Blast2GO PRO Plug-in User Manual

CLC bio Genomics Workbench
and Main Workbench
Version 1.4, September 2015

BioBam Bioinformatics S.L.
Valencia, Spain

Contents

Introduction

Support: **pluginsupport@blast2go.com**
Website: `https://www.blast2go.com`

Blast2GO [**?**] is a bioinformatics platform for high-quality functional annotation and analysis of genomic datasets. It allows analyzing and visualizing newly sequenced genomes by combining state-of-the-art methodologies, standard resources and algorithms. Blast2GO allows to gain biological insights fast and easy even for completely novel genomes. Perform out-of-the-box the entire workflow of functional annotation of your transcriptomic datasets including its analysis and biological interpretation.
Blast2GO allows the functional annotation of (novel) sequences and the analysis of annotation data. Its main function is to assign information about the biological function of gene or protein sequences by making use of diverse public resources like comparison algorithms and databases. The software identifies already characterized similar sequences, and transfers its functional labels to the uncharacterized sequences. In this manner, it is possible to obtain functional information for a whole dataset much faster than through experimentation. Apart from mere "in-silico" functional sequence characterization, the software suite has many other functions including joined data-visualization and statistical analysis procedures. These features help in the process of functional interpretation of experimental datasets.
The method uses local sequence alignments (BLAST) to find similar sequences (potential homologs) for one or several input sequences. The program extracts all GO terms associated to each of the obtained hits and returns an evaluated GO annotation for the query sequence(s). Enzyme codes are obtained by mapping from equivalent GOs while InterPro motifs can directly be queried at the InterProScan web service.
A basic annotation process with Blast2GO consists of 3 steps: blasting, mapping and annotation. These steps will be described in this manual including further explanations and information on additional functions. [**?**]

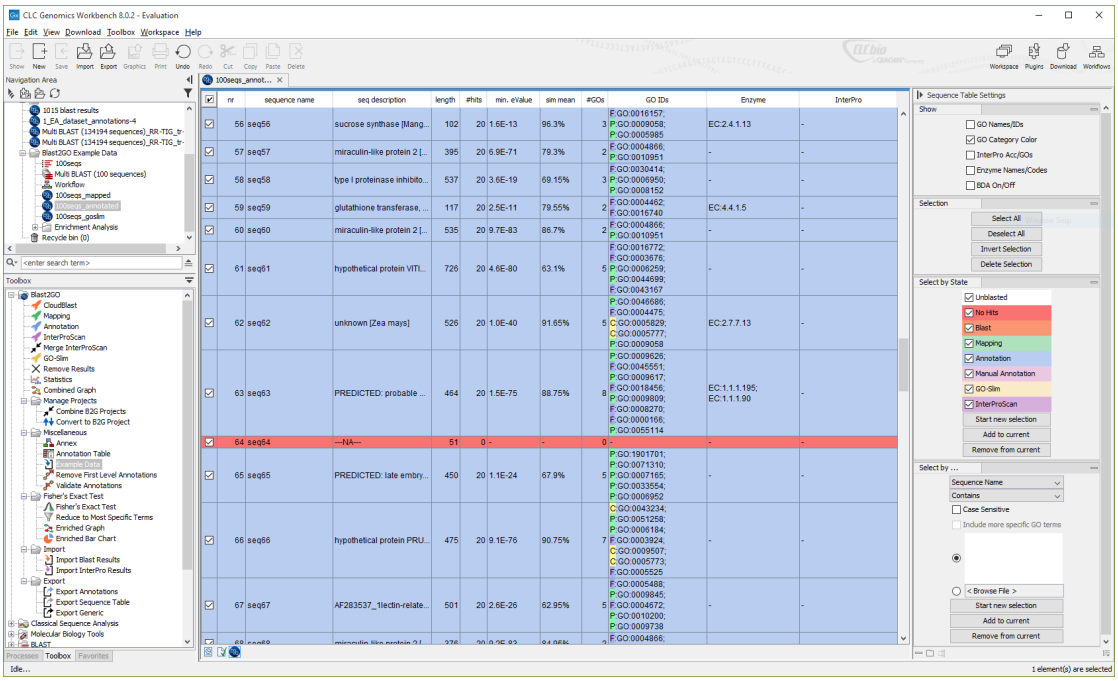


Figure 1.1: Main Sequence Table of the Blast2GO PRO Plugin

Quick-Start

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.

1. To start an annotation process load a Fasta sequence file:
   Import ⟩ Standard Import ⟩ File Type: Fasta File You can also add an example dataset to your Navigation Area from: Toolbox ⟩ Blast2GO ⟩ Miscellaneous ⟩ Example Data This dataset contains 100 sequences as plain sequences.

2. Blast your sequences:
   You have basically 3 options here.

   (a) Blast2GO's CloudBlast feature from the Blast2GO Toolbox. You first need to convert your FASTA information into a Blast2GO Project: Toolbox ⟩ Blast2GO ⟩ Manage ⟩ Project ⟩ Convert to Blast2GO Project

   (b) Genomics Workbench built-in blast functionalities. Please see: *BLAST at NCBI* in the Workbench Help

   (c) Import Blast results with Toolbox ⟩ Import ⟩ Import Blast Results

   Important:
   The GO 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.

3. Convert Blast results into a Blast2GO Project (skip this if you already have a Blast2GO Project e.g. if you used CloudBlast):
   Go to Toolbox ⟩ Blast2GO ⟩ Manage ⟩ Project ⟩ Convert to Blast2GO Project to convert your Blast results into a Blast2GO Project. Here you can select Multi Blast results created with the CLC Workbench.

4. Perform Gene Ontology Mapping:
   Go to Toolbox ⟩ Blast2GO ⟩ Mapping to start the process. Mapped sequences will turn green, visualize your results once it has finished: Toolbox ⟩ Statistics ⟩ Mapping

5. Annotation:
   Go to Toolbox ⟩ Annotation to run the annotation step. Don't change any parameters and sequence that can be annotated will turn blue.

6. Generate Statistics Charts:
   Once the annotation process is finished we can generate all the different statistics charts from: Toolbox ⟩ Statistics

7. Modify Annotations:
   To manually change the annotations click on one of the sequences form the Blast2GO sequence table with the right mouse button and select Edit Sequence. To summarize the functional content of a dataset run a GO-Slim reduction (Toolbox ⟩ GO-Slim).

8. InterProScan:
   To complement the blast-based annotations with domain-based annotations run an InterProScan Search. Go to Toolbox ⟩ InterProScan. This step is recommended to improve the annotation outcome. Once InterProScan results are retrieved use Toolbox ⟩ Merge InterProScan to add the GO terms obtained through motifs/domains to the current/existing annotations. In order to perform the InterProScan, you need to enter a valid email address.

9. Export Results:
   The Blast2GO plugin offers various possibilities to export data through the workbenches Export and Graphics function. Some of the most important are:

   - 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.
   - dat-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.
   - GAF 2.0: A tab-separated text file of the funtional 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.

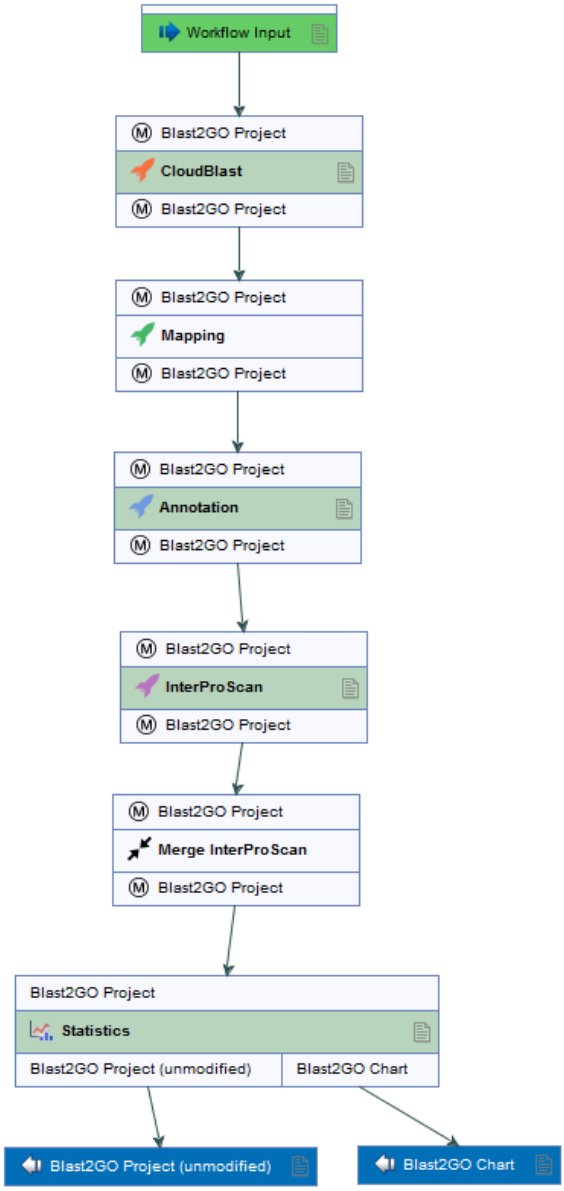


Figure 2.1: The work-flow from the example data shows a similar scenario as described above. The work-flow accepts as input a Blast2GO project generated from via a Blast XML file import or a Multi-Blast CLC Object. It proceeds than with mapping, annotation, InterporScan, merges the annotations obtained through Blast and Domain searches and generates several charts on the way. The Blast2GO Project is saved after each step.

User Manual

1 Introduction

The plugin integrates seamlessly the Blast2GO methodology for high-quality functional annotation into the CLC bio Workbench. It allows to perform all necessary step to functionally annotate un-characterized sequences and to functionally analyse the corresponding dataset. The annotation strategy proceeds via a basic workflow of BLAST, Gene Ontology mapping and the annotation prediction. The plugin offers further analysis option to improve the annotation quality. Many statistical charts allow to interactively run through the annotation process and to provide a quick but comprehensive summary of the whole process. Analysis options like GO-Slim, Gene Ontology graph visualization as well as Enrichment Analysis allows to summarize, describe and extract relevant function information from the whole or part of a dataset (e.g. whole genome, groups of sequences, etc.).

1.1 Main Plugin Features

- Sequence annotation data presented in spread-sheet format.
- Handle tens of thousand of sequences in one project.
- Functional annotation is done in 3 steps: BLAST to find homologus sequences, Gene Ontology mapping to retrieve GO terms and Annotation Prediction to select reliable functions.
- Different annotation databases are supported: GO, Enzyme Codes and InterPro
- Configurable annotation settings to adapt to your data.
- Statistical charts to monitor your annotation progress.
- Graphical display of annotation data through GO graphs, pie and bar charts.
- Select sequences based on keywords and functional information
- Functional Enrichment Analysis
- Completely workflowable
- Access to CloudBlast

1.2 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 solutions 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 distribution partners around the world. For more information about BioBam please visit: `https://www.biobam.com`.

1.3 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. The amount of required memory depends on the amount of sequences to be analysed within one project. We recommend a minimum of 1 GB RAM for any scenario.

1.4 Installation

The plugin can be installed via the Plugin Manager from within the CLC bio Workbench (Help Menu). From the *Download Plugin* tab choose the Blast2GO Plugin and click on *Download and Install*. In order to download the plugin you have to provide you credentials. Once installed the workbench will automatically inform about new updates. Regarding licencing option and quotes please contact the CLC bio sales team: sales@clcbio.com

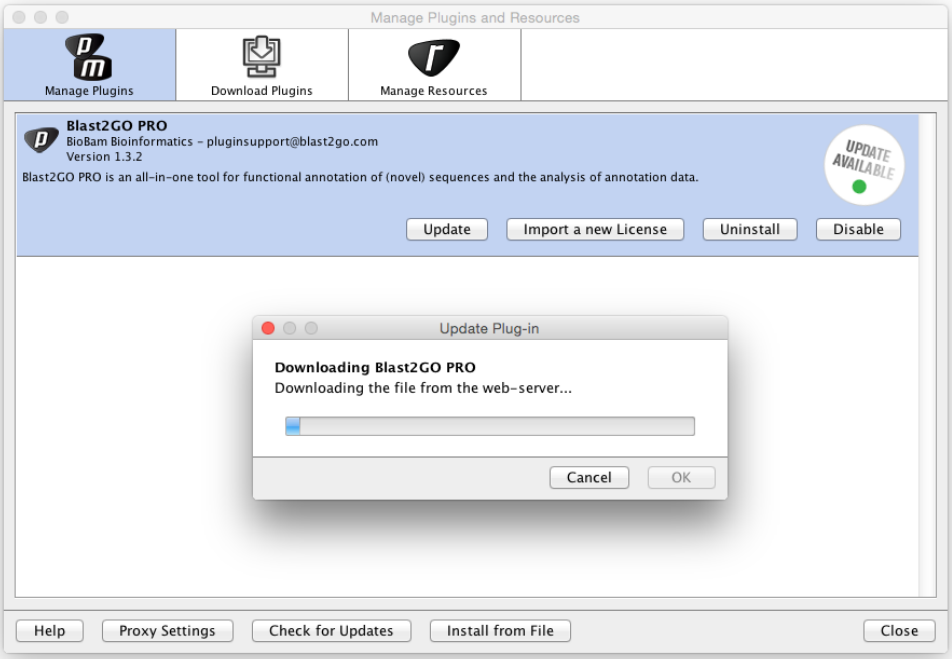


Figure 3.1: Download and update the plugin

1.5 Support

Our support and development team works close together to take care of all the issues you might have, does not matter if rather technical, 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.

2 User Interface

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

2.1 Sequence Table

The sequence table is the main view of a Blast2GO Project. 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: 3.2). 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.).

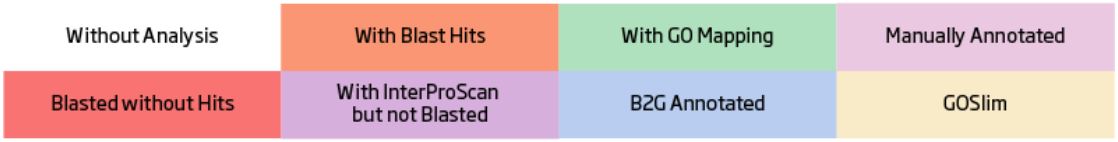
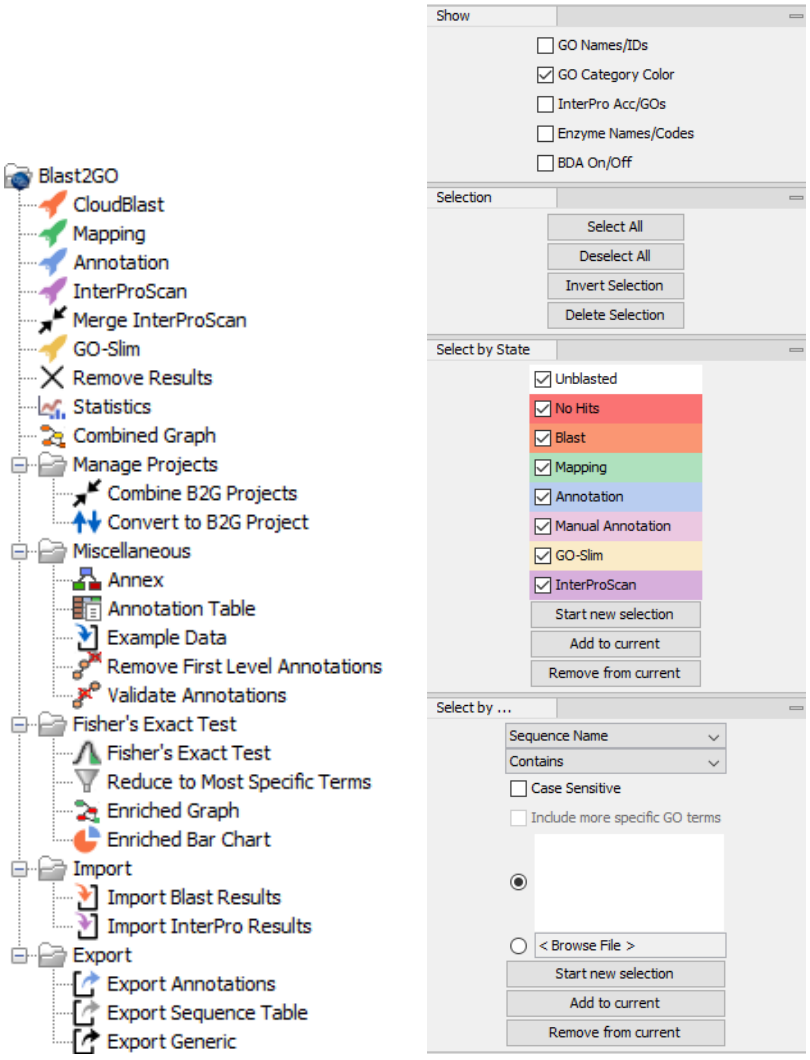


Figure 3.2: Different colour codes indicating the status of the sequences

2.2 Sequence Table Side Panel

The Sequence table offers a side panel which allows to select, filter and search sequences within a Blast2GO project. Additionally, the side-panel allows to change the view of several columns of the sequence table.

- Show: Allows to change the visualization of several columns of the sequence table. It is possible to switch between GO IDs or GO names, show or hide the GO categories of each GO term, show InterproScan Accession or the corresponding GO IDs, choose between Enzyme codes or names. The user can also switch between Blast Description Annotator and normal Blast view (see BDA below).
- Selection: Allows to select, deselect, invert and delete a given selection.
- Select by State: Allows to make a selections based on the sequence status (colors).
- Select by: Allows to select sequences based on their name, function (GO terms or IDs), description, 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. Finally we have to decide if the search result has do be added or removed form the actual selection (select or deselect the sequences which match the criteria). The search can be started by clicking the corresponding buttons.
- Blast Description Annotator (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.



(a) Toolbox area (b) Sequence editor side-panel

Figure 3.3: User Interface: The Blast2GO Toolbox and the Main Sequence Side-Panel

3 Main Analysis Options

3.1 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 to 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 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 ComputationUnits.

To get an approx. idea of the consumption of these units here and example: With 3.000.000

ComputationUnits 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 datasets - 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.

3.2 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 and 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 than 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.

3.3 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 3.4.
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

$$DT = max(similarity * EC_{weigth})$$
$$AT = (\#GO - 1) * GO_{weight}$$
$$AR : lowest.node(AS(DT + AT)) \geq threshold$$

Figure 3.4: Annotation Rule

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.

3.4 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.

3.5 GO-Slim

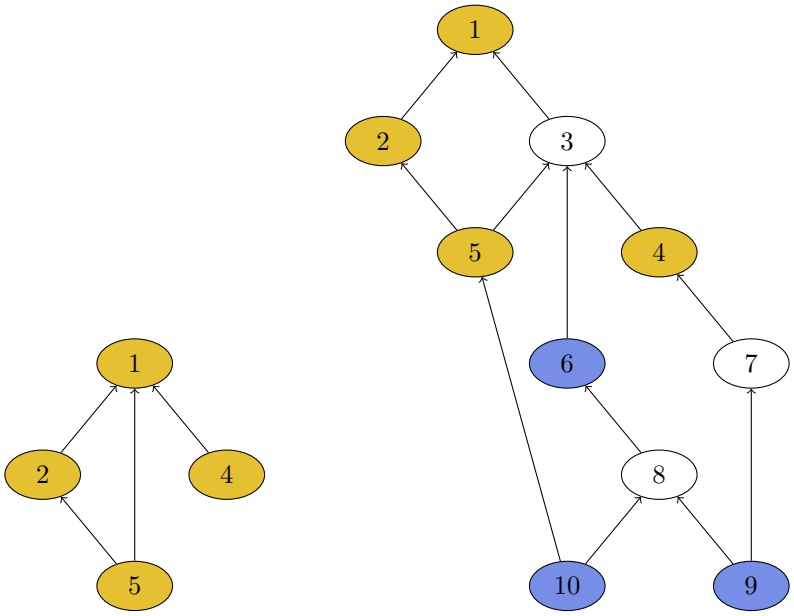
What is a GO Slim?
(Ref: Gene Ontology website, `http://geneontology.org/page/go-slim-and-subset-guide`)

> 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 organism-specific 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 3.5) is a small visual example.
Imagine figure 3.5(a) to be the subset of GO terms called GO Slim, figure 3.5(b) shows a data-set with GO 6,9 and 10 annotated. The GO Slim methodology will pull up the 3 annotated GOs as follows:

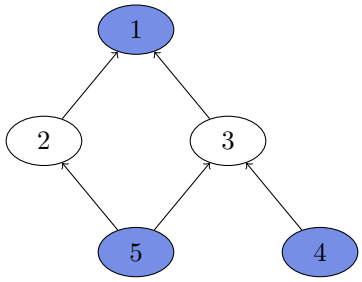
- $6 \rightarrow 1$
- $9 \rightarrow 4$
- $10 \rightarrow 5$

The result is shown in figure 3.5(c). 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.
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.



(a) GO Slim subset.

(b) Whole data-set with GOs 6,9 and 10 annotated.



(c) The final annotation of the data-set after applying the GO Slim.

Figure 3.5: 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.

4 Statistics (Charts)

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 quantitative/statistical charts in Blast2GO

- Project
  - 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.

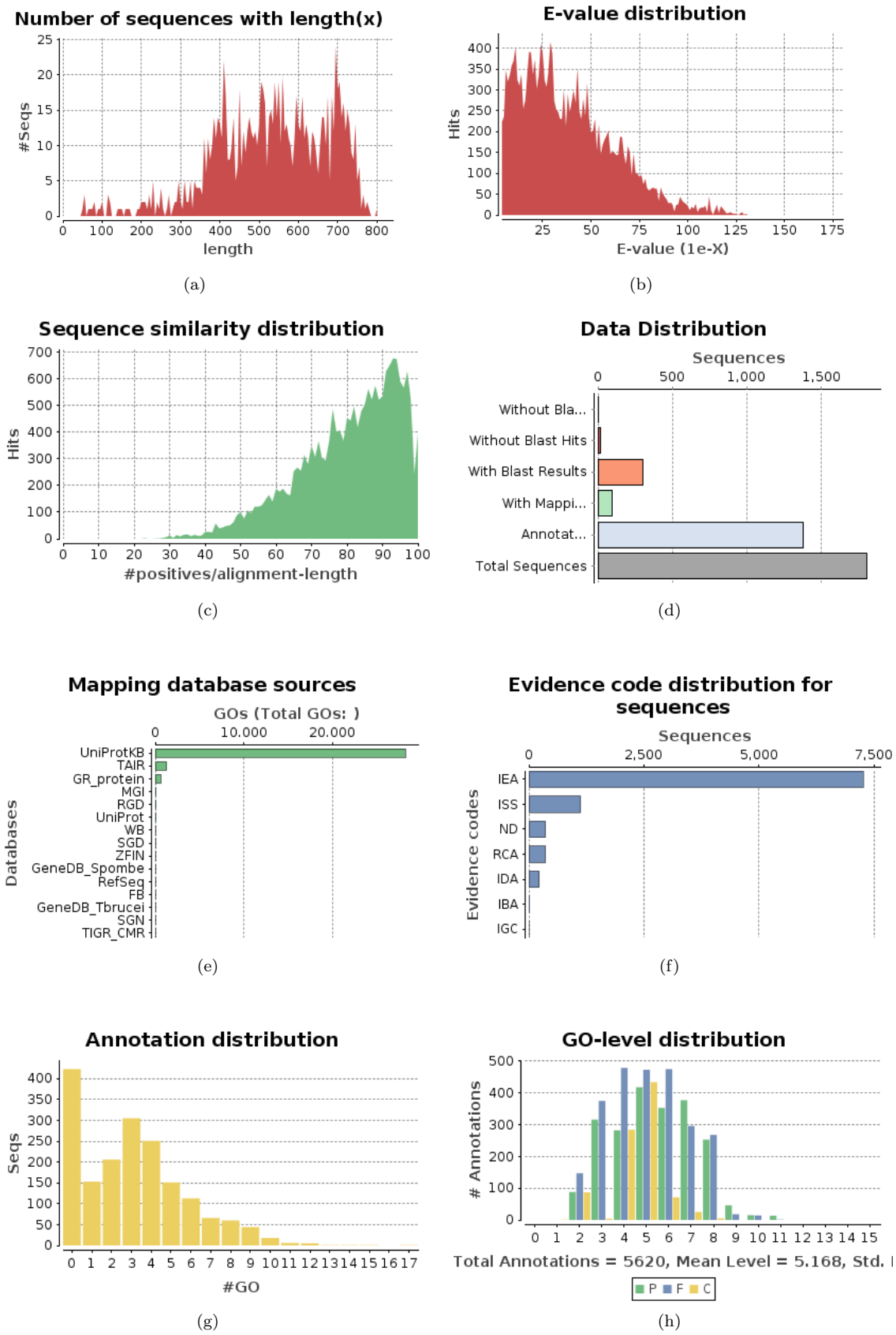


Figure 3.6: A collection of different Blast2GO Charts

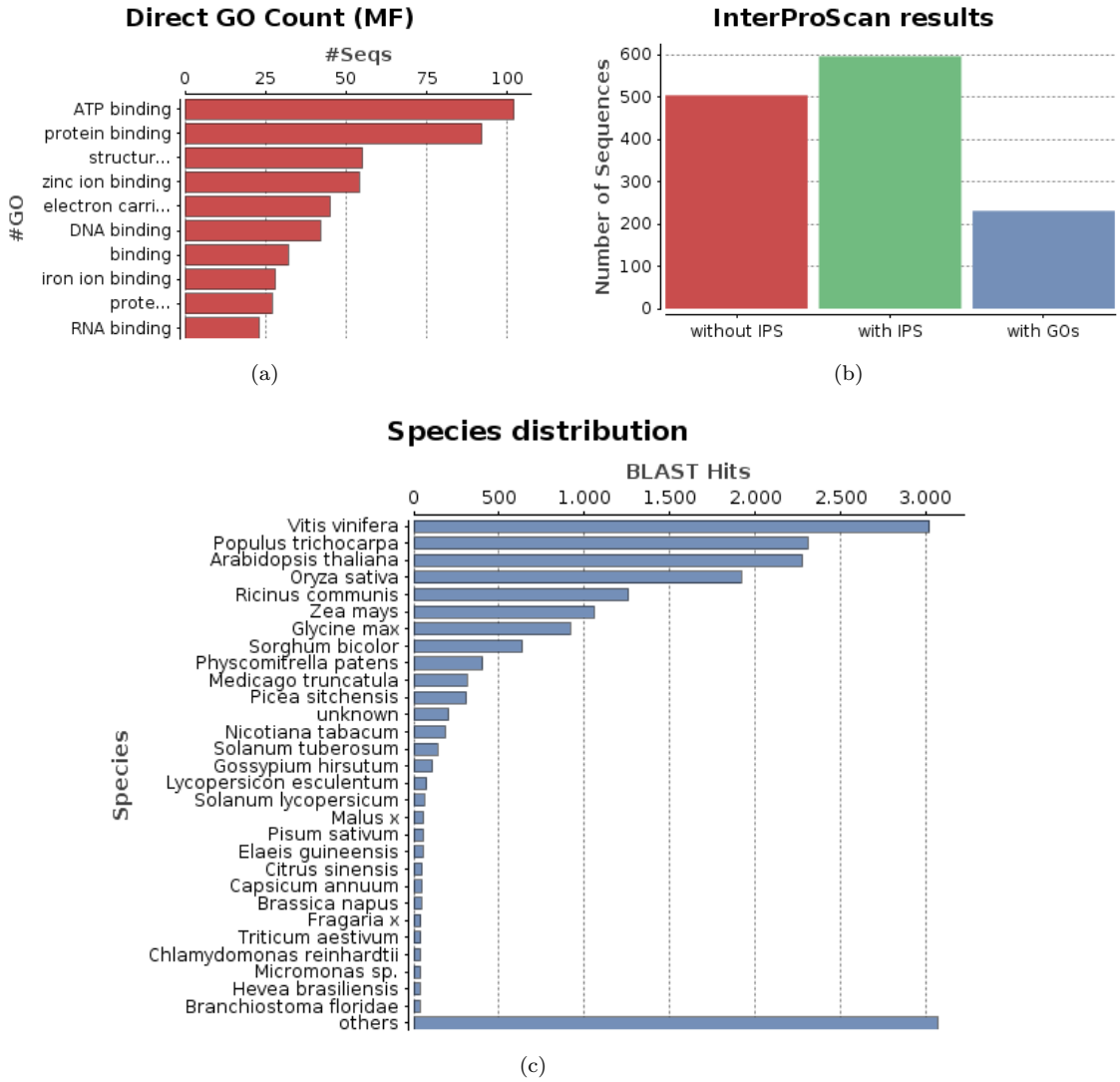


Figure 3.7: A collection of different Blast2GO Charts

- Analysis Progress:
  Gives an overview about the current analysis progress of the data-set.

- 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 Evalue cutoff in the annotation step.
  - 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.
  - HSP/HIT coverage:
    This chart shows a distribution of percentages. The percentages represent the cov-

erage between the HSPs and its corresponding HITs. This chart helps to get an understanding of the effect of this annotation parameter.

- 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.

- Mapping
  - 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.
  - 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.
- 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:
    TA 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 abode 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.
- InterProScan Statistics:
  This chart shows the effect of adding the GO-terms retrieved though the 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.

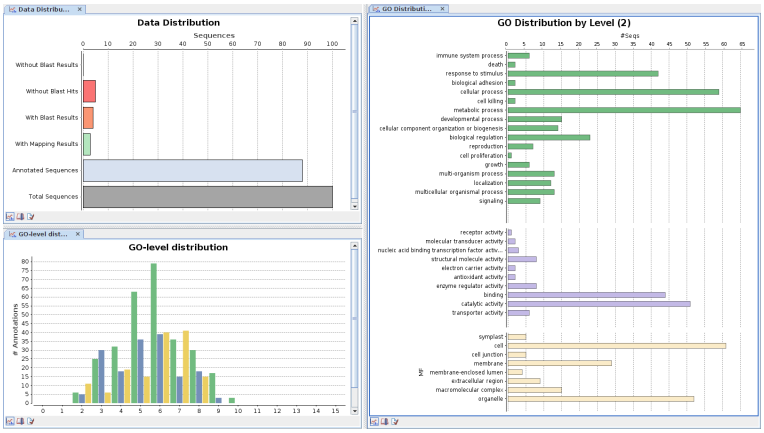


Figure 3.8: Different types of charts open next to each other in the workbench

5 Gene Ontology Graph Visualization

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 visualising 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". The 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 colour 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.

5.1 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 α [0,∞] to the power of the distance between the term and the term of direct annotation (see Equation 3.1 for a mathematical expression. 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 α. This compensates for the drawback of the earlier described method of simply counting the number of different sequences assigned to each GO-term. The α parameter allows this behaviour to be further adjusted. A value of zero means no propagation of information and can be increased by rising α.

$$score(g) = \sum_{g_a \in desc(g)} gp(g_a) \cdot \alpha^{dist(g,g_a)} \tag{3.1}$$

where:

- desc(g) represents all the descendant terms for a given GO term g
- dist(g,g$_a$) is the number of edges between the GO term g and the GO term g$_a$
- g is an element of the GO where GO is the overall set of all GO terms

- gp(g) is the number of gene products assigned to a given GO term g

5.2 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 loosing 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 GoSlim definitions.

In Figure 3.9, 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 GoSlim 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 α 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.

5.3 Combined Graph Editor Sidepanel

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.

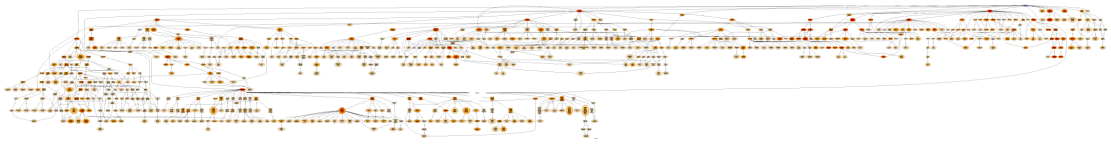
-  Cuts through the graph at a specific level and generates a pie representation of the number of sequences per GO node.
-  Allows to select a minimum filter value in order to include only GO nodes with a higher Node-Score or sequence count in the resulting pie chart.
-  Same as the first one but in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

All Charts will open in the **Geneious B2G Window** and can be saved in different file formats.

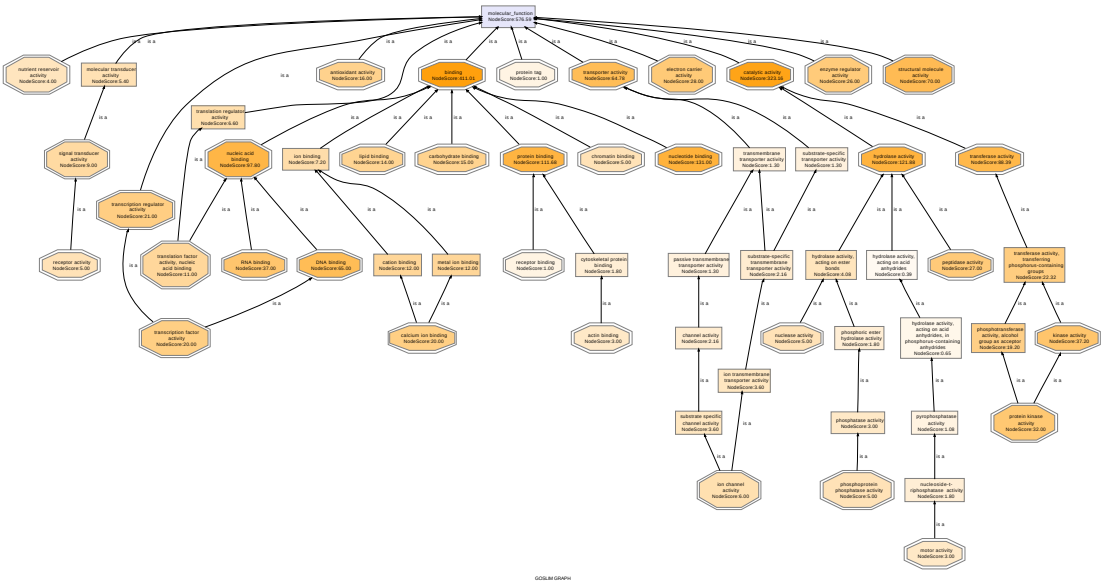
Graph Legend

The GO Graphs are displayed in different shapes (Figure 3.10).

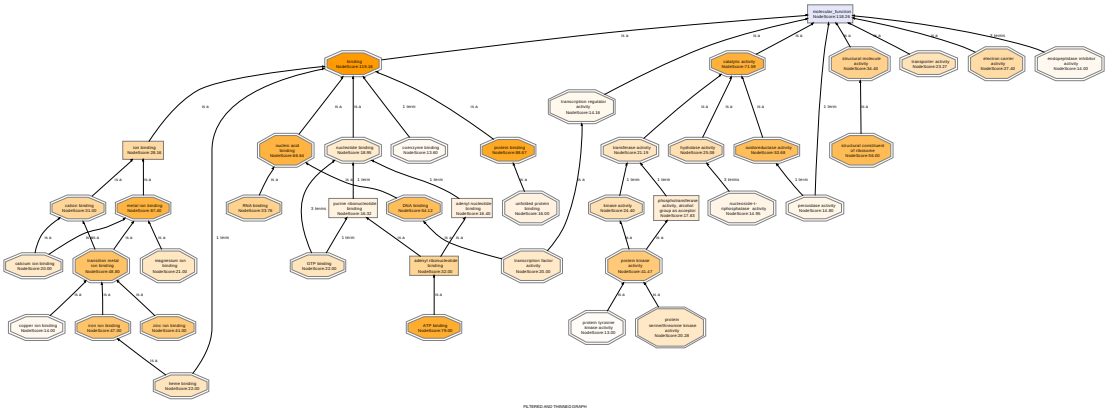
- **octagon** - Annotated GO Terms
- **square** - Intermediate GO Terms
- **ellipsis** - GO Terms linked to a Blast Hit



(a) Unfiltered Graph



(b) Filtered Graph 1



(c) Filtered Graph 2

Figure 3.9: The molecular functions of 1000 sequences visualized in 3 different ways: The first graph is unfiltered, the second graph shows the functional information after having applied a GoSlim reduction and 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 (tip nodes) and all intermediate terms with a node-score smaller than 12 (with α=0.4) were removed.

6 Data Import and Export Options

The import and export possibilities are grouped into two, standard and toolbox imports. 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.
It is important to notice that the .b2g format (available in the Blast2GO standalone version) is not yet compatible with this plugin. To interchange data between the Blast2GO Desktop

Figure 3.10: Graph Legend that shows the graph shapes

application and the plugin, please use the .dat format as before.

- Standard Import: The import via Import » Standard Import . . . allows to import .annot or .dat files.
- Toolbox Import: Blast and InterProScan xml 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 .dat, .annot, GAF, GFF, GeneSpring, GoStat and in WEGO format.
- 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 Sequence 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.

7 Fisher's Exact Test (Enrichment Analysis)

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 asses the functional differences between two sets of functional annotations (e.g. GO function of two groups of genes). This anaylsis 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 Grapf or as a bar chart, always colouring statistically significant terms in red (over-represented) and green (under-represented).

7.1 Run a Fisher's Exact Test

To perform the test we need to have a Blast2GO Project (or various) which contains the functional information of all seuquenes/genes to be included in the statistical test. In a second and third step we will select subsets of the first selection. The the second step the test-set is selected. This can be done by choosing a Blast2GO Project already loaded in the workbench or via a text file which contains the coresponding sequence names of gene-ids of the test set (one each line). In the third step we can than define a reference set. This step is optional and if no reference is selected the dataset in the first step minus the sub-set selected in the seconf step is choosen as a background or refenrece dataset.
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.

7.2 Parameters

- Step 1: Select one or more Blast2GO Projects
  Select one or more Blast2GO Projects which together contains the functional information of all sequences/genes included in the statistical test.

- Step 2: Select one or more Test-Sets from the navigation area
  The selected datasets will be combined to one. Please note that the given IDs have to match the sequence names of the Blast2GO Project selected in the first step. It is allowed to select Blast2GO Projects or sequence/gene ID lists in plain text format. In the case of the plain text format, please make sure to have only one ID per line. An example can be found in the Blast2GO example datasets.

- Step3: Select one or more Reference-Sets from the navigation area
  This is optional and the whole set selected in the first step will be used otherwise. The selected datasets will be combined to one. Please note that the given IDs have to match the sequence names of the Blast2GO Project selected in the first step. It is allowed to select Blast2GO Projects or sequence/gene ID lists in plain text format. In the case of the plain text format, please make sure to have only one ID per line. An example can be found in the Blast2GO example datasets.

- Step 4: Configure Parameters

  - 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.

7.3 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.

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.

7.4 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.

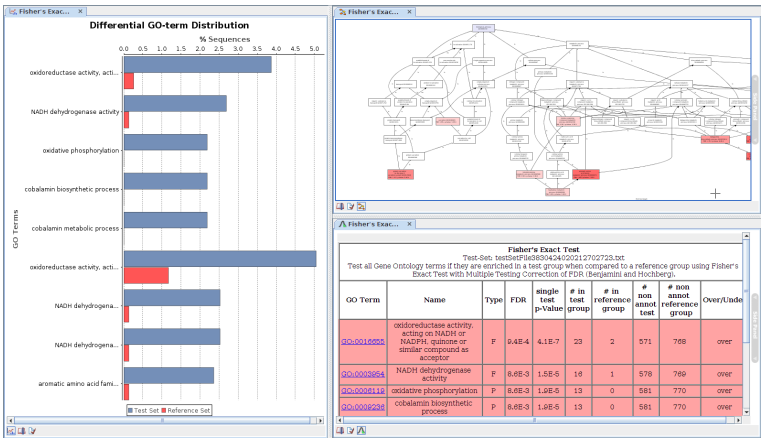


Figure 3.11: Different types of Fisher's Exact Test results

8 Manage Projects

8.1 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 with skip ids" is the same as "Project 2, Project 1 with overwrite ids".

8.2 Convert Data to Blast2GO Project

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

- Nucleotide Sequence(s)
- Protein Sequence(s)
- BLAST-Results

9 Miscellaneous

9.1 Validate Annotation

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.

9.2 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.

9.3 Create Annotation Table

This function allows to create an CLC-bio Annotation Table containing the Gene Ontology terms generated with Blast2GO .

9.4 Annex Annotation Augmentation (legacy)

Annex [**?**] 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 Bast2GO. Generally, between 10% and 15% extra annotation is achieved and around 30% of GO term confirmations are obtained through the Annex data-set.

9.5 Create Blast2GO Example Dataset

This functions allows to add several small example data-set to the Navigation Area in the Workbench. Each file contains just 10 sequences which allows to easily explore the different possibilities of the plugin.

10 Workflows

All major Blast2GO plugin functions are workflowable. 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 and for more detailed information please refer to the ClC bio Workbench user manual. The Blast2GO example dataset (available form 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.*
   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 3.12.

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 when ever 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 sicne 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.

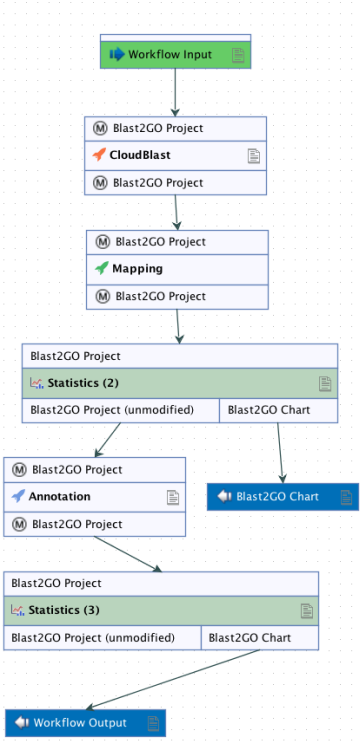


Figure 3.12: Basic workflow example which performs a blast, mapping and annotation step and generates some basic summary statistics.

Please Cite

- A. Conesa, S. Götz, J. M. Garcia-Gomez, J. Terol, M. Talon and M. Robles. **"Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research"**, Bioinformatics, Vol. 21, September, 2005, pp. 3674-3676.
- A. Conesa and S. Götz. **"Blast2GO: A Comprehensive Suite for Functional Analysis in Plant Genomics"**, International Journal of Plant Genomics, Vol. 2008. 2008, pp. 1-13.
- S. Götz et al. **"High-throughput functional annotation and data mining with the Blast2GO suite"**, Nucleic Acids Research, Vol. 36, June, 2008, pp. 3420-3435.
- S. Götz et al. **"B2G-FAR, a species centered GO annotation repository"**, Bioinformatics, Vol. 27 (7), 2011, pp. 919-924.