

# TreeHugger: A New Test for Enrichment of Gene Ontology Terms

Daniel Jupiter

Postdoctoral Research Associate  
Department of Systems Biology And Translational Medicine  
Texas A&M Health Science Center

# Methods for Testing GO Term Significance

1. For each term, does that term annotate more differentially expressed genes than expected? (hypergeometric test)
2. For each term, are the annotated and unannotated genes mutually differentially expressed? (GSEA)
3. For a given term of interest, are the genes annotated with that term predictive of clinical outcome? (global test)

Most tests do not take the structure of the ontology into account. This causes problems both in terms of understanding test results, and multiple test correction.

We focus on the first question, and build a test taking structure into account.

# TreeHugger: Idea

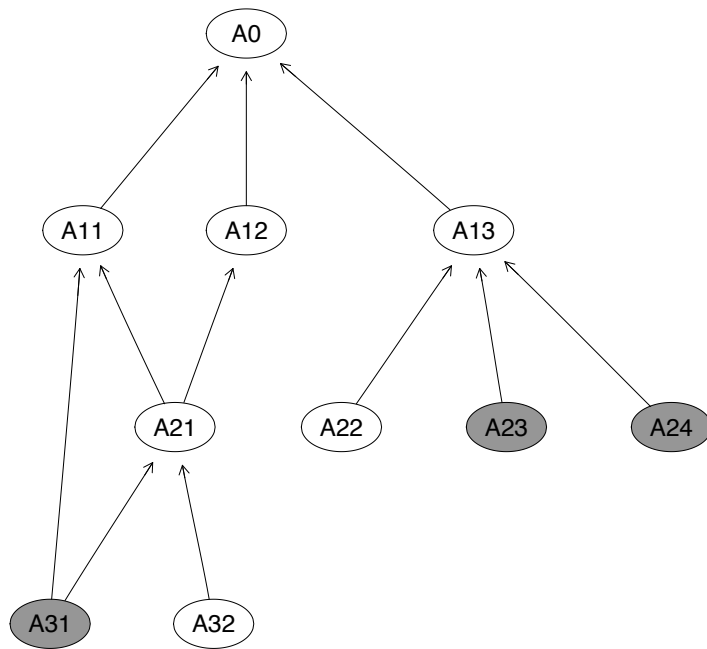
Use the structure of the GO graph directly in building the test.

1. Begin by studying genes, not terms. (gene by gene, not term by term)
2. For a given gene, terms that directly annotate the gene are given high score.
3. For a given gene, terms whose descendants annotate the gene are given lower scores. (Thus the root terms are given very low scores.)
4. Precalculate all term scores for all genes.
5. Given gene sets of interest, and a term of interest, compare the mean scores for that term between the various gene sets (t-test).

Specifically:

1. A directly annotating term is scored as 1.
2. A term whose children are annotated is scored the average of its children's scores.
3. Any other term (no annotation direct or indirect) is scored 0.

# Example Scoring for a Hypothetical Gene



Term	Traditional Score	TreeHugger Score
A23	1	1
A24	1	1
A22	0	0
A13	1	2/3
A31	1	1
A32	0	0
A21	1	1/2
A11	1	.75
A12	1	1/2
A0	1	.639

# Results

We generated simulated annotated gene sets based on the DAG in the previous slide.

We heavily weighted a specific term within the gene set of interest, and down weighted that term in the complementary set.

Both TreeHugger and the hypergeometric test flagged our term >90% of simulations.

TreeHugger rarely (<1%) flagged the parent of that term, or the root.

The hypergeometric flagged the parent in >50% of simulations and the root in >80% of simulations.

On a cystic fibrosis related data set, TreeHugger was again more conservative than traditional tests.

TreeHugger flagged 72 terms, while the hypergeometric flagged 481 terms.

In the biological process ontology TreeHugger flagged 37 terms, while the hypergeometric flagged 309.

However, the biological significance of the terms flagged was essentially the same, with TreeHugger tending to flag more specific terms.

Changing significance levels of the tests does not change the story: p-values of the two tests are not correlated.

# Implementation and Availability

1. GO terms and the hierarchical relationships were obtained from GO.
2. Annotations were obtained from NCBI.
3. Much of this data was placed into a MySQL database for easy manipulation. It is then exported in a format useful for us.
4. The tree structures and the annotations were then used to build score files: for each gene, all the associated term scores.
5. At our lab website users choose a species of interest (human, rat, mouse, etc.), and uploads gene lists, both interesting genes and their complement.
6. Enrichment is then computed.

The building of scores is done with Perl and MySQL.

The website is written in HTML with Perl/CGI.

The actual enrichment is done with R.

<http://vanburenlab.medicine.tamhsc.edu/treehugger.html>