Approaches for comparing Gene Ontologies

Sam Bassett

Developmental and Stem Cell Biology Lab,
Victor Chang Cardiac Research Institute,
Darlinghurst, Sydney, Australia

January 2, 2014

Abstract

Introduction

Gene Ontology was conceived as the "tool for the unification of biology", since it provides a vocabulary for gene function that is applicable to all eukaryotes [1]. Genes have terms assigned to them in three broad categories, namely biological process, molecular function and cellular component, and can have varying degrees of specificity - from the very broad, such as "cell", right down to the very specific. While this is very useful for inferring possible functional differences corresponding to different gene expression profiles, the ability to compare gene ontology profiles between different datasets and visualise these differences can give a much clearer idea of both their similarities and differences.
Many online tools have emerged in order to help perform both gene ontology annotation and enrichment analysis, then subsequently visualise the results. They all do similar things, but each has their own advantages and downsides. Here several such tools are compared, along with two methods of visually comparing the results of GO enrichment analysis created in-house.

DAVID

Web Interface

DAVID, coupled with their excellent R package RDAVIDWebService, fulfills most functional annotation requirements. It adopts the core strategy of systematically mapping a large number of interesting genes in a list to the associated biological annotation, then statistically highlight the most enriched annotations [2]. From this list of annotations, it

can subsequently undergo further analysis such as gene funcitonal classification, funcitonal annotation charts and functional annotation tables. It can also provide a pathway view of enriched terms [2].
When comparing gene lists, however, the web interface falls slightly short. The best method for ontological comparison found was to use one gene list as the test list and one as the background. This produced a list of terms more significantly enriched in the test list, but not the background. The process had to be run again with the gene lists swapped in order to reciprocally determine enrichment, and this didn't necessarily give similar results.

R package - RDAVIDWebService

The RDAVIDWebService R package bypassed this restriction by giving fine-grain control over the annotation process. Gene lists are uploaded to the server via R, and raw data such as functional annotation tables are returned. This means that instead of having to reciprocally compare two gene sets with each other as background, the entire GO enrichment can be done against the organism's genome. These data can then be directly compared. This is where the tools developed in-house can be used: RDAVIDWebService provides a method for plotting a directed acyclic graph (DAG) based on GO terms, but it can only do one set at a time. Using this code as a base, an extension of the method for two graphs or datasets was developed. Node colour was used to visualise similarities and differences between the sets.
The only downsides of this package are that a knowledge of R is required in order to appreciate its flexibility, and registration is required in order to use it.

WebGestalt

WebGestalt is similar to DAVID's web interface, in that it provides an interface to a server which is capable of several different types of analysis. This includes some that DAVID does not provide, such as protein interaction network analysis. It gives finer control over the statistical tests done to produce the enrichment as well. It does not have a corresponding R package, nor is it as extensible as DAVID. The results produced are similar, but not exactly the same.

GOEAST

GOEAST adopts the same core strategy as discussed previously with some extra choice in terms of advanced statistical parameters - it is the only tool that includes Alexa's improved weighting algorithm, a statistical test that is designed to correct for additional (incorrect) overrepresentation of neighbouring GO terms introduced by the hierarchical-dependent relationships of said terms [3]. The process of GO enrichment using GOEAST is very slow

compared to the other tools examined here, instead of taking several seconds to process a list of 1500 genes GOEAST can take up to 20 minutes. This is a significant barrier to entry, especially when compared to the other tools.
The benefit of using GOEAST is their comparison tools. A program called Multi-GOEAST is provided on the website which allows up to three GOEAST result files to be uploaded and compared with one another, where they can be visualised as a DAG with node colours corresponding to amount of enrichment in each of the three sets.

clusterProfiler

ClusterProfiler is an R package with no associated web interface which permits both functional enrichment and visualisation. The first thing to note is that the documentation is woeful. Some things just don't work, necessitating a dive into the R source (on GitHub, mind you, since most of the R functions themselves are obfuscated) in order to figure out how to use them. The package was used initially because it has a nice way of visualising n gene sets at once, the percentage enrichment of each term and their p values. However, in practice, this visualisation technique only worked for data generated by the package itself (so DAVID results cannot be visualised in this way, for example) and either showed only the five broadest terms or an impossibly long list of terms that don't really show anything meaningful.
The function provided would be incredibly helpful for comparing gene lists if it worked correctly, since it also allows the user to provide the comparison function. However, this isn't currently feasible.

Comparison Table

Name	Pros	Cons
DAVID Web Interface	Great for generating single annotations, does everything in one place. Fast and easy to use, frequently cited. Input genes can be in a variety of different formats.	Not designed for comparing multiple gene lists.
DAVIDWebService R package	Easily extensible, everything can be adjusted. Lends itself to novel analysis. Same flexibility of input genes.	Requires registration, knowledge of R. Queries can sometimes be slow. Each list must be uploaded to the webserver before proceeding.
WebGestalt	Extra features, such as further control over statistics and the ability to generate protein interaction networks. Similar flexibility of input gene set	Also not designed to compare multiple gene lists.
GOEAST	Further statistical control such as Alexa's algorithm Specific GO comparison tool	*Very* slow. Only MGI IDs can be used as input, so generally preprocessing is required before submitting a query.
clusterProfiler	Potentially very useful functions. Simple, requires only a gene list and it will do the annotation without further input.	Functions don't work as documented. Inflexible, only EntrezGene IDs can be used as input. Functions aren't editable, so the benefits of using R are lessened.

References

[1] The Gene Ontology Consortium, *Gene Ontology: tool for the unification of biology.* Nature Genetics, volume 25 May 2000.

[2] Huang, D.W., Sherman, B. & Lempicki, R., *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources*. Nature Protocols, volume 4 December 2008.

[3] Zheng, Q., Wang, X., *GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis*. Nucleic Acids Research, volume 36 May 2008.