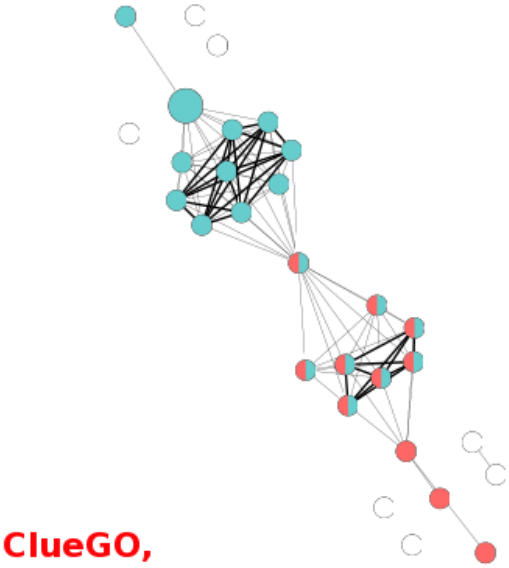
ClueGO

Documentation



Integrative Cancer Immunology Team

INSERM U872, Cordeliers Research Center, Paris, France

Institute for Genomics and Bioinformatics

Graz University of Technology, Graz, Austria

Contents

Installation

System Requirements:

- Windows, Linux, Unix or MacOS operating system.
- 512MB RAM needed, 1024 MB RAM recommended. Hard disk with at least 100MB free (for example files).
- Java 1.5+ needed, Java 1.6 recommended.
- Cytoscape 2.6.+ installed, with -Xmx1024m option in the start up file (.sh/.bat).

For downloading ClueGO, please visit: http://www.ici.upmc.fr/cluego/cluegoDownload.shtml.
ClueGO is a Cytoscape plugin, thus it is necessary to copy the ClueGOPlugin_v1.x.jar in the plugin folder of Cytoscape.
At the first startup, a folder containing the precompiled ClueGO files and sample files will be created in the user home folder (.cluegoplugin). If this folder is removed or the content is damaged, it will be recreated automatically at the next startup.
If a new version of the ClueGO plugin is downloaded, please delete the old version from the Cytoscape plugin folder.

Documentation

Generally, the biological interpretation of large gene clusters derived from high-throughput experiments is a real challenge. Many ontology sources exist in order to capture biological information in a meaningful way.
The Gene Ontology (GO) projects [1] aims to capture the increasing knowledge on gene function in a controlled vocabulary applicable to all the organisms. GO describes gene products in terms of their associated biological processes, cellular components and molecular functions. The terms are structured in a hierarchical relationship (parent-child). Due to the complexity of the hierarchical structure (directed acyclic graph), the terms can be in several different levels.

The specificity of the terms varies along the tree: from very general terms (in first levels of GO) to very specific ones.

Kyoto Encyclopedia of Genes and Genomes (KEGG) [2] is a database of biological systems that integrates genomic, chemical and systemic functional information.

BioCarta (http://www.biocarta.com/) provides usefull pathway information.

For a complete view on the studied process, several ontology sources should be consulted in order to integrate their complementary information. In each source, for each gene, there is a large amount of information. This makes the analysis of the relationship between genes and terms very difficult to represent and comprehend. Also, for close related terms, a high degree of redundancy of their associated genes exists.

To answer to this problematic, we developed ClueGO, an open-source Java tool that extracts the non-redundant biological information for large clusters of genes, using GO, KEGG and BioCarta. ClueGO is integrated in Cytoscape [3] as a plug-in and it is taking advantage of its complex visualization environment.

ClueGO features:

- ClueGO allows analysis of a single gene set (cluster) or cluster comparison.
- Different filter criteria can be applied to the terms.
- Fusion of the related terms that have similar associated genes.
- Functional grouping of the terms based on GO hierarchy or based on kappa score.
- Visualize the selected terms in a functionally grouped network.
- Charts presenting the specific terms and groups for the clusters compared.
- Statistical significance for the terms and for the groups.
- ClueGO can be used in combination with GOlorize.
- Easy updatable.
- Easy extendable.

Selection Panel

After starting Cytoscape, ClueGO can be found in the Plugins menu. Once selected, it will display the ClueGO selection panel on the left side in Cytoscape (see Figure 1):

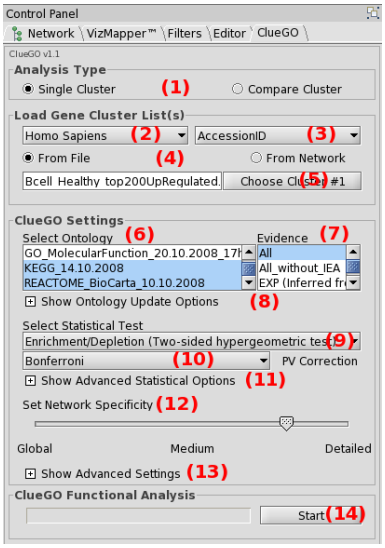


Figure 1: *ClueGO selection Panel*

1. Select the type of analysis. Besides analysis of a single gene set (cluster), ClueGO performs comparison of clusters, underlying the specificity and also the common aspects of their functionality.

2. Select organism. For the moment, ClueGO supports Arabidopsis thaliana, Bos taurus, Caenorhabditis elegans, Danio rerio, Dictyostelium discoideum, Drosophila melanogaster, Escherichia coli, Galus galus, Homo sapiens, Magnaporthe grisea, Mus musculus, Oryza sativa, Rattus norvegicus, Saccharomyces cerevisiae.

3. Select identifiers type (AffymetrixID, AccessionID, SymbolID).

4. The identifiers can be uploaded from text files or interactively from a Cytoscape sub-network. ClueGO provides sample files containing specific genes (AccessionID) for Human B and NK cells. A Saccaromyces network sample (galFiltered.cys) can be found in the Cytoscape sampleData folder.

5. Choose the file with identifiers.

6. Select the ontology/ontologies. To make ClueGO faster, we use precompiled files based on GO, KEGG and BioCarta. The precompiled files are included in the ClueGOPlugin_v1.x.jar. Perform point 8. before selecting the ontologies if necessary.

7. Select the evidence code, option available for the GO based files. For more details, see http://www.geneontology.org/GO.evidence.shtml.

8. Update the ontology (see ClueGO updates)

9. The significance of each term or group can be calculated with hypergeometric test (enrichment, depletion, two sided). Standard is hypergeometric test two sided.

10. Several methods for PValue correction are proposed: Bonferroni, Bonferroni step-down and Benjamini-Hochberg.

11. Advanced statistical options

12. Select network type. We predefined selection criteria leading to a "Global" network, a "Medium" network or a "Detailed" network.
    The "Global" network displays GO terms found in the GO levels 1-3, with a high number of genes associated and a small percentage of uploaded genes found. Those terms provide a general biological information. At the opposite, the "Detailed" network shows very specific terms placed in GO levels 9-14, with only few associated genes but a high percentage of the uploaded genes found. Those terms are more specific and informative underlying particular aspects of the studied gene product.
    The predefined settings provided by ClueGO should be combined with a customized selection for an analysis adapted to the biological question addressed (general/specific investigation) and to the number of genes involved.

13. Advanced Settings.

14. Start the analysis.

ClueGO result

A functionally grouped annotation network, a ClueGO information table, charts with terms and groups and the logging information are the result of ClueGO analysis. The created networks and analysis results can be saved in a specified project folder and used for further analysis. If the analysis is not needed anymore, it is recommended to close the project.

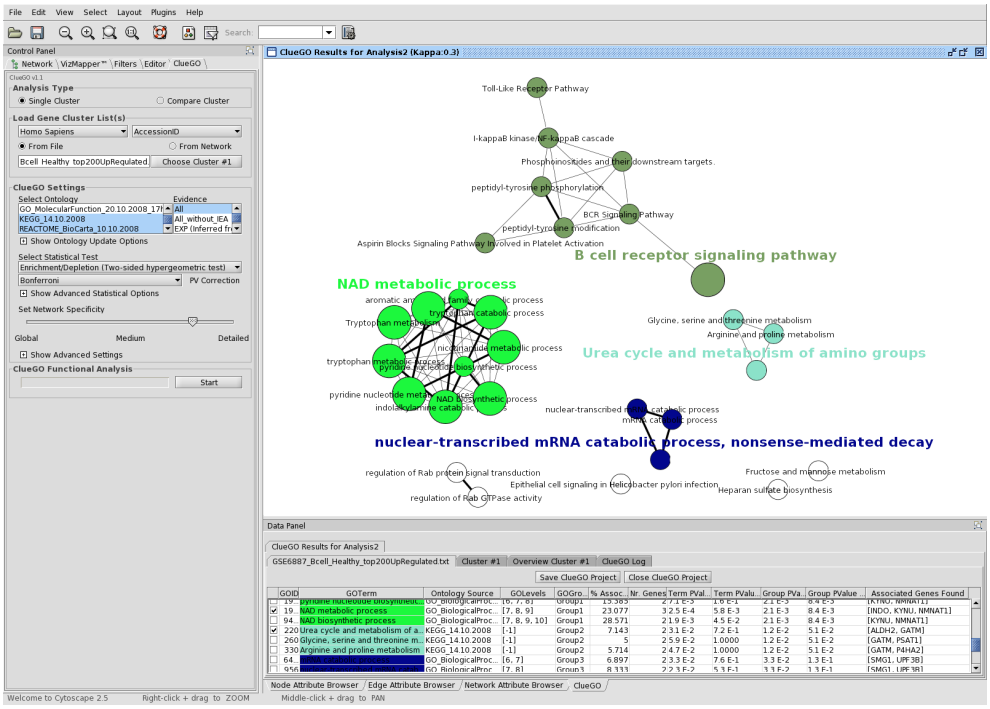


Figure 2: *The biological role of B cell specific genes visualized with ClueGO.*

Functionally Grouped Annotation Network

ClueGO visualizes the selected terms in a functionally grouped annotation network (see Figure 2) that reflects the relationships between the terms based on the similarity of their associated genes. The size of the nodes reflects the statistical significance of the terms (see Statistics). The degree of connectivity between terms (edges) is calculated using kappa statistics, in a similar way as described in [4]. The calculated kappa score is also used for defining functional groups. A term can be included in several groups. The reocurrence of the term is shown by adding "_n". The not grouped terms are shown in white color. Predefined, the group leading term is the most significant term of the group. The network

integrates only the positive kappa score term associations and is automatically laid out using Organic layout algorithm supported by Cytoscape.

Information table

ClueGO Information table provides information about the selected terms. From left to right:

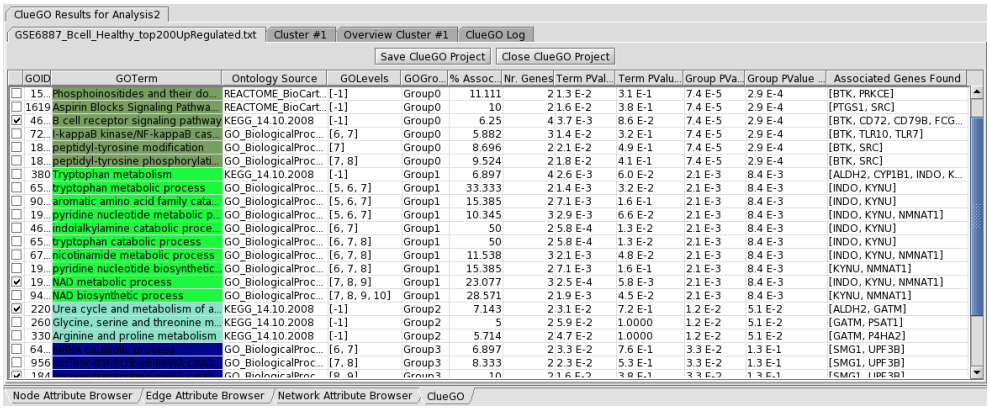


Figure 3: *ClueGO information table, detailed information about the selected terms and their associated genes*

1. Group leading term. There are several ways of defining the leading term of the group (see Advanced Settings). Predefined, the leading term has the highest significance in the group.

2. GOID.

3. GO Term.

4. Ontology Source.

5. GO levels. Due to the complex structure of GO tree (directed acyclic graph), the GO terms can be placed in several levels. In case of using sources without hierarchical structure (KEGG, BioCarta), the level it is assigned as -1.

6. The group or the groups that include the term.

7. The percentage of the genes from the uploaded cluster that were associated with the term, compared with all the genes associated with the term.

8. The number the genes from the uploaded cluster that were associated with the term.
9. Term significance (PValue).
10. Term significance (corrected PValue).
11. Group significance (PValue).
12. Group significance (corrected PValue).
13. The genes from the uploaded cluster that were associated with the term.

Histogram with terms

The chart presents the specific terms for the user genes and information related to their associated genes. The display can be customized (see Advanced Settings). Predefined, the bars represent the number of the genes from the analyzed cluster found to be associated with the term, and the label displayed on the bars is the percentage of found genes compared to all the genes associated with the term. Term significance information is included in the chart (see Statistics).

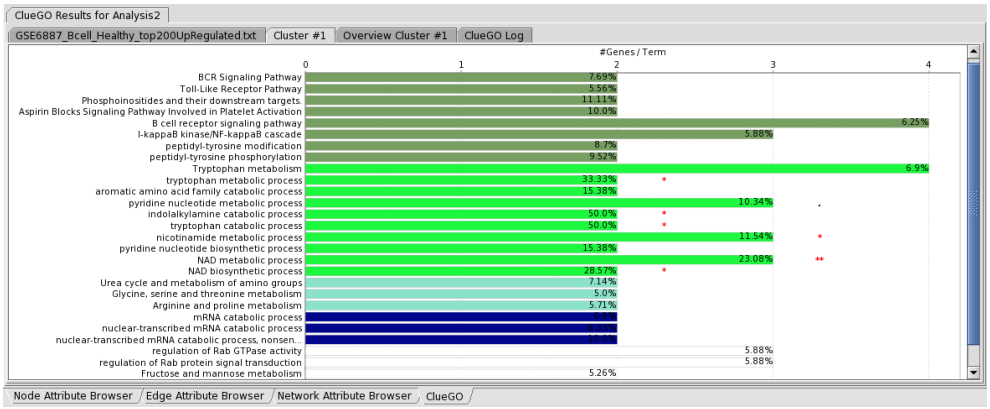


Figure 4: *Functional terms in ClueGO Chart*

Overview Chart with functional groups

The overview chart presents functional groups for the user genes. The name of the group is given by the group leading term (e.g. the most significant term in the group). The group sections correlate with the number of the terms included in group. The position of the sections can be changed (right click).

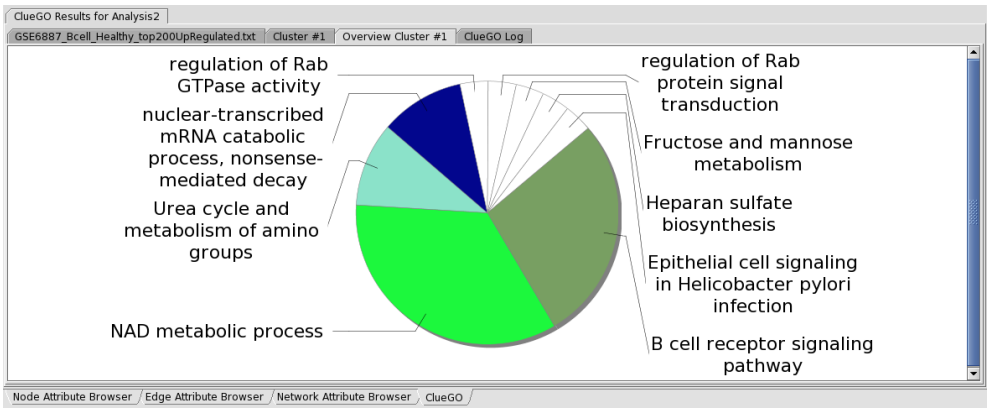


Figure 5: *Functional groups in ClueGO Overview*

Logging information

Complete information on the applied selection criteria for each analysis can be visualized in the logging information panel.

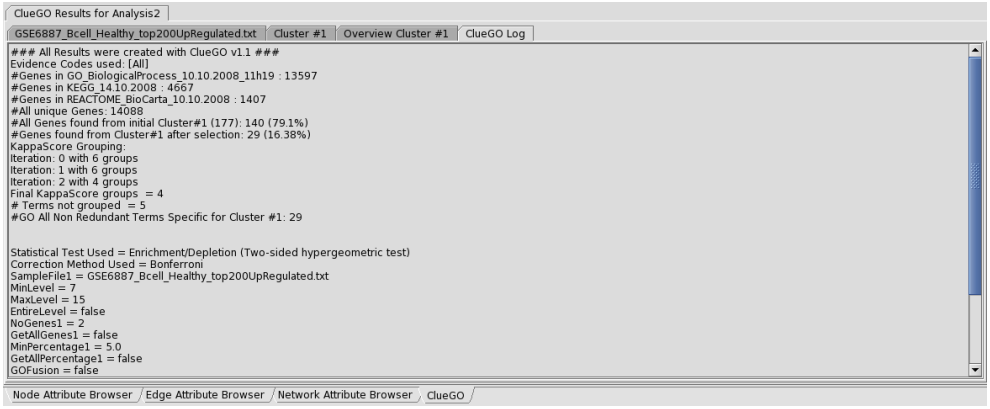


Figure 6: *ClueGO Logging Information*

Advanced Settings

ClueGO advanced settings allow for a precise adjustment of the selection criteria.

1. GO tree level. The identifiers can be mapped in the entire ontology tree or in a GO level interval (between Min and Max levels). If selected the first GO levels (1-3), the terms will be very general. Those terms have many associated genes, so, there will be a low percentage of found genes from the user genes. The very specific terms (e.g. level 12) have associated a small number of genes.

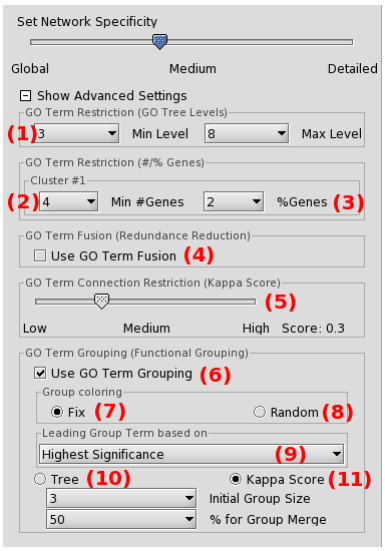


Figure 7: *Advanced Settings*

In this case, the percentage of found genes it is higher. Anyway, one has to consider the complex hierarchical GO structure that makes the terms to be in several levels. This feature is availabe while using a hierarchically structured ontology source (e.g. GO).

2. Minimum number of identifiers from the uploaded cluster found to be associated with a term

3. Minimum percentage of mapped identifiers present per term compared to the total of its associated genes

4. Fusion. The terms in parent-child relationship that share similar genes (identical, or with one gene difference) are assessed and the most representative parent or child term is retained.
   Note: This is an optional feature. The fusion strongly reduces the redundancy.

5. Kappa Score shows the relationships between the terms based on their overlapping genes. It is used for creating the network and for creating the groups. Since the term-term matrix is of categorical origin, kappa statistic was found to be the most suitable method. Initially, a term-gene matrix containing the selected terms and their associated genes is created. Based on this, a term-term similarity matrix is calculated using kappa statistics to determine the association strength between the terms (kappa score). High score = visualize in the network only the connections between close related terms, with very similar associated genes (high stringency). Low score = allows to visualize in the network the connections between less related terms (low stringency).

6. Create functional groups. The terms can be associated in functional groups using the ontology's hierarchy or the kappa score.

   When displaying the functional groups on the network, one has to consider that the network structure is calculated with kappa score. The kappa score grouping of the terms will correspond with the network structure that is calculated in the same manner. If the hierarchical grouping is used, it is possible to be a difference between the networks structure (based on the associated genes) and the functional groups defined using GO hierarchy.

7. Use fix coloring for the groups

8. Use random coloring for the groups

9. The most representative term in a group is giving the name of the group. It can be considered having:

   a) the highest number of the found genes per term

   b) the highest percentage of found genes per term

   c) the highest percentage of found genes per total number of found genes

   d) the most significant PValue (see Statistics)

   This selection determines the ClueGO charts display: the number/percentage/significance of the terms. If the group leading term is the term with the higher number of genes or the most significant term from the group, a label with the percentage of the found genes compared with all the associated genes with the term is displayed on the bars of the chart. Reversely, if the group leading term is representing the percentage of genes (compared to all the genes associated with the term or with all the genes found from the uploaded cluster), on the label will be displayed the number of the genes found for this term.

10. The hierarchy based grouping consists in analyzing the selected terms through the perspective of their parent terms. The size of the groups can be defined using the number of common and different parents, that is depending on the level of the term in GO hierarchy. Since KEGG and BioCarta don't have a hierarchical structure, the terms selected from those souces will not be grouped.

11. For grouping the terms based on the kappa score, it is necessary to set the size of the initial group (e.g. 3) and the percentage (overlaping terms/group) for group merge (e.g. 50 percent).

For more details see [4]. All the possible initial groups with the terms showing a kappa score equal or above the predefined threshold (e.g.<= 0.3) are created. Further, the initial groups are iterratively compared and merged if they are overlapping in the defined percentage. A term can be part of several groups. The reocurrence of the term is shown by adding "_n", where n is the number off additional occurences. If the terms have very similar associated genes, it is possible that the iterative merging of the created groups to take a longer time in the attempt of merging all the groups. In this case, it is recommnended to increase the kappa score level or to increase the percentage for group merge.

Advanced Statistical Options

1. Select mid-P-value for a less conservative hypergeometric test [5].
2. Select Doubling for a less conservative two tailed hypergeometric test [5].

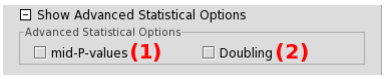


Figure 8: *Advanced Statistical Options*

Cluster Comparison

Once selected the option: ”Compare clusters”, two file choosers are made available. The selection of the ontology terms is following the same steps as in the analysis for a single cluster.The difference refers to the number and the percentage of the genes per term.

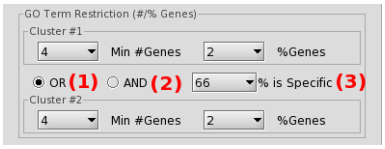


Figure 9: *ClueGO comparison, settings.*

Those selection criteria can be flexible: Cluster1 OR Cluster2 options (1) or applied strictly for both clusters: Cluster1 AND Cluster2 options (2). It is recommended to have comparable number of identifiers in the analysed clusters. If one cluster is much larger, the selection criteria should be applied more strictly (e.g. increase the number of genes per term).

Further, we analyze the terms trough the perspective of their associated genes. It is possible that genes from both clusters to be associated with a term, but in different proportions. We defined a term as specific for one of the clusters if the percentage of associated genes from this cluster is higher than the selected threshold (e.g. 66 percent) (3). As result, charts with specific terms for each cluster are provided. The common terms are included in a separate chart.

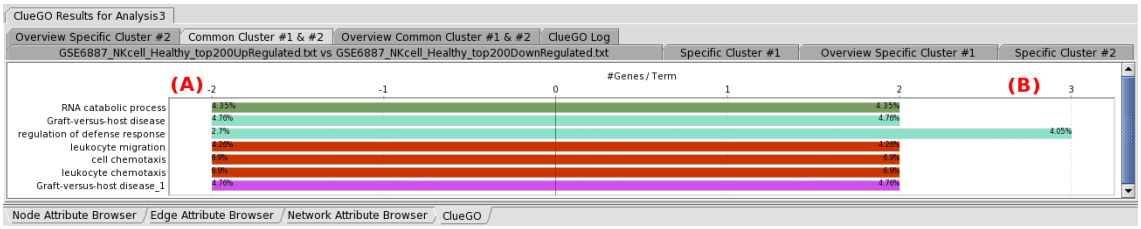
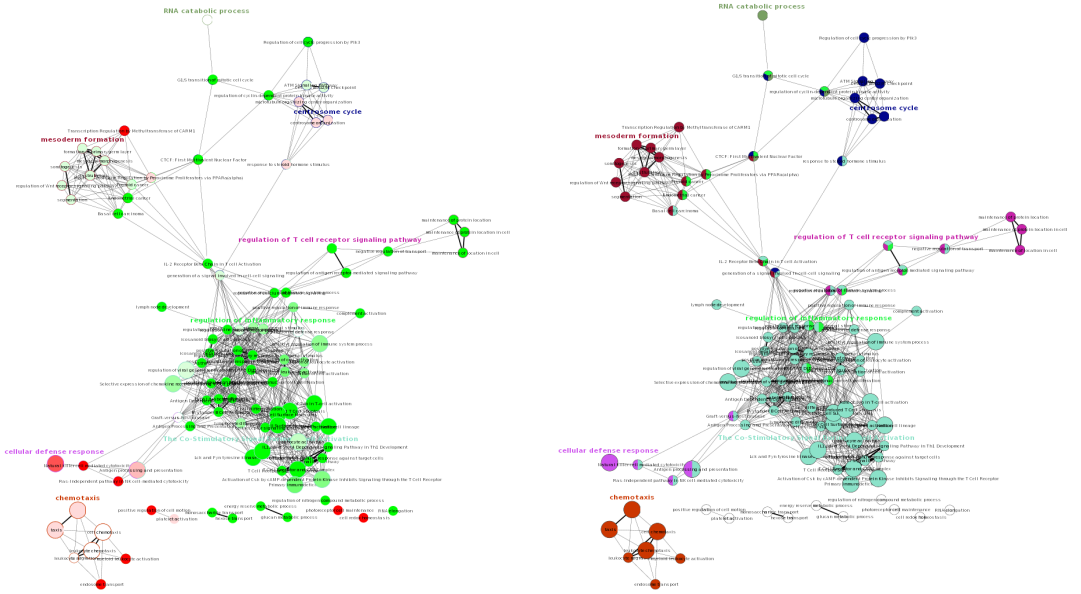


Figure 10: *Common terms for the analyzed clusters. A) The number of cluster1 genes associated with the term. B) The number of cluster2 genes associated with the term.*

On the network, the different proportion of the genes from the analyzed clusters is represented with a color gradient from green, for the first cluster genes, to red for the second cluster. The visualization of the groups on the network can be switched with the one of the uploaded clusters distribution on the selected terms.



(a) Functionally grouped network (b) Distribution of the two clusters on the network

Figure 11: *The biological role (GO, KEGG, BioCarta) of NK cell up and down regulated genes visualized with ClueGO (kappa score: ≥ 0.3). (a) Groups network. Not grouped terms are shown in white. (b) Cluster distribution network. Terms with up/down regulated genes are shown in red/green, respectively. The color gradient shows the gene proportion of each cluster associated with the term. Equal proportions of the two clusters are represented in white.*

Result folder

In the result folder are included (from up to down):

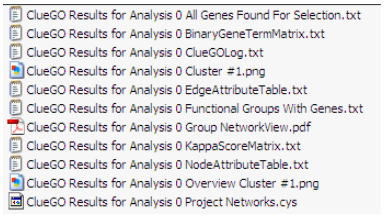


Figure 12: *ClueGO result folder*

1. The list of genes that were found associated with the selected terms (under the selection criteria applied).

2. The binary gene-term matrix.

3. The logging information.
4. The terms (image, png).
5. Network edges information: the kappa score value for each connection.
6. The groups and the corresponding genes (txt).
7. The network (image, png).
8. The kappa score matrix (see Figure 17(a)).
9. The ClueGO browser information (GOID, GOTerm, Levels, Groups, Associated Genes, Significance).
10. The groups (image, png).
11. The network (.cys). It is used for recreating the network.

If a cluster comparison is performed, charts with specific and common terms and groups are provided.

Statistics

ClueGO calculates automatically the significance of the terms and groups.
The resulting PValues are included in the network visualization (the size of the nodes) and are also displayed on the charts. If a PValue correction method is selected in ClueGO selection panel, then on the network and on the charts the corrected PValue will be represented. The terms and groups significance can be found in the ClueGO browser.
In the charts, we mark the level of the significance for terms and groups using:
a) **: if the term/group is oversignificant, PValue<0.001
b) *: if the term/group is significant, 0.001<=PValue<0.05
c) . (dot): 0.05<=PValue<0.01
By default, the PValue is calculated with a two-sided minimal-likely-hood test on the hypergeometric distribution, equivalent to a classical Fisher's exact test [5]. Additional options provided are the left-sided (Enrichment) and right-sided (Depletion) hypergeometric tests. We also implemented the possibility to calculate mid-P-values and to apply doubling to two-sided hypergeometric tests. These features are a

possibility to deal with the conservatism and discreteness of the hypergeometric distribution [5].

Several methods for PValue correction are proposed: Bonferroni, Bonferoni step-down and Benjamini-Hochberg.

We consider as reference the total number of the genes associated with all the terms included in the ontology source used. If several ontology sources are used, we consider all the unique genes found in all of the sources. The number of reference genes is mentioned in the logging information.

For the group significance we consider the number of unique genes found from the uploaded genes as associated with the terms included in the group and the total number of unique genes associated with those terms.

We provide a sample of PValue calculation for the term GO:0030837 (negative regulation of actin filament polymerization) using the classical Fisher Exact Test (equivalent to a two-sided hypergeometric test). The ontology used was: GO, Human, Biological Process (version from 26.09.2008). This ontology has 13592 associated genes. We did not apply any restriction regarding the evidence code of those genes. In the GO tree, genes with an evidence code inferred from electronic associations (IEA) are not displayed as associated with the terms in the tree.

The uploaded gene list had 300 of affymetrix identifiers (B cell sample file). 181 genes without any selection criteria of the term corresponding to these IDs were found in GO. From those genes, 2 genes were associated with the considered term.

The calculated PValue was 4.6E-3, and corrected PValue 9.8E-2 (Bonferroni).

	in term	NOT in term	Sum
user genes	2	179	181
reference genes	6	13586	13592
Sum	8	13765	13773

Table 1: Example of calculating the term significance

Update

ClueGO allows an easy integration of the most recent version of the GO and KEGG. The ontology and annotation source files are automatically downloaded and based on this, new ClueGO precompiled files are created.

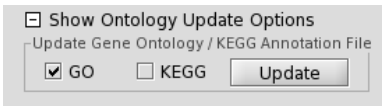


Figure 13: *ClueGO update*

GO ontology (OBO file) and annotation files are directly downloaded from the FTP site of GO. The name of the GO based precompiled files contains the creation date of the ontology and is taken from the OBO file. The usage of multiple GO based precompiled files (e.g. BiologicalProcess, MolecularFunction, CellularComponent) is possible for files having the same data included in their name. GO has daily content changes.

ClueGO allows also KEGG updates via SOAP API. ClueGO KEGG based files will have in the name the date of the download. The KEGG update rate: http://www.genome.jp/en/info_dbget.html.

Since Biocarta is not frequently changing and only supports human and mouse, updates will be supplied with each new release of ClueGO. BioCarta source files are taken from REACTOME via SOAP API.

Source files and interfaces used:

1. GO Ontology, in OBO format: ftp://ftp.geneontology.org/pub/go/ontology/gene_ontology_edit.obo
2. GO Annotation files: ftp://ftp.geneontology.org/pub/go/gene-associations/
3. KEGG API: http://www.genome.jp/kegg/soap/
4. Biocarta API: http://www.reactome.org/download/index.html

ClueGO releases will be available for download at: http://www.ici.upmc.fr/cluego/cluegoDownload.shtml. Note that the update speed depends on the user's internet connection. The KEGG update can fail in case of high web traffic. The ClueGO precompiled files not needed anymore can be easely removed (right click). However, the initial ontology files will be recreated again at the next startup.

Include your organism of interest in ClueGO

ClueGO allows to add new organisms, not included in our initial selection.

For each new organism, a keyIdentifier file: organism.gene2accession.txt (e.g. Oryza Sativa.gene2accession) has to be created. In this file, the first column ("UniqueID#EntrezGeneID") will contain the keyID (e.g. EntrezGeneID) that is used for mapping. Next column, "SymbolID" contains the official gene symbol

or official symbol and synonyms separated by "|". Other types of identifiers can be included in the next columns.

The second file needed is organism.properties (e.g. Oryza Sativa.properties). This file contains the organism name, the keggID, the url of the GO association file, the taxonomy id. The same url can be used for other organisms after changing the name of the annotation file. In most of the cases, there are no special files required for the update (false).

Here is the content of the .properties file for Oryza sativa:

organism.name = Oryza sativa

organism.kegg.name = osa

organism.go.url = ftp://anonymous:anonymous@ftp.geneontology.org/pub/go/gene-associations/gene_association.gramene_oryza.gz

organism.taxid = 39946

organism.update.needs.special.file = false

Those files have to be added in a new folder (e.g. Organism_Oryza sativa) created in UserHomeFolder, .cluegoplugin, currentVersion, ClueGOFiles, ClueGOSourceFiles.

Remark: At the addition of a new organism, make sure that the ids provided by KEGG are the same as the UniqueIDs from organism.gene2accession.txt.

Map other ID types than the IDs supported by ClueGO

Since ClueGO is supporting a limited number of identifiers type, we made available a feature that allows the mapping of other types of identifiers, not included in our selection.

For each supported organism, a keyIdentifier file: organism.gene2accession is provided (e.g. HomoSapiens.gene2accession). In this file, the first column ("UniqueID#EntrezGeneID") contains the keyID (e.g. EntrezGeneID) that is used for mapping. The type of keyID differs from one specie to another due to the most frequent usage of a certain type of ID for different species (e.g. for Drosophila is FBgnID).

If the identifiers of interest are not supported, a new text file (tab delimited) containing the keyID (from the provided files) as first column and the corresponding new ids in the next colums can be created. The name of the first column has to be "UniqueID#keyID". If several identifiers correspond to a keyID, they have to be separated with "|". This file is added in the UserHomeFolder, ClueGOFiles folder, ClueGOSourceFiles, OrganismOfInterest.

Analysis from Cytoscape network

In ClueGO, the identifiers can be uploaded interactively from a Cytoscape network.

If selected the option ”From Network” (1) in the ClueGO Selection Panel, new features are made available. After selecting the genes from the network, their attributes have to be loaded (2: ”Load Attributes” button). Then, from the available attributes (3) is selected the attribute to display (e.g. SGD ID). This attribute must correspond to the selected type of identifier (e.g. SGDID). The selected attribute list is uploaded in ClueGO (4: ”Choose Cluster” button). The number of the selected nodes will be displayed in a text field (5). If the type of identifier is not the same with the selected attribute node, the analysis cannot be performed. In this case, after selecting a new type of attribute, make sure that the new selected attribute list is uploaded in ClueGO (4).

The comparison of two clusters is also available.

A sample Cytoscape network for Saccaromyces (galFiltered.cys) can be found in Cytoscape, sampleData folder. The node attributes can be visualised in Cytoscape in ”Node Attribute Browser”.

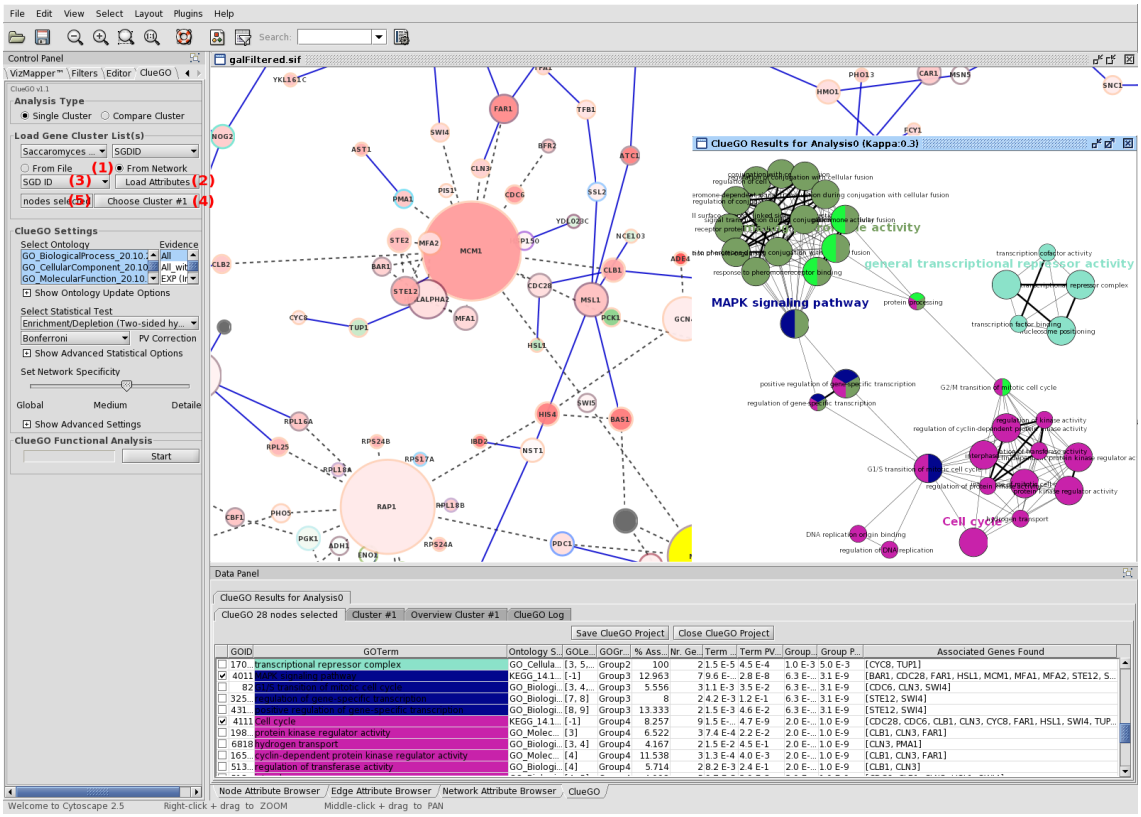


Figure 14: *ClueGO analysis with genes uploaded interactively from Cytoscape network*

ClueGO and GOlorize

ClueGO can be used in combination with GOlorize ([6]).

If both plugins are present in Cytoscape plugin folder, the ClueGO data panel is automatically transformed into ClueGOlorize panel. This panel combines the functionality of both, ClueGO and GOlorize panel.

From a Cytoscape network the genes of interest are selected. With ClueGO, biological functions are associated to those genes. In ClueGOlorize panel, the terms to be mapped on the network can be selected (1: ”Select All”). Using GOlorize, the genes will be coloured on the network with ”Validate”(2).

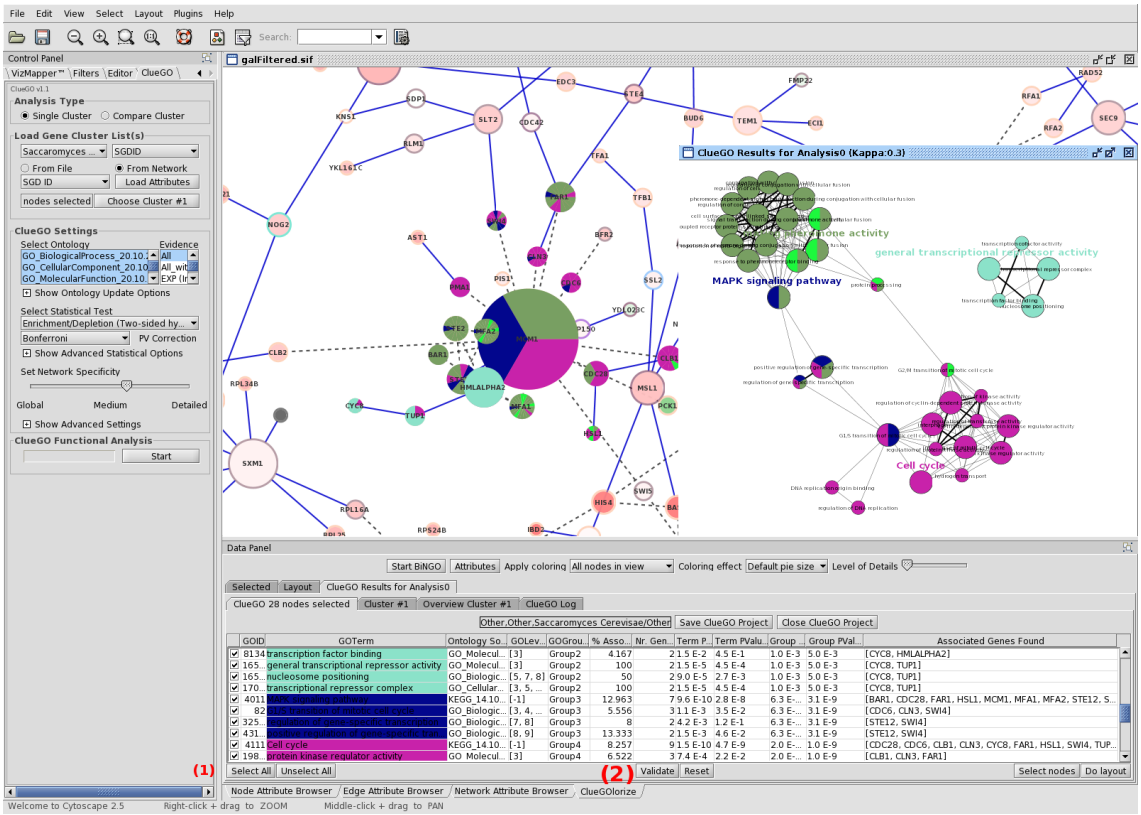


Figure 15: *ClueGO used in combination with GOlorize*

Sample data

From Gene Expression Omnibus we downloaded the dataset GSE6887, a gene expression profile of peripheral blood lymphocytes of melanoma patients and healthy controls. From this dataset we selected the expression profiles NK cells in healthy controls and calculated the mean over all replicates. The genes were sorted based on their expression and the top 200 most expressed genes were included in the list with up regulated genes and the top 200 less expressed genes in the list with down regulated genes. The same steps we applied to the B cell data. The reference used in this dataset was a pool of all immune cell types.

We use as ClueGO sample files (Human, AccessionID) the lists with the 200 up and the 200 down regulated genes for B cells and NK cells. The sample lists can be found in .cluegoplugin folder, in the user's home folder.

Changing the ClueGO network and charts

- In ClueGO, the colors attributed to the groups are created automatically each time the analysis is performed.

  One can change the color of the groups using the Cytoscape feature: "Node color", available under "VizMapper". If a comparison analysis is performed, in order to change the colors of the groups make sure that the groups are displayed on the network (option "Show Groups"). If the "Show Difference" option is active in ClueGO, under "Node color" feature will be displayed the color gradient used for showing the percentage of the genes found from the analysed clusters.

- The two colors defining the gradient displayed under "Show Difference" can be changed by using the color choosers available in the ClueGO selection panel, next to the file choosers.

- Cytoscape Layout features allow modifications of the network (e.g. rescaling, rotating, apply different layout patterns).

- The size of the displayed text on the network can be modified in Cytoscape VizMapper under "Node Font size". By default, the font size of the term names is 12 and of the group names 20. If the font size is set to a small value (e.g. 0.1), the names will not be anymore visible on the network.

- One can sort the data from ClueGO browser and select the terms of interest (e.g. the terms significant after PValue correction). Then, in Cytoscape, with File, New, Network, From selected nodes (Ctrl N) a new network is created. For the display, we use the Organic layout.

Running time

The running time depends on the number of genes uploaded but also on the selection criteria.
A longer time will take the analysis of a large list of genes mapped in the entire GO with a small minimum number of genes found per term, small percentage of those genes and small kappa score. If the percentage for group to merge is increased, there will be a high number of groups in the end.
The status of the analysis can be followed in ClueGO selection panel, under Current Action. If the number of the groups visualized for different iterations of an analysis is very high, it is recommended to review the selection criteria:

- increase the minimum number of the genes found from the list for a term to be selected
- increase the minimum percentage of those genes for a term
- increase the kappa score
- decrease percentage for group to merge
- restrict the number of the terms based on GO levels

Also, if a comparison of cluster is performed, it is recommended to have a similar number of genes in each cluster. If the cluster have different number of genes, for the one with more genes more strict selection criteria should be applied.
To note is that certain parts of GO are more developed than others, due to an increased interest from the scientific community in that specific area. This could also impair the analysis. The number of iteration and the number of groups of the current iteration can be seen above the progression panel in the left side of Cytoskape. The number of iterations depends on the group-merging options, and on the fact that the algorithm converges. If the number of groups per iteration exceeds 500, the algorithm stops and applies the grouping with minimal number of groups found. An iteration with around 300 groups can take around 30 - 50 seconds on a 2.4GHz computer.

ClueGO examples

Fusion example

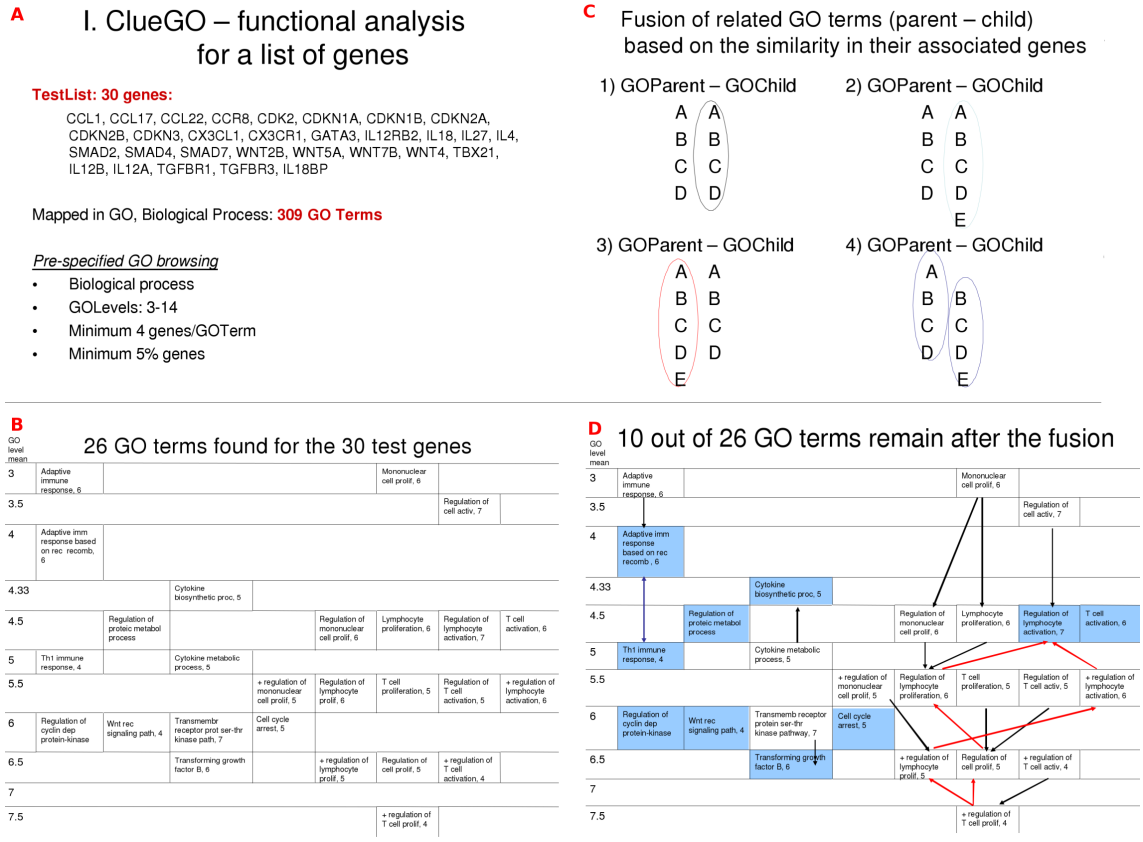
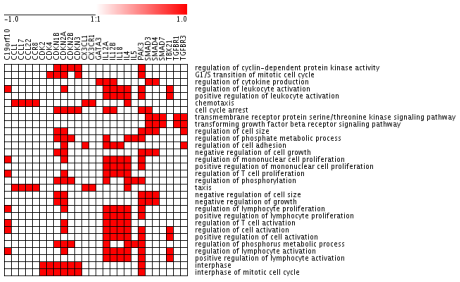


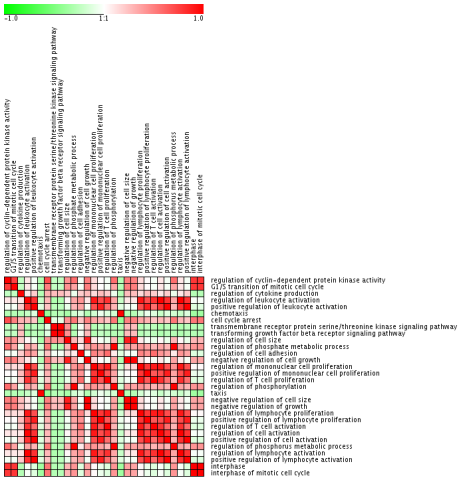
Figure 16: *Fusion example. A) The test gene list and the total number of corresponding GO terms. Selection criteria applied in ClueGO. B) GO terms selected under the specified selection criteria. C) Fusion criteria. It applies to GO terms in parent-child relation, with similar associated genes. The most representative parent or child term is preserved. The terms are organized based on their mean level specificity (up: general terms with low specificity, down: specific terms). D) GO terms mentained after fusion (colored in blue). Colored arrows show the fusion criteria used (concordance with C).*

Kappa statistics in ClueGO

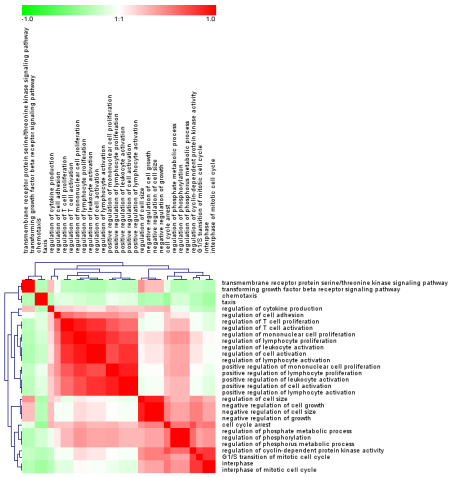
Steps in creating functional groups and the network.



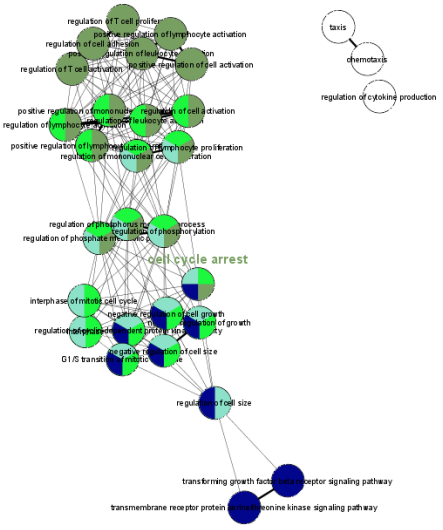
(a) Binary gene-term matrix



(b) Term-term similarity matrix



(c) Functional groups (Genesis)



(d) ClueGO network

1. Initially a binary gene-term matrix is created.

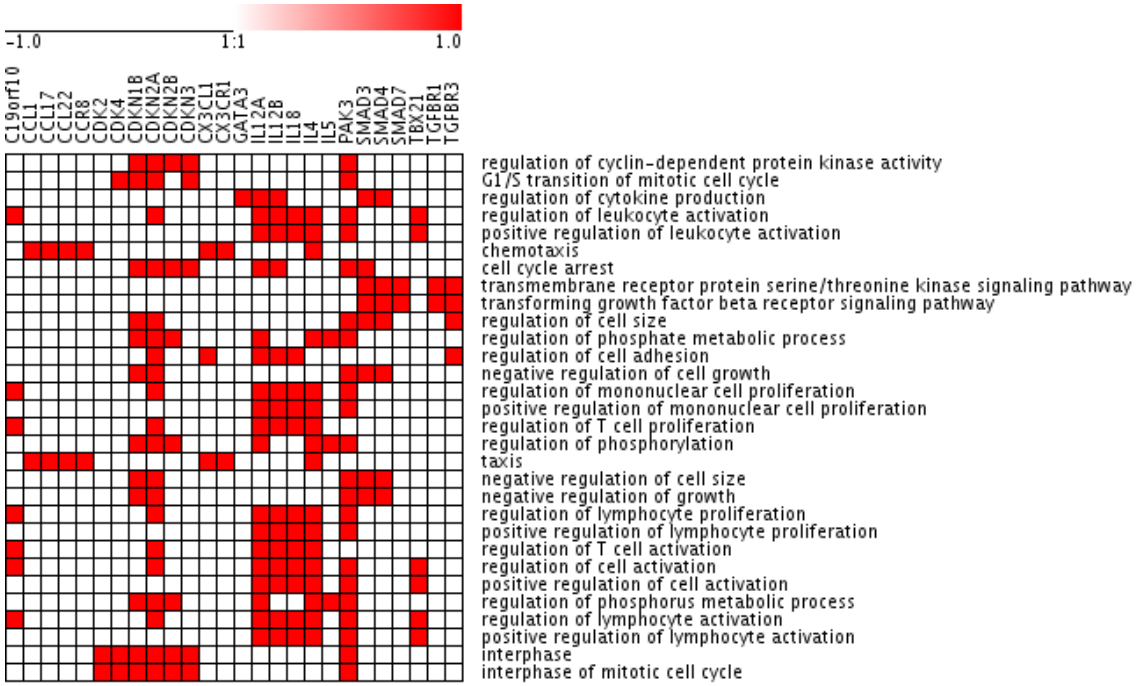


Figure 17: *Gene-term matrix*

2. Based on the gene-term matrix (step 1) and using kappa statistics, a term-term similarity matrix is calculated. Since the term-term matrix is of categorical origin, kappa statistics was found to be the most suitable method [7]. The method used for defining functional groups is similar with the one described in [4].

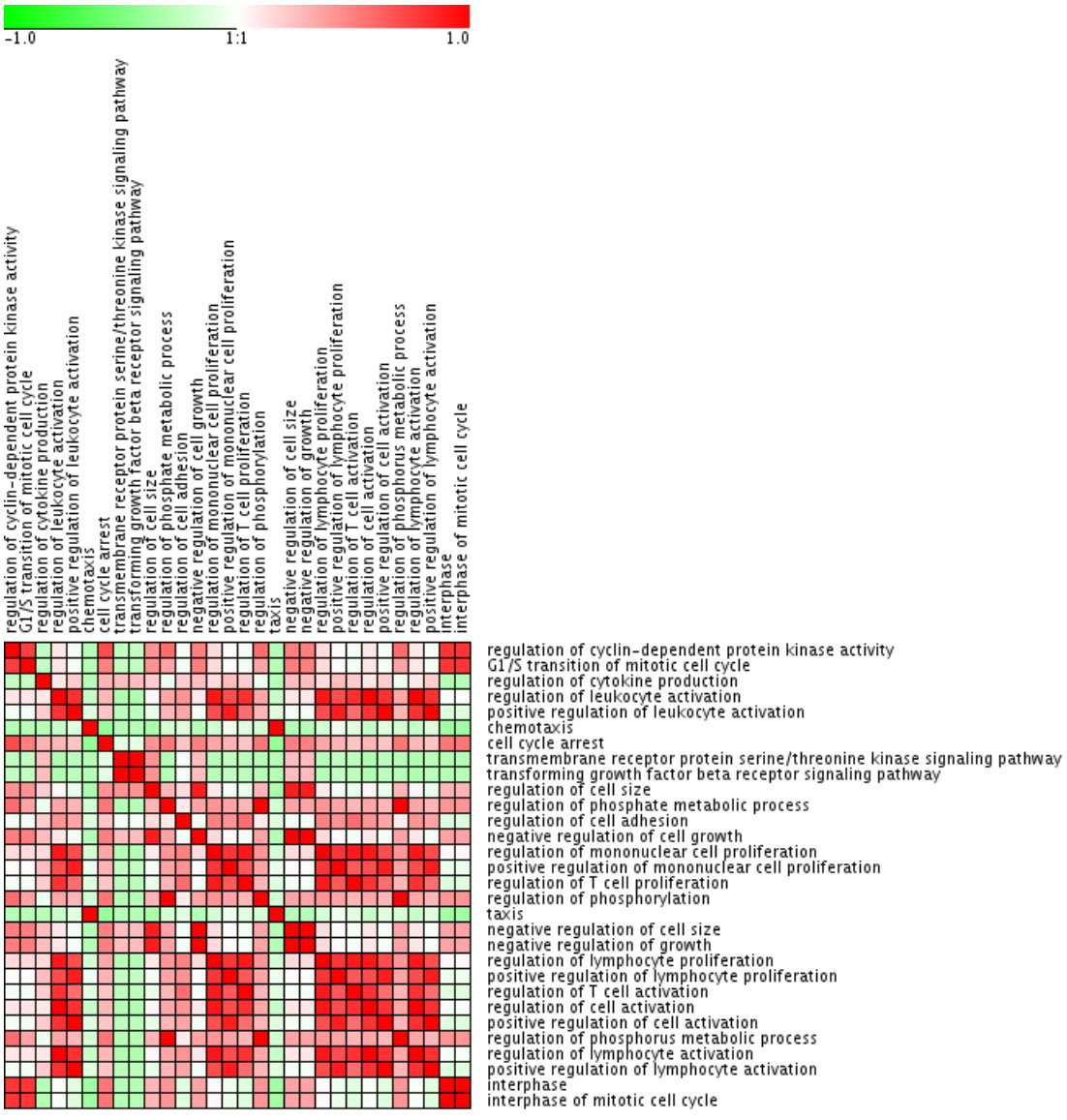


Figure 18: *Kappa score matrix*

3. Functional groups are created.

Initial groups are defined as having a minimum number of terms (e.g 3 terms/group) with a similarity of associated genes (kappa score) bigger than a predefined threshold (e.g. 0.03).

In the next step, the initial groups are iteratively merged if they overlap with a defined percentage (e.g 60). The iteration ends when the merging is not possible anymore.

Figure 19: *Kappa score matrix (visualized in Genesis)*

4. The network with terms (nodes) linked based on the predefined kappa score level is created. The edges with a kappa score smaller than 0.03 are not displayed on the network. E.g. the link between regulation of cell size and regulation of leukocyte activation with kappa 0.029.

The groups are visualized with different colors on the network.

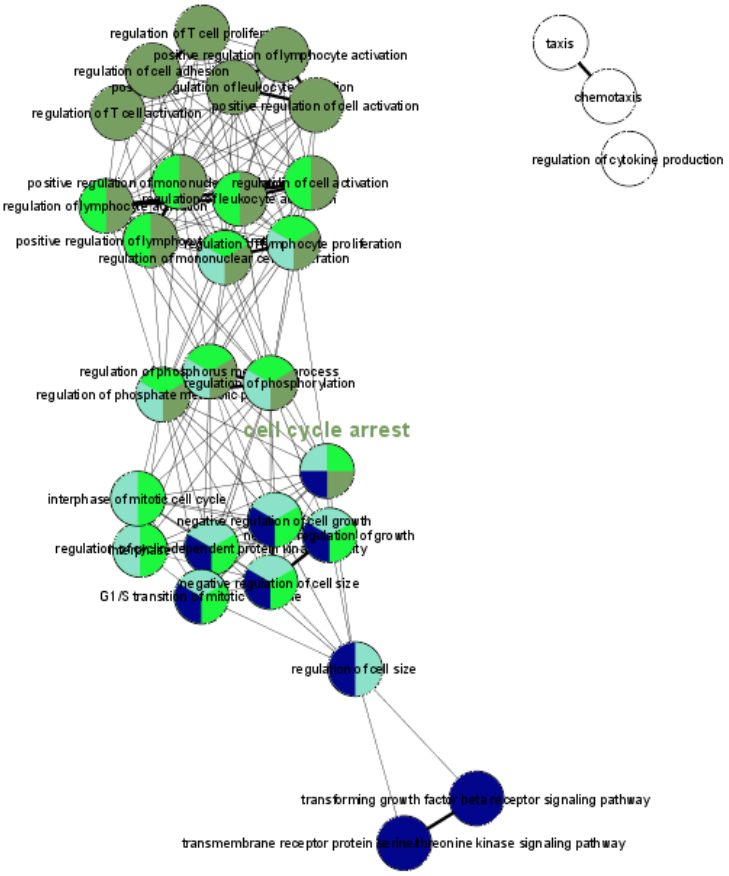


Figure 20: *ClueGO network*

Walk-through example

ClueGO settings in Figure 1 (paper)

- set the type of analysis: Compare Cluster
- select the organism: Homo Sapiens and the the type of ids used: AccessionID
- load sample gene lists:
  choose Cluster #1: select GSE6887_NKcell_Healthy_top200UpRegulated.txt
  choose Cluster #2: GSE6887_NKcell_Healthy_top200DownRegulated.txt
- select the Ontologies:
  GO_BiologicalProcess_10.10.2008 (All Evidence codes),
  KEGG_Pathways_14.10.2008 and
  REACTOME_BioCarta_10.10.2008
- select the statistical test: Enrichment/Deplection (Two sided hypergeometric test), FisherExactTest
- select the correction method: Bonferroni
- click Show Advanced Settings
- set GO Tree Level: Min 4 and Max 5
- set the selection criteria for the terms that have associated genes from cluster 1: min 2 genes/term and minimum 4% from all the Genes associated with the term
- select OR (e.g. min 2 genes from cluster #1 or min 2 genes from cluster #2)
- set is specific to 66% (if 66% or the genes associated with the term are from cluster #1, the term is considered specific for this cluster)
- set the selection criteria for the terms that have associated genes from cluster 2: min 2 genes/term and minimum 4% from all the Genes associated with the term
- select Use GO Term Grouping
- select Fix Group coloring

- select Leading Group Term based on Highest Significance
- select Kappa Score grouping with 3 terms in initial group and 50% overlap for groups to merge
- select ShowDifference
- Start

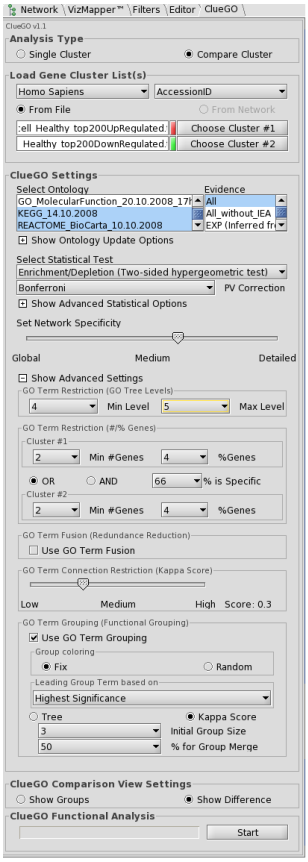


Figure 21: *ClueGO settings example*

Customize the network using Cytoscape features

- select VizMapper (Cytoscape Control Panel)
- select Node Font Size
- set the value of FALSE (size for the name of the terms) to 0.001 and press Enter

- set the value of TRUE (size for the name of groups) to 26 and press Enter
- select Layout (Cytoscape menu bar)
- set scale to 1/3
- zoom the image
- change the position of the leading terms to make visible the name of the group
- change the possition of notgrouped terms if due to rescaling they are too close to a group
- for visualizing groups press Show Groups

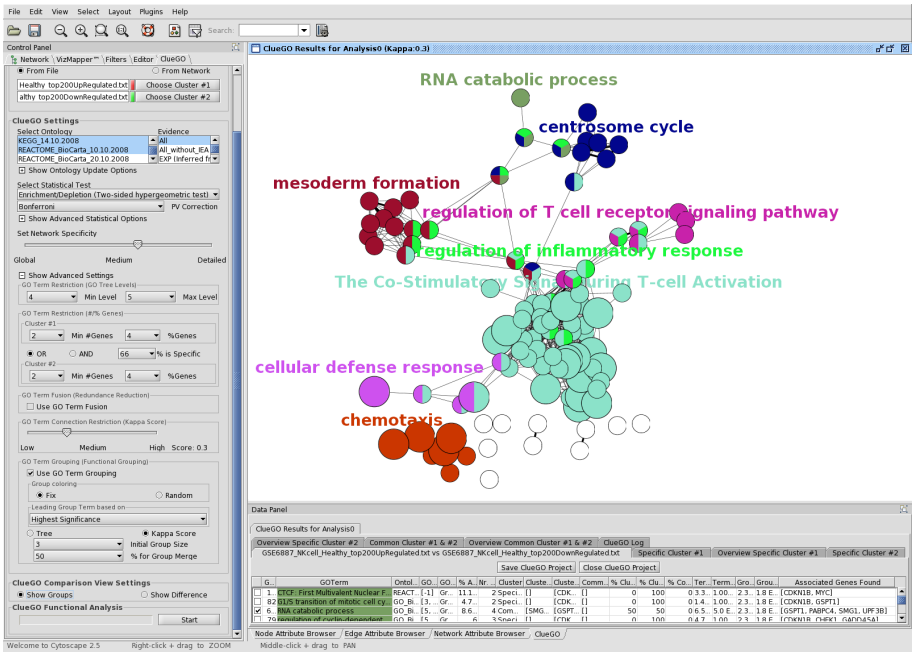


Figure 22: *ClueGO customized figure*

Bibliography

[1] M Ashburner, C A Ball, J A Blake, D Botstein, H Butler, J M Cherry, A P Davis, K Dolinski, S S Dwight, J T Eppig, M A Harris, D P Hill, L Issel-Tarver, A Kasarskis, S Lewis, J C Matese, J E Richardson, M Ringwald, G M Rubin, G Sherlock. Gene ontology: tool for the unification of biology The Gene Ontology Consortium. *Nat Genet*, 25:25–29, 2000.

[2] M Kanehisa, S Goto, S Kawashima, A Nakaya. The KEGG databases at GenomeNet. *Nucleic Acids Res*, 30:42–46, 2002.

[3] P Shannon, A Markiel, O Ozier, N S Baliga, J T Wang, D Ramage, N Amin, B Schwikowski, T Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, 13:2498–2504, 2003.

[4] da W Huang, B T Sherman, Q Tan, J R Collins, W G Alvord, J Roayaei, R Stephens, M W Baseler, H C Lane, R A Lempicki. The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol*, 8:R183–R183, 2007.

[5] I Rivals, L Personnaz, L Taing, MC Potier. Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics*, 23:401–7, 2007.

[6] O Garcia, C Saveanu, M Cline, M Fromont-Racine, A Jacquier, B Schwikowski, T Aittokallio. GOlorize: a Cytoscape plug-in for network visualization with Gene Ontology-based layout and coloring. *Bioinformatics*, 23:394–396, 2007.

[7] J Cohen. A coefficient of agreement for nominal scales. *Educ Psychol Meas*, 20:37–46, 1960.