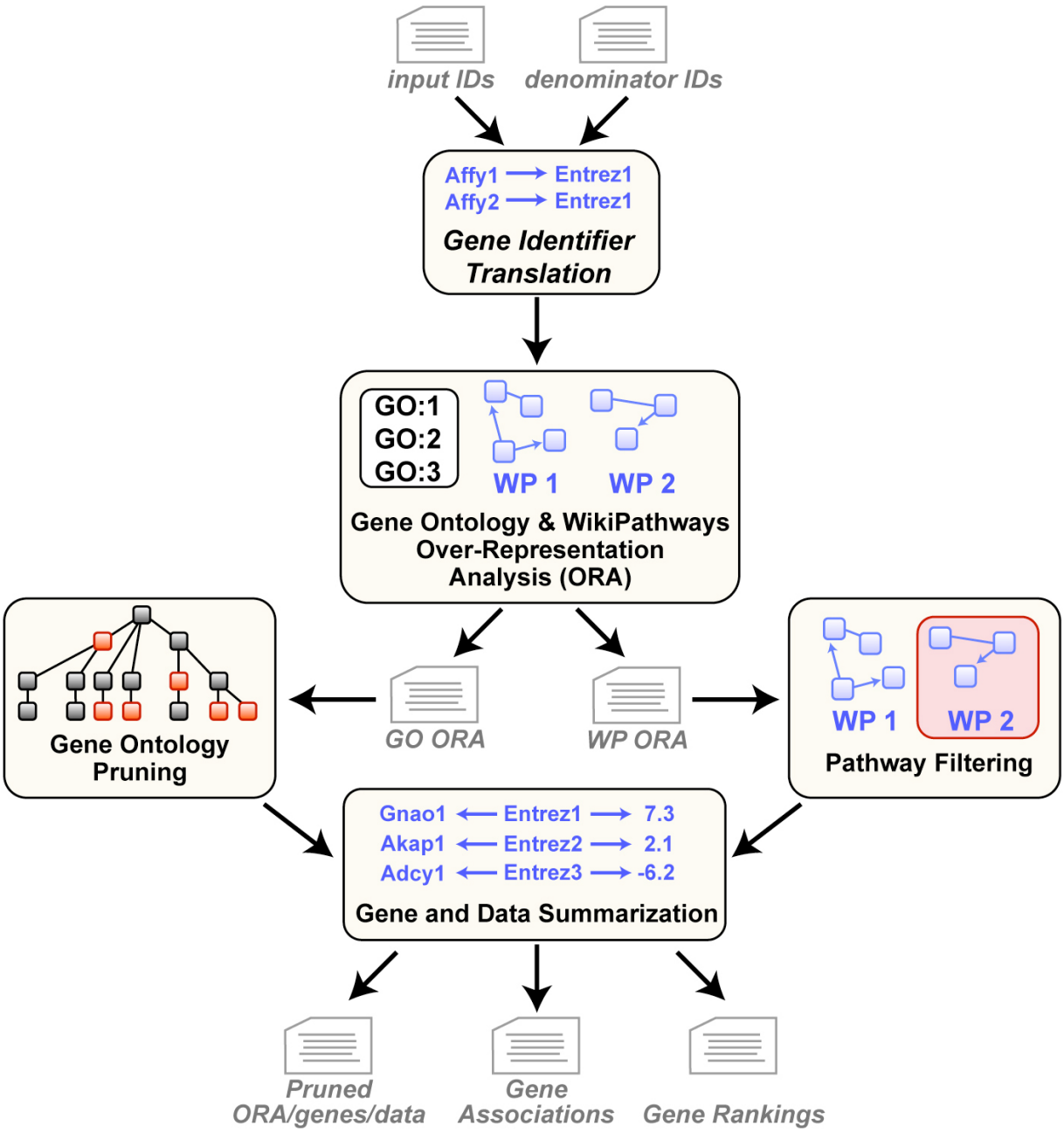
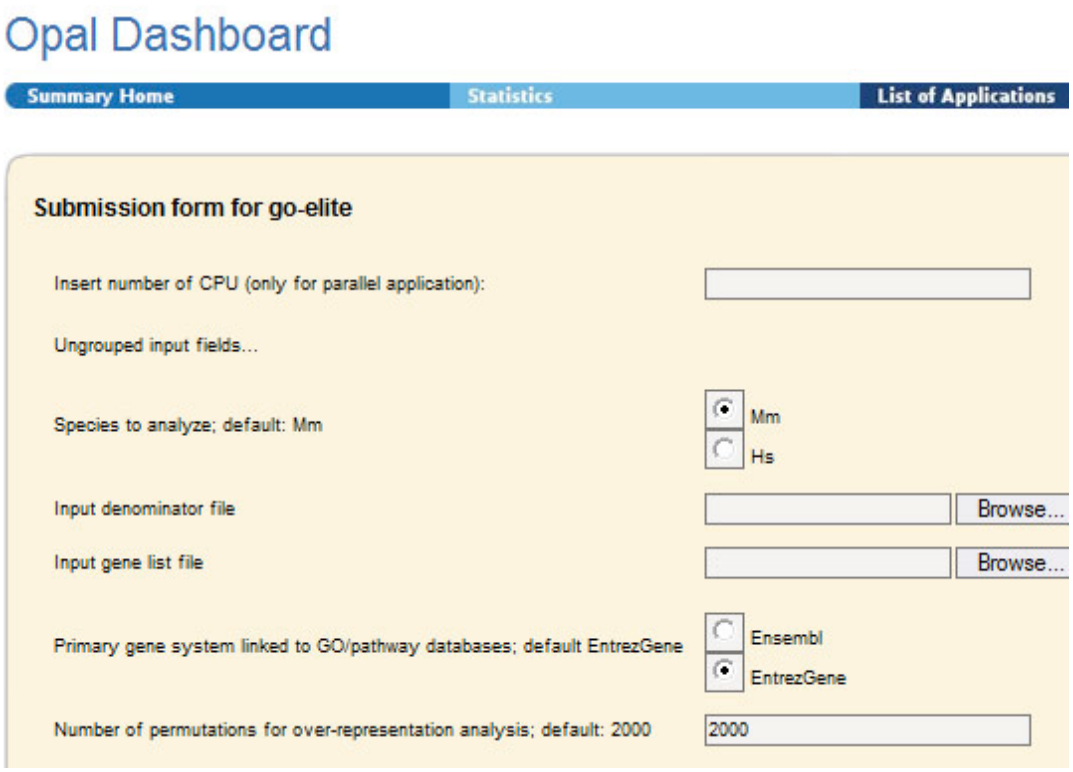


Program Overview - When analyzing user data, GO-Elite begins by reading in at least one input and one denominator file. The input file contains all gene IDs to be examined for over-representation (e.g., Affymetrix probesets, Ensembl or EntrezGene) along with the gene system code, whereas the denominator contains all gene IDs examined (e.g., all array probesets). The input file can also contain any number of additional columns with numeric data for pathway level summarization. These lists are submitted to GO-Elite for over-representation analysis (ORA) along with parameters for downstream filtering and pruning of the resulting GO-terms and pathways.

Sample Input Gene List

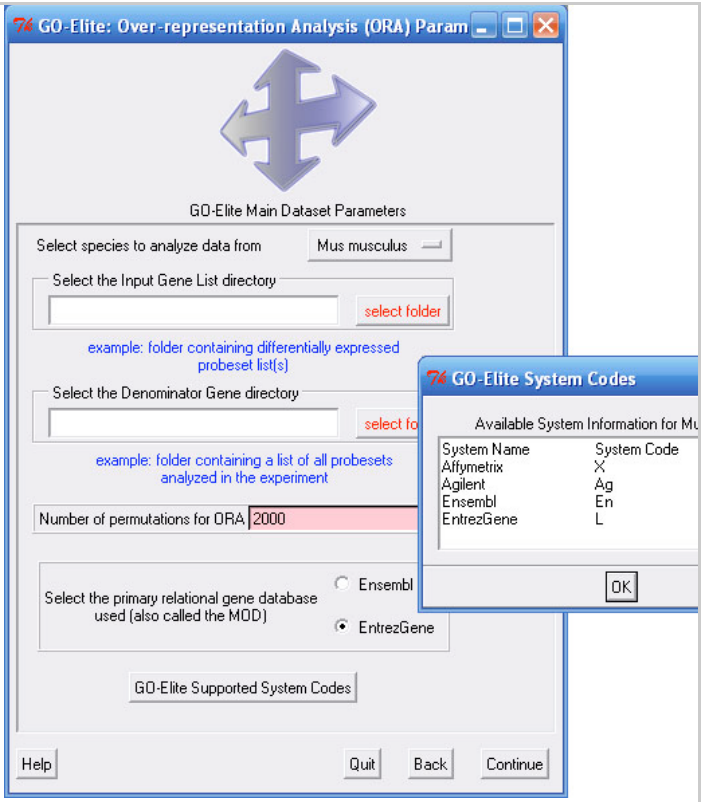
Probeset (REQUIRED)	SystemCode (RECOMMENDED)	Fold TP1 (OPTIONAL)	Fold TP2 (OPTIONAL)
j05479_s_at	X	1.23	2.31
L49502_s_at	X	-1.92	-1.85
Msa.33069.0_s_at	X	-2.41	-1.33
Msa.37566.0_s_at	X	3.03	0.25
AF028071_s_at	X	-1.91	0.85
Aa462409_s_at	X	0.25	4.35

Sample Input - The input list is a subset of the denominator list which consists of gene IDs or probesets that are highlighted from a user analysis (e.g., up and downregulated genes). Both input and denominator lists consist of a column of gene IDs (column 1), system code for each ID (column 2) and any other data you may wish to summarize at the pathway level (e.g. fold change). Commonly used system codes are “X” for Affymetrix, “L” for EntrezGene and “En” for Ensembl. If your array is not one of the supported array types (Affymetrix, Agilent, Codelink and Illumina), the system "Ma" for Miscellaneous Array may contain your array system (see http://www.ensembl.org/biomart/martview). Currently, GO-Elite can only accept one gene system per file. If the system code column is not present, GO-Elite will may be able to guess what type of ID system it is, but this is not recommended. For the denominator file, only the first two columns are used (gene IDs and system code). If you have multiple input files in a single directory that corresponds to different denominator files, GO-Elite will properly match these up if you place a unique number, letter or name before the name of each input file, separated by a period (e.g. **exp1**.input.txt and **exp2**.input.txt) and denominator file that matches it (e.g., **exp1**.denominator.txt and **exp2**.denominator.txt).



Web Interface - In addition to the stand-alone GO-Elite program, a simple query interface is available over the internet that does not require the user to download any software. This interface has all basic analysis features for GO-Elite, however, it is not possible to modify or add new database information (e.g., new gene systems or species support). The online interface supports a limited number species and ID systems (typically Affymetrix probeset IDs, EntrezGene and Ensembl). If performing different analyses from these, we recommend downloading the GO-Elite software to your computer.

Introduction Window

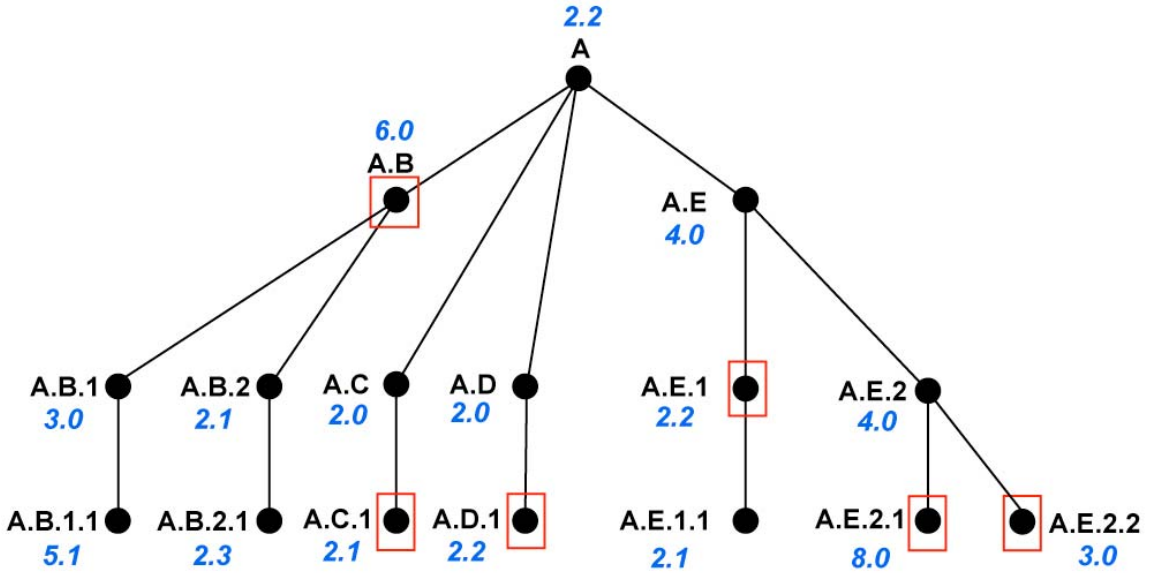


Over-Representation Analysis

Graphical User Interface - Unlike the web interface, using the installed versions of GO-elite you can:

1) Analyze Gene Lists
2) Prune Existing Results
3) Update or Add Databases/Pathways

This is useful for when your species or gene system is not supported by the web interface. Over 60 species are supported along with over 60 different gene identifier systems. For large jobs, such as the analysis of many independent lists of gene lists ov vary sizes, this interface makes allows you to select of folder of input and denominator files and assess the status of each job as they are run, unlike the web interface. Users can also modify GO-Elite's gene database and add support for any new species and new gene systems.



Step 1) Build all possible parent-child relationships.
Step 2) Find parents from this list more significant (see score options) than all of their children
Step 3) Find the most significant child terms (downstream of the last bifurcation)
Step 4) Eliminate terms from step 3 that are children of any other term from step 3
Step 5) Report the most signficant parent OR child terms

Pruning of Gene Ontology Results - The Gene Ontology (GO) is composed of a tree of highly related terms, beginning with very broad biological groups and ending with many highly descriptive biological functions, processes and cellular compartments. For GO over-representation results, after GO terms are initially filtered based on user defined statistics (permutation p-value, number of genes changed and z-score), GO-Elite will prune out GO-terms that are found to be “significant in the user analysis, but that have a lower “score” than a related term. This allows GO-Elite to report a relatively non-redundant set of GO-terms. Since GO-terms can be replicated on different branches of the GO-hierarchy, this process also ensures that redundant terms are removed from all branches. A visual representation of this pruning strategy is shown for a theoretical set of parent-child relationships with corresponding z-scores above.

Advanced Results - If your input gene file contains numeric data from your experiment, this will also be summarized in the GO-Elite output files. For example, if there are fold changes for each probeset in the input file, all fold changes for probesets associated with genes in a particular pathway will be averaged for that pathway, likewise for GO-terms. If your input file contains such data, the “pruned-results” file, will contain similar fold changes to those found in the input gene list, but summarized for each gene (where there are multiple probesets associating) and each pathway. These expression results can be graphed in a graphing program or clustered in an expression clustering and visualization program like Cluster and TreeViewer (see above).