annotation package for some organism (for example, "org.Hs.eg.db" (>=2.4.1) for human (See Input of the package for other organisms)). Before installing Rgraphviz package, the users should firstly install graphviz tool which can be downloaded from the website http://www.graphviz.org/pub/graphviz/stable/. Note: the version of graphviz has to match the version used to build Rgraphviz. Meanwhile, graphviz have to be on the path. We suggest using Rgraphviz 1.26.0 in R 2.11.1, which is corresponding to graphviz 2.20.3. We provide this graphviz version in our web. The installation process of the above packages is as follows.

>source("http://www.bioconductor.org/biocLite.R")

>biocLite()

>biocLite("graph")

>biocLite("Rgraphviz") #before installing this package, the user has to install the graphviz

>biocLite("SparseM")

### 2.2 Input

After building up the basic environment mentioned above, the users can install GO-function package and run this package for the DE gene example.

```
> library("GOFunction")
> data(exampledata)
```

```
> sigTerm <- GOFunction(interestGenes, refGenes, organism="org.Hs.eg.db",
+             ontology="BP", fdrmethod="BY", fdrth=0.05, ppth=0.05, pcth=0.05,
+             poth=0.05, peth=0.05, bmpSize=2000, filename="sigTerm")

Finding statistically significant terms...
Treating for local redundant terms...
Treating for global redundant terms...
Visualizing the GO DAG...

********************  Results  ********************
The number of annotated interesting genes: 7928
The number of annotated reference genes: 15403
The number of statistically significant terms: 474
The number of terms after treating local redundancy: 123
The number of terms after treating global redundancy: 97
Please see details about the significant terms in the files sigTerm.csv and sigTerm.bmp!
```

Here is a description of all the arguments needed to get the enrichment results.

1. *interestGenes* are the interesting genes and *refGenes* are the reference genes for a dataset. For the gene expression data, the differentially expressed genes are the interesting genes and all scanned genes in the profile are the reference genes. *interestGenes* and *refGenes* should be denoted using the Entrez gene ID.

2. GO-function can be currently applied to analyse data for 18 organisms and the user should install the corresponding gene annotation package when analysing data for these organisms. The 18 organisms and the corresponding packages are as follows: Anopheles "org.Ag.eg.db", Bovine "org.Bt.eg.db", Canine "org.Cf.eg.db", Chicken "org.Gg.eg.db", Chimp "org.Pt.eg.db", E coli strain K12 "org.EcK12.eg.db", E coli strain Sakai "org.EcSakai.eg.db", Fly "org.Dm.eg.db", Human "org.Hs.eg.db", Mouse "org.Mm.eg.db", Pig "org.Ss.eg.db", Rat "org.Rn.eg.db", Rhesus "org.Mmu.eg.db", Streptomyces coelicolor "org.Sco.eg.db", Worm "org.Ce.eg.db", Xenopus "org.Xl.eg.db", Yeast "org.Sc.sgd.db", Zebrafish "org.Dr.eg.db". The default is "org.Hs.eg.db" for human.

3. The default *ontology* is "BP" (Biological Process). The "CC" (Cellular Component) and "MF" (Molecular Function) ontologies can also be used.

4. GO-function provides three $p$ value correction methods: "bonferroni" [10], "BH" [11] and "BY" [12]. The default is "BY".

5. *fdrth* is the fdr cutoff to identify statistically significant GO terms. The default is 0.05.

6. *ppth* is the significant level to test whether the remaining genes of the ancestor term are enriched with interesting genes after removing the genes in its significant offspring terms. The default is 0.05.

7. *pcth* is the significant level to test whether the frequency of interesting genes in the offspring terms are significantly different from that in the ancestor term. The default is 0.05.

8. *poth* is the significant level to test whether the overlapping genes of one term is significantly different from the non-overlapping genes of the term. The default is 0.05.

9. *peth* is the significant level to test whether the non-overlapping genes of one term is enriched with interesting genes. The default is 0.05.

10. *bmpSize* is the width and height of the plot of GO DAG for all statistically significant terms. GO-function sets the default width and height of the plot as 2000 pixels in order to clearly show the GO DAG structure. If the GO DAG is very complexity, the user should increase *bmpSize*. Note: If there is an error at the step of "bmp(filename, width = 2000,...)" when running GO-function, the user should decrease *bmpSize*.

11. *filename* is the name of the files saving the table and the GO DAG of all statistically significant terms.

# 3 Output

There are two types of result output of GO-function. The first type is that GO-function saves a table to a CSV file (e.g. "sigTerm.csv") in the current working folder. As exemplified in Table 1, the table contains seven

columns: goid, name, refnum (the number of the reference genes in a GO term), interestnum (the number of the interesting genes in a GO term), pvalue, adjustp (the corrected $p$ value by the fdr control) and FinalResults. The "FinalResults" contains three types of terms: (1) "Local" represents terms removed after treating for local redundancy, (2) "Global" represents terms removed after treating for global redundancy, and (3) "Final" represents the remained terms with evidence that their significance should not be simply due to the overlapping genes. For the example of the gene expression profile, using GO-function, 108 GO terms are identified as statistically significant terms with a 5% FDR. After treating for local redundancy, 33 terms are remained. Finally, 25 terms are remained by GO-function after global redundancy treatment. Table1 shows three terms for each of the three types of terms and all terms can be found in Figure 1.

| goid | name | refnum | interestnum | pvalue | adjustp | FinalResult |
|---|---|---|---|---|---|---|
| GO:0000087 | M phase of mitotic cell cycle | 267 | 199 | 7.11E-15 | 3.15E-11 | Local |
| GO:0000278 | mitotic cell cycle | 449 | 323 | <2.20E-16 | <2.20E-16 | Local |
| GO:0000279 | M phase | 371 | 258 | 7.74E-13 | 2.16E-09 | Local |
| GO:0000226 | microtubule cytoskeleton organization | 156 | 107 | 1.07E-05 | 8.89E-03 | Global |
| GO:0007346 | regulation of mitotic cell cycle | 141 | 99 | 4.67E-06 | 4.24E-03 | Global |
| GO:0032268 | regulation of cellular protein metabolic process | 471 | 292 | 2.29E-06 | 2.29E-03 | Global |
| GO:0006260 | DNA replication | 222 | 166 | 7.72E-13 | 2.16E-09 | Final |
| GO:0007059 | chromosome segregation | 85 | 65 | 1.96E-06 | 2.0E-03 | Final |
| GO:0007067 | mitosis | 257 | 193 | 5.0E-15 | 2.51E-11 | Final |

**Table 1:** *An example of the output table by GO-function.*

GO-function also saves the structure of GO DAG for all statistically significant terms into a plot (e.g. "sigTerm.bmp") in the current working folder. As shown in Figure 1, "circle", "box" and "rectangle" represent the "Local", "Global" and "Final" terms in the table respectively. The different color shades represent the adjusted p values of the terms.

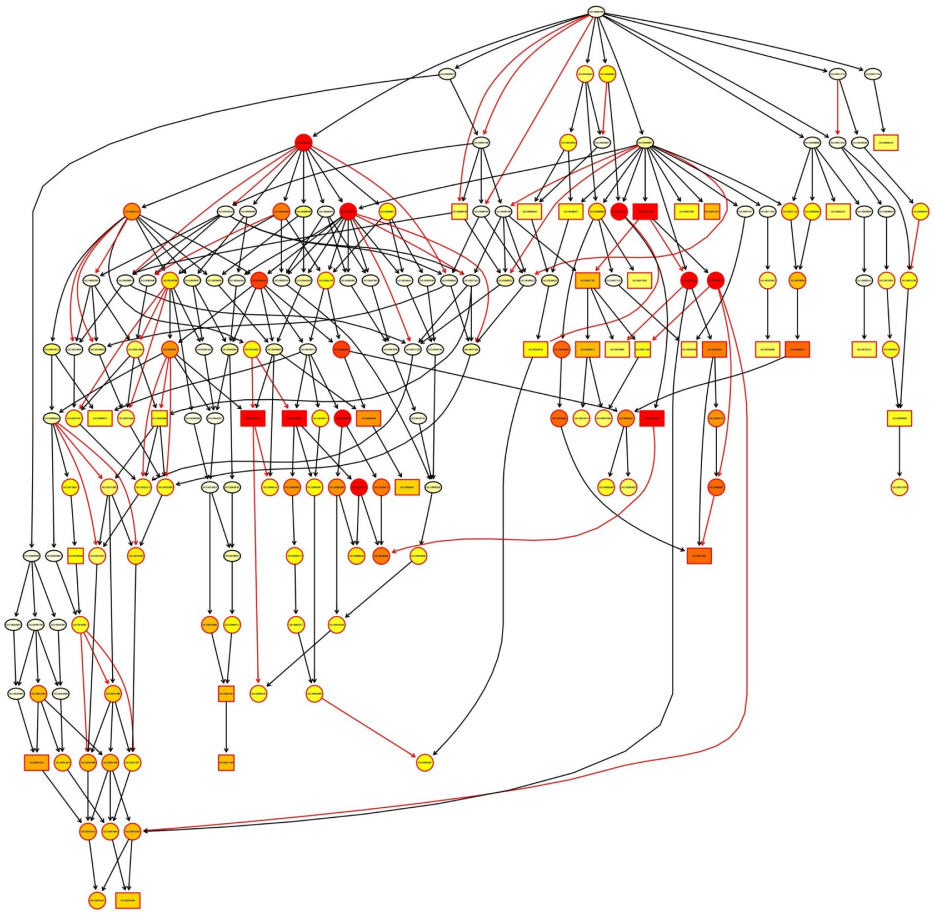

**Figure 1:** *An example of the output plot by GO-function*

## References


[1] The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):1600–1607, 2000.

[2] Alexa A., Rahnenfuhrer J., and Lengauer T. Improved scoring of functional groups from gene expression data by decorrelating go graph structure. *Bioinformatics*, 22(13):1600–1607, 2006.

[3] Grossmann S., Bauer S., Robinson P.N., and Vingron M. Improved detection of overrepresentation of gene-ontology annotations with parent child analysis. *Bioinformatics*, 23(22):3024–3031, 2007.

[4] Lu Y., Rosenfeld R., Simon I., Nau G.J., and Bar-Joseph Z. A probabilistic generative model for go enrichment analysis. *Nucleic Acids Research*, 36(17):e109, 2008.

[5] Bauer S., Gagneur J., and Robinson P.N. Going bayesian: model-based gene set analysis of genome-scale data. *Nucleic Acids Research*, 38(11):3523–3532, 2010.

[6] Sabates-Bellver J., Van der Flier L.G., de Palo M., Cattaneo E., Maake C., Rehrauer H., Laczko E., Kurowski M.A., Bujnicki J.M., Menigatti M., Luz J., Ranalli T.V., Gomes V., Pastorelli A., Faggiani R., Anti M., Jiricny J., Clevers H., and Marra G. Transcriptome profile of human colorectal adenomas. *Mol Cancer Res*, 5(12):1263–1275, 2007.

[7] Irizarry R.A., Hobbs B., Collin F., Beazer-Barclay Y.D., Antonellis K.J., Scherf U., and Speed T.P. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, 2003.

[8] Diehn M., Sherlock G., Binkley G., Jin H., Matese J.C., Hernandez-Boussard T., Rees C.A., Cherry J.M., Botstein D., Brown P.O., and Alizadeh A.A. Source: a unified genomic resource of functional annotations, ontologies, and gene expression data. *Nucleic Acids Research*, 31(1):219–223, 2003.

[9] Tusher V.G., Tibshirani R., and Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A*, 98(9):5116–5121, 2001.

[10] Bland J.M. and Altman D.G. Multiple significance tests: The bonferroni method. *British Medical Journal*, 310:170, 1995.

[11] Benjamini Y. and Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B. Methodological*, 57:289–330, 1995.

[12] Benjamini Y. and Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, 29(4):1165–1188, 2001.