Getting started with *goTools* package

Agnes Paquet[1] and (Jean) Yee Hwa Yang[2]

August 24, 2004

1. Functional Genomics Core Facility, University of California, San Francisco,
apaquet@medsfgh.ucsf.edu
2. Department of Medicine, University of California, San Francisco,
http://www.biostat.ucsf.edu/jean

Contents

1 Getting started

This document provides a tutorial for the *goTools* package, which allow graphical comparisons of functional groups between two sets of genes.

Installing the package: To install the *goTools* package, go to the Bioconductor installation web site http://www.bioconductor.org/faq.html for more detailed instructions.

Help files: As with any R package, detailed information on functions, classes and methods can be obtained in the help files. For instance, to view the help file for the function `ontoCompare` in a browser, use `help.start()` followed by `?ontoCompare`.
We demonstrate the functionality with a randomly selected set of probe IDs from both Affymetrix hgu133a chip (`affylist`)and Operon version 2 Human oligos (`operonlist`). To load the `probeID` dataset, use `data(probeID)`, and to view a description of the experiments and data, type `?probeID`.

Sweave: This document was generated using the `Sweave` function from the R *tools* package. The source file is in the **/inst/doc** directory of the package *goTools*.

To begin, let's load the package and the probeID datasets into your R session.

```
> library("goTools", verbose = FALSE)
> data(probeID)
> ls()

[1] "affylist"   "operonlist"
```

As shown below, `affylist` is a list containing 3 vectorsof probe ids from Affymetrix hgu133a chip, and `operonlist` is a list of 2 vectors of probe ids from Operon Version 2 Human oligos.

```
> class(affylist)

[1] "list"

> length(affylist)

[1] 3

> affylist[[1]][1:5]

[1] "215828_at"   "201849_at"   "219719_at"   "213690_s_at" "203172_at"

> length(operonlist)

[1] 2

> operonlist[[1]][1:5]

[1] "H200000481" "H200012124" "H200016088" "H200001913" "H200016976"
```

2 Graphical comparisons of two sets of genes

Gene Ontology is a Direct Acyclic Graph (DAG) that provides three structured networks of defined terms to describe gene product attributes: Molecular Function (MF), Biological Process (BP) and Cellular Component (CC). A gene product has one or more molecular functions and is used in one or more biological processes; it might be associated with one or more cellular components. To learn more about GO and DAG, please refer to Gene Ontology web site (http://www.geneontology.org/).

We have created a set of R functions that use GO structure to describe and compare the composition of sets of genes (or probes). We use the following algorithm:

1. Read in a **list of sets of probe id you want to compare.**

2. Map each probe id to corresponding ontologies in the GO tree, if any.

3. Create the set of GO ids of interest used to compare your datasets (endnode). The function EndNodeList() will create a set of nodes of the DAG located one level under MF, BP or CC, but you can use any sets of GO ids.

4. For each GO id, go up the GO tree until reaching the nodes in endnode. Search may be limited to MF, BP or CC if specified in goType.

5. Compute the percentage of direct children found under each node in endnode.

6. Return the results. Plot them if plot=TRUE.

2.1 How to use goTools

The main function that we provide is ontoCompare. It takes as argument **a list** of probe ids. Their type must be specified in the argument probeType. For more details about it,you can refer to the corresponding help file by typing: ?ontoCompare.

2.1.1 Examples

The following examples demonstrate how to use ontoCompare on the probeID dataset.

```
> library(GO)
> library(annotate)

Loading required package: Biobase
Welcome to Bioconductor
        Vignettes contain introductory material.  To view,
        simply type: openVignette()
        For details on reading vignettes, see
        the openVignette help page.

> res <- ontoCompare(affylist, probeType = "hgu133a")

[1] "Starting ontoCompare..."
[1] "Number of lists > 1"
[1] "Using method: TGenes"

> res[, 1:2]

   antioxidant activity apoptosis regulator activity
L1                 0.01                     0.01000000
L2                 0.00                     0.01111111
L3                 0.00                     0.01428571

> res2 <- ontoCompare(operonlist["L1"], probeType = "operon", method = "TIDS")
```

```
[1] "Starting ontoCompare..."
Downloading Oligo 2 GO annotation table ...done.
[1] 41396      2
[1] "Number of lists = 1"
[1] "Using method: TIDS"
```

2.1.2 Methods for computing percentages

ontoCompare allows you to choose from 3 different methods to estimate the percentage of probes under each element of endnode. The default method is TGenes.

1. TGenes: for each end node, return the number of direct children found / total number of probe ids.
   This includes oligos which do not have GO annotations.

2. TIDS: for each end node, return the number of direct children found / total number of GO ids describing the list.

3. none: for each end node, return the number of direct children found.

2.2 Plotting the results

The plots are produced using the function ontoPlot. It is called by ontoCompare when you set plot=TRUE. You can also call it directly, passing as argument ontoCompare results. If only one set of genes is passed to ontoCompare, ontoPlot will return a pie chart. In other cases, it will return a bargraph. You can modify ontoPlot layout parameters using usual R graphics layout parameters. For more details, type ?par.

```
> res <- ontoCompare(affylist, probeType = "hgu133a", plot = TRUE)

[1] "Starting ontoCompare..."
[1] "Number of lists > 1"
[1] "Using method: TGenes"
```

The folowing R code generates Figure 1b. It uses the list L1 from the dataset operonlist, and the percentages are calculated using the TIDS method. As is only one list is given as input argument, the call to ontoPlot returns a pie chart.

```
> par(mar = c(5, 8, 5, 8))
> res2 <- ontoCompare(operonlist["L1"], probeType = "operon", method = "TIDS",
+      plot = FALSE)

[1] "Starting ontoCompare..."
[1] 41396      2
[1] "Number of lists = 1"
[1] "Using method: TIDS"

> ontoPlot(res2, cex = 0.7)
```

See Figure 1

3 How to set up the "end nodes"

3.1 Default list

The default end nodes list is defined by a call to the function `EndNodeList`. It contains all children of MF(GO:0003674), BP(GO:0008150) and CC (GO:0005575).

```
> EndNodeList()

 [1] "GO:0003674" "GO:0005575" "GO:0008150" "GO:0008435" "GO:0016172"
 [6] "GO:0016209" "GO:0016329" "GO:0005488" "GO:0003824" "GO:0005194"
[11] "GO:0003754" "GO:0030188" "GO:0003793" "GO:0030234" "GO:0019833"
[16] "GO:0005554" "GO:0003774" "GO:0045735" "GO:0008369" "GO:0017028"
[21] "GO:0008638" "GO:0004871" "GO:0005198" "GO:0019214" "GO:0015070"
[26] "GO:0030528" "GO:0045182" "GO:0005215" "GO:0030533" "GO:0005623"
[31] "GO:0008372" "GO:0005576" "GO:0019814" "GO:0008370" "GO:0005941"
[36] "GO:0019012" "GO:0007610" "GO:0000004" "GO:0009987" "GO:0007275"
[41] "GO:0008371" "GO:0007582" "GO:0016032"
```

3.2 Customized end node list

If you want to use more ontologies to describe your set of genes, you can use the function `CustomEndNodeList(id,rank)` to create a bigger set of end nodes. It returns all GO ids children of `id` up to `rank` levels below `id`.

```
> MFendnode <- CustomEndNodeList("GO:0003674", rank = 2)

[1] "rank= 1"
[1] "rank= 2"
```

Finally, the code below shows you how to use a custom end node list, and also how to modify the `goType` argument to select only Molecular Function (MF) ontologies.

```
> res <- ontoCompare(affylist, probeType = "hgu133a", endnode = MFendnode,
+     goType = "MF")

[1] "Starting ontoCompare..."
[1] "Number of lists > 1"
[1] "Using method: TGenes"
```
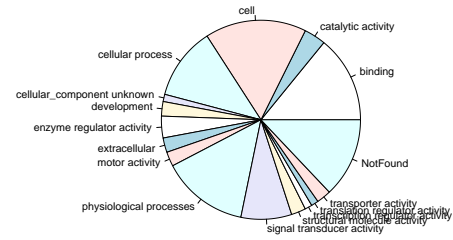
You can also create a list of GO ids of nodes of interest and pass it directly to the `endnode` argument in `ontoCompare`. GO ids must be in the following format: "GO:XXXXXXX."

References
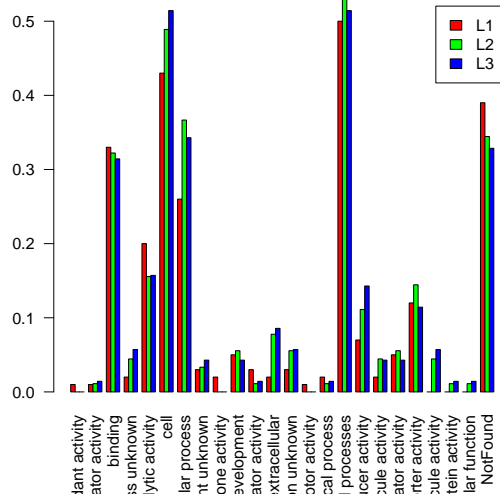
Figure 1: Plots obtained using `ontoCompare` on `probeID` dataset