Approaches for comparing Gene Ontologies

Sam Bassett

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

Abstract

Finding a good method for comparing gene ontology (GO) enrichments between lists of genes is an important problem for functional comparison. Here, several tools are presented that accomplish this task, with their respective benefits and downsides examined. An in-house method of comparing enrichment analyses between sets, CompGO, is also developed to accommodate enhancements that benefit bioinformatics research at the Victor Chang Cardiac Research Institute. CompGO draws on and enhances methodologies implemented in the existing R/BioConductor packages, a standard and promoted practice in this community. Specifically, CompGO enables both similarities and differences between sets to be brought forth that would otherwise not be immediately obvious from performing and contrasting individual gene ontology enrichments.

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 - GO:1901544, for example, which is the biological process term for "positive regulation of ent-pimara-8(14),15-diene biosynthetic process"; or GO:0033593 which is the cellular compartment term for "BRCA2-MAGE-D1 complex". The benefits to using this approach are mainly concerned with creating a standard, structured and consistent vocabulary for describing gene products. This provides a method for representing and communicating both knowledge about a topic and relationships between topics that is not only standardised, but also easy to interpret using computational methods. As such, Gene Ontology enrichment analysis has emerged as a standard methodology for understanding the relevance of gene list output from Bioinformatics analyses.
Many online tools have therefore emerged with the following common strategy: first, to extract gene ontology annotations from a list of genes; second, to perform some form of statistical enrichment analysis; finally, to meaningfully visualise results. They all perform similar functions, but each has their own advantages and downsides (as reviewed below). For example,

some tools (such as topGO) require that the data originate from an Affymetrix or similar gene chip as a first step of analysis and provide no alternative entry point; such tools have not been reviewed here.
While this process of enrichment analysis is very useful for inferring the biological meaning of gene lists, the ability to compare gene ontology profiles between different datasets and visualise these differences can give a much clearer idea of functional similarity and/or difference. Currently, few tools are available that are able to compare enrichments from multiple gene sets, so the R package CompGO was developed in order to streamline and meaningfully visualise these relationships. In addition to simply comparing gene lists, CompGO also includes functions to streamline the processing of DNA binding data when provided in standard BED format.

Existing bioinformatics tools

DAVID

The Database for Annotation, Visualisation and Integrated Discovery, or DAVID, adopts the core strategy of systematically mapping a large number of interesting genes in a list to the associated biological annotation, then statistically highlighting the most enriched annotations [2]. DAVID is available through both a web interface (http://david.abcc.ncifcrf.gov/) and a number of R packages which provide programmatic interfaces. The web interface enables users to determine gene functional classification, functional annotation charts and functional annotation tables from a list of genes (or other supported identifiers). It also provides a pathway view of enriched terms via the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database. However the web interface has no options for performing or visualising comparative analyses. Direct ontological comparison between two lists is only possible if one gene list is used as the test list and the other as the background. This produces a list of terms more significantly enriched in the test list, but not the background - the cumbersome process has to be run again with the gene lists swapped in order to reciprocally determine enrichment, and this may not be the best way to determine functional differences.

The RDAVIDWebService R package bypasses this restriction by giving fine-grain control over the enrichment 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. The package then provides methods for visualisation, most significantly, the ability to generate graphs from GO terms (including their counts and significance). The only downsides of this package are that knowledge of R is required in order to appreciate its flexibility and registration is required in order to use it. RDAVIDWebService forms the base for CompGO, appreciating the community aspect of R/BioConductor, building upon its functionality in order to compare sets of genes.

WebGestalt

WebGestalt [4] 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 performed to determine enrichment as well. It does however not have a corresponding R package, nor is it as extensible as DAVID - the options are more limited for gene ontology analysis and visualisation, plus a lack of an R package means it is much more difficult to use in a pipeline or interface with programmatically. The results produced are similar, but seem to be pruned at a significance level that the user cannot view or modify.

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.
GOEAST does however provide 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 - the colour saturation of a node corresponds to its p-value, with more significant nodes having more saturated colours, while the colour of the node itself shows which sets it can be found in. When comparing two sets, for example, terms found only in set A are red, terms in set B are green, and those in both are yellow. However, the user does not have much control over this visualisation process, and the resultant graphs are also heavily pruned when compared with DAVID.

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.

CompGO

All of the currently available tools have limitations when it comes to comparing gene sets. Whether they don't offer much control over the comparison process (GOEAST) or don't natively support comparison at all (the other tools discussed above), it's clear that a method for gene set comparison is currently lacking. To this end, we present CompGO. CompGO is an R pipeline built on DAVID's web service which provides both gene annotation from .bed file coordinates and, importantly, several methods for gene ontology enrichment comparison. First, this method allows much finer control over the data generation and visualisation when compared to the web interfaces; it is also open-source, so it can be easily modified or extended.

This is already a real improvement over the existing methods, but in addition, CompGO contains two methods of visualisation that aid the interpretation of data generated.
First, a plot of p values from two different enrichment tables can be performed. Each GO term shared by both tables is plotted based on its p value, and the statistical correlation between the two is computed. The Jaccard coefficient of each GO term, which represents the proportion of shared genes compared with the overall set of genes, can also be plotted as colour on the same graph; this gives additional power to the analysis as terms which may be skewed by a few extraneous genes can easily be picked out. Fine control over this plot also is given in the form of function parameters.
The other method, adapted from code sourced in the RDAVIDWebService package, is plotting a DAG where individual nodes (representing GO terms) are given colour based on the gene set from which they came. Again, the plotting function is flexible, and this approach aids in seeing specifically which enriched terms come from which set, and which ones are common to both.

Conclusion

The ability to compare gene ontologies between different experiments or organisms is both useful and currently nascent. We present a package, CompGO, which builds upon current approaches used by DAVID to perform enrichment analysis on single gene sets and extends them to compare gene sets pairwise, then visually compare the results.

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.

[4] Zhang, B., Kirov, S., Snoddy, J., *WebGestalt: an integrated system for exploring gene sets in various biological contexts.* Nucleic Acids Research, volume 33 June 2005.