GOstat - Find statistically overrepresented Gene Ontologies within a group of genes

FIAT·LUX

Tim Beissbarth and Terry Speed
Bioinformatics - Walter and Eliza Hall Institute of Medical Research (WEHI)

2. Background

Ontologies are a widely used concept to create a controlled vocabulary to communicate and annotate knowledge. The Gene Ontology Consortium defines GO as an international standard to annotate genes [1]. GO has a hierarchical structure starting with top level ontologies for molecular function, biological process and cellular component. The GO database consists of two essential parts the current ontologies, which define the vocabulary and structure, and the current annotations, which create a link between the known genes and the associated gene ontologies that define their function. Currently many groups are working on the development of the ontologies and annotations for different organisms. All the information can be downloaded from the web-site: *http://www.geneontology.org*.

Figure 1:
Structure of the Gene Ontology DB [1]

Directed Acyclic Graph (DAG)

Top Level:
- **Molecular Function**
- **Biological Process**
- **Cellular Component**
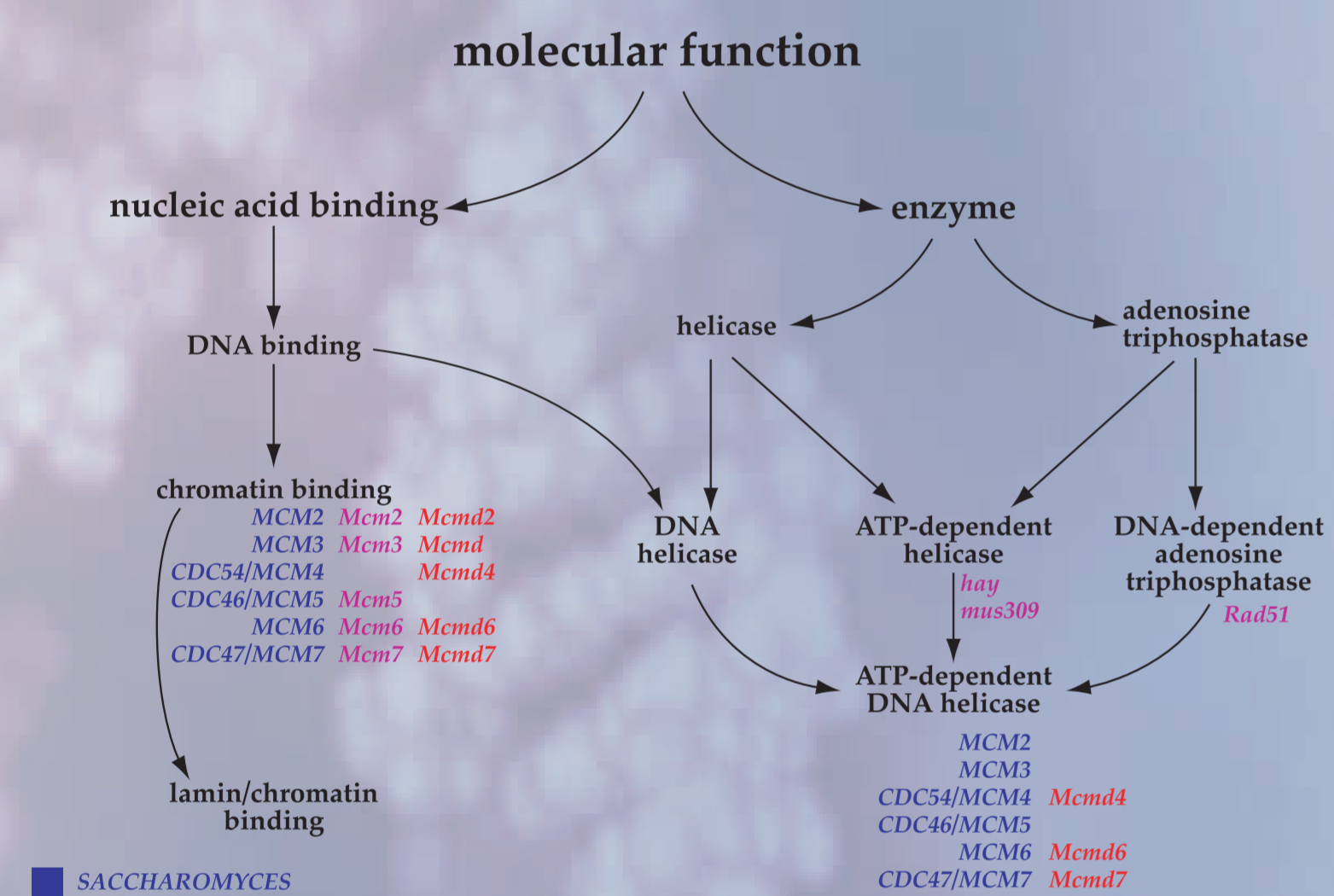
Annotations:
Mouse: 12325 genes
Human: 21521 genes
Yeast: 6910 genes
Drosophila: 7536
Etc. ...

Here we would like to make use of the annotations and structure of the gene ontologies in order to understand the biological processes being present in a large data set of genes. Each gene can have several associated GO terms. Further, due to the hierarchical structure of the gene ontologies each GO term can be connected to several other GO terms higher in the GO hierarchy and therefore associated with the gene as well (Figure 2). We call the list of GO terms that are in between a top level and the annotated GO term its path. In fact, several such paths might lead to an individual GO term. Each GO term in the path we call a split. So in the end a list of 100 genes will usually have many hundreds of associated GO terms and several thousand associated splits.

Figure 2: Structure of GO annotation

Annotated GO: GO:0006917

ⓘ denotes an 'is-a' relationship
ⓟ denotes a 'part-of' relationship
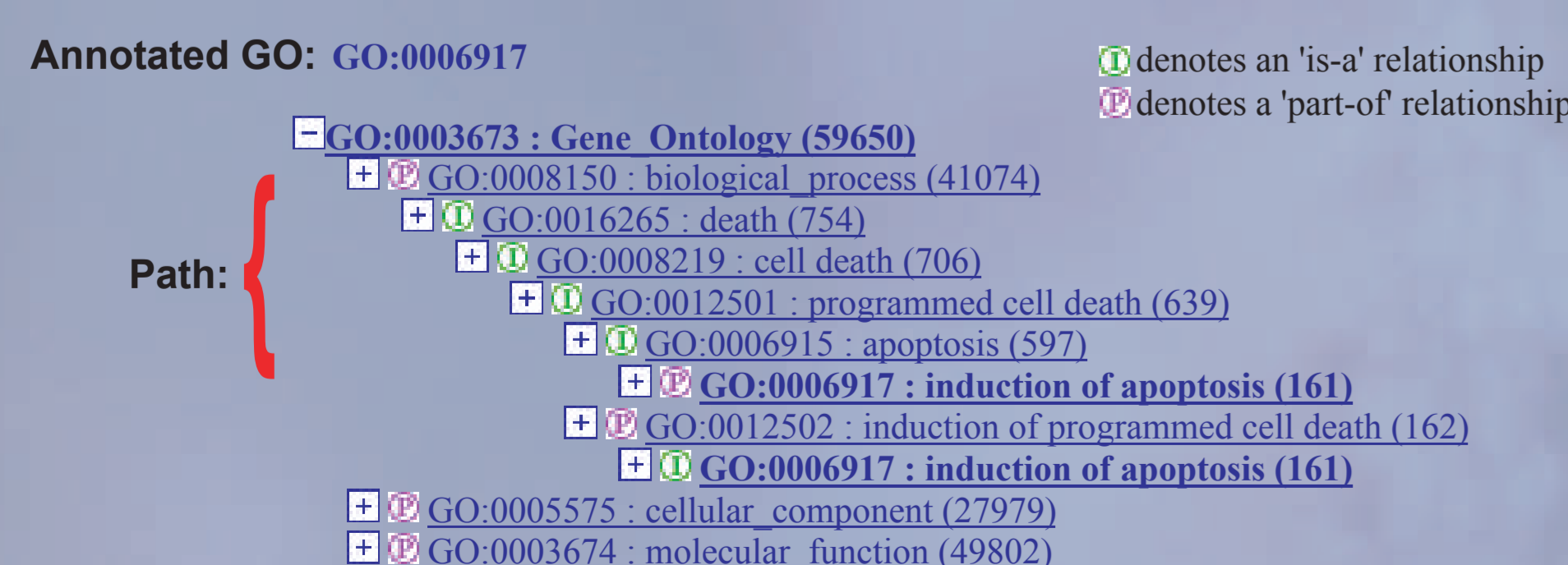
Each gene can have several annotated GOs and each GO can have several splits. Example:
DNA topoisomerase II alpha
- 8 GO annotations
- 11 splits

1. Abstract

Modern experimental techniques, as for example DNA microarrays, as a result produce usually a long list of genes, which are potentially interesting in the analyzed process. In order to gain biological understanding from this type of data, it is necessary to analyze the functional annotations of all genes in this list. The Gene-Ontology database (GO) provides a useful tool to annotate and analyze the functions of a large number of genes. Here we introduce a tool that utilizes this information to obtain an understanding of which annotations are typical for the analyzed list of genes. This program automatically obtains the GO annotations from a database and generates statistics of which annotations are overrepresented in the analyzed list of genes. This results in a list of GO terms sorted by their specificity. Our program GOstat is accessible via the Internet at

http://gostat.wehi.edu.au

3. Input

GOstat requires a list of gene identifiers, that specify the group of genes of interest. The program uses several synonyms, each of which is sufficient to identify a gene. These synonyms are derived from the release of the GO database as well as from Unigene [2]. GO databases for several organisms (human, mouse, drosophila, yeast, arabidopsis thaliana, etc.) are provided. In order to find GO terms that are statistically significant within the group a second set of genes needs to be used to obtain a total count for the occurrence of the GO term. For this, either the complete database of annotated genes can be used or one of several subsets can be selected that are commonly used on widely available microarrays. Alternatively a second list of gene identifiers can be passed to the program. In this case, the second list is used as a reference to search for GO terms, which are significantly more represented in the first list than in the second.

4. Determination of statistically significant GO terms

For all of the genes analyzed, GOstat will determine the annotated GO terms and all splits. The program will then count the number of appearances of each GO term for the genes in the group as well as in the reference group. For each GO term, a p-value is calculated representing the probability that the observed numbers of counts could have resulted from randomly distributing this GO term between the tested group and the reference group. A chi-square test is used in order to approximate this p-value. If the expected value for any count is below 5 the chi-square approximation is inaccurate. Therefore, we use Fisher's Exact Test in these cases. The resulting list of p-values is sorted. The GO terms which are most specific for the analyzed list of genes will have the lowest p-values.

Figure 3: GO statistics

Contingency Table

count genes with GO in group → 51

51	416	467
125	8588	8713
173	9004	9177

count genes without GO in group → 125

count in group (e.g. differential genes)
reference group (e.g. all genes on array)

P-value
8e-52

Fisher's Exact Test or Chi-Square Test

Hypergeometric Distribution

$$P(X) = \frac{\binom{N_1}{X}\binom{N_2}{n-X}}{\binom{N_T}{n}}$$

5. Output

GOstat will result in a list of p-values, which state how specific certain GO terms are for a given list of genes (Figure 4). The output is sorted by the p-value and can be limited by various cutoff values. It is possible to display the over- or underrepresented terms only. P-values of GO terms which are overrepresented in the dataset are typeset in green, p-values of underrepresented GO terms are colored red. GO terms that are annotated in more or less the same subsets of genes can be grouped together. GOstat will also output the complete list of the associations for the supplied genes to the annotated GO terms. The GO IDs in the output are linked to AmiGO, a visualization tool for the hierarchy in the GO database (*http://www.godatabase.org*). It is possible to format the output in HTML or as a tabular text.

Figure 4: GOstat output

List of significant GO terms

List of genes and GO associations

6. Conclusions

GOstat provides a useful tool in order to find biological processes or annotations characteristic for a group of genes. This is greatly helpful in analyzing lists of genes resulting from high throughput screening experiments, such as microarrays, for their biological meaning. It is also possible to compare two sets of genes in order to find the differences between the biological processes that characterize them. The program can be used via the Internet at http://gostat.wehi.edu.au.

As the number of GO terms we test significance for is large, the computed p-values have to be corrected in order to control the rate of errors we expect with multiple testing [3]. Two methods for correcting the p-value are offered in GOstat. The Holm correction controls the familywise error rate, e.g. selecting genes with a p-value below 0.1 we expect a 10% chance that any of the selected GO terms are not specific. The Benjamini and Hochberg correction controls the false discovery rate, e.g. selecting genes with a p-value below 0.1 we expect that 10% of the selected GO terms are not specific. However, there are dependencies between various GO terms in the resulting list. Frequently, genes share more or less the same set of annotations, as several GO terms are indicative for the same process. Also, GO terms that are within one path have strongly correlated results. In order to make the resulting list of GO terms more interpretable, GOstat has the option to cluster the GO terms. In this process, GO terms which are annotated in the same set of genes, or where one set of genes is a subset of the other, are grouped.

8. References

[1] The Gene Ontology Consortium. 2000. Gene Ontology: tool for the unification of biology. Nature Genetics 25: 25-29.
[2] Boguski & Schuler. 1995. ESTablishing a human transcript map. Nature Genetics 10: 369-371.
[3] Shaffer. 1995. Multiple Hypothesis Testing. Annu. Rev. Psychol 46: 561-584.