GO-elite ReadMe File

Section 1) Introduction

1.1 Description

GO-Elite is a software tool designed to identify a minimal non-redundant set of Gene Ontology (GO) biological terms or pathways to describe a particular set of genes. This application is able to calculate advanced over-representation analysis (ORA) statistics from user gene lists, determine the minimal set of biologically distinct GO terms and pathways from these results and summarize these results at multiple levels. GO-Elite version 1 Beta is currently provided as a cross-platform stand-alone application which can be run as compiled Windows executable file (GO_Elite.exe), Mac OS X application (GO_elite.app) or as cross-platform source code in python in any OS. To download the program and to get documentation, go to: http://www.genmapp.org/go_elite/.

The two most common use cases for GO-Elite are:

1) Summarize biological processes regulated in a high throughput experiment for supported or new species/gene ID systems.
2) Quickly identify genes associated with such processes or pathways.
3) Summarize gene expression changes for whole pathways along multiple time-points.

More information and updated documentation, including bug fixes can be found at our Google Groups page at: http://groups.google.com/group/go-elite

1.2 Installation

Three installers are available for GO-Elite:

1) Windows PC installer
2) Mac OS X installer
3) Cross-platform python source code compressed zip archive.

Install/extract the GO-elite bundle archive to any desired directory on your computer. If installed for Mac OS X, installation in the Applications or Desktop directories is recommended.

1.3 Requirements

When installing the compiled OS specific installers, no additional software is required (python components are bundled with the application). If running GO-Elite directly from the source code (e.g., on Linux or Unix), Python 2.3 or greater and Tkinter are required, however, these are almost always installed by default on Linux and Mac OS X.

1.4 Updates

The GeneOntology tree structure can be easily updated using the Update DBs feature from the main GO-Elite interface menu (see below) or manually by downloading files from http://www.geneontology.org/GO.downloads.ontology.shtml. These files are saved directly to

the OBO directory in the GO-Elite folder, with the indicated names. Gene databases can also be updated from the UpdateDB interface, which automatically downloads and installs necessary files or updated manually by the user, see the file ReadMe/How_to_Make_or_Update_Species_Databases.rtf or online documentation.

Section 2) Preparing data

2.1 GO-Elite Input Files

The user is only required to provide two files: (A) input gene lists and (B) denominator list (containing all genes examined, including all input genes). Alternatively, you can supply an existing MAPPFinder file, such as a GenMAPP results file but is no longer recommended, since GO-Elite has databases that are more current or often are more up-to-date than GenMAPP. This option also works with recently generated GO-Elite mappfinder results.

2.2 Preparing Gene Lists

To perform ORA in GO-Elite, you need to provide at least one input gene list and denominator list. 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 - look up in the folder Databases/source_data.txt) and any other data you may wish to summarize at the pathway level (e.g. fold change), for desired number of additional columns. Currently, GO-Elite will use only one system code per file. If the system code column is not present (only the first column is absolutely required), GO-Elite will likely be able to guess what type of ID system it is, but this is typically not recommended. For the denominator file, only the first two columns are used (gene IDs and system code).

Figure 1.1 - 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
Msa.2129.0_s_at	X	-1.71	-2.54

These files can be saved anywhere on your hard-drive. If you want to save time in the future, however, you can save your input gene lists to the folder named input>GenesToQuery>*species* and your denominator files to the folder DenominatorGenes in the same directory as your input gene files. If you have more than one type of denominator (e.g. two types of Affymetrix arrays) place a unique number before the name of each input file (e.g. genelistA.txt -> 1.genelistA.txt) and denominator file that matches it. This will instruct GO-Elite on how to match up your gene input set to which proper denominator set.

Section 3) Running GO-Elite

3.1 Using GO-Elite on Different Operating Systems

To run GO elite, follow the appropriate instructions depending on if you are starting with gene lists or existing GO or pathway results.

- **PC** – Double click on the file "GO_elite.exe".
- **Mac** – Double click on the file "GO_elite".
- **Linux** – Installation of python is required, but is typically present with most Linux installations or can be installed via Yum update. To run, open a terminal window and go to the GO-Elite main folder (e.g. cd GO-Elite_119beta). Once in this directory typing "python GO_elite.py" in the terminal window will begin to run GO-Elite (you will see the below interface screen).

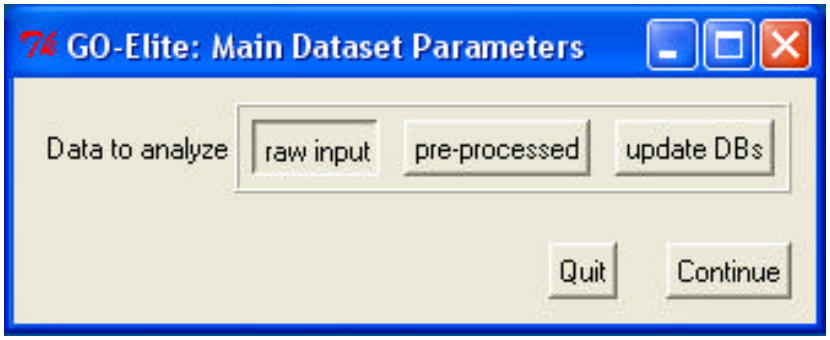
3.2 Analysis Options

GO-elite has a simple graphic and intuitive user interface intended to be run by novices and experts alike. When opened the user is presented with a introduction screen, which they can use to get version information or begin the analysis. After selecting "Begin Analysis" the user is presented with three options:

1) Calculate MAPPFinder results from input gene lists prior to GO-Elite (raw input)
2) Run GO-Elite on pre-run MAPPFinder results (pre-processed)
3) Build new gene-association files (update DBs)

Figure 3.1 – GO-Elite Main Winodws

Introduction Window



Main Data Analysis Window

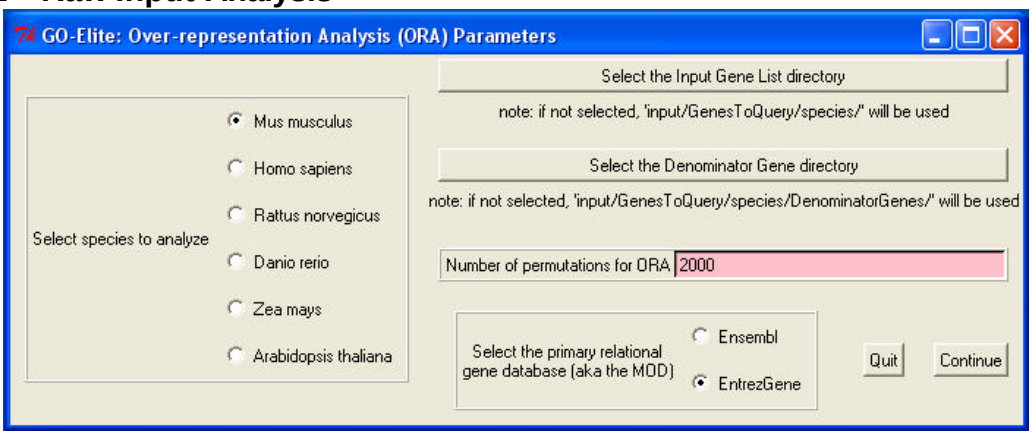
Selecting **"raw input"** is used when starting with an input and denominator gene list, whereas **"pre-processed"** is only necessary when starting with existing GO or pathway results (see GO-Elite white paper or manuscript), which can include previously GO-Elite outputs. Selecting **"update DBs"** is used if you have a currently unsupported species and wish to quickly build new databases for current or new gene ID systems. This option can be very useful, when dealing with a species that has available Affymetrix array annotation files

or is a species with GO information at NCBI (see How_to_Make_or_Update_Species_Databases.rtf for details).

3.3 Raw Input ORA (*aka* mappfinder)

When selecting the analysis option **"raw-input" (select Continue after selection)**, the user will be presented with a series of options for performing an over-representation analysis (ORA) on your gene/array ID lists. All options have selected defaults (defined in Config/options.txt), thus if the user simply clicks **"Continue"** at each step, GO-Elite will run with those default options. This analysis computes an over-representation z-score and permutation p-value for each GO-term/pathway similar to the GenMAPP program MAPPFinder 2.0. Unlike MAPPFinder the Benjamini-Hochberg false discovery rate method is used to calculate adjusted p-values based on multiple hypothesis correction, as opposed to the Westfall-Young method. The following is description of the analysis options available:

Figure 3.2 – Raw Input Analysis



1. **Selecting species to analyze**: Tells GO-Elite what species your gene lists correspond to.
2. **Select the Input Gene List directory**: Tells GO-Elite where your input gene list(s) are. Here you are selecting a directory rather than a file, since GO-Elite can handle many files at once from a single directory. When this button is selected a folder selection menu pops-up. From this menu, double-click on the folder with the input gene files. If this button is not selected GO-Elite defaults to the designated folder in the GO-Elite program folder (input/GenesToQuery/*species*/).
3. **Select the Denominator Gene List directory**: Tells GO-Elite where your denominator gene list(s) are.
4. **Select the number of permutations for ORA**: Indicates the number of permutations you want to run for ORA. With increased permutations, you have more statistical confidence but also much longer run-times. The default of 2000 is recommended for most analyses.
5. **Select the primary relational gene database**: This option tells GO-Elite which gene ID system to use when linking GO terms and pathways. If starting with gene IDs from Ensembl or EntrezGene, simply select the appropriate option, but if working with other gene IDs (such as Affymetrix), you have an option. Typically, EntrezGene has more genes per GO term and pathways associated and can produced slightly different

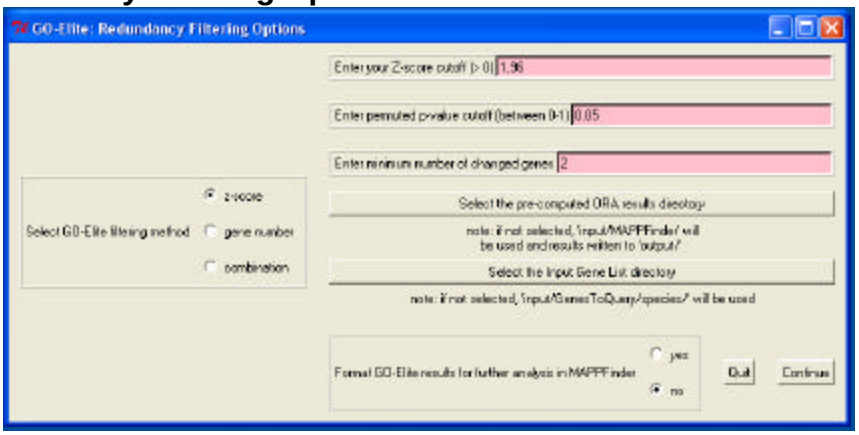
results from Ensembl, but if running for the first time it is recommended you try both in separate runs.

Once these options are established, GO-Elite will prompt the user on which filtering statistics to apply to the computed ORA results (**see section 3.4**).

3.4 Analyzing Computed ORA results (raw-input OR pre-processed)

If analyzing input gene lists, the "Redundancy Filtering Options" window appears after designating ORA options. If choosing "pre-processed" option from the main menu, this window appears directly and has an additional option for selecting the folder containing input gene lists that corresponds to the processed ORA files being analyzed. This window establishes criterion for filtering pre-processed ORA results from GO-Elite or GenMAPP's MAPPFinder program. Filtering is based on the relationship of GO-terms in the GO-hierarchy, relative ORA scores of such GO-terms and redundancy in gene content for both GO-terms and pathways (see GO-Elite white paper). Below are a description of the available options:

Figure 3.3 - Redundancy Filtering Options

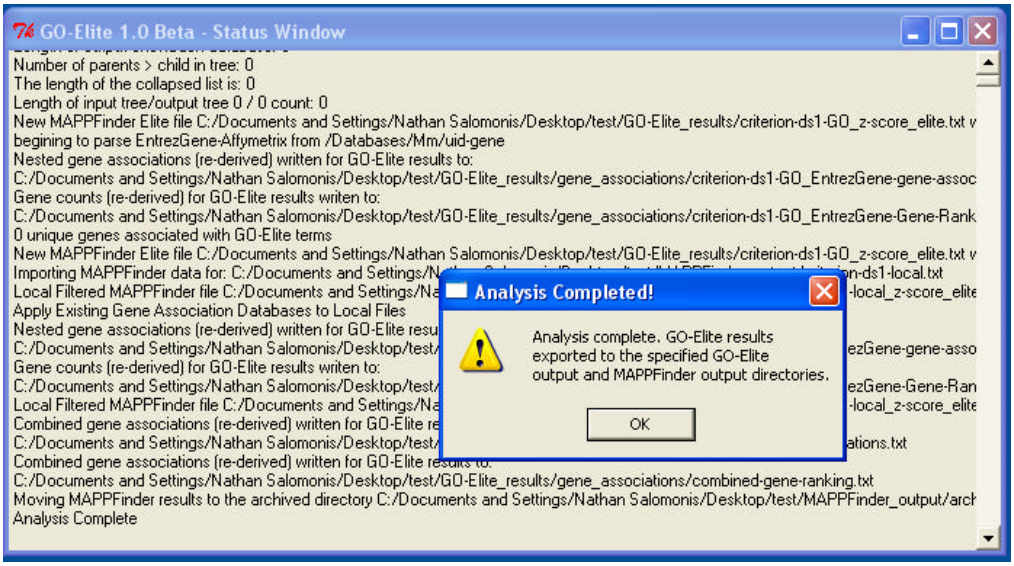


1. **Select GO-Elite filtering options**: This option instructs GO-Elite which pruning option to use on the pre-processed ORA data (this is the Elite step). The options are z-score, gene number and combination. These comparison statistics are used to compare related GO terms that have **a minimum z-score**, **maximum permuted p-value** and **minimum number of genes** designated by the user in the adjacent entry fields. Only those GO-terms or pathways which meet these filters are used filtering in GO-Elite. The z-score option will instruct the program to compare related GO terms based on their relative over-representation z-score, which is an indicator of degree of over-representation. Gene number compares the number of genes changed in the pathway and combination is the z-score weighted by the log2 number of genes changed. See the GO-Elite methods paper to see the description for each of these options (combination additionally weights z-score results based on the number of genes changed for each term). In each case, when two or more related GO terms are compared, GO-Elite chooses the GO-term with the highest scoring statistic based on the relative position of the terms in the GO hierarchy. The user is encouraged to try different methods and compare these.

2. **Select the pre-computed ORA results directory**: Selecting this button opens a folder selection menu, where the user can choose which directory to save your GO-Elite pruned results to. The results will be stored in a new sub-directory of this folder named “GO-Elite_results”. If using the “raw input” option, the users ORA results will also be saved to this folder in the directory named “MAPPFinder_output”. You will notice the result files will be saved to a sub-directory of “MAPPFinder_output”, named “archived-*time-stamp*” where the time-stamp indicates the date and time the analysis was run. If the user chooses not to set the pre-computed output directory, the folder “output” in the GO-Elite program folder will be used instead.
3. **Select the Input Gene List directory**: This option is only present if using the “pre-processed” option. Selecting this button allows the user to select which directory contains a set of input gene lists that corresponds to the ORA results file. The array IDs/genes in this file will be linked to the results file and used to annotate the results based on which actual genes are changed. The name of the input files should be the same as the ORA results file, where the ORA results may also be proceeded by “-GO.txt” or “-local.txt”.
4. **Format the GO-Elite Results for further analysis in MAPPFinder**: This option is useful for uses with pre-processed results from GenMAPP who wish to re-view the filtered results in that program.

Once these options are selected GO-Elite will generate ORA statistics and/or filtered GO-Elite summary results. Processing can take several minutes to 45 minutes per input gene list, if running the “raw input” option for > 4000 array IDs/genes with 2000 permutations. While your data is being processed a status window with reported progress is displayed. When finished a pop-up window will indicate that the analysis is finished and return you to the main menu.

Figure 3.4 – Status Windows



These summary result files include two GO-Eslite summary result files for each input or pre-processed list corresponding to GO and local associations (local is typically composed of

pathways automatically downloaded from Wikipathways – see How_to_Make_or_Update_Species_Databases.rtf). A third file is a combination of all GO-Elite results built or analyzed in a single run (can include multiple input files run at once). These sets of files can include gene symbols associated with each GO term or pathway and even summarized gene expression values present with the input gene IDs. In addition, there are gene level annotation files saved to the folder "Gene_Associations" in the output directory, useful for seeing which genes and input IDs are linked to each GO term along with gene annotations. The gene ranking file provides a means to see which genes are most highly represented in which GO term or pathways.