UNIVERSITÄT BERN

Bulk RNA-seq

Gene ontology (GO) enrichment analysis

Interfaculty Bioinformatics Unit, University of Bern

03.06.2021

SIB Swiss Institute of Bioinformatics

The Gene Ontology (GO)

http://geneontology.org/

UNIVERSITÄT BERN

The Gene Ontology project provides controlled vocabularies of defined terms representing gene product properties. These cover three domains:

1) Cellular component (CC):
   These terms describe a component of a cell that is part of a larger object, such as an anatomical structure (e.g. rough endoplasmic reticulum or nucleus) or a gene product group (e.g. ribosome, proteasome or a protein dimer).

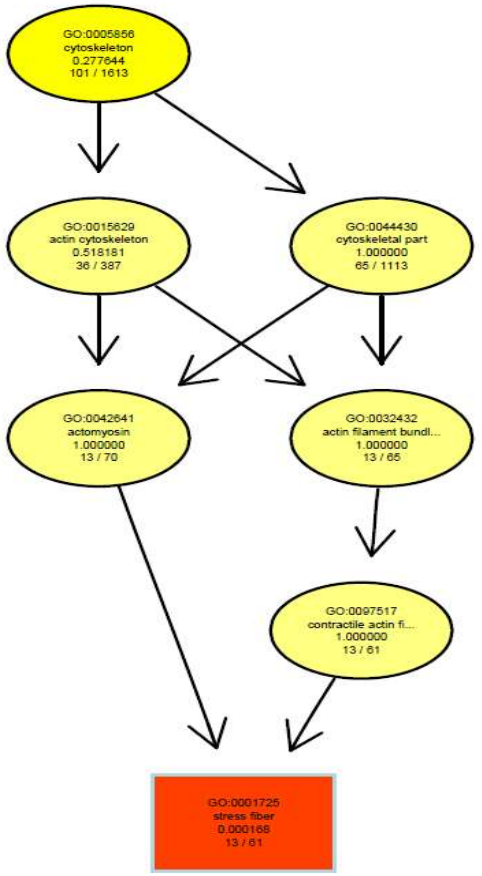
2) Biological Process (BP): → Often tends to be the most interesting category
   A biological process term describes a series of events accomplished by one or more organized assemblies of molecular functions. Examples of broad biological process terms are "cellular physiological process" or "signal transduction". Examples of more specific terms are "pyrimidine metabolic process" or "alpha-glucoside transport". The general rule to assist in distinguishing between a biological process and a molecular function is that a process must have more than one distinct steps.

3) Molecular Function (MF):
   Molecular function terms describes activities that occur at the molecular level, such as "catalytic activity" or "binding activity". GO molecular function terms represent activities rather than the entities (molecules or complexes) that perform the actions, and do not specify where, when, or in what context the action takes place. Molecular functions generally correspond to activities that can be performed by individual gene products, but some activities are performed by assembled complexes of gene products. Examples of broad functional terms are "catalytic activity" and "transporter activity"; examples of narrower functional terms are "adenylate cyclase activity" or "Toll receptor binding".

The GO as a graph

The structure of the GO can be described in terms of a graph (see example on right), where each GO term is a node, and the relationships between the terms are edges between the nodes. GO is loosely hierarchical, with 'child' terms being more specialized than their 'parent' terms, but unlike a strict hierarchy, a term may have more than one parent term.

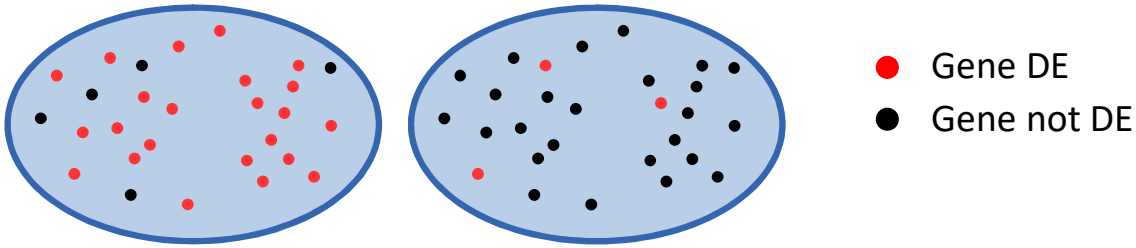
http://geneontology.org/page/ontology-structure



GO enrichment analysis

Goal: Detect which processes are affected by the experiment

Principle: Test if differentially expressed (DE) genes are significantly overrepresented within a particular GO term



A) Genes that belong to GO term

B) Genes that DO NOT belong to GO term

Is the proportion of red genes higher in A than in B?

GO enrichment analysis

UNIVERSITÄT BERN

There are many different tools to perform GO enrichment analysis, and currently there is no consensus in the literature as to which one is the best. It is also very well possible that there is no single tool that always performs best in all datasets.

Your results were produced with **topGO** which is a widely used R Bioconductor package. An advantage of topGO is that it can take into account the hierarchical structure of the GO (i.e. the parent-child relationships).

BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 13 2006, pages 1600–1607
doi:10.1093/bioinformatics/btl140

Gene expression

Improved scoring of functional groups from gene expression data by decorrelating GO graph structure

Adrian Alexa*, Jörg Rahnenführer and Thomas Lengauer
Max-Planck-Institute for Informatics, Stuhlsatzenhausweg 85, D-66123 Saarbrücken, Germany

Received on September 28, 2005; revised on March 30, 2006; accepted on April 4, 2006
Advance Access publication April 10, 2006
Associate Editor: Martin Bishop

http://bioconductor.org/packages/release/bioc/html/topGO.html

topGO results file

UNIVERSITÄT BERN

For each pairwise comparison between experimental groups, you will receive a text file. The file name is set up like this: **Condition1.Condition2.topGoResults.txt**

Identifier and description of the GO term

Position the GO term would have if table were ranked by classic.Fisher P-value rather than weight01.Fisher

See slide 2

GO.ID	Term	Annotated	Significant	Expected	Rank in classic.Fisher	weight01.Fisher	weight01.KS	classic.Fisher	ontology
GO:0008201	heparin binding	118	61	28.63	5	1.10E-10	2.60E-09	1.10E-10	MF
GO:0008013	beta-catenin binding	75	39	18.2	10	2.00E-07	5.00E-06	2.00E-07	MF
....									
GO:0090090	negative regulation of canonical Wnt sig...	92	49	22.42	174	2.20E-09	2.00E-09	2.20E-09	BP
GO:0002053	positive regulation of mesenchymal cell ...	32	23	7.8	213	2.00E-08	1.90E-07	2.00E-08	BP
....									

Total number of genes assigned to GO term and actually detected in our dataset

Number of these genes that are detected as differentially expressed in our DE analysis (adjusted-P<0.05)

Number of differentially expressed (DE) genes we would expect to see if DE genes were randomly distributed across GO terms

P-values from three different ways to perform the enrichment test → see next slide for details

The results are ranked by weight01.Fisher and the **top 50** terms are output for each of the three subontologies, **provided at least one of the three P-values is below 0.05. No correction for multiple testing is applied** as it is not clear how to correctly do this (See Section 6.2. of http://bioconductor.org/packages/release/bioc/vignettes/topGO/inst/doc/topGO.pdf)

topGO results file

UNIVERSITÄT BERN

The table contains P-values from 3 different enrichment tests. This lets you assess how consistently a particular term is detected as significant.

weight01.Fisher	weight01.KS	classic.Fisher
1.10E-10	2.60E-09	1.10E-10
2.00E-07	5.00E-06	2.00E-07

The three analyses differ in

1) Whether or not they consider the **hierarchical structure of the GO**

   weight01: Considers the GO graph and tries to find the most interesting term in a particular region of the graph (see Alexa et al. 2006 for details). It tends to prioritise more specific terms (i.e. children) over more general terms.

   classic: Performs a separate test for each GO term, ignoring the overlap between terms. It tends to favour larger terms.

2) How they **rank the genes** within a GO term

   Fisher: Performs a Fisher's exact test which compares the proportion of differentially expressed (DE) genes among all genes assigned to the GO term and all other genes (→ Slide 4). This approach relies on a fixed threshold that determines if a gene is considered DE or not.

   KS (Kolmogorov-Smirnov test): Orders all genes by P-value and tests if the genes assigned to a particular GO term are enriched at the top or the bottom of this table (Subramanian et al. 2005 PNAS).

topGO results file

UNIVERSITÄT BERN

For each pairwise comparison between experimental groups, you will also receive three pdf files which show how the detected terms are distributed across the GO graph. These plots include the **TOP 10** terms based on weight01.Fisher, **without applying a P-value cut-off.**

The file name is set up like this: **Condition1.Condition2.subontology_weight01_10_all.pdf**. Subontology is one of CC, MF, BP.

The top 10 terms are shown as rectangles, all other terms as ovals. The colour of the box indicates the relative significance (red=most significant).