GOTM Manual

GOTM (Gene Ontology Tree Machine) is a web-based platform for interpreting microarray data or other interesting gene sets using Gene Ontology hierarchies. GOTM currently works with human, mouse, rat and fly. GOTM is free to academic users after registration. GOTM can be accessed from http://genereg.ornl.gov/~gotm.

Key features:

- User friendly web-based interface
- Expandable tree for browsing the GO hierarchy, Fixed tree as HTML output for archive, Bar chart for publication
- Statistic analysis indicating GO terms with relatively enriched gene numbers and suggesting biological areas that warrant further study. Sub-tree and DAG visualizing enriched GO categories
- Retrieving subset of genes by GO term or keyword searching.

Input

Figure 1 shows the input interface of GOTM. The description of the input form is given below.

1. **Give a name for this analysis.** You need to give a unique name for each analysis. Ordinarily, the result of your analysis will be stored for 1 month. During this period of time, you can come back and retrieve you result using the unique name. HOWEVER, PLEASE BE AWARE THAT THIS IS NOT A PLACE FOR SAVING YOUR RESULTS, YOUR RESULTS MAY BE DELETED DURING OUR PERIODICALLY SYSTEM UPGRADING WITHOUT PRIOR NOTICE**.** Anyway, the old results will be out of date after system upgrading and it's better to generate new trees.
2. **Select the ID type in your file.** You can choose from LocusID, Gene Symbol, Unigene ID, Swiss-Prot ID, Ensembl ID, Affymetrix probe set ID.
3. **What kind of analysis you want to do.** You can choose either single gene list analysis or interesting gene list vs. reference gene list analysis. For single gene list analysis, you only need to upload the file of interesting gene list (step 4). And you will get a GO Tree for this gene list. For interesting gene list vs. reference gene list analysis, you need to upload the file of interesting gene list (step 4), and choose an existing reference gene list (step 5) or upload the file of reference gene list (step 6). You will get a GO Tree for the interesting gene list, and GO terms with relatively enriched gene numbers in the interesting gene list comparing to the reference gene list.
4. **Upload your file of interesting gene list.** The file should include the appropriate ID (required) and corresponding microarray ratio (optional),

seperated by tab. One ID per row. **Don't use excel files directly. If you have an excel file of gene list, save as tab delimited text file first. GOTM only accept plain text file.**
Example:
4509 3.5
4578 -1.2
If you are doing a single gene list analysis, you can press the MAKE TREE button now.

5. **Choose an existing reference gene list** If you are using commercial arrays whose information is available in our database, or if your reference gene list has been choosed to be stored in our database, you can choose from this list. This will save you some time. However, we need to mention that an impropriate reference gene set will lead to possible false positives and negatives. Unless the user can find the right reference gene set from our stored data, uploading an appropriate reference gene set for the analysis is always suggested.
6. **Or upload your file of reference gene list.** If your reference gene list is not in the above list, you have to upload your file of reference gene list. For the statistic analysis, the reference list should have all the genes in the interesting gene list. The file should include appropriate IDs, one ID per row. **Use plain text file only**

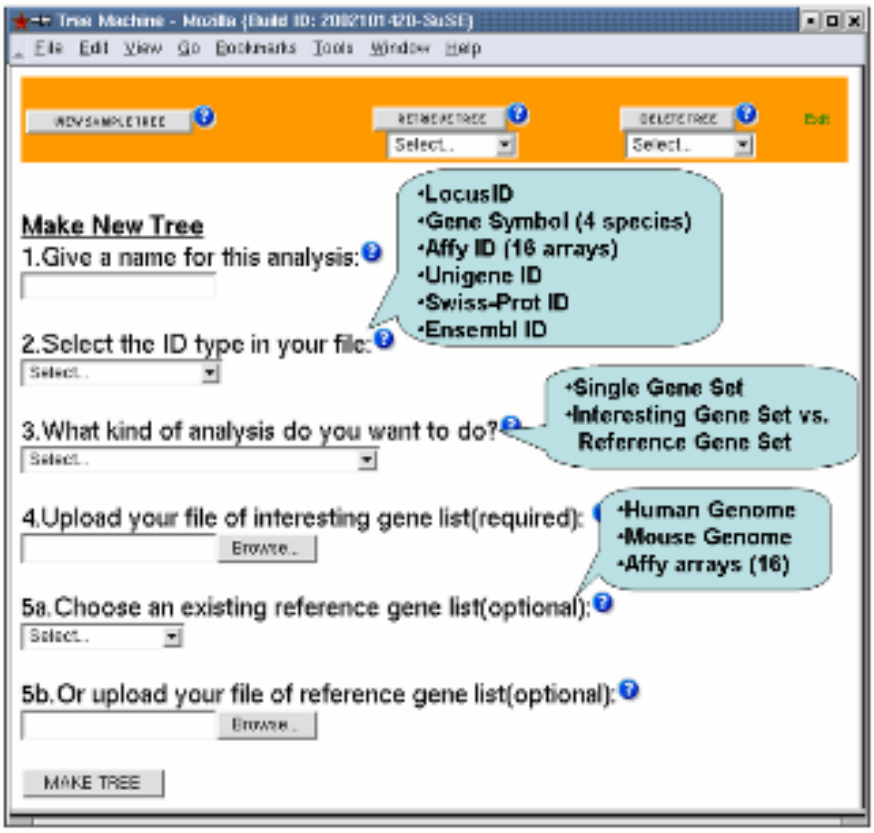

Figure 1. Input user interface of the GOTM

Statistics

Identifying GO categories with significantly enriched gene numbers in the interesting gene set compared to a reference gene set will allow the user to focus on biological areas that are most important for the interesting gene set. In order to identify GO categories with significantly enriched gene numbers, we need to compare the distribution of genes in the interesting gene set in each GO category to those in the reference gene set. A reference gene set could be all genes in a genome or another appropriate reference gene set (e.g. the list of genes on the array). We need to repeat that an impropriate reference gene set will lead to possible false positives and negatives. Unless the user can find the right reference gene set from our stored data, uploading an appropriate reference gene set for the analysis is always suggested. Suppose n genes were identified as interesting genes based on a microarray experiment (such as responsive, up-regulated or down-regulated genes) using an array with N genes. For a given GO category X, a gene is either in the category or not in the category. Suppose further that K out of the N reference genes and k out of the n interesting genes are in category X. If the n interesting genes were effectively a random sample uniformly selected from the reference gene set, the expected value of k would be $k_e=(n/N)K$. If, on the other hand, k exceeds the above expected value, category X is said to be enriched, with a ratio of enrichment (R) given by $R=k/k_e$. As the hypergeometric test is appropriate for sampling without replacement, it is used by GOTM. GOTM reports only those enrichments that are statistically significant as determined by the hypergeometric test. The significance of enrichment (P) for a given category is determined by $P=\sum_{i=k}^{n}\frac{\binom{N-K}{n-i}\binom{K}{i}}{\binom{N}{n}}$. As GO is organized based on the three relatively independent categories, biological process, molecular function and cellular component, GOTM calculates P value for each of the categories separately.

Output

Figure 2 shows the output user interface of GOTM. The output view is divided into 3 windows. The upper-left window is the GOTree window, the upper-right window is the gene/category list window and the bottom window is the gene information window. The expandable GOTree will be shown in the GOTree window. The user can browse the tree by clicking the “+”symbol. For single gene set analysis, the number of genes in each GO category will be given. If the interesting gene set vs. reference gene set analysis is selected, three parameters will be given for each GO category: O (observed gene number in the category), E

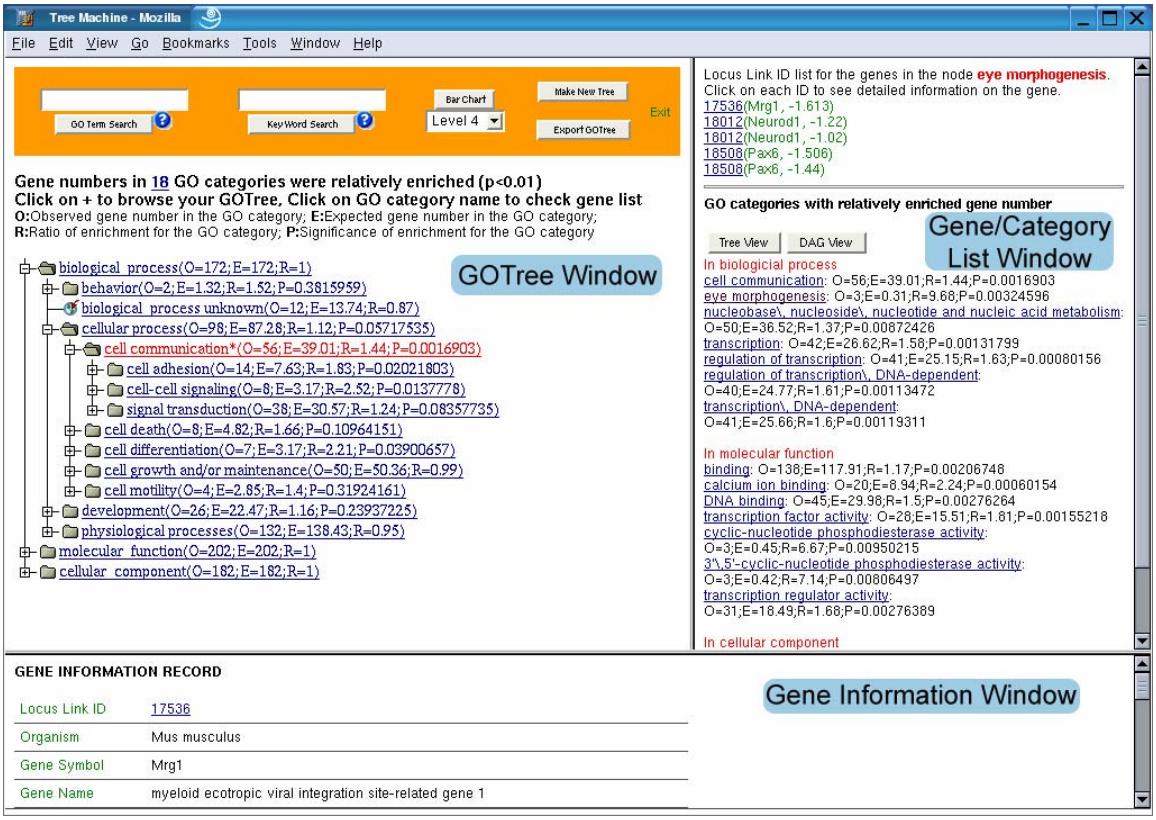

Figure 2. Output user interface of the GOTM

(expected gene number in the category), R (ratio of enrichment for the category). For those GO categories with R>1, the fourth parameter P indicating significance of enrichment will be given. GO categories with significantly enriched gene numbers (P<0.01) will be colored red. By clicking on individual GO categories, the genes in the category will be shown in the gene/category list window. It might be sometimes difficult for a user to browse and find the GO category the user is interested in. In this case, the user can do an exact search for a GO category using “GO Term Search” or a fuzzy key word search using “Keyword Search” at the top of the GOTree window. The returned GO categories and genes inside each category will be shown in the gene/category list window. The number of GO categories with enriched gene numbers will also be shown in the GOTree window. By clicking on the number, the names of enriched GO categories will be shown in the gene/category list window. GOTree provides comprehensive classification of the genes in a hierarchical structure, however, due to the complex structure, it’s not easily publishable. After browsing the GOTree, the user may pick appropriate annotation levels and get corresponding bar charts for publication using the Bar Chart button (for an example, see figure 3). GOTree can also be exported and locally stored in html format using the Export GOTree button. Enriched GO categories are colored red, and genes in each category are also included in the exported GOTree (for an example, see Figure 4).

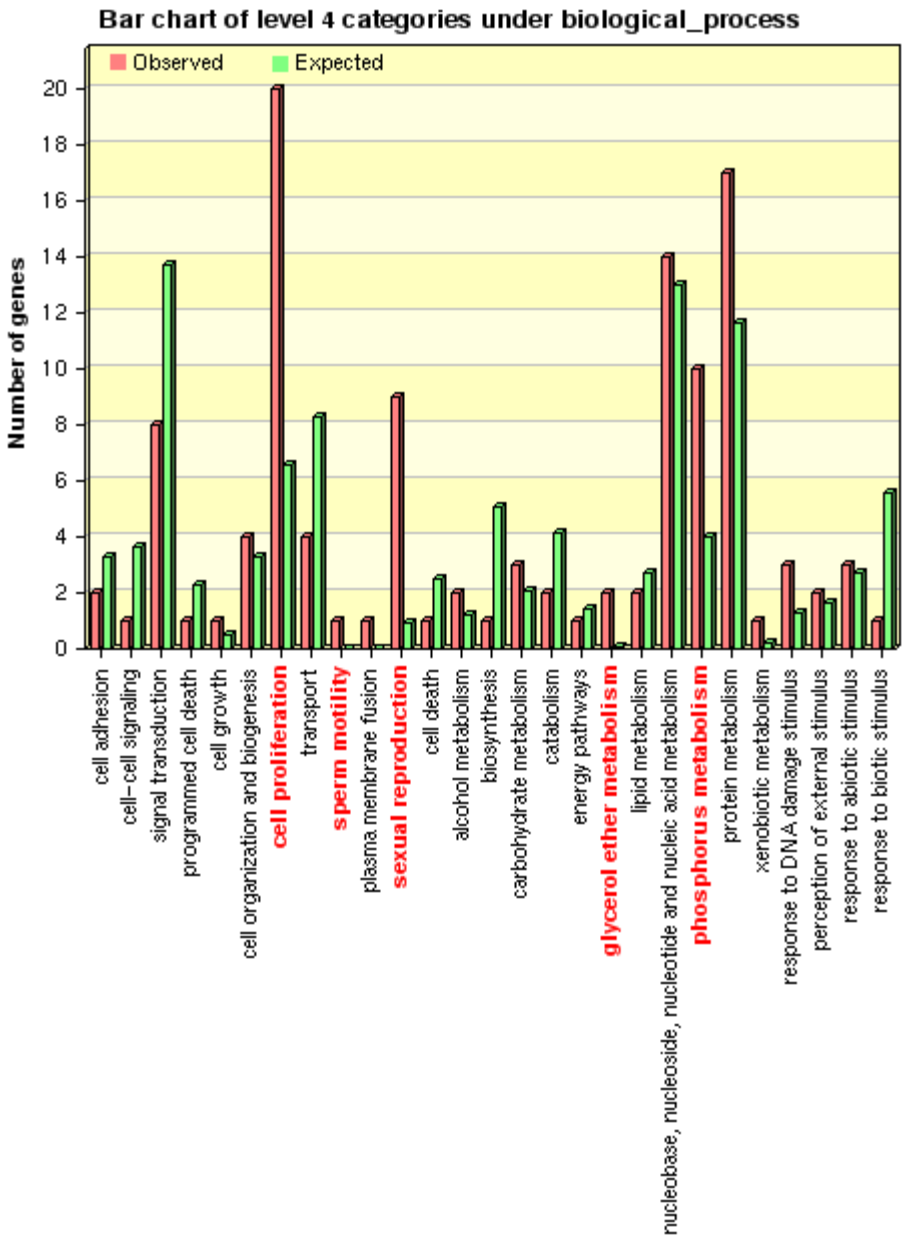

Figure 3. Bar chart generated by the GOTM

The gene/category list window shows the genes in a selected GO category, and enriched GO categories in the three main GO categories, biological process, molecular function and cellular component respectively. Each gene is represented by a LocusID, followed by the input ID. In addition, the ratio in the microarray experiment is shown if that information was included in the input file. Up-regulated genes are colored red while down-regulated genes are colored green. A flat view of enriched GO categories doesn't reveal the relationship among the GO categories. When tens or hundreds of GO categories are identified as significantly enriched, it becomes difficult for users to interpret the

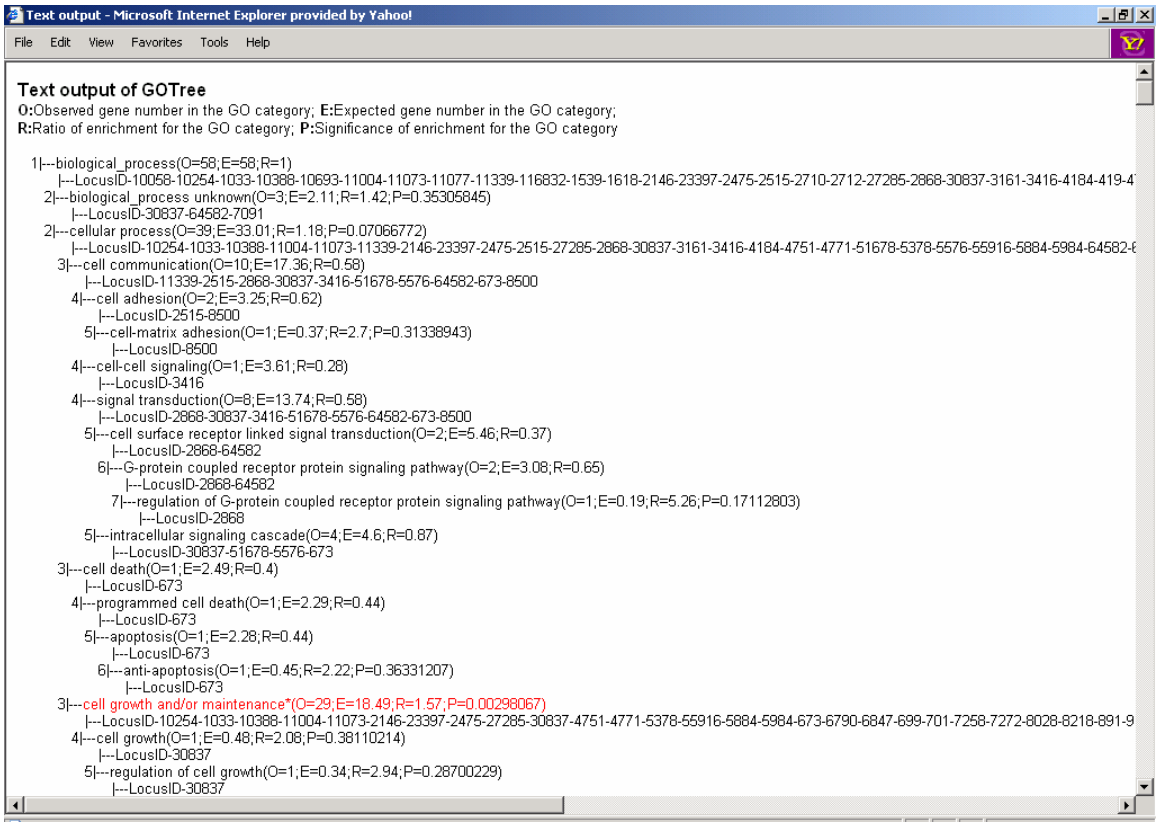

Figure 4. GOTree can be exported as html format for archive

results. In this case, the user can press the TREE VIEW button to get a sub-tree (Figure 5) or press the DAG VIEW button to get a DAG (Figure 6) for the enriched GO categories in a new window. The GO categories in red in the sub-tree or the DAG are the enriched GO categories while the black ones are their non-enriched parents. The sub-tree or the DAG group related enriched GO categories together. These visualization tools are very useful in helping the user to find important biological areas that are worth further studying. By clicking on individual LocusIDs in the gene/category list window, related information for the genes will be queried from GeneKeyDB and shown in the gene information window.

The gene information window shows the gene information record for the selected gene, which includes LocusLink ID, organism, gene symbol, gene name, map location, homolog, GO terms, function summary, GRIF (Gene Reference Into Function), phenotype, PubMed record, OMIM (Online Mendelian Inheritance in Man) record, KEGG (Kyoto Encyclopaedia of Genes and Genomes) Map etc. A link is given to external databases such as PubMed, OMIM, KEGG etc when available.

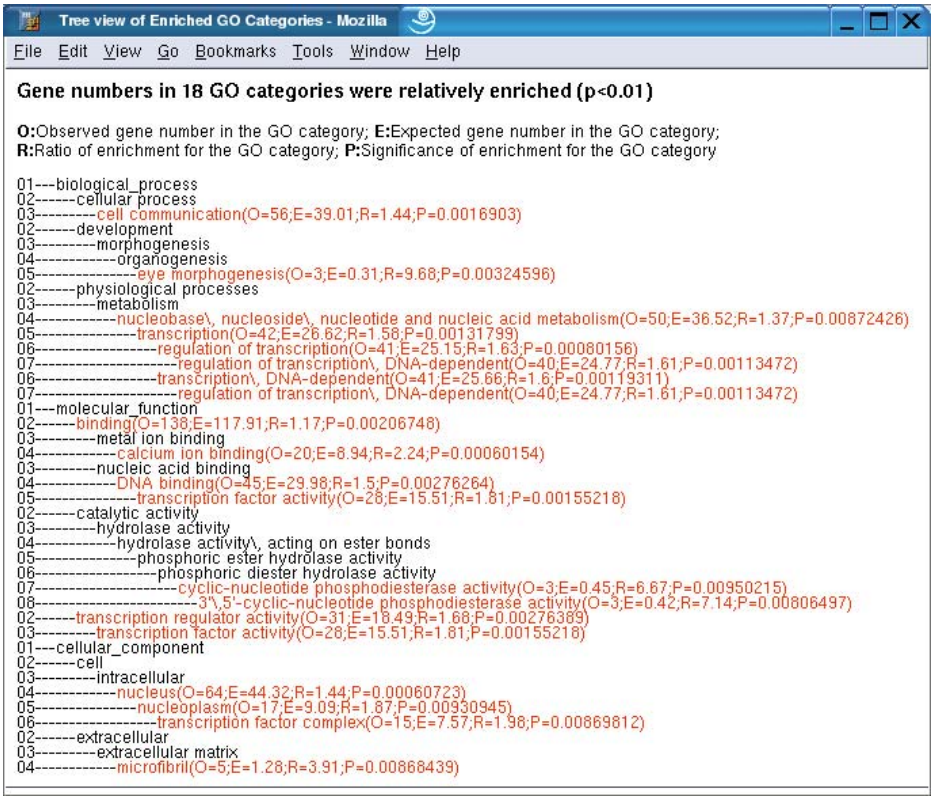

Figure 5. The enriched GO categories are brought together and visualized as a sub-tree.

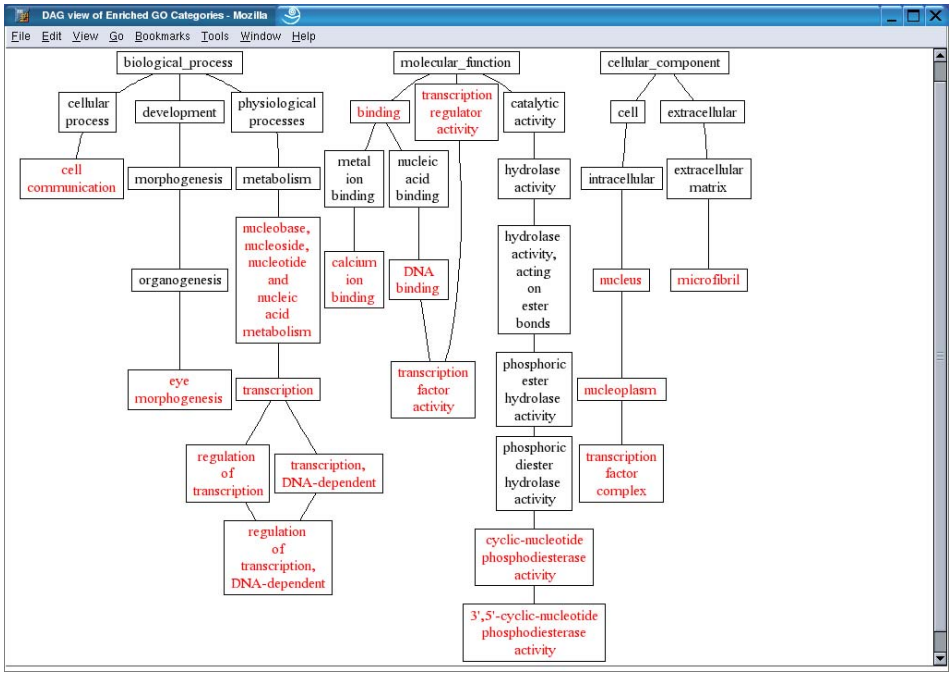

Figure 6. The enriched GO categories are brought together and visualized as a Directed Acyclic Graph (DAG).