Using \textit{clusterProfiler} to identify and compare functional profiles of gene lists

Guangchuang Yu
School of Biological Sciences
The University of Hong Kong, Hong Kong SAR, China
e-mail: guangchuangyu@gmail.com

February 26, 2015

Contents

1 Introduction ........................................2
2 Citation .............................................2
3 Supported organisms .................................2
4 Gene Ontology Classification .......................3
5 Enrichment Analysis .................................4
   5.1 Hypergeometric model ..............................4
   5.2 Gene set enrichment analysis .....................4
   5.3 GO enrichment analysis ...........................5
   5.4 KEGG pathway enrichment analysis ..............6
   5.5 DO enrichment analysis ...........................8
   5.6 Reactome pathway enrichment analysis ..........8
   5.7 Function call ....................................8
   5.8 Visualization ....................................8
      5.8.1 barplot ....................................8
      5.8.2 enrichMap ..................................10
      5.8.3 cnetplot ..................................10
      5.8.4 gseaplot ..................................13
      5.8.5 pathview from pathview package ..........13
6 Biological theme comparison .........................14
7 Session Information ..................................16
1 Introduction

In recent years, high-throughput experimental techniques such as microarray, RNA-Seq and mass spectrometry can detect cellular molecules at systems-level. These kinds of analyses generate huge quantities of data, which need to be given a biological interpretation. A commonly used approach is via clustering in the gene dimension for grouping different genes based on their similarities [1].

To search for shared functions among genes, a common way is to incorporate the biological knowledge, such as Gene Ontology (GO) and Kyoto Encyclopedia of genes and Genomes (KEGG), for identifying predominant biological themes of a collection of genes.

After clustering analysis, researchers not only want to determine whether there is a common theme of a particular gene cluster, but also to compare the biological themes among gene clusters. The manual step to choose interesting clusters followed by enrichment analysis on each selected cluster is slow and tedious. To bridge this gap, we designed clusterProfiler [2], for comparing and visualizing functional profiles among gene clusters.

2 Citation

Please cite the following articles when using clusterProfiler.


3 Supported organisms

At present, clusterProfiler about 20 species as shown below:

- Arabidopsis
- Anopheles
- Bovine
- Canine
- Chicken
- Chimp
- E coli strain K12
These species are all supported by GO and KEGG analyses. GO analyses also support *Coelicolor* and *Gondii*.

### 4 Gene Ontology Classification

In *clusterProfiler*, `groupGO` is designed for gene classification based on GO distribution at a specific level.

```r
require(DOSE)
data(geneList)
gene <- names(geneList)[abs(geneList) > 2]
head(gene)
## [1] "4312"   "8318"   "10874"  "55143"  "55388"  "991"

ggo <- groupGO(gene = gene, organism = "human", ont = "BP",
               level = 3, readable = TRUE)
head(summary(ggo))
## ID                Description                          Count GeneRatio
## GO:0019953    sexual reproduction                     10 10/207
## GO:0019954   asexual reproduction                     0  0/207
## GO:0032504   multicellular organism reproduction     11 11/207
## GO:0032505 reproduction of a single-celled organism   0  0/207
```
5 Enrichment Analysis

5.1 Hypergeometric model

Enrichment analysis [3] is a widely used approach to identify biological themes. Here we implement hypergeometric model to assess whether the number of selected genes associated with disease is larger than expected.

To determine whether any terms annotate a specified list of genes at frequency greater than that would be expected by chance, clusterProfiler calculates a p-value using the hypergeometric distribution:

\[
p = 1 - \sum_{i=0}^{k-1} \frac{M \choose i \cdot (N-M) \choose n-i}{N \choose n}
\]

In this equation, \( N \) is the total number of genes in the background distribution, \( M \) is the number of genes within that distribution that are annotated (either directly or indirectly) to the node of interest, \( n \) is the size of the list of genes of interest and \( k \) is the number of genes within that list which are annotated to the node. The background distribution by default is all the genes that have annotation.

P-values were adjusted for multiple comparison, and q-values were also calculated for FDR control.

5.2 Gene set enrichment analysis

A common approach in analyzing gene expression profiles was identifying differentially expressed genes that are deemed interesting. The enrichment analysis we demonstrated previous were based on these differential expressed genes. This approach will find genes where the difference is large, but it will not detect a situation where the difference is small, but evidenced in coordinated way in a set of related genes. Gene Set Enrichment Analysis (GSEA) [4] directly addresses this limitation. All genes can be used in GSEA; GSEA aggregates the per gene statistics across genes within a gene set, therefore making it possible to detect situations where all genes in a predefined set change in a small but coordinated
way. Since it is likely that many relevant phenotypic differences are manifested by small but consistent changes in a set of genes.

Genes are ranked based on their phenotypes. Given a priori defined set of gens $S$ (e.g., genes sharing the same GO or KEGG category), the goal of GSEA is to determine whether the members of $S$ are randomly distributed throughout the ranked gene list ($L$) or primarily found at the top or bottom.

There are three key elements of the GSEA method:

- **Calculation of an Enrichment Score.**
  The enrichment score ($ES$) represent the degree to which a set $S$ is over-represented at the top or bottom of the ranked list $L$. The score is calculated by walking down the list $L$, increasing a running-sum statistic when we encounter a gene in $S$ and decreasing when it is not. The magnitude of the increment depends on the gene statistics (e.g., correlation of the gene with phenotype). The $ES$ is the maximum deviation from zero encountered in the random walk; it corresponds to a weighted Kolmogorov-Smirnov-like statistic [4].

- **Estimation of Significance Level of $ES$.**
  The $p$-value of the $ES$ is calculated using permutation test. Specifically, we permute the gene labels of the gene list $L$ and recompute the $ES$ of the gene set for the permutated data, which generate a null distribution for the $ES$. The $p$-value of the observed $ES$ is then calculated relative to this null distribution.

- **Adjustment for Multiple Hypothesis Testing.**
  When the entire GO or KEGG gene sets is evaluated, clusterProfiler adjust the estimated significance level to account for multiple hypothesis testing and also $q$-values were calculated for FDR control.

### 5.3 GO enrichment analysis

```r
ego <- enrichGO(gene = gene, universe = names(geneList),
                organism = "human", ont = "CC", pvalueCutoff = 0.01,
                readable = TRUE)
head(summary(ego))
```

<table>
<thead>
<tr>
<th>#</th>
<th>ID</th>
<th>Description</th>
<th>GeneRatio</th>
</tr>
</thead>
<tbody>
<tr>
<td># G0:0005819 G0:0005819</td>
<td>spindle</td>
<td>24/196</td>
<td></td>
</tr>
<tr>
<td># G0:0015630 G0:0015630</td>
<td>microtubule cytoskeleton</td>
<td>37/196</td>
<td></td>
</tr>
<tr>
<td># G0:0005876 G0:0005876</td>
<td>spindle microtubule</td>
<td>10/196</td>
<td></td>
</tr>
<tr>
<td># G0:0000793 G0:0000793</td>
<td>condensed chromosome</td>
<td>16/196</td>
<td></td>
</tr>
<tr>
<td># G0:0000779 G0:0000779</td>
<td>condensed chromosome, centromeric region</td>
<td>12/196</td>
<td></td>
</tr>
<tr>
<td># G0:0044430 G0:0044430</td>
<td>cytoskeletal part</td>
<td>43/196</td>
<td></td>
</tr>
<tr>
<td>#</td>
<td>BgRatio</td>
<td>pvalue</td>
<td>p.adjust</td>
</tr>
</tbody>
</table>
```
5.4 KEGG pathway enrichment analysis
```r
ck <- enrichKEGG(gene = gene, organism = "human", pvalueCutoff = 0.01, 
readable = TRUE)
head(summary(ck))

## ID Description GeneRatio BgRatio
## hsa04110 hsa04110 Cell cycle 11/74 128/5894
## hsa04114 hsa04114 Oocyte meiosis 10/74 114/5894
## hsa03320 hsa03320 PPAR signaling pathway 7/74 70/5894
## hsa04914 hsa04914 Progesterone-mediated oocyte maturation 6/74 87/5894
## hsa04062 hsa04062 Chemokine signaling pathway 8/74 189/5894
## hsa04060 hsa04060 Cytokine-cytokine receptor interaction 9/74 265/5894

pvalue p.adjust qvalue
## hsa04110 4.31e-07 3.02e-06 4.54e-07
## hsa04114 1.25e-06 4.38e-06 6.59e-07
## hsa03320 2.35e-06 5.49e-05 8.25e-06
## hsa04914 7.21e-04 1.26e-03 1.90e-04
## hsa04062 2.37e-03 3.32e-03 5.00e-04
## hsa04060 5.58e-03 6.51e-03 9.79e-04

geneID Count
## hsa04110 CDC45/CDC20/CCNB2/CCNA2/CDK1/MAD2L1/TTK/CHEK1/CCNB1/MCM5/PTTG1 11
## hsa04114 CDC20/CCNB2/CDK1/MAD2L1/CALML5/AURKA/CCNB1/PTTG1/ITPR1/PGR 10
## hsa03320 MMP1/FADS2/ADIPQ/PCK1/FABP4/HMGCS2/PLIN1 7
## hsa04914 CCNB2/CCNA2/CDK1/MAD2L1/CCNB1/PGR 6
## hsa04062 CXCL10/CXCL13/CXCL11/CXCL9/CCL8/CXCL14/CX3CR1 8
## hsa04060 CXCL10/CXCL13/CXCL11/CXCL9/CCL18/IL1R2/CCL8/CXCL14/CX3CR1 9

ck2 <- gseKEGG(geneList = geneList, organism = "human", 
nPerm = 100, minGSSize = 120, pvalueCutoff = 0.01,
verbose = FALSE)
head(summary(ck2))

## ID Description setSize
## hsa04062 hsa04062 Chemokine signaling pathway 166
## hsa04510 hsa04510 Focal adhesion 193
## hsa03013 hsa03013 RNA transport 124
## hsa04060 hsa04060 Cytokine-cytokine receptor interaction 233

enrichmentScore pvalue p.adjust qvalues
## hsa04062 0.383 0 0 0
## hsa04510 -0.446 0 0 0
## hsa03013 0.427 0 0 0
## hsa04060 0.339 0 0 0
```
5.5 DO enrichment analysis

Disease Ontology (DO) enrichment analysis is implemented in *DOSE*, please refer to the package vignettes. The `enrichDO` function is very useful for identifying disease association of interesting genes, and function `gseAnalyzer` function is designed for gene set enrichment analysis of *DO*.

5.6 Reactome pathway enrichment analysis

With the demise of KEGG (at least without subscription), the KEGG pathway data in Bioconductor will not update and we encourage user to analyze pathway using *ReactomePA* which use Reactome as a source of pathway data. The function call of `enrichPathway` and `gsePathway` in *ReactomePA* is consistent with `enrichKEGG` and `gseKEGG`.

5.7 Function call

The function calls of `groupGO`, `enrichGO`, `enrichKEGG`, `enrichDO` and `enrichPathway` are consistent. The input parameters of `gene` is a vector of entrezgene (for human and mouse) or ORF (for yeast) IDs, and `organism` should be supported species (please refer to the manual of the specific function).

For gene set enrichment analysis, the function of `gseGO`, `gseKEGG`, `gseAnalyzer` and `gsePathway` need extra parameter `nPerm` to specify the permutation number.

For GO analysis, `ont` must be assigned to one of "BP", "MF", and "CC" for biological process, molecular function and cellular component, respectively. In `groupGO`, the `level` specify the GO level for gene projection.

In enrichment analysis, the `pvalueCutoff` is to restrict the result based on their pvalues and the adjusted p values. *Q-values* were also calculated for controlling false discovery rate (FDR).

The `readable` is a logical parameter to indicate the input gene IDs will map to gene symbols or not.

5.8 Visualization

The output of `groupGO`, `enrichGO` and `enrichKEGG` can be visualized by bar plot, enrichment map and category-gene-network plot. It is very common to visualize the enrichment result in bar or pie chart. We believe the pie chart is misleading and only provide bar chart.

5.8.1 barplot
barplot(ggo, drop = TRUE, showCategory = 12)

barplot(ego, showCategory = 8)
5.8.2 enrichMap

Enrichment map can be visualized by enrichMap, which supports results obtained from hypergeometric test and gene set enrichment analysis.

enrichMap(ego)

Figure 1: enrichment map of enrichment result

enrichMap(ego2)

5.8.3 cnetplot

In order to consider the potentially biological complexities in which a gene may belong to multiple annotation categories and provide information of numeric changes if available, we developed cnetplot function to extract the complex association.
Figure 2: enrichment map of gsea result

cnetplot(ego, categorySize = "pvalue", foldChange = geneList)
cnetplot(kk, categorySize = "geneNum", foldChange = geneList)
5.8.4 gseaplot

Running score of gene set enrichment analysis and its association of phenotype can be visualized by gseaplot.

```
gseaplot(kk2, geneSetID = "hsa04145")
```

5.8.5 pathview from pathview package

`clusterProfiler` users can also use `pathview` from the `pathview` [5] to visualize KEGG pathway.

The following example illustrate how to visualize "hsa04110" pathway, which was enriched in our previous analysis.
Figure 3: plotting gsea result

```r
require(pathview)
hsa04110 <- pathview(gene.data = geneList, pathway.id = "hsa04110",
    species = "hsa", limit = list(gene = max(abs(geneList)),
    cpd = 1))

## Info: Downloading xml files for hsa04110, 1/1 pathways..
## Info: Downloading png files for hsa04110, 1/1 pathways..
## Info: Working in directory /tmp/RtmpA6HVmO/Rbuild6f9e5c4df50c/clusterProfiler/vignettes
## Info: Writing image file hsa04110.pathview.png

For further information, please refer to the vignette of `pathview` [5].

6 Biological theme comparison

`clusterProfiler` was developed for biological theme comparison, and it provides a function, `compareCluster`, to automatically calculate enriched functional categories of each gene clusters.

```r
data(gcSample)
ck <- compareCluster(geneCluster = gcSample, fun = "enrichKEGG")
plot(ck)
```
Figure 4: visualize KEGG pathway using pathview
By default, only top 5 (most significant) categories of each cluster was plotted. User can changes the parameter `showCategory` to specify how many categories of each cluster to be plotted, and if `showCategory` was set to `NULL`, the whole result will be plotted.

The dot sizes were based on their corresponding row percentage by default, and user can set the parameter `by` to "count" to make the comparison based on gene counts. The parameter `by` can also set to "rowPercentage" to normalize the dot sizes, since some categories may contain a large number of genes, and make the dot sizes of those small categories too small to compare. The default parameter `by` is setting to "geneRatio", which corresponding to the "GeneRatio" column of the output. To provide the full information, we also provide number of identified genes in each category (numbers in parentheses) when `by` is setting to "rowPercentage" and number of gene clusters in each cluster label (numbers in parentheses) when `by` is setting to "geneRatio", as shown in Figure 3. If the dot sizes were based on "count", the row numbers will not shown.

The p-values indicate that which categories are more likely to have biological meanings. The dots in the plot are color-coded based on their corresponding p-values. Color gradient ranging from red to blue correspond to in order of increasing p-values. That is, red indicate low p-values (high enrichment), and blue indicate high p-values (low enrichment). P-values and adjusted p-values were filtered out by the threshold giving by parameter `pvalueCutoff`, and FDR can be estimated by `qvalue`.

User can refer to the example in [2]; we analyzed the publicly available expression dataset of breast tumour tissues from 200 patients (GSE11121, Gene Expression Omnibus) [6]. We identified 8 gene clusters from differentially expressed genes, and using `compareCluster` to compare these gene clusters by their enriched biological process.

Another example was shown in [7], we calculated functional similarities among viral miRNAs using method described in [8], and compared significant KEGG pathways regulated by different viruses using `compareCluster`.

The comparison function was designed as a general-package for comparing gene clusters of any kind of ontology associations, not only `groupGO`, `enrichGO`, and `enrichKEGG` this package provided, but also other biological and biomedical ontologies, for instance, `enrichDO` from `DOSE` and `enrichPathway` from `ReactomePA` work fine with `compareCluster` for comparing biological themes in disease and reactome pathway perspective. More details can be found in the vignettes of `DOSE` and `ReactomePA`.

## 7 Session Information

The version number of R and packages loaded for generating the vignette were:

- R version 3.1.2 (2014-10-31), x86_64-unknown-linux-gnu
• Locale: LC_CTYPE=en_US.UTF-8, LC_NUMERIC=C, LC_TIME=en_US.UTF-8, LC_COLLATE=C, LC_MONETARY=en_US.UTF-8, LC_MESSAGES=en_US.UTF-8, LC_PAPER=en_US.UTF-8, LC_NAME=C, LC_ADDRESS=C, LC_TELEPHONE=C, LC_MEASUREMENT=en_US.UTF-8, LC_IDENTIFICATION=C

• Base packages: base, datasets, grDevices, graphics, methods, parallel, stats, stats4, utils

• Other packages: AnnotationDbi 1.28.1, Biobase 2.26.0, BiocGenerics 0.12.1, DBI 0.3.1, DOSE 2.4.0, GO.db 3.0.0, GenomeInfoDb 1.2.4, IRanges 2.0.1, KEGGgraph 1.24.0, RSQLite 1.0.0, S4Vectors 0.4.0, XML 3.98-1.1, clusterProfiler 2.0.1, graph 1.44.1, knitr 1.9, org.Hs.eg.db 3.0.0, pathview 1.6.0

• Loaded via a namespace (and not attached): Biostrings 2.34.1, DO.db 2.8.0, GOSemSim 1.24.1, KEGG.db 3.0.0, KEGGREST 1.6.4, MASS 7.3-39, Rcpp 0.11.4, Rgraphviz 2.10.0, XVector 0.6.0, codetools 0.2-10, colorspace 1.2-4, digest 0.6.8, evaluate 0.5.5, formatR 1.0, ggplot2 1.0.0, grid 3.1.2, gtable 0.1.2, highr 0.4, htr 0.6.1, igraph 0.7.1, labeling 0.3, munsell 0.4.2, plyr 1.8.1, png 0.1-7, proto 0.3-10, qvalue 1.40.0, reshape2 1.4.1, scales 0.2.4, stringr 0.6.2, tools 3.1.2, zlibbioc 1.12.0

References


