

Blast2GO for CLC User Manual

Version 6

March, 2017



BioBam Bioinformatics S.L.

Blast2GO for CLC Documentation

The Blast2GO PRO Plugin makes the most popular Blast2GO features directly available from within the workbench (Main and Genomics). It allows you to extend and integrate your bioinformatics data analysis with cutting-edge functional genomics tools. All features are seamlessly integrated and provide a professional solution for the functional analysis of novel sequence datasets. Genome-wide functional annotation and interpretation can be executed as workflows directly on your existing datasets. The plugin allows to speed-up your sequence alignments with CloudBlast to find homologous sequences, extract GO terms and create high-quality Gene Ontology functional annotations. Additional classifications can be performed through functional enrichment analysis, Gene Ontology summaries and rich graph visualizations. Download the plugin to get a free trial.

Table of Contents

- [Plugin Introduction](#)
- [Quick Start](#)
- [User Interface](#)
- [Main Analysis Options](#)
- [Statistics](#)
- [Gene Ontology Graph Visualization](#)
- [Data Import and Export Options](#)
- [Fisher's Exact Test \(Enrichment Analysis\)](#)
- [Manage Projects](#)
- [Miscellaneous](#)
- [Workflows](#)
- [Orthologous Groups \(NOG/COG\)](#)
- [PSORTb](#)
- [Rfam](#)



Plugin Introduction

Content of this page:

- [Main Plugin Features](#)
- [Developed by](#)
- [System Requirements](#)
- [Installation](#)
- [Support](#)

The Blast2GO PRO Plugin makes the most popular Blast2GO features directly available from within the workbench (Main and Genomics). It allows you to extend and integrate your bioinformatics data analysis with cutting-edge functional genomics tools. All features are seamlessly integrated and provide a professional solution for the functional analysis of novel sequence datasets. Genome-wide functional annotation and interpretation can be executed as workflows directly on your existing datasets. The plugin allows to speed-up your sequence alignments with CloudBlast to find homologous sequences, extract GO terms and create high-quality Gene Ontology functional annotations. Additional classifications can be performed through functional enrichment analysis, Gene Ontology summaries and rich graph visualizations. Download the plugin to get a free trial.

Main Plugin Features

- Access to the High-Performance CloudBlast for fast NCBI Blast+
- Gene Ontology Mapping of Blast Results
- Best-Practice Functional Annotation
- InterProScan Domain Searches
- Functional Enrichment Analysis
- Rich Graph Visualization and GO-Slim Reduction
- Clusters of Orthologous Groups (COG/NOG)
- Subcellular localization with PSORTb
- Identification of ncRNA with Rfam
- Many Statistics Charts
- Multiple Data Import and Export Options

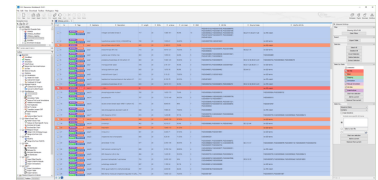


Figure 1: Main Sequence Table of the Blast2GO PRO Plugin

Developed by

The Blast2GO PRO Plugin is developed and maintained by BioBam Bioinformatics. BioBam, founded in 2011 and located in Valencia, Spain is dedicated to creating user-friendly software for the scientific community. With Blast2GO it provides an all-in-one solution for functional genomics (functional annotation and analysis of genomic datasets) especially popular in non-model organism research. Blast2GO counts with over 3000 scientific citations and is used by top private and public research institutions worldwide. BioBam is collaborating with leading world-class bioinformatics companies like e.g. in this case with CLC bio/Qiagen as well as a growing number of distribution partners around the world. For more information about BioBam please visit <https://www.biobam.com>.

System Requirements

The plugin is available for CLC bio Main-, Genomics- and Biomedical Workbench. In general there are no plugin-specific requirements other than the ones of the CLC bio Workbench itself. Since several features depend on online resources we strongly recommend a working network/internet connection when working with Blast2GO.

Installation

The plugin can be installed via the **Plugin Manager** from within the CLC bio Workbench. From the **Manage Plugin** tab choose the **Blast2GO PRO** plugin and click on download and Install. Once installed the workbench will automatically inform about new updates. Regarding licensing options and quotes please contact the CLC bio sales team: sales@clcbio.com

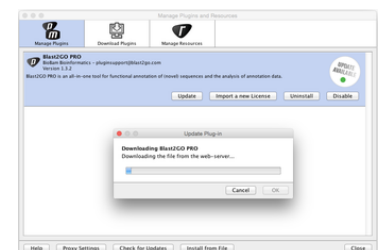


Figure 2: Download and update the plugin

Support

Our support and development team works close together to take care of all the trouble you might have. It does not matter if the issue is rather technical or about bioinformatics or biology. You can reach us at pluginsupport@blast2go.com and we will do our best to answer all requests within one day. All comments and suggestions regarding our services are most welcome and a valuable source of information for us.

Support: pluginsupport@blast2go.com

Website: <https://www.blast2go.com>

This section provides a quick run-through of a basic functional annotation process done with Blast2GO. More detailed descriptions of the different analysis steps and more advanced features are described in the remaining sections of this documentation.

- Important:**

- **annot-file:** The annot file is the standard format to export GO annotations. It is a tab-separated text file, each row contains one GO term.
- **b2g-file:** The standard Blast2GO project file. This file can also be opened with the standalone Blast2GO application.
- **Sequence Table:** A tab-separated text file containing all the information given in the Blast2GO sequence table, available from the sequence editor's sidepanel.
- **GAF 2.0:** A tab-separated text file of the functional information in the Gene Ontology annotation file format. The content of this format can also be viewed within the workbench via the **Annotation Table** function from the toolbox.

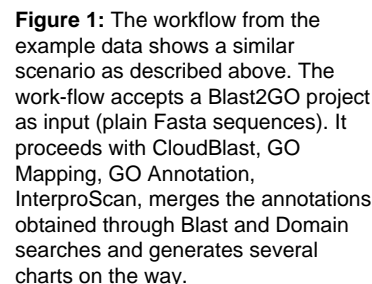


Figure 1: The workflow from the example data shows a similar scenario as described above. The work-flow accepts a Blast2GO project as input (plain Fasta sequences). It proceeds with CloudBlast, GO Mapping, GO Annotation, InterProScan, merges the annotations obtained through Blast and Domain searches and generates several charts on the way.

User Interface

Blast2GO functions are seamlessly integrated within the workbench. All major functions can be accessed via the Blast2GO section in the Toolbox.

Content of this page:

- [Sequence Table](#)
- [Sequence Table Context Menu](#)
- [Filter out Sequences](#)
- [Hide Columns](#)
- [Side Panel](#)

Sequence Table

The sequence table is the main view of a Blast2GO Project, where each row represents a sequence. Colors indicate the status of each sequence, starting from white (without analysis results), to orange, green, blue, etc. (see figure below). The sequence table also offers a context menu with several additional options for each individual sequence, like (blast result details, sequence information, manual annotation modifications, etc.).

Without Analysis	With Blast Hits	With GO Mapping	Manually Annotated
Blasted without Hits	With InterProScan but not Blasted	B2G Annotated	GOSlim

Figure 1: Different color codes indicating the status of the sequences

Sequence Table Context Menu

- **Show Sequence:** Allows to see the sequence information.
- **Show Blast Result:** Revise the Blast results in detail: hits, species, identifiers, etc.
- **Show InterProScan Result:** Revise the InterProScan results in detail: databases, GOs, etc.
- **Show Mapping Result:** Revise the retrieved candidate GO terms in detail: Accessions, Evidence Codes, Databases.
- **Show GO Description:** Review information about annotated GOs.
- **Change Annotation and Description:** Allows to manually change/add annotations.
- **Extract Selection to new Tab:** Create a new project from the marked sequences (Shift or Ctrl).
- **Copy Selection to Clipboard (tabular format):** Copies the marked sequence to the clipboard in tabular format for further processing in a spreadsheet editor.
- **Copy Content of Column: SeqName to Clipboard:** The content of a specific column will be copied to the clipboard.
- **Create ID List of Column: Sequence:** Allows to create an ID list of a specific column which can then be used for Fisher's Exact Test or Selection of sequences.
- **Create ID Value-List of SeqName and Length:** Allows to create a list with two specific columns e.g. SeqName and Length.
- **Create Category Chart of Column: Length:** Create a category chart of a certain column e.g. sequence length.
- **Create Distribution Chart of Column: InterProScan IDs:** Create a distribution chart of a certain column.

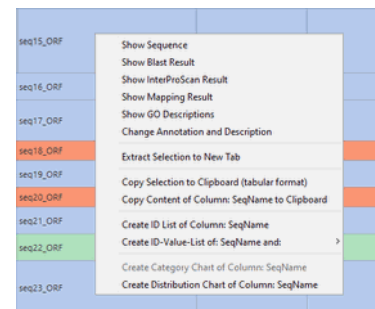


Figure 2: Sequence Context Menu

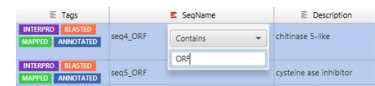


Figure 3: Filter Criteria

Filter out Sequences

The Blast2GO table allows to filter out rows, depending on different search criteria for each column. Each column header show a small icon which opens a context menu when left-clicked.

- Filters can be applied in various columns and are joined via an AND condition.
- Different data-types allow different filter settings (e.g. numbers allow greater than).
- When a filter is applied on a column the filter icon turns red - double clicking the icon will remove the filter.
- The side panel on the right-hand side shows how many rows are currently visible.
- The button below it, **Clear Filters**, removes all current filters and show all rows.

All the algorithms, blast, mapping, annotation, etc., work on the selected sequences and not only on the filtered ones. This means if one has a filter but there are some sequences selected on your project and one runs e.g. remove blast results, it will work on all selected sequences and not only on the ones you see on the table.

Hide Columns

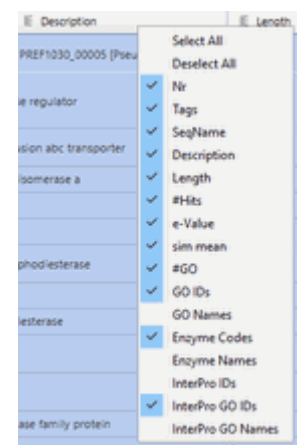


Figure 4: Hide Columns

This feature allows to hide the columns of the sequence table.

By right clicking on a column and a menu will be displayed and one can select those columns to hide from the table. In combination with **Export Table** from the side panel, this can be used to customize the output.

Side Panel

The Sequence table offers a side panel which allows to select, filter and search sequences within a Blast2GO project. Filter: Blast2GO table allows to filter sequence out. Here one can see the number of sequences that are on the table and also to reset the filter options.

- Selection: Allows to select, deselect, invert, delete and extract a given selection.
- Select by State: Allows to make a selection based on the sequence status (colors).
- Select by: Allows to select sequences based on their name, description, species, function (GO terms or IDs), description, gene name, enzyme code or InterPro ID. The selection-type, exact search (whole word (important for IDs) has to match) and case sensitivity can be chosen. A search criteria can be provided via a search field. Alternatively a list of sequence names or GO functions can be loaded via a text file or an **ID List**. The 3 buttons let you choose between starting a new selection, adding rows to the current selection and removing rows from the current selection.

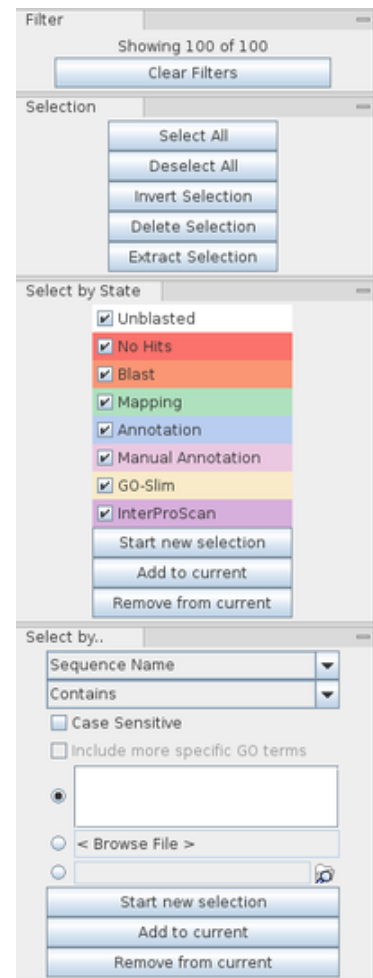


Figure 5: Side-Panel

Main Analysis Options

Content of this page:

- [CloudBlast](#)
- [Mapping](#)
- [Annotation](#)
- [InterProScan](#)
- [GO-Slim](#)

CloudBlast

CloudBlast is a cloud-based Blast2GO PRO Community Resource for massive sequence alignment tasks. It allows you to execute standard NCBI Blast+ searches directly from within Blast2GO PRO in a dedicated computing cloud. CloudBlast is a high-performance, secure and cost-optimized solution for your analysis. This is a Blast service totally independent from the NCBI servers to provide fast and reliable sequence alignments. CloudBlast offers a highly optimized, self-sustained HPC solution to address a very specific need of the Blast2GO PRO community. It consists of a high performance computing cluster dedicated exclusively to Blast searches. To make use of this resource, so-called ComputationUnits are required. CloudBlast allows you to perform standard Blast searches for tens of thousands of sequences within a few days against a large collection of protein databases. Each sequence alignment performed in the system consumes a certain amount of computation time (translated into Computation Units) depending on the sequence length, the used blast algorithm (blastx, blastp) and parameters used. The smaller the database you blast against the more sequences you can analyze with a certain amount of Computation Units.

To get an idea of the consumption of these units here is an example: With 3.000.000 Computation Units you would be able to blastx close to 500.000 sequences against the vertebrate NR-subset. A Blast search against the entire NR database, the largest protein database available, should allow you to process approx. 35.000 sequences (with an average length of 800nt per sequence). The difference in consumption can be explained due to the different size of the protein target databases - which result in a significant reduction of the amount of computation time required for a given amount of sequences - which speeds-up your overall analysis. These numbers are only orientative and change over size due to the increase of sequences available in public database collections.

Mapping

Mapping is the process of retrieving GO terms associated to the hits obtained after a BLAST search. To run mapping, select one or various data-sets, which contain blasted sequences and execute the mapping function. When a BLAST result is successfully mapped to one or several GO terms, these will come up at the GOs column of the Main Sequence Table. Assigned GOs to hits can be reviewed in the BLAST Results Browser. Successfully mapped sequences will turn green.

Blast2GO performs different mapping steps to link all BLAST hits to the functional information stored in the Gene Ontology database. Therefore Blast2GO uses different public resources provided by the NCBI, PIR and GO to link the different protein IDs (names, symbols, GIs, UniProts, etc.) to the information stored in the Gene Ontology database - the GO database contains several million functionally annotated gene products for hundreds of different species. All annotations are associated to an Evidence Code which provides information about the quality of this functional assignment.

1. BLAST result accessions are used to retrieve gene names or Symbols making use of two mapping files provided by NCBI. Identified gene names are then searched in the species specific entries of the GO database.
2. BLAST result GI identifiers are used to retrieve UniProt IDs making use of a mapping file from PIR including PSD, UniProt, Swiss-Prot, TrEMBL, RefSeq, GenPept and PDB.
3. BLAST result accessions are searched directly in the GO database.

Important:

The mapping only works if the sequences have been blasted with an adequate blast program. Be sure to run blastx or blastp, since you need to get protein IDs. This is because GOs are assigned to proteins only.

Annotation

This is the process of selecting GO terms from the GO pool obtained by the Mapping step and assigning them to the query sequences. GO annotation is carried out by applying an annotation rule (AR) on the found ontology terms. The rule seeks to find the most specific annotations with a certain level of reliability. This process is adjustable in specificity and stringency. For each candidate GO an annotation score (AS) is computed. The AS is composed of two additive terms. The first, direct term (DT), represents the highest hit similarity of this GO weighted by a factor corresponding to its EC.

The second term (AT) of the AS provides the possibility of abstraction. This is defined as annotation to a parent node when several child nodes are present in the GO candidate collection. This term multiplies the number of total GOs unified at the node by a user defined GO weight factor that controls the possibility and strength of abstraction. When GO weight is set to 0, no abstraction is done. Finally, the AR selects the lowest term per branch that lies over a user defined threshold. DT, AT and the AR terms are defined as given in the following figure.

$$DT = \max(similarity \times EC_{weight})$$

$$AT = (\#GO - 1) \times GO_{weight}$$

$$AR: lowest.node(AS(DT + AT)) \geq threshold$$

Figure 1: Annotation Rule

To better understand how the annotation score works, the following reasoning can be done:

When EC-weight is set to 1 for all ECs (no EC influence) and GO-weight equals zero (no abstraction), then the annotation score equals the maximum similarity value of the hits that have that GO term and the sequence will be annotated with that GO term if that score is above the given threshold provided. The situation when EC-weights are lower than 1 means that higher similarities are required to reach the threshold. If the GO-weight is different to 0 this means that the possibility is enabled that a parent node will reach the threshold while its various children nodes would not.

The annotation rule provides a general framework for annotation. The actual way annotation occurs depends on how the different parameters at the AS are set.

1. E-Value Hit Filter. This value can be understood as a pre-filter: only GO terms obtained from hits with a greater e-value than given will be used for annotation and/or shown in a generated graph (default=1.0E-6).
2. Annotation Cut-Off (threshold). The annotation rule selects the lowest term per branch that lies over this threshold (default=55).
3. GO-Weight. This is the weight given to the contribution of mapped children terms to the annotation of a parent term (default=5).
4. Hsp-Hit Coverage CutOff. Sets the minimum needed coverage between a Hit and his HSP. For example a value of 80 would mean that the aligned HSP must cover at least 80% of the longitude of its Hit. Only annotations from Hit fulfilling this criterion will be considered for annotation transference.
5. EC-Weight. Note that in case influence by evidence codes is not wanted, you can set them all at 1. Alternatively, when you want to exclude GO annotations of a certain EC (for example IEAs), you can set this EC weight at 0.

A detailed explanation of the GO-Evidence-Codes can be found here: <http://www.geneontology.org/GO.evidence.shtml>.

Successful annotation for each query sequence will result in a color change for that sequence from light-green to blue at the Main Sequence Table, and only the annotated GOs will remain in the GO IDs column. An overview of the extent and intensity of the annotation can be obtained from the Annotation Distribution Chart, which shows the number of sequences annotated at different amounts of GO-terms.

InterProScan

The functionality of InterPro annotations in Blast2GO allows to retrieve domain/motif information in a sequence-wise manner. The processed sequence have to contain a valid sequence string. This is not the case when your Blast2GO project has been created via blast result import. Many InterProScan families are directly related to certain biological functions and linked to the corresponding Gene Ontology terms.

Functional information obtained via the algorithm that form part of the InterProScan family can in a subsequent step be added to the information already available for your sequence data. To merge domain based GO terms to the once obtained via the blast based annotation step the "Merge InterProScan Results" function has to be called. If this step is omitted the GO terms obtained via the InterPro are not added and combined with the already existing annotations. Result details can be viewed through the Single Sequence Menu.

Important:

You have to provide a valid email address to be able to run the InterProScan at EBI.

GO-Slim

What is a GO Slim?

(Ref: Gene Ontology website, <http://geneontology.org/page/go-slim-and-subset-guide>)

GO slims are cut-down versions of the Gene Ontology containing a subset of the terms in the whole GO. They give a broad overview of the ontology content without the detail of the specific fine grained terms. GO slims are particularly useful for giving a summary of the results of GO annotation of a genome, microarray, or cDNA collection when broad classification of gene product function is required. GO slims are created by users according to their needs, and may be specific to species or to particular areas of the ontologies. GO provides a generic GO slim which, like the GO itself, is not species specific, and which should be suitable for most purposes. Alternatively, users can create their own GO slims or use one of the model organismspecific slims integrated into the GO flat file. Please email the GO helpdesk for more information about creating and submitting your GO slim.

To get a better understanding of what GO Slim does in practice and how it works, here (Figure 2) is a small visual example.

Imagine figure 2a to be the subset of GO terms called GO Slim, figure 2b shows a data-set with GO 6,9 and 10 annotated. The GO Slim methodology will pull up the 3 annotated GOs as follows:

- 6 > 1
- 9 > 4
- 10 > 5

The result is shown in figure 2c. Keep in mind that this would be a data-set containing various sequences, because one sequence that has annotated GO 1 and 4 would remain only with GO 4 because of the true-path rule.

In the application our GO Slim subset is represented by a file with the extension .obo, this file contains all GO nodes and their hierarchical structure. The Gene Ontology Consortium provides various GO Slims that can be used and accessed directly from within the application. To select a predefined GO Slim, select Obo file from GO-Website and select your preferred file, it will then be used in combination with the currently selected obo file under **Edit > Preferences > General > Blast2GO Data Access Settings > Change Settings > Obo File** selection at the bottom. The latter file contains the whole set of Gene Ontology terms.

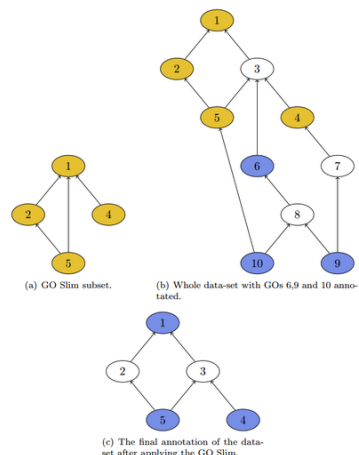


Figure 2: This shows an example of GO Slim in practice, each node represents one GO. White stands for normal, yellow for GO Slim and blue for directly annotated.

If the user wants to experiment and to try something separate, he can go for Custom obo files and select the two obo files by hand. Keep in mind that the GO Slim file has to contain a real subset of GOs, otherwise the result is undefined.

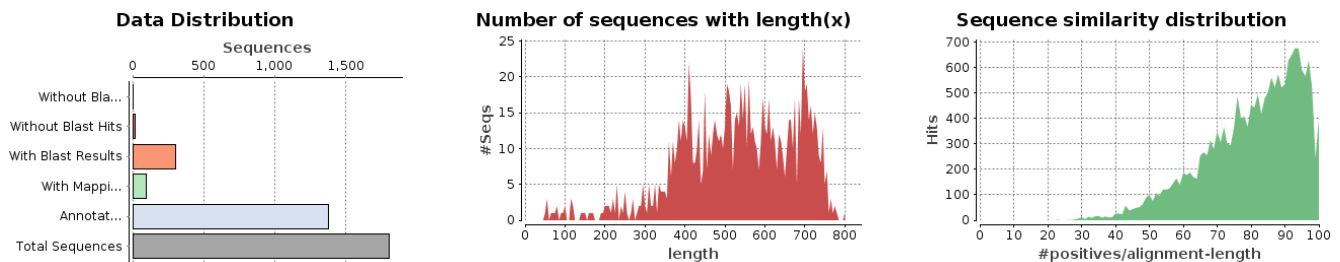
Statistics

The Statistics wizard allows to select and generate all available charts in one run. Statistical charts are available to provide direct feedback about data composition. Charts such as mean sequence length, involved species distribution, BLAST e-value distribution or the standard deviation of GO level annotation distribution, allow the visualisation of intermediate and final result summaries. These charts are especially helpful to validate the results of each analysis step and to re-adjust or determine the parameters of subsequent processing. In this interactive manner the annotation process can be adjusted to specific data-set and user requirements.

List of all available statistical charts in Blast2GO

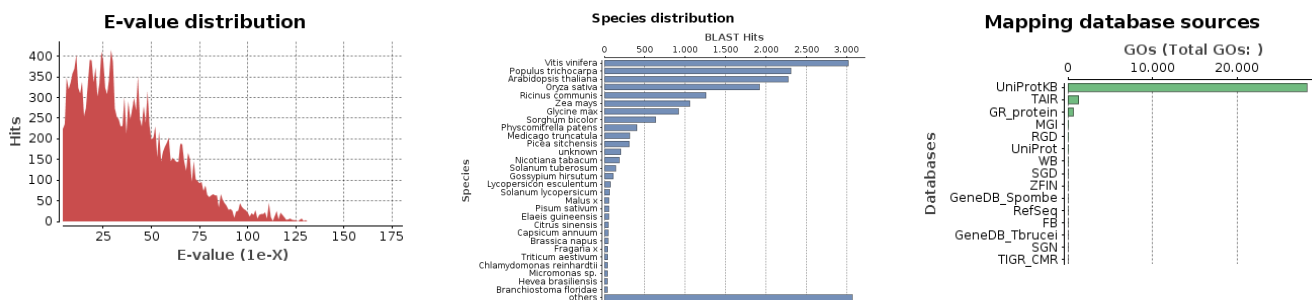
Project

- **Analysis Progress:**
Gives an overview about the current analysis progress of the data-set.
- **Data Distribution:**
This bar chart shows the distribution of un-blasted, blasted, mapped and annotated sequences over the whole data-set.
- **Data Distribution (pie):**
The same as the former but pie-style.
- **Sequence Length:**
Plots the sequence length for all sequences.



Blast

- **E-value distribution:**
This chart plots the distribution of E-values for all selected BLAST hits. It is useful to evaluate the success of the alignment for a given sequence database and help to adjust the Evaluate cutoff in the annotation step.
- **Hit Distribution:**
Shows the distribution of hits for each sequence (Blast Result).
- **Hsp Distribution:**
This bar chart shows the distribution of hsps per hit.
- **Hsp/Seq Distribution:**
This chart shows a distribution of percentages which represents the coverage between the hsps and their corresponding sequences.
- **Hsp/Hit Distribution:**
Same as above but for hits instead of sequences.
- **Sequence similarity distribution:**
This chart displays the distribution of all calculated sequence similarities (percentages), shows the overall performance of the alignments and helps to adjust the annotation score in the annotation step.
- **Species distribution:**
This chart gives a listing of the different species to which most sequences were aligned during the BLAST step.
- **Top-Blast Species distribution:**
This chart gives the species distribution of the Top-BLAST hits.



Mapping

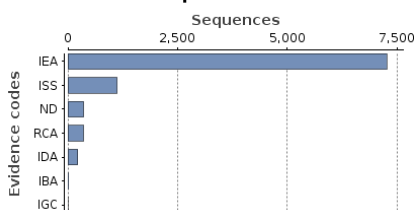
- **GO Mapping Distribution:**
Shows the distribution of the amount of Gene Ontology candidate terms assigned to each sequences during the GO Mapping step.

- **DB-source of mapping:**
This chart gives the distribution of the number of annotations (GO-terms) retrieved from the different source databases like e.g. UniProt, PDB, TAIR etc.
- **EC Distribution for Blast Hits:**
Same as above but per Blast hit.
- **Evidence Code distribution:**
This chart shows the distribution of GO evidence codes for the functional terms obtained during the mapping step. It gives an idea about how many annotations derive from automatic/ computational annotations or manually curated ones.

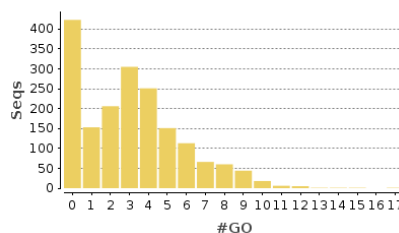
Annotation

- **Annotation distribution:**
This chart informs about the number of GO terms assigned per sequence.
- **Annotation Score distribution:**
A chart that shows the number of sequences per annotation score.
- **GO Annotation Level distribution:**
A bar chart which shows all GO terms for all 3 categories for a given GO level taking into account the GO hierarchy (parent-child relationships).
- **GO Distribution Level:**
A bar chart which shows all GO terms for all 3 categories for GO level 2, taking into account the GO hierarchy.
- **Direct GO Count MF:**
A chart for the Molecular Function GO category, which shows the most frequent GO terms within a data-set without taking into account the GO hierarchy.
- **Direct GO Count BP:**
Same as above but for Biological Process.
- **Direct GO Count CC:**
Same as above but for Cellular Component.
- **Number of GOs/Seq-Length:**
Shows the relation between sequence length and number of GOs.
- **Annotated Seqs/Seq-Length:**
Shows the relation between amount of annotated sequences and sequence lengths.

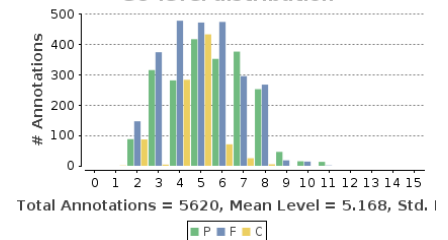
Evidence code distribution for sequences



Annotation distribution



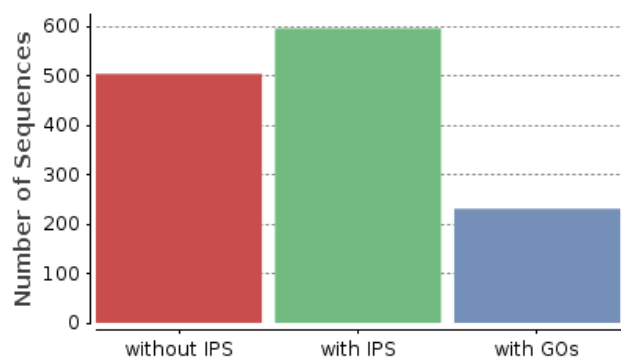
GO-level distribution



InterProScan

- **InterProScan Results:**
This chart shows the effect of adding the GO-terms retrieved through the InterProScan results.
- **InterProScan Families Distribution:**
Distribution of IPS results by families.
- **InterProScan Domains Distribution:**
Distribution of IPS results by domains.
- **InterProScan Repeats Distribution:**
Distribution of IPS results by repeats.
- **InterProScan Sites Distribution:**
Distribution of IPS results by sites.
- **InterProScan IDs Distribution:**
Shows the number of results per IPS ID.
- **InterProScan IDs by Database:**
Shows the number of results per IPS ID per database.

InterProScan results

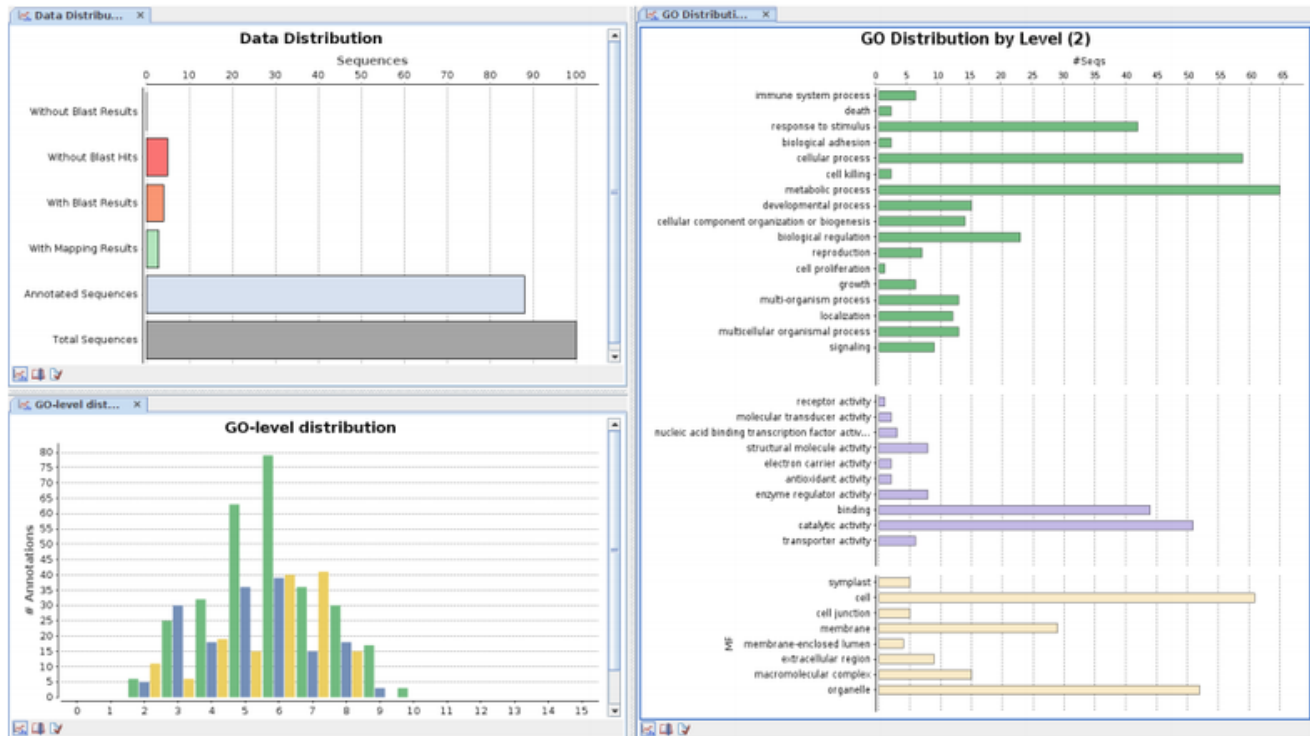


Enzyme

- **Main Enzyme Classes:**
Shows the distribution of the 6 main enzyme classes over all sequences.
- **Second Level Classes:**
Same as above but for the corresponding subclass.

Annex

- This chart shows the performance of the Annex annotation augmentation step. It shows the number of GO terms which were confirmed, replaced or removed through this method.



Gene Ontology Graph Visualization

Content of this page:

- [Node Score](#)
- [Graph Term Filtering](#)
- [Combined Graph Editor Sidepanel](#)
 - [General](#)
 - [Find](#)
 - [Node Info](#)
 - [Layout](#)
 - [Options](#)
 - [Charts](#)
 - [Graph Legend](#)

Visualization is a helpful component in the process of interpreting results from high-throughput experiments, and can be indispensable when working with large data-sets. Within the GO, the "natural" visualization format is the Direct Acyclic Graph of a group of annotated sequences. In the DAG, each node represents a GO term. Arcs represent the relationships between the biological concepts. A problem when visualizing GO functional information of genomic data-sets is that these graphs can become extremely large and difficult to navigate when the number of represented sequences is high.

One of the functions of Blast2GO is the ability to display the annotation result of one or several sequences in the same GO graph. Within Blast2GO these graphs are called "Combined Graphs". This function generates joined GO DAGs to create overviews of the functional context of groups of annotations and sequences. Combined Graph nodes are highlighted through a color scale proportional to their number of sequences annotated to a given term. This confluence score (from now on denoted "node-Score") takes into account the number of sequences converging at one GO term and at the same time penalizes by the distance to the term where each sequence was actually annotated. Assigned sequences and scores can be displayed at the terms level.

Node Score

The node score is calculated for each GO term in the DAG and takes into account the topology of the ontology and the number of sequences belonging (i.e. annotated) to a given node (i.e. GO term). The score is the sum of sequences directly or indirectly associated to a given GO term weighted by the distance of the term to the term of "direct annotation" i.e. the GO term the sequence is originally annotated to. This weighting is achieved by multiplying the sequence number by a factor alpha [0, infinity] to the power of the distance between the term and the term of direct annotation (equation below). In this way, the node score is accumulative and the information of lower-level GO-terms is considered, but the influence of more distant information (i.e. annotations) is suppressed/decreased depending on the value of alpha. This compensates for the drawback of the earlier described method of simply counting the number of different sequences assigned to each GO-term. The alpha parameter allows this behavior to be further adjusted. A value of zero means no propagation of information and can be increased by raising alpha.

A Score is computed at each node according to the formula on the right-hand side, where seq is the number of different sequences annotated at a child GO term and dist the distance to the node of the child. GO term Coloring by Score will highlight areas of high annotation density.

$$score = \sum_{GOs} seq \times \alpha^{dist}$$

Graph Term Filtering

Combined graphs can become extremely large and difficult to navigate when the number of visualized sequences is high. Additionally, the relevant information in these cases is frequently concentrated in a relatively small subset of terms. We have introduced graph-pruning functions to simplify DAG structures to display only the most relevant information. In the case of the Combined Graph function, a cutoff on the number of sequences or the node-score value can be set to filter out GO terms. In this case the size of a graph is reduced without losing the important information (i.e. hiding tip and intermediate low informative nodes).

This approach of graph-filtering and trimming is based on a combination of different scoring schemes. On the one hand, graph filtering can be based on the number of sequences assigned to each node, and on the other hand, a graph can be "thinned out" by removing intermediate nodes that are below a given cutoff. The latter approach allows a certain level of details to be maintained while drastically reducing the size of the graph by removing "unimportant" intermediate graph elements. In this way, any large GO graph can be reduced by abundance and information content instead of simply "cutting through" the Gene Ontology at a certain hierarchical level or by the use of GO-Slim definitions.

Below the molecular functions of 1000 sequences are visualized in 3 different ways. The first graph is unfiltered, the second graph shows the functional information after having applied a Go-Slim reduction. The third graph is filtered and thinned according to the number of sequences belonging to each GO-term and the node-score. All GO terms with less than 10 sequences were removed (tip nodes) and all the nodes with a node-score smaller than 12, applying an alpha of 0.4, were removed (intermediate nodes). This strategy allows the removal of terms that are less significant to a particular data-set while at the same time it maintains frequently present terms at lower levels of specificity.

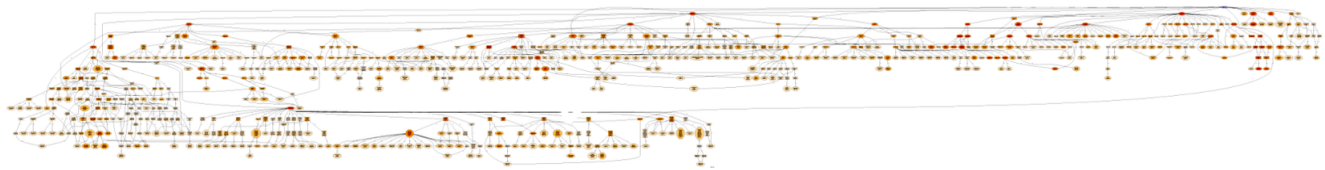


Figure 1: Unfiltered Graph

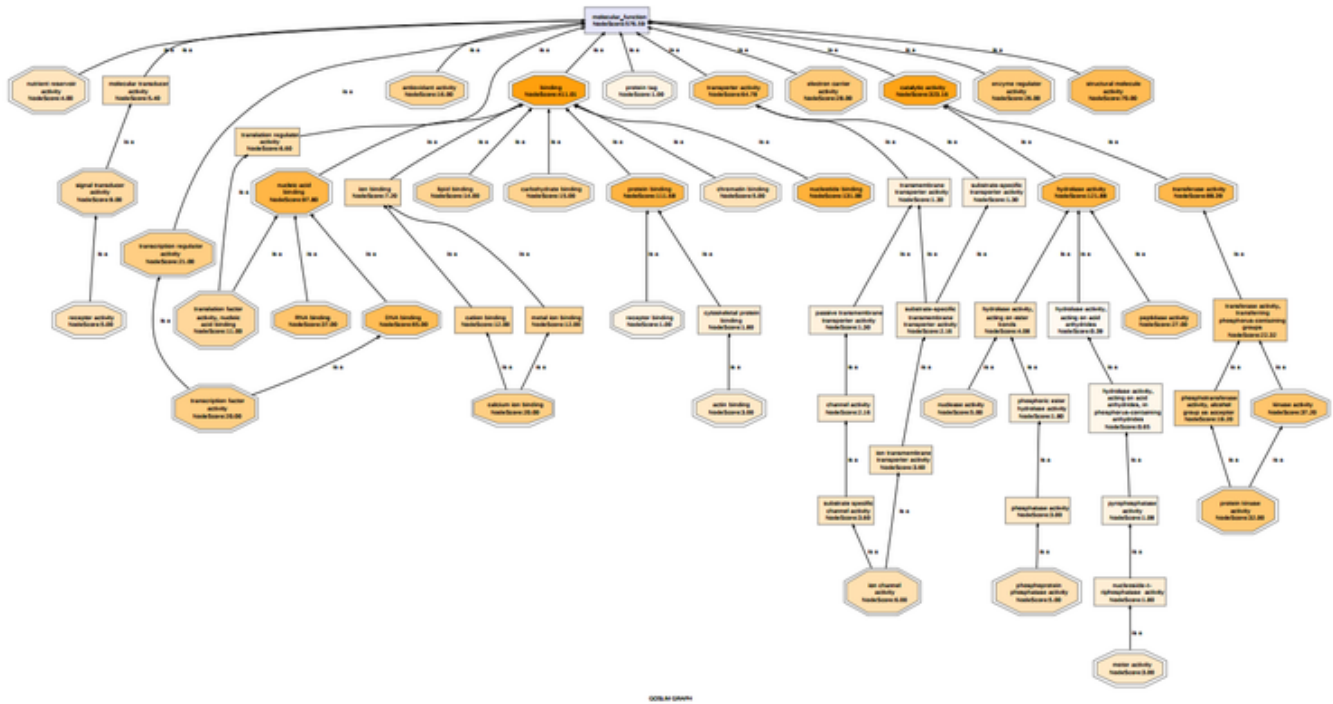


Figure 2: Filtered Graph 1

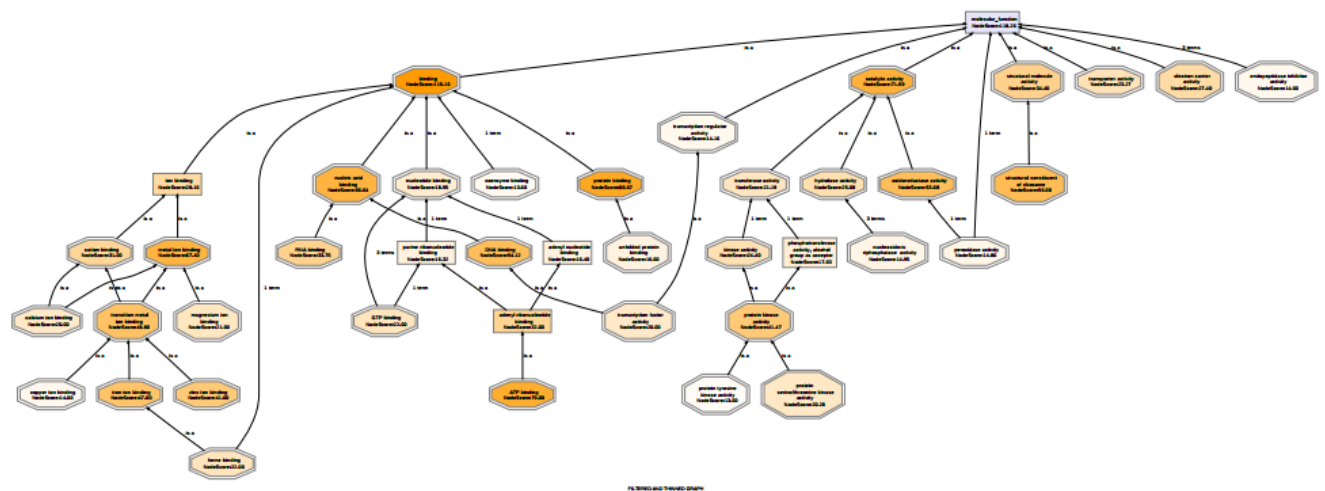


Figure 3: Filtered Graph 2

Combined Graph Editor Sidepanel

General

This section controls the graph visualization within its area.

- Zoom
- Collapse All: The nodes will collapse and only the root will be visible.
- Expand All: The graph will expand to its entire size.
- Re-Layout: The whole graph will be re-scaled to adjust to the visualization area.

Find

Allows to search for GO IDs/ Terms/ Description in the Graph.

Node Info

Adjust the visible information in each node.

- GO ID: If checked the GO ID will be included in the node.
- GO Name: The GO Names are shown in the node.
- GO Description: The GO Description will be included in the node.
- Node score: The node score will be shown in the node.
- Sequence Names: The names of the sequences annotated with this GO are shown (max. 15).
- Sequences: The number of sequences annotated with that particular GO will be displayed in the node.

Layout





- Edge Labels: When checked, the labels at the edges will be shown.
- Expand/Collapse Icon: If checked the icons that represent expand/collapse on the node are displayed.
- Only **is a** Relations: Only the **is a** relations between nodes will be displayed if the box is checked.
- Color
 - Ontology: All nodes will be colored according to the ontology category, Biological Process - green; Molecular Function - blue; Cellular Component - yellow.
 - White: The nodes will turn white.
 - By Node score: The higher the node score, the more intense is the color (orange).
 - By Sequence Count: Node color intensity will be proportional to the number of contributing sequences at the node.

Options

- Sequence Filter: The minimal number of sequences, a GO node must have assigned, to be displayed. This filter is used to control the number of nodes present in the graph. Depending on the result, adjust this value until you obtain a satisfactory graph. Start with 10% of your total number of annotated sequences.
- Node score Filter: The minimal node score that is necessary for a node to be displayed.
- Score alpha. The value for parameter alpha in the Score formula Node Score Filter. Only nodes with a Score value higher than the Filter will be shown. Use this parameter to thin out the GO-DAG for low informative nodes.
- Restore Defaults: All filters will be set to the default values.

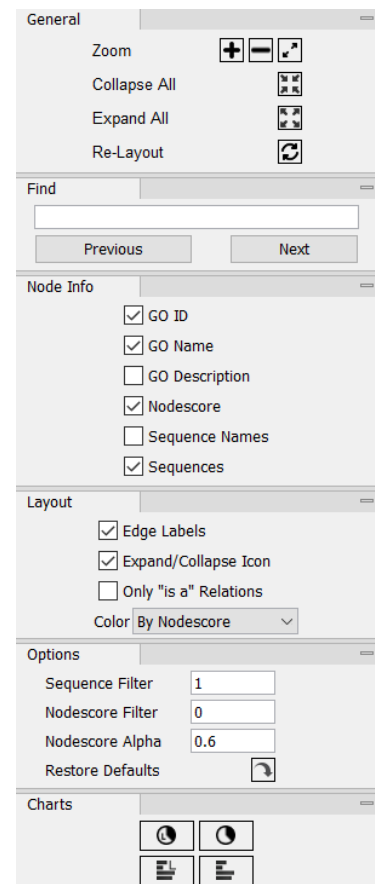
Charts

Analysis of GO term associations in a set of sequences can also be done with pie/bar charts. Once the graph is visible, the Charts area allows the creation of 4 different charts.

-  Cuts through the graph at a specific level and generates a pie representation of the number of sequences per GO node.
-  Allows to select a minimum filter value in order to include only GO nodes with a higher Node-Score or sequence count in the resulting pie chart.
-  Same as the first one but in bar chart style.
-  Will show a bar chart with the number of sequences that have been annotated with a specific GO Term.

Graph Legend

The GO Graphs are displayed in different shapes (Figure 5).



The screenshot displays the Blast2GO software interface with the following panels:

- General:** Contains controls for Zoom (plus, minus, reset), Collapse All, Expand All, and Re-Layout (refresh).
- Find:** Includes a search input field and Previous/Next navigation buttons.
- Node Info:** A list of checkboxes to toggle information shown in nodes: GO ID (checked), GO Name (checked), GO Description (unchecked), Nodescore (checked), Sequence Names (unchecked), and Sequences (checked).
- Layout:** Contains checkboxes for Edge Labels (checked), Expand/Collapse Icon (checked), and Only "is a" Relations (unchecked). It also features a Color dropdown menu currently set to "By Nodescore".
- Options:** Includes input fields for Sequence Filter (1), Nodescore Filter (0), and Nodescore Alpha (0.6), along with a Restore Defaults button.
- Charts:** Displays four icons representing different chart types: a pie chart, a bar chart, a pie chart with a legend, and a bar chart with a legend.

- Octagon - Annotated GO Terms
- Square - Intermediate GO Terms
- Ellipsis - GO Terms linked to a Blast Hit

Figure 4: Combined Graph Sidepanel**Figure 5:** Graph Legend that shows the graph shapes

Data Import and Export Options

The import and export options are distributed between standard im/export available from the main CLC Genomics Workbench toolbar and the Blast2GO toolbox. The im/exports available from the Blast2GO toolbox provide additional functionality which could not be provided using the standard import and are therefore separated from the rest.

- Standard Import: The import via **Import > Standard Import ...** allows to import **.annot**, **.dat** or **.b2g** files.
- Toolbox Import: **Blast** and **InterProScanxml**, as well as **.annot** files can be imported from here. To create an entirely new project from a file, simply skip the first wizard step (Select Blast2GO Project). Can be found via **Toolbox > Blast2GO > Import**.
- Standard Export: Blast2GO Projects can be exported as **.b2g**, **.annot**, **GAF**, **GFF**, **GeneSpring**, **GoStat**, **.dat** and in **WEGO** format. You should always prefer **.b2g** over **.dat**. Most Blast2GO datatypes can also be exported as plain text, this can be useful in order to create your own charts outside Blast2GO with the original data.
- Toolbox Export:
 - Export Annotations - A more refined version of the annotations export. Results of this export may not be used to import them again.
 - Export Table - Allows to export the data-set as seen in the sequence table editor.
 - Export Generic - An adaptation of the well-known "Generic Export" from Blast2GO standalone, which offers many possibilities to create very customized results.
- The Blast2GO Table Editor also allows for exporting the table data currently visible (Sidepanel).

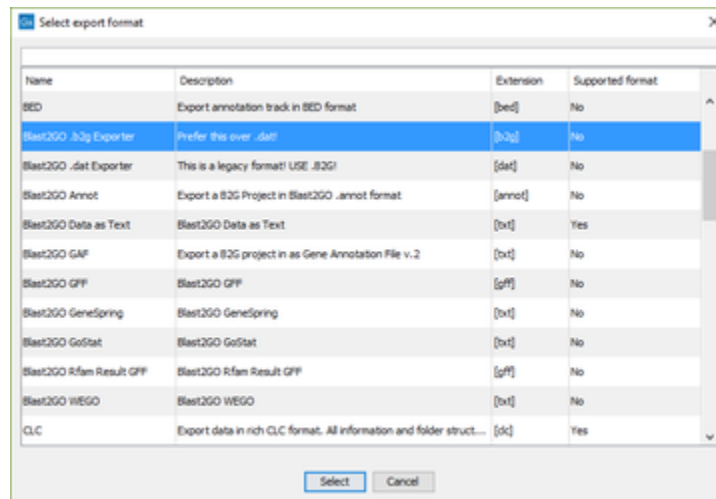


Figure 1: Standard Export options for Blast2GO datatypes.

Fisher's Exact Test (Enrichment Analysis)

Content of this page:

- [Run a Fisher's Exact Test](#)
- [Parameters](#)
- [Results](#)
 - [Table](#)
 - [Enriched Graph](#)
 - [Enriched Bar Chart](#)
- [Reduce to most specific terms](#)

Blast2GO offers the possibility of direct statistical analysis on gene function information. A common analysis is the statistical assessment of GO term enrichment in a group of interesting genes when compared to a reference group i.e. to assess the functional differences between two sets of functional annotations (e.g. GO function of two groups of genes). This analysis is typically performed by a Fisher's Exact Test in combination with a robust False Discovery Rate (FDR) correction for multiple testing. Fisher's exact test is a statistical significance test used in the analysis of contingency tables. Although in practice it is employed when sample sizes are small, it is valid for all sample sizes. It is named after its inventor, R. A. Fisher. The false discovery rate (FDR) control is a statistical method used in multiple hypothesis testing to correct for multiple comparisons. In a list of statistically significant findings FDR is used to control the expected proportion of incorrectly rejected null hypotheses ("false discoveries"). Here a Benjamini-Hochberg correction is used. The result is a list of statistically significant Gene Ontology terms ranked by their adjusted p-values. Results can be viewed in several different ways like tabular format, directly visualized on the Gene Ontology Graph or as a bar chart, always coloring statistically significant terms in red (over-represented) and green (under-represented).

Run a Fisher's Exact Test

To perform the test we need to have a Blast2GO Project which contains the functional information of all sequences/genes to be included in the statistical test. Now we need to select the test and reference set, this is done by indicating ID Lists that contain the sequence ids (gene). Providing a reference set is optional and if no reference is selected, the

The calculation of the p-values for all functions can take several minutes, depending on the size of the dataset and network connection speed. Once the This table lists the adjusted p-values of the Fisher's Exact Test for each GO term.

Parameters

- **Annotations**
Select a Blast2GO Project which contains the functional information of all sequences/genes included in the statistical test (test and/or ref-set).
- **Test and Reference-Set**
Select a Test-Set from the navigation area. Please note that the given IDs have to match the sequence names of the Blast2GO Project selected in **Annotations**. The most convenient way to create ID-lists is to open the **Annotations** project and to select the desired sequences by hand. Once selected, right-click into the **Sequence Name** column and select **Create ID-List of Column: SeqName**. The resulting list can be selected either as Test or Reference-Set An example can be found in the Blast2GO example data-sets.
The **reference-set** is optional and the whole annotations set selected in the first parameter will be used otherwise.
- **GO Categories**
Decide for which GO Category you want to perform this test.
- **Two Sided**
In statistical significance testing, a one-tailed test or two-tailed test are alternative ways of computing the statistical significance of a data set in terms of a test statistic, depending on whether only one direction is considered extreme (and unlikely) or both directions are considered extreme. This translates to over- and under-represented Gene Ontology functions in the test-set compared to a reference set. A two tailed test means therefore to test for over- and under-representation at the same time. Note: The correction for multiple testing (FDR) is higher in a two tailed test and therefore it is less likely to detect significant results since the number of performed test is doubled.
- **Remove Double IDs**
This options allows you to automatically remove all sequences/gene-ids which are present in the test-set and in the reference set at the same time. By default double/common IDs are only removed from the reference set.

Results

Table

Blast2GO offers several options to view the results of an Enrichment Analysis. The table format shows a list of all the terms which add been included in the analyse. With the side-panel we can filter the results and can only visualize e.g. statistically significant results with a FDR p-value smaller than 0.05.

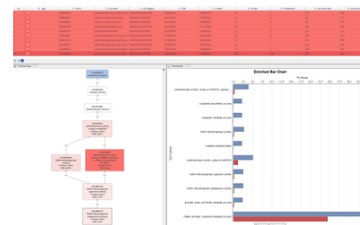


Figure 1: Different types of Fisher's

Enriched Graph

The same results can also be visualized in form of a Enriched GO Graph. The Enriched Graph shows the Gene Ontology graph of the significant terms with a node-coloring which is proportional to the significance value (p-value). This type of graphical representations helps to understand the biological context of the functional differences and to find pseudo-redundancies in the parent-child relationships of significant GO term. A node filter value can be set for the p-value or adjusted FDR p-value. In this way intermediate GO terms are not shown in the graph which reduced the overall size of the graph and graphs can be thinned out deleting these terms. A node filter value determines the p-value for the lowest nodes to be included in the graph. GO-Terms with a value higher than the given filter are not shown. To perform an Enriched GO Graph a Fisher's Exact Test result is necessary.

Enriched Bar Chart

An Enrichment Bar Chart shows for each significant GO term the amount (percentage) of sequences annotated with this term. The Y-axis shows significantly enriched GO terms and the X-axis gives the relative frequency of each term. Red bars correspond to the sequences of the test-set and blue bars correspond to the reference or background dataset (e.g. a whole genome). To perform an Enrichment Bar Chart a Fisher's Exact Test result is necessary.

Reduce to most specific terms

This function allows to reduce the size of the result-set of over-represented GO terms; useful in case of a very large list of enriched GO terms. In many cases, reported enriched functions have a parent-child relationship and therefore these terms represent the same functional concept but at different levels of specificity. In case of large result sets it can be convenient to filter the results by removing parent terms of already existing, statistically significant, child GO terms. In this way only a reduced list of the most specific information is reported.

Manage Projects

Content of this page:

- [Combine Blast2GO Projects](#)
- [Convert Data to Blast2GO Project](#)

Combine Blast2GO Projects

This function allows to combine two or more Blast2GO projects with each other. The options "skip id" and "overwrite id" allows the user to decide on how to treat duplicated sequence names/IDs. The order in which the projects are selected is important, the user would achieve the same result with the following two scenarios: "Project 1 with Project 2 and skip ids" is the same as "Project 2, Project 1 with overwrite ids".

Convert Data to Blast2GO Project

This function allows to convert various data-types to a Blast2GO projects.
Supported data-types are:

- Nucleotide and Protein Sequences (Imported Fasta)
- Multi-Blast Results (Obtained with GWB tools)
- Older Blast2GO projects created with former versions of the plugin, e.g. **pre version 6**

Miscellaneous

Content of this page:

- [Example Data](#)
- [Annex \(legacy\)](#)
- [Annotation Table](#)
- [Remove First Level Annotations](#)
- [Validate Annotations](#)
- [Find Duplicates](#)
- [Set to Sense](#)
- [Translate Longest ORF](#)
- [Batch Rename](#)
- [BDA](#)
- [Retrieve Blast Top-Hit](#)

Example Data

Add some Blast2GO example data to the **Navigation Area** to play around with.

Annex (legacy)

Annex [Myhre et al.,2006] was developed by the Norwegian University of Science and Technology and is essentially a set of manually curated relationships between the three different Gene Ontology categories. The approach uses uni-vocal relationships between GO terms to add implicit annotation. The Annex dataset consists of 6000+ manually reviewed relations between molecular function terms which are "involved in" biological processes and molecular function terms "acting in" cellular components. Annex-based GO term augmentation can be run on any annotation loaded in Blast2GO. Generally, between 10% and 15% extra annotation is achieved and around 30% of GO term confirmations are obtained through the Annex data-set.

Annotation Table

This function allows to create a CLC-bio Annotation Table containing the Gene Ontology terms generated with Blast2GO and can be used in combination with **Add Annotations** in order to perform **Hypergeometric Tests on Annotations** and **Gene Set Enrichment Analysis (Tool box > Microarray and Small RNA Analysis > Annotation Test)**.

Remove First Level Annotations

This function removes for each sequence the three main (root or top-level) GO terms (molecular function, biological process and cellular component), if present since they do not provide any relevant information.

Validate Annotations

This function validates the annotation result and removes redundant GOs from the dataset. It assures that only the most specific annotations for a given sequence are saved. In this way this function prevents that two or more GO terms lying on the same GO branch are assigned to the same sequence. The Gene Ontology "true path rule" assures that all the terms lying on the branch or route from a term up to the root (top-level) must always be true for a given gene product. Therefore, any term is considered as redundant and is removed if a child term coexists for the same sequence.

This function can be run independently, however Blast2GO applies this method automatically always after a modification is made to an existing annotation, such as merging GO terms from InterProScan search, after Annex augmentation or upon manual curation.

Find Duplicates

This function helps to identify (and optionally delete) sequences from a data-set which are 100% identical.

Set to Sense

Converts sequences with a negative reading frame Top-Blast-Hit to anti-sense i.e. query sequences will be translated to its reverse complement (e.g. ATTG > CAAT).

Also adds "_antisense" to the sequence name, use **Batch Rename** to revoke the name change afterwards.

Translate Longest ORF

Converts nucleotide sequences to its longest open reading frame.

Also adds "_ORF" to the sequence name, use **Batch Rename** to revoke the name change afterwards.

Batch Rename

Renames all selected sequences by either adding text or by replacing text patterns (regular expressions can be used). Also allows to convert the sequence names to lower or upper case.

BDA

The primary goal of Blast2GO is to assign functional labels in form of GO-terms to nucleotide or protein sequences. However, not only functional labels but also a meaningful description for novel sequences is desired. A common approach is to directly transfer the "Best-BLAST-hit description to the novel sequence. It is frequent that best-hit descriptions are of low-informative text such as "unknown", "putative" or "hypothetical" while descriptions of other Blast hits of the same sequence do contain informative keywords. For this reason, a text-mining functionality has been included in Blast2GO. It analyses a set of sequence descriptions of a given BLAST result. The feature is called the BLAST Description Annotator (BDA). Depending on the frequency of occurrence and the information content, the most suitable description is selected out of the collection of words. In this way, this simple approach avoids sequence descriptions like for example "hypothetical", "putative" or "unknown protein" in the case that a more informative and representative description is available. These descriptions are only of exploratory nature and do not have the same weight of evidence as the functional labels.

Retrieve Blast Top-Hit

This feature uses Blast result information to search the top-hit at NCBI, Ensembl or Uniprot (via web services). It is then possible to replace the original query sequence with its Blast top-hit or to extract the information to a new project (various scenarios are possible).

A possible use case scenario would be a so called "Double-Blast" (Figure 1): The blast results of a first run are used to replace the sequence data for a second run against a different set of query sequences. Imagine an RNA-seq data-set with a high percentage of sequences without any alignments against a protein database (e.g. blastx against NR). This feature could be used to select and extract the sequences without hits (red ones) into a new project. These sequences could be blasted (blastn) against a set of EST sequences. The initially unaligned sequences are now replaced with the ESTs (**Retrieve Blast Top-Hit**). In the last step, these sequences are blasted against NR and will hopefully return valid protein hits to follow the functional annotation pipeline with GO Mapping and GO Annotation.

For each Top-Hit (first significant alignment from an already performed BLAST), apply the filters (bottom part of the dialog) and search them in the corresponding database (online).

It is possible to either replace the sequence from your data-set or to extract them into a new data-set (Action option). You can also decide whether you want to keep the original sequence names or if you want to rename them to the downloaded sequences names. The latter will add a small note to the sequence description, telling you the original name.

The last remaining option allows you to decide whether you want to replace your sequences with the downloaded ones (default) or if you just want to retrieve their name.

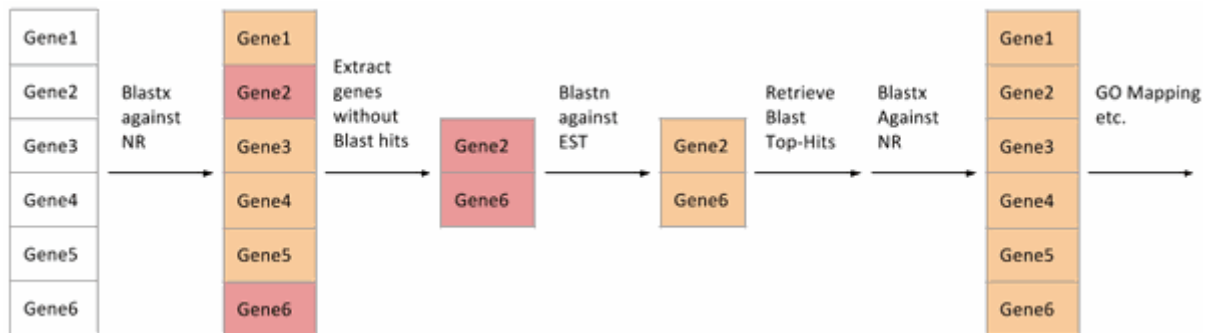


Figure 1: Example workflow for "Double-Blast", using Retrieve Blast Top-Hit.

Workflows

All major Blast2GO plugin functions are workflow-able. This allows us to create an annotation pipeline with only a few mouse-clicks. We describe here only the basic steps to get you started with workflows, for more detailed information please refer to the CLC bio Workbench user manual. The Blast2GO example dataset (available from the Toolbox) also contains an example workflow.

How to create a basic Blast2GO workflow:

1. Create a new workflow.
Go to: **Workflows > New Workflow...**
2. Add the desired functions with **right-click > Add Element > Blast2GO**, or simply drag and drop them from the Toolbox.
We add CloudBlast, Mapping, Annotation and two Statistics boxes.
3. The selected functions now appear in the workflow area, we can arrange them to graphically form the pipeline shown in figure 1.
4. Now we connect all the available outputs with the logical proceeding inputs. Apart from that all functions that create a result that you want to save to disk, have to be connected to a so-called workflow output. To achieve this, we **right-click** on the desired functions outputs and select **Use as Workflow Output**. We must not forget to connect the workflow input to the CloudBlast, which will be our entrance point of the pipeline.
5. The next step would be to configure a few parameters (Configurable functions are indicated by a little notepad symbol). To set the parameters of a function, we double-click on it to show a wizard similar to the ordinary one. We can activate the Data Distribution chart in both statistic steps. With this we can examine the success-rate of the mapping step, while the annotation step is still running.
6. After configuring the functions as desired, we save the workflow. The workflow can now be executed.

It is important to understand that a Blast2GO Project has no attribute which indicates the status of a project (e.g. project is mapped or annotated). The workbench is therefore not able to verify if the processed project is annotated, mapped or has only blast results. Therefore whenever we need to choose input data or connect algorithms in the workflow we have to verify this ourselves and check that all steps are connected in the right order; e.g. the mapping step has to be placed before the annotation. Otherwise we end up with a mapped project without annotations since the annotation step needs the information from the mapping. However we will not receive any error messages or similar, because of the above mentioned reason.

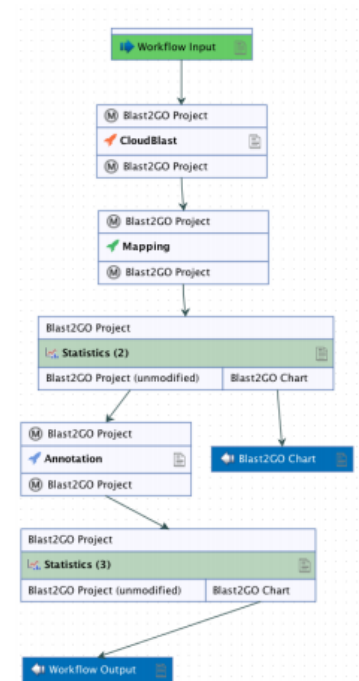


Figure: Basic workflow example which performs a blast, mapping and annotation step and generates some basic summary statistics.

Orthologous Groups (NOG/COG)

Content of this page:

- [Clusters of Orthologs](#)
- [Orthologous Group Annotation Tool](#)
 - [EggNOG](#)
 - [Orthologous Group Annotation: Wizard](#)
 - [Orthologous Group Annotation: Algorithm](#)
 - [Orthologous Group Annotation: Results](#)
 - [Reorder Results Table](#)
- [Merge GOs from COG](#)
 - [Merge Orthologous Group GOs to Annotation: Results](#)

Clusters of Orthologs

The growth of the annotation databases, has opened the path for multiple bioinformatic algorithms to search for homologies between sequences. Similarity information has multiple uses, such as sequence annotation or evolutionary inference.

A Cluster of Orthologous Group (COG) corresponds to a group of proteins that share a high level of sequence similarity. Sequence similarity, in the vast majority of the cases, can be associated to evolutionary convergence. All sequences contained in an COG presumably derive from the same ancestor sequence, which has diverged into the different members of the orthologous group via speciation (orthologous) and duplication (paralogous) events.

Orthologous Group Annotation Tool

With this tool, we intend to provide a method to annotate the orthologous group of a sequence within the Blast2GO annotation pipeline. Since the sequences from an orthologous group share many distinctive features (e.g. functional annotation, phylogenetics), the orthologous group annotation can be used to infer properties that can improve the Blast2GO sequence characterization.

To this extent, we made use of the EggNOG database (Evolutionary genealogy of genes: Nonsupervised Orthologous Groups) to annotate any sequence present in the database with its corresponding orthologous group.

The Orthologous Group Annotation Tool can be found in **Toolbox > Blast2GO > Orthologous Groups > Obtain COG Ortholog Groups**. **Blast** and **GO Mapping** steps are needed to be completed beforehand in order to perform the Orthologous Group assignment. The Orthologous Group Annotation will have a better performance if the Blast is executed against the UniProtKb database.

In case the Blast is performed against the nr database, the result of the Orthologous Group Annotation will be satisfactory, but the quality will still be lower than if UniProtKb was used as the Blast database.

EggNOG

EggNOG (Evolutionary genealogy of genes: Non-supervised Orthologous Groups) is a “graphbased unsupervised clustering algorithm extending the COG methodology”. It provides an orthology classification method, based on sequence similarity. The EggNOG database is built collecting genomes from public datasets, and performing an all-against-all pairwise similarity matrix. Such matrix is stored in a relational database, in which the high-similarity sequences are grouped together.

The clusters classification takes its basis in the Clusters of Orthologous Groups (COG), those clusters that have been described in this manually-curated orthology classification will be correspondingly annotated. On the other hand, the clusters that have not been described in COG, are defined *de novo*, and functionally annotated using GO, KEGG pathways and SMART/PFAM protein domains.

Orthologous Group Annotation: Wizard

The parameters used in the Orthologous Group Annotation Wizard are based on the Blast results. As it was mentioned above, **Blast** and **GO Mapping** are required to perform this analysis.

- **e-Value:** Select an e-Value limit to which the sequence will be included in the analysis. The e-Value is assigned in the Blast step. Select one of the multiple options available in this widget (Default: 1E-3).
- **Filter by Similarity (%):** Specify a minimum similarity percentage from which the input sequences are filtered out. The similarity is gathered from the Blast result, it is obtained dividing the positive matches of the alignment and the Hsp length (Default: 50%).
- **Hsp/Hit coverage filter:** Establish a Hsp/Hit coverage to filter those results that cover less hit length than the minimum specified. The coverage is a percentage, therefore it must be a number between 0-100, 0 is selected to disable this filter option (Default: 0).

Orthologous Group Annotation: Algorithm

Blast and GO Mapping results are required to execute the Orthologous Group Annotation. When this feature is launched, all the Blast2GO-project selected sequences are iterated. Those sequences that do not have Blast/Mapping Annotation will be skipped, while the mapping results of the annotated sequences are extracted. The program iterates over the mapping results, if the Blast parameters pass the filters set in the Orthologous Group Annotation Wizard (previous section), the method identifies the Orthologous Group annotation of each mapping result (if it has been described).

Orthologous Groups are assigned to the project using the EggNOG database. Since EggNOG does not provide an API to assign the annotation directly from their REST service, we use UniProt RESTful service API, which contains the information of the Orthologous Group via EggNOG. The information regarding the Orthologous Group Description, Category, or Gene Ontology is gathered from the EggNOG RESTful API. Such information is stored in the b2gFiles folder locally, and loaded to a Blast2GO Object, which is visualized in a Table Viewer.

The Blast2GO project sequence is annotated with the "Top-Hit" Orthologous Group, which corresponds to the mapping with a higher score in the Blast search that can be assigned an Orthologous Group via EggNOG. If such mapping can be assigned to more than one Orthologous Group, we assign all of them. 'COG' groups have been manually-curated, 'KOG' (euKaryotes Clusters of Orthologs) are manually-curated multi-domain proteins, and 'ENOG' are computationally obtained in an all-vs-all homology search.

Normally, all the mapping results are annotated within the same Orthologous Groups.

Orthologous Group Annotation: Results

The information gathered from the previous section is retained in a Blast2GO object and visualized in a Blast2GO Table by Sequence. Figure 1 shows the default viewer for the Orthologous Group Blast2GO object.

SeqNames	NOG IDs	NOG Description	OG Categories	GO IDs
...	...	transcription, 5S2 related protein homolog
...	...	transcription, 5S2 related protein homolog
...	...	transcription, 5S2 related protein homolog
...	...	transcription, 5S2 related protein homolog
...	...	transcription, 5S2 related protein homolog
...	...	transcription, 5S2 related protein homolog

Figure 1: Ortholog Group Annotation Main Table Results. Results are ordered by sequence

The Annotation Results are visualized in eight columns:

- **Tag:** Describes if there is a NOG assigned, and if it has been manually-curated (COG/KOG) or unsupervised (ENOG).
- **SeqNames:** Name of the sequence (as in the Blast2GO Project).
- **Nog IDs:** Identifier for the Orthologous Groups assigned to the given sequence, they are comma-separated.
- **Nog Description:** Description of the Orthologous Groups, separated by ';'.
 • **OG Categories:** Categories to which the Orthologous Group corresponds to, they are comma-separated. There are a total of 23 categories, which are more general than the Orthologous Groups.
- **OG Categories Description:** Description of the Orthologous Group Categories separated by ';'.
 • **GO IDs:** Gene Ontology described for the annotated OG Categories, separated by ';'.
 • **GO Names:** Gene Ontology IDs described for the annotated OG Categories, separated by ';'. By default this column is not shown, it can be activated by right-clicking the column headers.

Reorder Results Table

The results can be visualized by Sequence ID, as it is shown in Figure 1, or can also be displayed reordering the Table by NOG ID or by OG Category. The Orthologous Group object can be opened in different formats from the table side panel.

Opening the Orthologous Group Table by NOG ID (**Figure 2**), allows the user to determine which sequences are present in each Orthologous Group. A new column, which shows the number of sequences present in each Orthologous Group, is added to the Table Viewer.

When the Orthologous Group Table is opened by Category (**Figure 3**), the NOG IDs are grouped in their Orthologous Group Categories. Also, sequences that correspond to the NOG IDs grouped, are combined to allow the recovery of sequences by Category. Here, we add two new columns: one corresponding to the number of sequences per Category, and another corresponding to the number of NOG IDs per category. The column corresponding to NOG Description is not present.

NO	Term	NO Categories	NO Description	NOG ID	NOG number	Synonyms	Number of sequences	GO	GO IDs
1	ENR0000002	A	DNA recombination	ENR0000002	1	ENR0000002	1	Enzymatic modification of protein structure, recombination	GO:0006969 GO:0006970
2	ENR0000003	A	DNA replication	ENR0000003	2	ENR0000003	2	Enzymatic modification of protein structure, recombination	GO:0006969 GO:0006970
3	ENR0000004	A	DNA replication	ENR0000004	3	ENR0000004	3	Enzymatic modification of protein structure, recombination	GO:0006969 GO:0006970
4	ENR0000005	A	DNA replication	ENR0000005	4	ENR0000005	4	Enzymatic modification of protein structure, recombination	GO:0006969 GO:0006970
5	ENR0000006	A	DNA replication	ENR0000006	5	ENR0000006	5	Enzymatic modification of protein structure, recombination	GO:0006969 GO:0006970
6	ENR0000007	A	DNA replication	ENR0000007	6	ENR0000007	6	Enzymatic modification of protein structure, recombination	GO:0006969 GO:0006970
7	ENR0000008	A	DNA replication	ENR0000008	7	ENR0000008	7	Enzymatic modification of protein structure, recombination	GO:0006969 GO:0006970
8	ENR0000009	A	DNA replication	ENR0000009	8	ENR0000009	8	Enzymatic modification of protein structure, recombination	GO:0006969 GO:0006970
9	ENR0000010	A	DNA replication	ENR0000010	9	ENR0000010	9	Enzymatic modification of protein structure, recombination	GO:0006969 GO:0006970
10	ENR0000011	A	DNA replication	ENR0000011	10	ENR0000011	10	Enzymatic modification of protein structure, recombination	GO:0006969 GO:0006970

Figure 2: Ortholog Group Annotation Table results ordered by NOG ID

All the results can be exported by selecting clicking **Export Table** in the sidepanel. The Table will be exported in the format selected when the Export is requested.

NO	Term	NO Categories	NO Description	NOG ID	NOG number	Synonyms	Number of sequences	GO	GO IDs
1	ENR0000002	A	Energy production and conversion	ENR0000002	1	ENR0000002	1	Enzymatic modification of protein structure, recombination	GO:0006969 GO:0006970
2	ENR0000003	A	Energy production and conversion	ENR0000003	2	ENR0000003	2	Enzymatic modification of protein structure, recombination	GO:0006969 GO:0006970
3	ENR0000004	A	Energy production and conversion	ENR0000004	3	ENR0000004	3	Enzymatic modification of protein structure, recombination	GO:0006969 GO:0006970
4	ENR0000005	A	Energy production and conversion	ENR0000005	4	ENR0000005	4	Enzymatic modification of protein structure, recombination	GO:0006969 GO:0006970
5	ENR0000006	A	Energy production and conversion	ENR0000006	5	ENR0000006	5	Enzymatic modification of protein structure, recombination	GO:0006969 GO:0006970
6	ENR0000007	A	Energy production and conversion	ENR0000007	6	ENR0000007	6	Enzymatic modification of protein structure, recombination	GO:0006969 GO:0006970
7	ENR0000008	A	Energy production and conversion	ENR0000008	7	ENR0000008	7	Enzymatic modification of protein structure, recombination	GO:0006969 GO:0006970
8	ENR0000009	A	Energy production and conversion	ENR0000009	8	ENR0000009	8	Enzymatic modification of protein structure, recombination	GO:0006969 GO:0006970
9	ENR0000010	A	Energy production and conversion	ENR0000010	9	ENR0000010	9	Enzymatic modification of protein structure, recombination	GO:0006969 GO:0006970
10	ENR0000011	A	Energy production and conversion	ENR0000011	10	ENR0000011	10	Enzymatic modification of protein structure, recombination	GO:0006969 GO:0006970

Figure 3: Ortholog Group Annotation Table results ordered by OG Category

Merge GOs from COG

With the Merge GOs from COG feature, the GOs assigned to each sequence in the Orthologous Group Annotation can be merged to the Mapping GO annotation. Since the Blast2GO project only displays the most specific gene ontologies, this function only adds those GOs that are more specific than the ones being displayed in the project; it uses the Blast2GO DAG feature to achieve this.

Merge Orthologous Group GOs to Annotation: Results

When the GOs are merged into the Blast2GO project, the GO annotation is changed as more specific GOs are added into the project and less specific are removed. **Figure 4** shows an statistical evaluation after the GO merge has been performed. It includes the following information:

- **GOs Before Merge:** Number of GOs present in the Blast2GO project before it is merged with the Orthologous Group Annotation GOs.
- **GOs After Merge:** Number of GOs present in the Blast2GO project after it is merged with the Orthologous Group Annotation GOs.
- **Confirmed GOs:** Number of GOs that were present in the Blast2GO project before the merge has been performed, and are also present in the Orthologous Group Annotation GOs. These GOs do not add new information, apart from confirming the previous knowledge.
- **Too General GOs:** Number of GOs that are more general than the Blast2GO project GO Annotation.
- **New GOs:** Number of GOs from the Orthologous Groups Annotation that are more specific than the GOs present in the Blast2GO project.

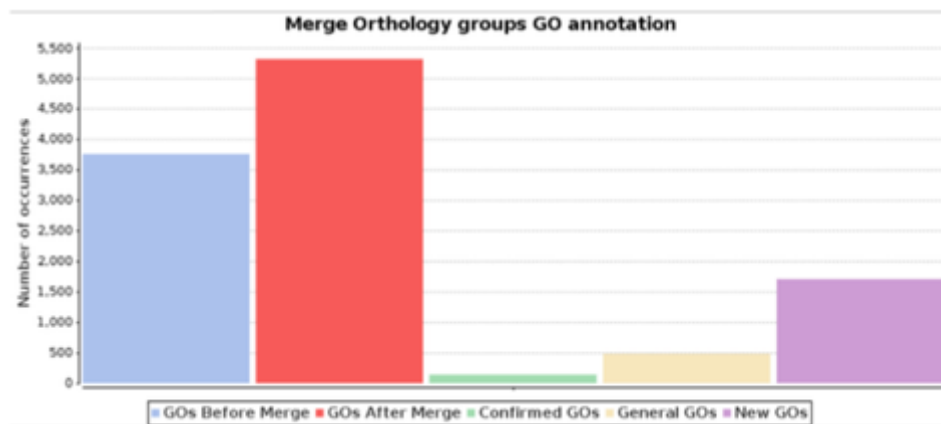


Figure 4: Merge Orthologous Group GOs result. Shows the statistics of: number of GOs before and after the merge is performed, confirmed GOs, too general GOs and new GOs

We recommend to run this analysis once the annotation step has been performed.

PSORTb

Content of this page:

- [Introduction](#)
- [Parameters and Execution](#)
- [Results](#)
- [Merge GO information](#)

Introduction

The PSORT principle uses the amino acid sequence information to generate an overall prediction of the protein localization sites. This rules are derived from experimental observations. For example, when analysing a gram negative organism, possible localization sites are: cytoplasm, cytoplasmic membrane, periplasm, outer membrane and extracellular space.

Blast2GO allows to assign sub-cellular localization sites to proteins based on their amino acid sequence via PSORTb. PSORTb is an algorithm which can be applied to bacteria or archaea protein sequences and uses a probabilistic system to predict the most probable localization. Once sites are predicted, its corresponding cellular component GO terms can be merged with the already existing Blast2GO annotations.

Parameters and Execution

Starting with a previously loaded Blast2GO project with protein sequences, the PSORTb tool can be found under **Toolbox > Blast2GO > Psortb > PSORTb**.

If the loaded project contains nucleotide sequences, the **Translate Longest ORF** tool can help to obtain the predicted protein sequences and be able to run PSORTb.

The wizard allows to adjust the algorithm parameters and it performs different analysis depending on the **Organism Type** and the **Gram Stain**. It can be used with bacteria positive and negative gram stains or archaea organism sequences. For more details of the core algorithm, visit <http://www.psortb.org>.

The algorithm returns score values between 0 and 10 for each localisation site, the **Cutoff** parameter allows to set a minimum value of each localization above which the value can be considered as possible localization.

Results

The tool will iterate over the input sequences and analyse each of them with PSORTb. The process will open a new tab and as the results come back, they are shown in a table format.

The table contains one row for each sequence, where the columns have the following meaning:

- **Sequence name:** shows each sequence identifier.
- **Final localization:** contains the the predicted localization name.
- **Final score:** represents the prediction score for the localization.
- **GO ID:** the Gene Ontology ID associated to the location.
- **Secondary Localization:** a possible secondary localization when there is more than one score above the cutoff.
- The next 6 columns, hidden by default, show the score for all possible localizations.

Merge GO information

The GO IDs from the prediction can be merged into the original Blast2GO project as cellular component characterization of the sequences. The merge wizard asks for the Blast2GO project file where to merge the GO results and will add the GO information to the project, matching the Sequence Name.

Rfam

Content of this page:

- Introduction
- Results Table
- Graphical Representation of the Results

Introduction

The Rfam database is a collection of RNA families, each represented by multiple sequence alignments, consensus secondary structures and covariance models (CMs). The families in Rfam break down into three broad functional classes: non-coding RNA genes, structured cis-regulatory elements and self-splicing RNAs. Typically these functional RNAs often have a conserved secondary structure which may be better preserved than the RNA sequence. The CMs used to describe each family are a slightly more complicated relative of the profile hidden Markov models (HMMs) used by Pfam. CMs can simultaneously model RNA sequence and the structure in an elegant and accurate fashion (Rfam description from: <http://rfam.xfam.org/>).

Please cite: Nawrocki, E. P., Burge, S. W., Bateman, A., Daub, J., Eberhardt, R. Y., Eddy, S. R., Floden, E. W., Gardner, P. P., Jones, T. A., Tate, J., et al. (2014). Rfam 12.0: updates to the rna families database. Nucleic acids research, page gku1063.

This functionality can be found under **Toolbox > Blast2GO > Rfam > Rfam**. It may take a while depending on the number of sequences and the EMBL-EBI servers.

Tags	Sequence	#Hits	Start	End	Score	Strand	e-Value
RFAM	comp66944_c0_seq1	1	337	337	73.9	-	1.1E-14
RFAM	comp16916_c0_seq1	3	531	531	102.5	+	5.5E-24
RFAM	comp16916_c0_seq4						1.1E-20
RFAM	comp52083_c0_seq1						6.5E-19
RFAM	comp135245_c0_seq1						5.2E-14
RFAM	comp183927_c0_seq1						1.5E-10
RFAM	comp5433_c0_seq1						1.7E-15
RFAM	comp6164_c0_seq1						9.5E-19
RFAM	comp16916_c0_seq9						1.4E-20
RFAM	comp6784_c0_seq2	2	1072	1072	102.4	+	3.4E-24

Figure 1: Rfam Table Results

Results Table

Once Rfam analysis has begun a table with the corresponding results will be displayed in a new tab. Sequences will turn red/orange depending if Rfam found hits for them (red if no hits were found, orange otherwise). White rows are sequences that have not been analysed yet. For each sequence it is possible to consult details about each one of their hits using the context menu (similar to consult Blast results).

The obtained results can be exported via the **GWB Standard Export** functionality in **Blast2GO GFF format**.

Graphical Representation of the Results

Multiple charts are available from the Rfam toolbox.

1. **Hit Distribution:** This chart shows a distribution chart of the number sequences with hits in the Rfam analysis.
2. **Biotypes Pie Chart:** This pie chart shows the distribution of the Rfam families of the sequences.
3. **Biotypes Distribution:** The same as the former but in a bar-style.
4. **E-Value Distribution:** This chart plots the distribution of E-values for the Rfam hits.

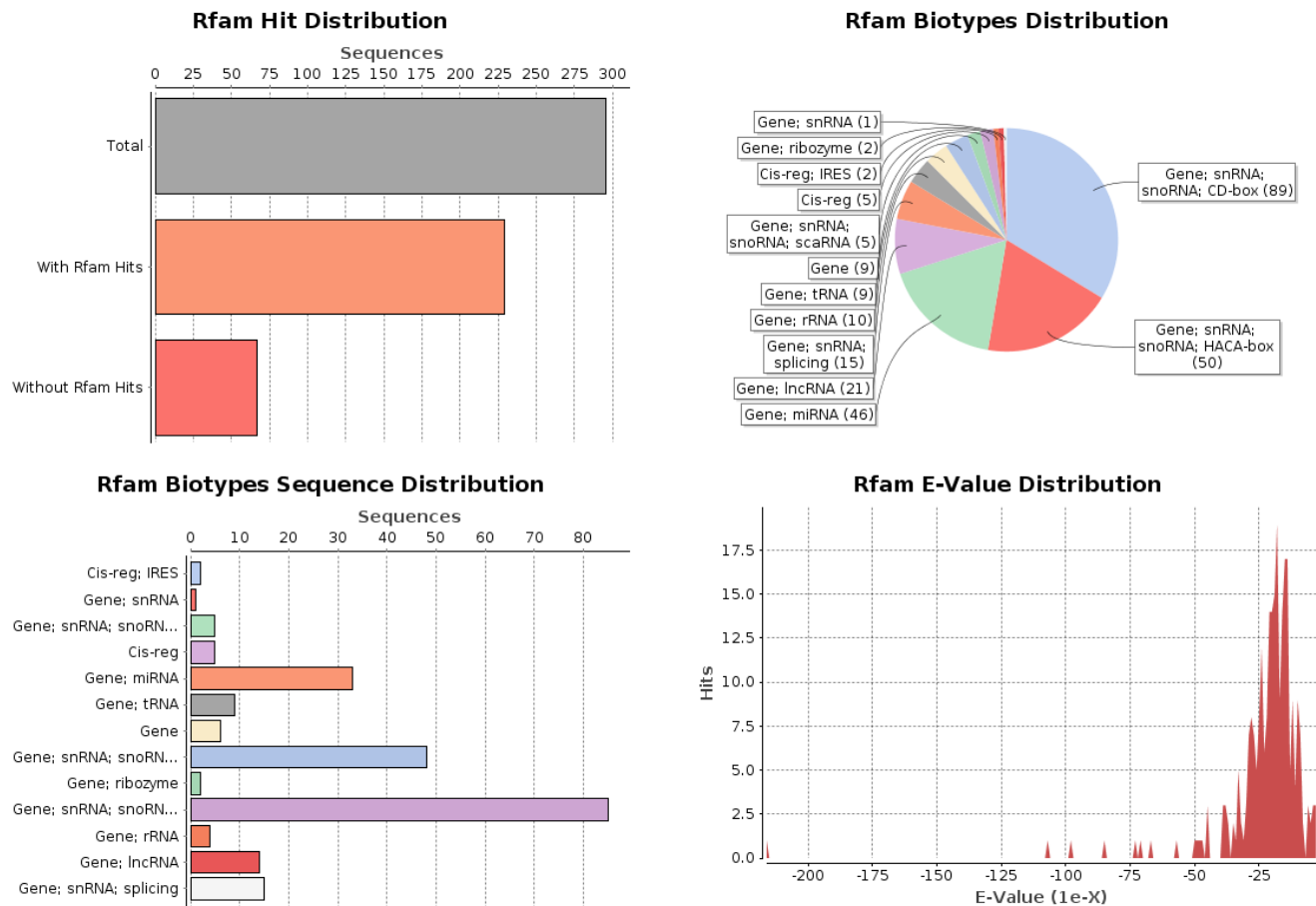


Figure 2: Rfam Statistics Graphs and Visualization