

Blast2GO Tutorial

Ana Conesa, Stefan Götz
June 2009



Bioinformatics and Genomics Department
Prince Felipe Research Center

Valencia
SPAIN

Contents

1	Introduction	2
2	Start Blast2GO	3
3	Blast2GO User Interface	4
4	Quick Start	5
5	Load Sequences	7
6	BLASTing	8
6.1	Export BLAST results	9
6.2	Import BLAST results	10
6.3	View BLAST results	10
7	Mapping	11
8	Annotation	12
8.1	GO annotation	12
8.2	InterPro annotation	13
8.3	Enzyme code annotation and KEGG maps	14
8.4	Modulate Annotation Intensity	14
8.5	Exporting Annotation	15
8.6	Importing Annotation	15
9	Single Sequence Menu	16
10	Visualization	18
10.1	Coloring on the Main Sequence Table	18
10.2	Directed Acyclic Graphs	18
10.3	Statistical charts	19
10.4	Pies and Bar Charts	19
11	Quantitative Analysis	20
11.1	Descriptive analysis. Combined Graph Function	20
12	Statistical Analysis	23
13	Other Functions	25
14	Blast2GO for advanced users	27

1 Introduction

Blast2GO (B2G) (Conesa et al., 2005) is a comprehensive bioinformatics tool for the functional annotation and analysis of gene or protein sequences. The tool was originally developed to provide a user-friendly interface for Gene Ontology The_gene_ontology_consortium (2008) annotation. Recent improvements have considerably increased the annotation functionality of the tool and currently Enzyme code (EC), KEGG Maps and InterPro motifs are also supported (Götz et al., 2008). Additionally the application offers a wide array of graphical and analytical tools for annotation manipulation and data mining.

The main concept behind the developments is the easiness for biological researchers: minimal setup requirements, automatic updates, simplicity in the usage and visual-oriented information display. Advanced functionalities are also available requiring a minimal computational background for setting up.

Basically, Blast2GO uses local or remote BLAST searches to find similar sequences to one or several input sequences. The program extracts the 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 are directly queried at the InterProScan web service. GO annotation can be visualized reconstructing the structure of the Gene Ontology relationships and ECs are highlighted on KEGG maps.

A typical use case of Blast2GO basically consists of 5 steps: BLASTing, mapping, annotation, statistics analysis and visualization. These steps will be described in this document including installation instructions, further explanations and information on additional functions.



Figure 1: Blast2GO (v.2)

2 Start Blast2GO

Blast2GO is an operating system independent Java Application made available via Java Web Start (JWS). By JWS technology Blast2GO can be started with a single click over the network. JWS ensures the most current version of the application as well as the correct version of the Java Runtime Environment (JRE). Blast2GO has therefore only a few requirements:

1. Internet connection
2. Java 1.5 JRE or higher (which includes the JWS application)
3. Open network port 3306 for a direct MySQL connection to the Blast2GO database. In case this is impossible for you due to local network security restrictions a local database installation is possible.

To start Blast2GO for the first time go to <http://www.blast2go.com> ->Start Blast2GO. If you do not have a Java Runtime Environment (JRE) on your machine, you can download it from the official JAVA web site at <http://java.sun.com>. Start and download the Blast2GO version that better fits the memory amount of your PC.

Direct Desktop Link to Blast2GO: After your first Blast2GO installation you will find the application in the JWS application interface. Here you have the possibility to create a Desktop Icon as shown in Figure 2.

Software updates: After the first, initial installation JWS will store the whole Blast2GO application on your system. When starting Blast2GO again JWS will check online for updates and download (only) them, if available. This guarantees that you are using always the latest version of Blast2GO.

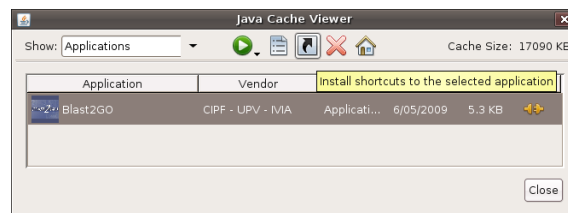


Figure 2: Java(TM) Web Start 1.6.0 Interface

3 Blast2GO User Interface

There are three basic sections in the Blast2GO main user interface (Figure 3):

1. Menu Bar: The main application menu contains 10 tabs:
 - File: Hosts functions for opening, saving and closing Blast2GO projects, and for data import/export in different formats.
 - BLAST: Contains functions for performing BLAST searches and resetting results.
 - Mapping: This function fetches GO terms associated to hit sequences obtained by BLAST.
 - Annotation: Includes different functions to obtain and modulate GO, computing GoSlim view, Enzyme Code annotation with KEGG maps and InterPro annotation.
 - Analysis: This tab hosts different options for the analysis of the available functional annotation. Includes graphical exploration through the Combined Graph Display and performing statistical analysis of GO distributions for groups of sequences
 - Statistics: This tab offers different descriptive statistics charts for the results of BLAST, mapping and annotation.
 - Select: Allows to make sequence selections based on the sequences status (colors), names, GO terms and descriptions as well as delete selected sequences from a project.
 - Tools: Miscellaneous tools to perform special annotation manipulations, a Java memory monitor, database configuration etc...
 - View: Switch the main table view from GO-IDs to term, show only selected sequences, show/hide GOs of InterProScan results and color highlight the different GO categories.
 - Info: Information about Blast2GO and on-line availability of documents and examples.
2. Main Sequence Table: This table shows the details and progression of the analysis for the loaded sequence data set. Each row represents a query sequence and has the following fields, which are filled with information as it is generated by the application:
 - Check box.
 - Sequence Name.
 - Hit description.
 - Sequence Length.
 - Number of hits.
 - E-value of the best BLAST hit.
 - Mean similarity value for the BLAST results. This value is computed as the average hsp-similarity value for all the hits of a given sequence.
 - Number of mapped GOs.
 - GO IDs associated to the sequence (GO Type plus GO ID).
 - Enzyme code (EC) associated to the sequence.
 - InterProScan results.
3. Result Tabs: This area of the application show results and messages of different analysis. There are five tabs:
 - Ontology Graphs: Hosts the display for single sequence, combined and enriched graphs.
 - Application Messages: Displays information on the progress of the analysis.
 - BLASTs Results: Hosts the display of the BLAST Browser for individual sequences.
 - Statistics: This tab shows the results of different statistical analysis performed by the application.
 - KEGG Maps. Here you can visualize KEGG maps associated to one or several sequences.

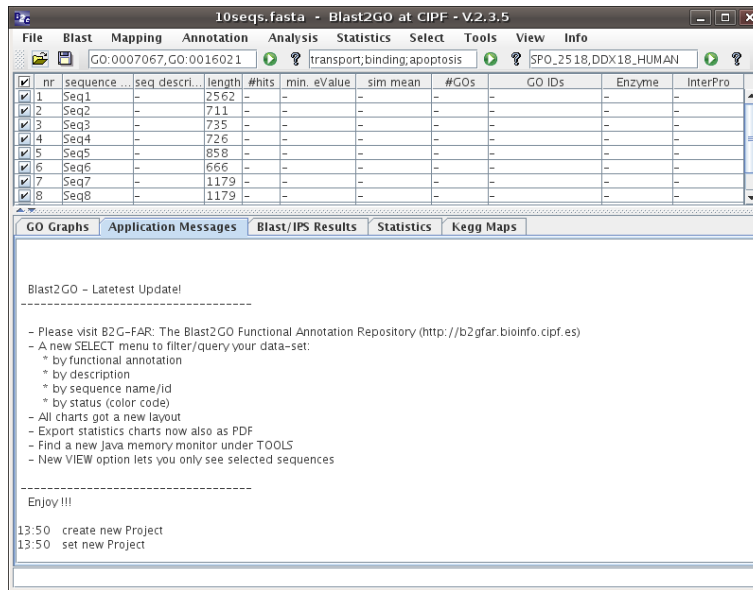


Figure 3: Blast2GO Interface

4 Quick Start

This section gives a quick survey on a typical Blast2GO usage. Detailed descriptions of the different steps and possibilities of the application are given in the remaining sections of this Tutorial.

1. Load data: Go to “File” -> “Load FASTA File” and select your *.fasta* file containing the set of sequences in FASTA format. Several example file can be downloaded from the Blast2GO site. Alternatively you can load 10 example sequences into Blast2GO choosing “Load 10 Example Sequences” in the “File” menu.
2. BLAST: Go to “Blast” -> “Make BLAST”. At the BLAST Configuration Dialog (Figure 4) select the type of BLAST mode which is appropriate for your sequence type (blastx for nucleotide and blastp for protein data) and click on the top arrow to start the BLAST search against NCBI’s non redundant NR database.
 - Once your BLAST analysis is finished visualize your results at “Statistics” -> “BLAST Statistics”.
 - On the Main Sequence Table, right-click on a sequence to open the Single Sequence Menu (Figure 15). Select Show BLAST Result to open the BLAST Browser for that sequence.
3. Mapping: Go to “Mapping” -> “Run GO-Mapping Step” and click on the top arrow to start mapping GO terms. Mapped sequences will turn **green**. Once Mapping is completed visualize your results at “Statistics” -> “Mapping Statistics”.
4. Annotation: Go to “Annotation” -> “Run Annotation Step” and click on the top arrow to start the annotation. Annotated sequences will turn **blue**.
 - Once Annotation is completed visualize your results at “Statistics” -> “Annotation Statistics”.
 - On the Main Sequence Table, right-click on a sequence to open the Single Sequence Menu. Select “Draw Graph of Annotations” to visualize the annotation on the GO DAG for that sequence.
 - If desired, modify the annotation by clicking with the left mouse button and select Change Annotation or change the extent of annotation by adding implicit terms (“Annotation” -> “Run ANNEX”) or reducing to a GO-Slim representation (“Analysis” -> “Run GO-Slim”).

- During the annotation process, Enzyme Codes will be also given when a GO-term/EC number equivalence is available.
 - Optionally, “InterProScan” -> “Run InterProScan” to obtain InterPro annotations is available and highly recommended to improve the annotation. Once InterProScan results are retrieved use “Merge InterProScan GOs to Annotation” to add GO terms obtained through motifs/domains to the current annotations.
5. Enrichment Analysis: Blast2GO provides tools for the statistical Analysis of GO term frequency differences between two sets of sequences. Go to “Analysis” -> “Enrichment Analysis” -> “Make Fisher’s Exact Test”. A new Dialog window is opened (Figure 25). Select a *.txt* file containing a sequence IDs list for a subset of sequences. A test-set example file can be downloaded from the Blast2GO web site. Select a second set of sequences as reference/background set if desired or skip this step and the whole actual loaded set of annotations present in Blast2GO will be used as reference. Click on run button to start the analysis. A table containing the results of this test will appear on the “Statistics Tab”.
- Click on “Make Enriched Graph” to visualize the results of the Fisher’s Exact Test on the GO DAG.
 - Click on “Bar Chart” to obtain a bar chart representation of GO term frequencies.
6. Combined Graph: Blast2GO can visualize the combined annotation for a group of sequences on the GO DAG. Select a group of sequences to generate their combined graph at “Tools” -> “Select Sequences by Names”. You can use the Demo Test Set used previously for this. Alternatively, you can select sequences manually using the sequence check boxes of the Main Sequence Table. Go to “Analysis” -> “Make Combined Graph”. Set on the Seq Filter of the Combined Graph Dialog the minimal number of sequences collected at a GO term for its node to be shown in the generated graph. If you are using the Blast2GO Test Set example, set this value to 100. Click on the top arrow to generate the graph.
7. Save Results:
- “File” -> “Save B2G-Project” saves the current Blast2GO project as *.dat* file.
 - “File” -> “Export” allows to export the generated data in many different formats. “Export Annotation” exports the actual annotation results as *.annot* file.
 - “Enrichment Analysis” -> Export Results: exports results of the Fisher’s Exact Test as a tabulator separated text file.
 - To save graph information, use the little icons/buttons on the corresponding graph windows. Graphs can be saved/exported as *.pdf*, *.png*, *.svg* and *.txt*.

5 Load Sequences

To start a new Blast2GO project you just have to load your sequence data from a file into blasy2GO. At the “File” menu, go to “Load FASTA File” and select the file containing your sequences. The application accepts text files containing one or more DNA or protein sequences in FASTA format. These files must have the extension *.fasta*, *.fnn*, *.faa*, *.fna* or *.ffn* to be accepted by the application. A sequence in FASTA format begins with a single-line description or header starting with a “>” character. The rest of the header line is arbitrary but should be informative. Subsequent lines contain the sequence, one character per residue. Lines can have different lengths. Be sure your file is in this format and avoid strange characters in the sequence header, such as ‘&’ or ‘

’ and use ‘N’ to denote in-determinations in the sequences.

An example for the FASTA format:

```
>gi|121664|sp|P00435|GSHC BOVIN GLUTATHIONE PEROXIDASE  
MCAAQRSAAALAAAAPRTVYAFSARPLAGGEPFNLSSLRKGKVLLENVASLUGTTVRDYTQMND  
LQRLGPRGLVVLGFPCNQFGHQENAKNEEILNCLKYVRPGGGF
```

In case you already have a project-file (*.dat*) just open it by selecting the file-type *.dat* in the “Load File” dialog.

6 BLASTing

Blast2GO uses the Basic Local Alignment Search Tool (BLAST) to find sequences similar to your query set. Please, refer to <http://www.ncbi.nlm.nih.gov/BLAST> for details on the BLAST function. Figure 4 shows the BLAST Configuration Dialog Window (Figure 4 that controls the BLAST step. Here, the user can specify the following parameters:

- BLAST server URL: The URL of the BLAST Server you wish to use. If you want to add here your own BLAST server direction edit the Blast2GO properties file.
- BLAST DB: The name of the database to search in (eg. nr, swissprot, pdb). To see a list of possible DBs at NCBI see http://www.ncbi.nlm.nih.gov/staff/tao/URLAPI/remote_blastdblist.html
- Number of BLAST hits: The number of alignments you want to achieve. (0-100)
- BLAST expect value: The statistical significance threshold for reporting matches against database sequences. If the statistical significance ascribed to a match is greater than the EXPECT threshold, the match will not be reported. Lower EXPECT thresholds are more stringent, leading to fewer chance matches being reported. Increasing the threshold shows less stringent matches.
- BLAST program: The algorithm you want to use:
 - blastp - Compares an amino acid query sequence against a protein sequence database.
 - blastn - Compares a nucleotide query sequence against a nucleotide sequence database.
 - blastx - Compares a nucleotide query sequence translated in all reading frames against a protein sequence database. Used to find potential translation products of an unknown nucleotide sequence
 - tblastn - Compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames.
 - tblastx - compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.
- BLAST mode: To make a BLAST search at NCBI use QBLAST (Queue BLAST is a service offered by NCBI to make multiple queries), to make a BLAST search on an own Server use WWW-BLAST.
- Your e-mail address in case you are using the NCBI BLAST web service.
- HSP length cutoff: A Cutoff value for the minimal length of the first hsp of a blast hit, used to exclude hits with only small local alignments from the BLAST result. The given length corresponds to amino-acids or nucleotides depending the type of performed BLAST.
- Low complexity filter: The BLAST programs employ the SEG algorithm to filter low complexity regions from proteins before executing a database search. Default is ON.
- Save results as: Choose a format type to additionally save your BLAST results. It is recommended to save your BLAST results as xml as this format is supported by the Blast2GO Import BLAST Results function.
- BLAST Description Annotator: The BDA finds the best possible description for a new sequence based on a given BLAST result.
- Try SIMAP first: The Simap database of sequence alignments (Rattei et al., 2008) maintains the worldwide biggest "all against all" matrix of protein sequence alignments and allows Blast2GO to obtain within seconds for all sequences from RefSeq, GenBank, UniProt, PDB and many more sources an exact sequence alignments. Sequences not found in Simap will be blasted as usual against the chosen server. The way to find the corresponding sequences is directly based on a MD5 hash checksum made from the sequence string.

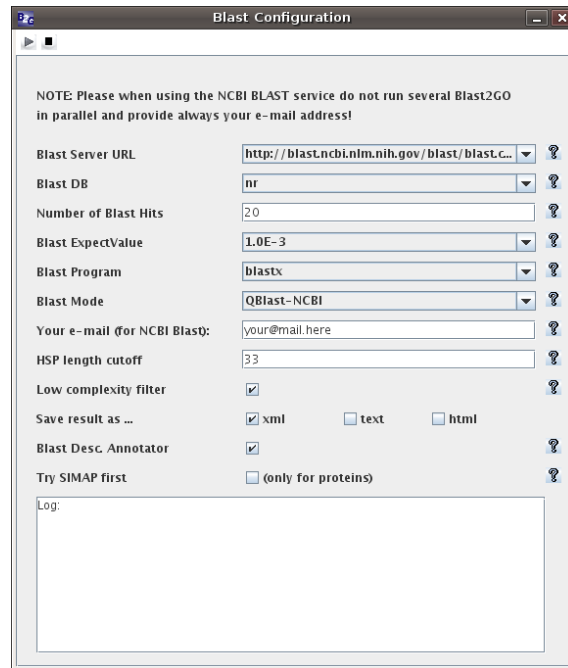


Figure 4: BLAST Configuration Dialog

BLAST in Blast2GO can basically be performed in two different fashions:

1. Qblast@NCBI. NCBI offers a public service that allows searching molecular sequence databases with the BLAST algorithm. The main advantages of making use of this service are its versatility and that no database maintenance is required. Therefore by selecting this option at Blast2GO no additional installations have to be done.
2. WWW-BLAST. Alternatively, BLAST can be done locally against a custom database. For this, you need to place a copy of your FASTA formatted custom DB plus a WWW-BLAST installation on a local BLAST server and indicate Blast2GO their location. This is done by editing your b2g.properties file (you will find this file e.g. under Linux at /home/yourname/blast2go/blast2go.properties) and adding both your local blast.cgi URL and your database name at the Blast.urls and Blast.Blastdbs lines, respectively (see Section 14). These options will then be available for selection at the BLAST Configuration Dialog. The WWW-BLAST program and several specific sequence databases can be found at NCBI.

6.1 Export BLAST results

The results of the BLAST queries can also be directly saved to a file in different formats by selecting the corresponding check boxes at the BLAST Configuration Dialog. If the chosen file already exists, upcoming results will be appended. As the BLAST search progresses, sequences with successful BLAST results change their color on the Main Sequence Table from white to **light-red** and the BLAST result related columns will be filled. In case no results could be retrieved for a given sequence, this row will turn **dark-red**.

At any point of the progress of the BLAST search, three different charts (Figure 5, 6 and 7) can be generated for a global visualization of the results. These charts provide a general view of the similarity of the query set with the selected databases and can be used to choose cut-off levels for the e-value, similarity and annotation threshold parameters at the annotation step. Additionally a BLAST hit species distribution chart is available.

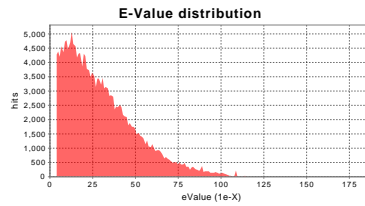


Figure 5: E-Value Distribution

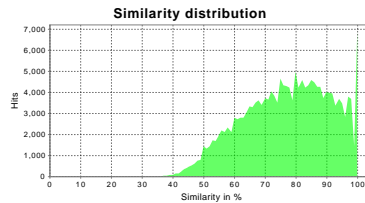


Figure 6: Similarity Distribution

6.2 Import BLAST results

If a BLAST result is already available in xml format, this can be directly loaded into Blast2GO by using “Import” -> “Import BLAST Results” in the File menu. You can choose here to import a whole directory containing a collection of BLAST XML file or to select a single XML file. The BLAST results will be added to your current Blast2GO session.

6.3 View BLAST results

By a mouse right-click at the Main Sequence Table, the Single Sequence Menu will appear (Figure 15). This menu provides some functions for sequences individually, i.e. will apply to the sequence at that position of the Table. Show BLAST Results will generate a table in the BLAST Results Browser containing information on the results of the similarity search of the selected sequence. Detailed description of this and other functions available in the Single Sequence Menu is given in Section 9.

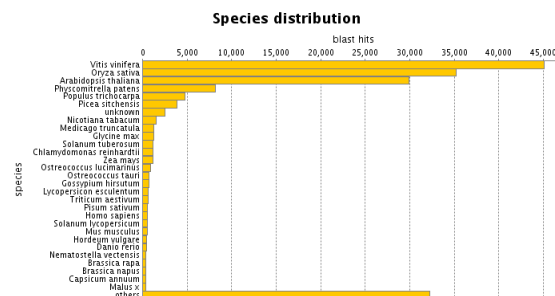


Figure 7: Species Distribution

7 Mapping

Mapping is the process of retrieving GO terms associated to the hits obtained after a BLAST search. Blast2GO performs four different mappings:

1. BLAST result accessions are used to retrieve gene names or Symbols making use of two mapping files provided by NCBI (gene_info, gene2accession). Identified gene names are then searched in the species specific entries of the gene-product table of the GO database.
2. BLAST result GI identifiers are used to retrieve UniProt IDs making use of a mapping file from PIR (Non-redundant Reference Protein Database) including PSD, UniProt, Swiss-Prot, TrEMBL, RefSeq, GenPept and PDB.
3. Accessions are searched directly in the dbxref table of the GO database.
4. BLAST result accessions are searched directly in the gene-product table of the GO database.

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 and the sequence row position will turn **light-green**. Assigned GOs to hits can be reviewed in the BLAST Results Browser (see Section 9 and Figure 16).

Three evaluation charts become available in the Statistics menu after the mapping step: The “DB resources of mapping” (Figure 10) shows from which database annotations had been obtained and the Evidence Code distribution for hits and sequences (Figure 8 and 9) indicated how EC associate in the obtained GO pool. Note that commonly IEA (electronic annotation) is overwhelmed in the mapping results. However, the contribution of this (and other) type of annotation to the finally assigned annotation to the query set can be modulated at the annotation step.

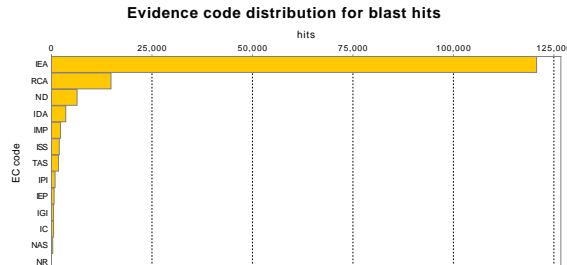


Figure 8: Evidence Code Distribution of BLAST hits

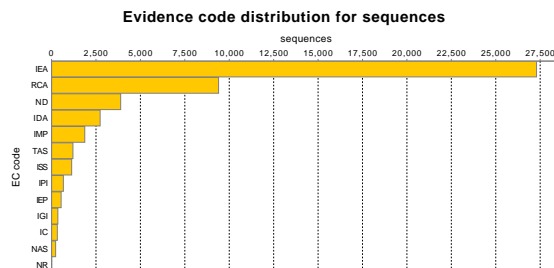


Figure 9: Evidence Code Distribution for sequences

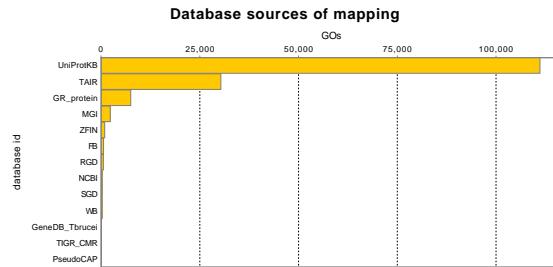


Figure 10: Source DBs urces of Mapping

8 Annotation

8.1 GO 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. In the current Blast2GO version this is the core type of functional annotation.

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 Figure 11.

$$\begin{aligned}
 DT &= \max(\text{similarity} \times EC_{\text{weight}}) \\
 AT &= (\#GO - 1) \times GO_{\text{weight}} \\
 AR &: \text{lowest.node}(AS(DT + AT)) \geq \text{threshold}
 \end{aligned}$$

Figure 11: 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. These can be adjusted in the Annotation Configuration Dialog (Figure 12) and in the Evidence Code Weight Configuration Dialog (Figure 13).

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-HitCoverage 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. EC code weights can be modified at Annotation ->Evidence Code Weights. 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.

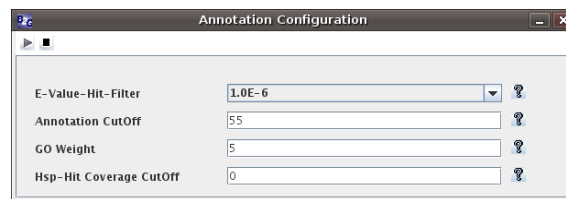


Figure 12: Annotation Configuration

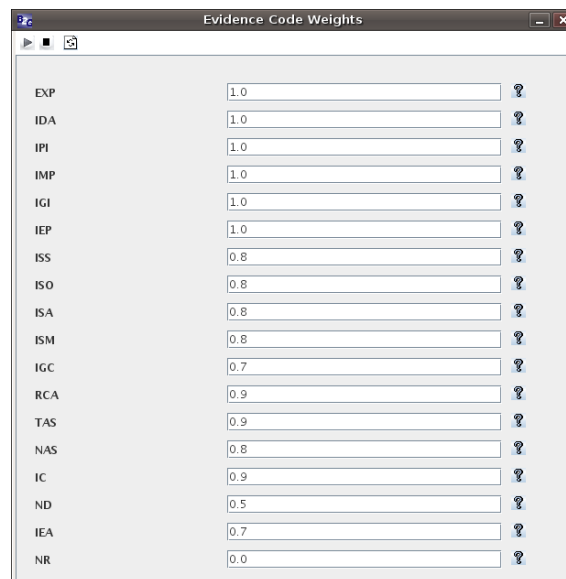


Figure 13: Evidence Code weight configuration

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 (Figure 14), which shows the number of sequences annotated at different amounts of GO-terms.

8.2 InterPro annotation

The functionality of InterPro annotations in Blast2GO allows to retrieved domain/motif information in a sequence-wise manner. Corresponding GO terms are then transferred to the sequences and merged with already existent GO terms. IPRscan results are saved in a given directory and can be viewed through the Single Sequence Menu.

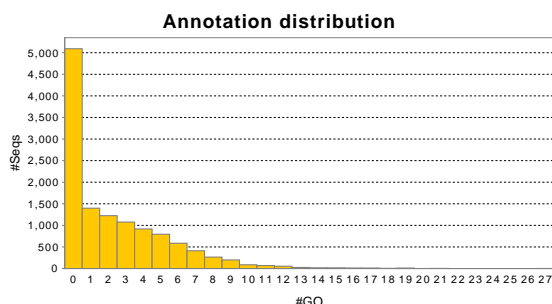


Figure 14: Annotation Distribution

8.3 Enzyme code annotation and KEGG maps

Blast2GO provides EC annotation through the direct GO ->EC mapping file available at the GO web site. This means that only sequences with GO annotations will eventually show also EC numbers and that the GO annotation accuracy can be made extensive to Enzyme annotations. Additionally, the KEGG map module allows the display of enzymatic functions in the context of the metabolic pathways in which they participate.

Select the sequences of your interest and go to “Analysis” -> “Enzyme Code and KEGG” -> “Load KEGG maps”. The application will search all KEGG maps containing the EC numbers of the selected sequences and make them available at the KEGG Maps tab. The list of found KEGG maps will appear at the upper left frame. By double-clicking on a given pathway, it will be loaded on the left graphical window. Sequences and EC codes contained in that pathway are shown in the lower frame, and highlighted with different colors (one color for each EC) in the pathway map.

8.4 Modulate Annotation Intensity

Blast2GO offers the possibility of refining GO annotation by two different methods:

8.4.1 Augment Annotation by Annex

Blast2GO integrates the Second Layer Concept developed by the Norwegian University of Science and Technology (Myhre et al., 2006) for augmenting GO annotation. Basically, this approach uses uni-vocal relationships between GO terms from the different GO Categories to add implicit annotation. In Blast2GO you can bind this option under the “Annotation” menu. For more details visit the Annex Project at <http://www.goat.no>.

8.4.2 Generate GO-Slims

GO-Slim is a reduced version of the Gene Ontology that contains a selected number of relevant nodes. The “Change to GO-Slim View” function (under the Annotation menu) generates a GO-Slim mapping for the available annotations. Different GO-Slims are available which are adapted to specific organisms. Blast2GO supports the following GO-Slim mappings: General, Plant, Yeast, GOA (GO-Association) and TAIR.

8.4.3 Other options

Annotation results for each sequence can also be visualized on the GO DAG by selecting “Draw Graph of Annotations” at the “Single Sequence Menu”. Additionally, the “Annotate Seq” and “Change Annotation and Description” options of this menu offer also the possibility to adjust annotations specifically for a single sequence (see Section 9).

8.5 Exporting Annotation

The annotation results can be exported in a variety of formats. This function is available at “File” -> “Export” -> “Export Annotations”.

1. *.annot*. This is the default option for Annotation export and the exchange annotation format in Blast2GO. Annotations are provided in a three-column fashion. The first column contains the sequence name, the second the annotation code and the third the sequence description. When multiple annotations for the same sequence are available, these come in subsequent rows. GO and EC annotations are exported jointly in the same format.
2. by Seq. One single row is given by sequence and all annotations are separated by commas.
3. by GO. One single row is given by GO term and sequences are separated by commas.
4. GoStats format. One single row is given by sequence and GO terms are only denoted by entire numbers (“GO:“ and left zero’s are skipped)
5. Genespring format. One single row is given by sequence are three different columns are provided for Molecular Function, Biological Process and Cellular Component. GO terms are denoted by their description rather than by their code.
6. etc.

8.6 Importing Annotation

Already made or existent annotation can be imported using the *.annot* format. For import purposes only, the *.annot* format allows also multiple annotations of the same sequence to be given in one single row, separated by commas, as shown above (Schema: Seq-Name <tab>GO(s) or EC(s) <tab>Sequence description):

```
Blast2GO Annotation File (.annot)
Seq1  GO:0001234          glycolipid transfer protein-like
Seq1  GO:0001264,G0:0004567,...
Seq1  GO:0034567
Seq1  EC:2.1.2.10
Seq2  GO:0001234,...      sorbitol transporter
Seq2  GO:0001244
Seq3  GO:0001234,G0:0004567,G0:0009123
Seq3  EC:1.2.4.1, EC:3.1
....
```

9 Single Sequence Menu

Blast2GO has a number of single-sequence based functionalities to allow curators an easy evaluation and refinement of individual annotations. The Single Sequence Menu is opened by a right-button mouse click on a given sequence row in the sequence table.

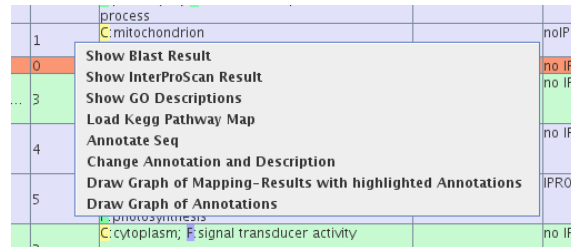


Figure 15: Single Sequence Menu

1. Show BLAST Result. A table on the BLAST Result Browser will be generated that displays information of the BLAST result for the selected sequence. For each of the obtained hits, the following information is given: Hit id and definition Gene name assigned to the hit by its accession e-value of the alignment Alignment length of the longest hsp Positive matches of the longest hsp Hsp similarity of hit: Number of hsp mapped GO-Terms with its evidence code UniProt codes of the hit sequences
2. Show GO annotations. GO ID, description, type and definition are given for all GO terms associated with the selected sequence. The GO ID is linked to the Amigo browser at the Gene Ontology site while the show option displays the DAG representation of the GO term.
3. Load KEGG map pathway. Shows the KEGG maps associated to the sequence (see section 8 for details)
4. Annotate seq. This function allows changing annotation parameters for the selected sequence and re-running automatic annotation.
5. Change annotation and description. This function edits the annotation of the selected and allows typing and deleting of annotation or sequence description. A manual annotation check-box is available for marking sequences with manual annotation. The sequence will get the violet label on the Main Sequence Table.
6. Draw Graph of mapping results with highlighted annotations. Displays the DAG for all GO terms related to that sequence, both the result of the mapping step as the derived annotations, the last ones being highlighted. At the Single Graph Drawing Configuration Dialog (Figure 13), the user can set the parameters for this graph (see section 10 for more details about Visualization in Blast2GO) Hit Filter. Nodes can be filtered out by number of hits: only nodes with more than a given number of BLAST-Hits will be shown in the graph. Pre-e-value-Hit-Filter. Only GOs obtained from hits with a greater e-value than the given value will be shown in the graph. Pre-Similarity-Hit-Filter. Only GOs obtained from hits with a greater similarity value than the given threshold will be shown in the graph. Graph coloring: By Annotation Score: Nodes with a higher annotation score will be more intensively colored. By hit count: Node color intensity will be proportional to the number of contributing hits.
7. Draw Graph of annotations. Display the DAG of all GO annotations of the sequence.

10 Visualization

Blast2GO aims to be a visual-oriented tool. This means that special attention is paid to show information through graphs, coloring and charts.

10.1 Coloring on the Main Sequence Table

The color shown by a sequence on the Main Sequence Table indicates the processing step reached by that sequence. The current coding is:

1. White: Non-processes sequence
2. Dark-red: Sequence blasted with a negative BLAST result
3. Light-red: Positive result obtained, no mapping available
4. Light-green: Mapping available
5. Blue: Annotation available
6. Violet: Manually annotated sequence
7. Yellow: GO-Slim view

10.2 Directed Acyclic Graphs

Blast2GO offers the possibility of visualizing the hierarchical structure of the gene ontology by directed acyclic graphs (DAG). This functionality is available to visualize results at different stages of the application and although configuration dialogs may vary, there are some shared features when generating graphs. 1. Software. Blast2GO integrates a viewer based on the ZVTM framework developed by Emmanuel Pietriga at the INRA (France) for graph visualization (Pietriga, 2005). This high-performing vectored visualization framework allows fast navigation and zooms on the GO DAG. A graph overview is permanently shown at the upper right corner of the graphical tab to easy follow exploring across the DAG surface. Zoom in/out is supported on the mouse wheel and fast zoom to readability is reached by double click on a DAG node. Information about the current node is given on the lower application bar 2. Parameters. Node Filters. A potential drawback during drawing Gene Ontology DAGs where numerous sequences are involved is the presence of an excessive number of nodes that would make the graph hard to visualize and will demand large memory resources. Blast2GO allows modulation of graph size by introducing node filters that depend of the type of graph considered. Additionally, there are a maximum possible number of nodes to be displayed. Coloring mode. Blast2GO highlights nodes proportionally to some parameter of the analysis which result is visualized on the DAG. By this intensity variation of node color relevant terms get more visual weight which is a useful way to guide visual inspection of the results.

10.2.1 Graph element legend

Gene Ontology term obtained by mapping which can directly be associated to one ore more BLAST hits. (GO-Accession, maximum hit e-value assigned, max. hit similarity assigned, number of hits belonging to this)

Non-annotated GO term node (GO term name, mean e-value of all hits contributing to this node, max. e-value, max. Similarity, number of Hits contributing to this node, Annotation Algorithm Score)

Annotated GO term node (GO term name, mean e-value of all hits contributing to this node, max. e-value, max. Similarity, number of Hits contributing to this node, Annotation Algorithm Score)

There exist two types of relationships between child and parent terms. Children that represent a more specific instance of a parent term have an 'instance of' or 'is a' relationship to the parent. Children that are a constituent of the parent term have a 'part of' relationship.

10.3 Statistical charts

The menu Statistics collects a number of charts that are generated during the BLAST, mapping and annotation processes. They are aimed to offer the researcher an overview of the results obtained in each step to facilitate decision for parameter choice in latter annotation steps. Some of these charts have been introduced previously in this Tutorial along with the functions they are related. In this section we will just summarize them:

1. BLAST statistics E-Value Distribution: Histogram of number of hits with a given e-value. Similarity Distribution. Histogram of number of hits with a given similarity value.
2. Mapping statistics Evidence Code Distribution: Histogram of the number of GO terms with a given Evidence Code. DB-resources of mapping: Histogram of the number of GO obtained from each possible DB source of annotations.
3. Annotation statistics Annotation Distribution: Histogram with the number of sequences having a given number of annotations. GO Annotation: Bar chart with the number of sequences having a given annotation.

10.4 Pies and Bar Charts

Some of the results obtained by the Data Mining tools present in the application (see Section 11) are displayed either as a Bar or Pie charts. Similarly to the DAGs, parameters for modulating the size of these graphs are available at their configuration menus. As these charts are very much related to the Data Mining functions they correspond, they will be explained together in the next section.

11 Quantitative Analysis

As a Data Mining tool, Blast2GO provides two basic ways for the joint analysis of a group of annotated sequences.

11.1 Descriptive analysis. Combined Graph Function

Blast2GO generates combined graphs where the combined annotation of a group of sequences is visualized together. This can be used to study the joined biological meaning of set of sequences. Combined graphs are a good alternative to enrichment analysis where there is no reference set to be considered or the number of involved sequences is low. This function is available under “Analysis ->Combined Graph”.

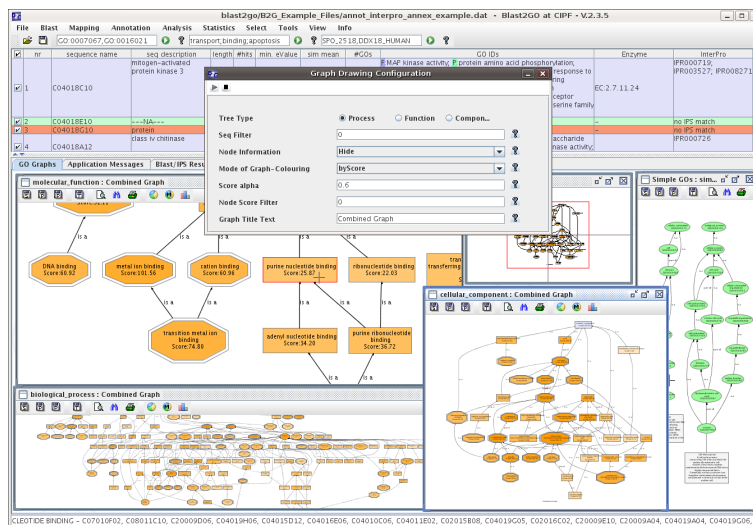


Figure 19: Combined graph visualization

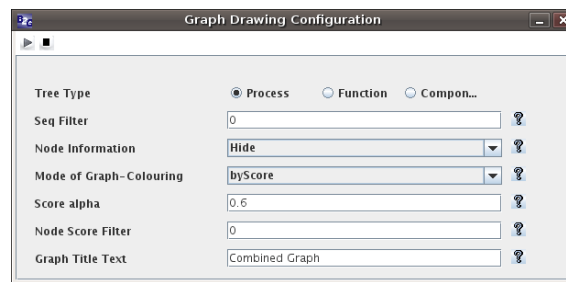


Figure 20: Combined Graph Drawing Configuration Dialog

Figure 20 shows the Combined Graph Drawing Configuration Dialog, where the following parameters are available:

- Tree Type. You can select each of the Gene Ontology Main Categories to be displayed
- Seq 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. It is recommended to start the analysis with a high number that, depending on the number of total sequences, is expected not to overload the graph. Depending in the result adjust this value until you obtain a satisfactory graph. Start with 10 your total number of sequences. Additionally, nodes can be filtered out by the Node Score Filter (see below)
- Node Information. This parameter controls the information shown at a node. Possible values are:

- Hide: Only GO Description and Score (see below) are shown
- Show: Displays annotation information (number of sequences, e-value, similarity and Score)
- With Seqs: The names of the sequences annotated at each GO are included in the node.
- Mode of Graph Coloring. Two possibilities:
 - By Score: A Score is computed at each node according to the formula:

$$score = \sum_{GOs} seq \times \alpha^{dist} \quad (1)$$
 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.
 - By Seq-Count: Node color intensity will be proportional to the number of contributing sequences at the node.
- 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.

11.1.1 Export Combined Graph Results

The information present in a Combined Graph can be also exported in a table format using the function Export Graph Information. This will generate a *.csv* file where all information related to each node of the plotted Graph is provided in different columns.

11.1.2 Charts

Analysis of GO Term associations in a set of sequences can also be done by Pie/Bar Charts. For this analysis, a Combined Graph must have been generated first. Once the graph is visible in the GO Graph panel you can find several icons to visualize the 3 different types of charts (see Figure 21).

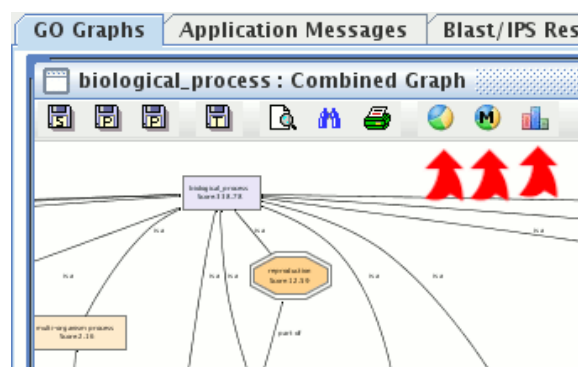


Figure 21: Combined Graph Pie and Bar-Charts

Three possibilities are available:

1. GO distribution by level: This function cuts the DAG at the given level and generated a Pie representation of the number of sequences at the nodes of that level. See Figure 22.
2. Seq distribution by GO (Bar-Chart): This function generates a Bar Chart of all annotated nodes, giving the number of annotations at each node. See Figure 23.

- Seq Distribution by GO (Multilevel Pie): This function generates a Pie with the lowest node per branch of the DAG that fulfills the filter condition., e.g. will find all the lowest nodes with the given number of sequences or Score value and will plot them jointly in a Pie representation. See Figure 24.

When any of these functions is called, a table of node counts is generated and displayed in the statistics tab.

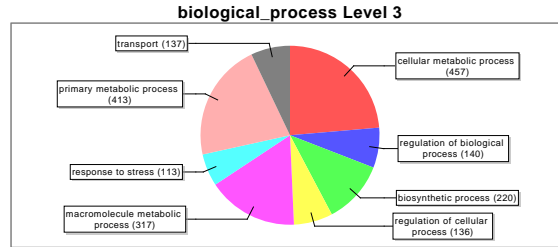


Figure 22: GO Distribution by Level

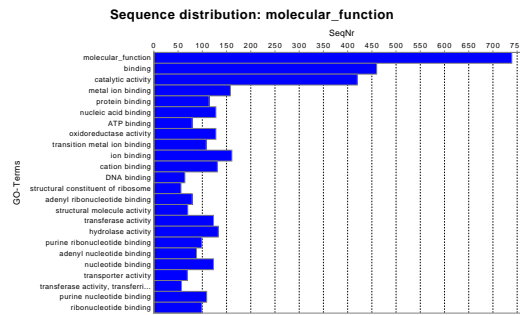


Figure 23: Sequence Distribution/GO as Bar-Chart

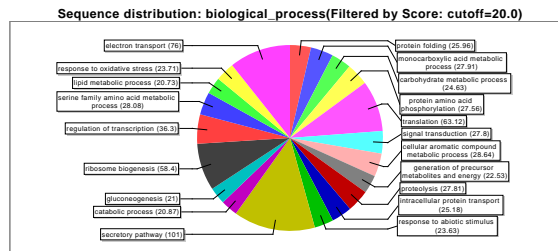


Figure 24: Sequence Distribution/GO as Multilevel-Pie (#score or #seq cutoff)

12 Statistical Analysis

Blast2GO has integrated the Gossip (Blüthgen et al., 2005) package for statistical assessment of annotation differences between 2 sets of sequences. This package uses the Fisher's Exact Test and corrects for multiple testing. For this analysis, the completion (but not exclusively) of the involved sequences with their annotations must be loaded in the application. This can either be the result of a Blast2GO annotation or the imported annotation by file (*.annot*), see Section 8 of this tutorial. This functionality can be found under "Analysis" -> "Enrichment Analysis" -> "Make Fishers Exact Test". A dialog screen appears. Test and Reference Sequences can be selected by uploading *.txt* files containing the lists of sequence IDs for the 2 groups (Figure 25).

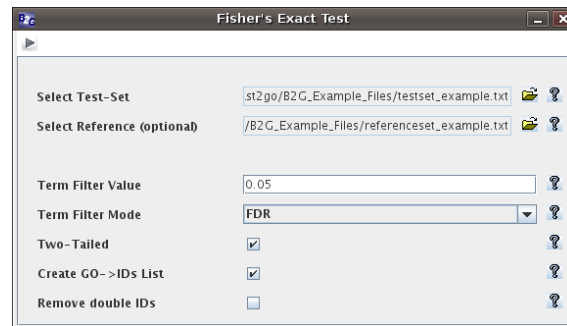


Figure 25: Fisher's Exact Test Dialog

When there is no reference set selected, the whole data set present in the project will be taken as reference. Click on the run button to start the analysis. This will send data to the Gossip web service where the actual computing is performed. This action can take several minutes, depending of the amount of data to be transferred and the speed of your Internet connection. After analysis completion, the results will be sent back to your Blast2GO session and a Gossip Results Table will appear in the Statistics Tab. This table lists the adjusted p-values of the Fisher's Exact Test for each GO term.

- FDR: corrected p-value by False Discovery Rate control.
- FWER: corrected p-value by Family Wise Error Rate.
- Single Test p-Value: p-Value without multiple testing corrections.

For further details please refer to the Gossip publication (Blüthgen et al., 2005).

The Enrichment Analysis Menu has two options for the visual display of the results:

1. Make Enriched Graph: Click here to generate a representation on the GO DAG (for an example see Figure 27): Nodes are color highlighted proportionally to their significance value. The user can choose which type of calculated p-value to use for highlighting and the threshold for filtering out nodes. Additionally, the "Thinned out Graph" Node Filter will hide nodes with a significance value higher that indicated value. Section 10 of this Tutorial gives further information on the graphical functions in Blast2GO.
2. Bar Chart: This option generates a bar display of the percentages of sequences at both, test and reference set, for GO terms having a significance value under the given threshold (Figure 28).

13 Other Functions

In this section we refer to additional functions which have not been considered in the previous sections.

Menu Bar

- Check box: A check box is available on the left side of each sequence for sequence selection/deselection. Blast2GO functions will only apply to the selected sequences!
- Draw GO terms: Individual GO terms can be visualized on the GO DAG using the single GO term drawing tool at the Menu Bar. Simply write or paste one or more (coma separated) GO number in the white box and click on the green arrow. The graph will appear in the Ontology Graphs tab.

File Menu

- Favorites. Blast2GO keeps a record of recent opened files which can be opened directly from Favorites.
- Save Project. At any moment you can save your project as a Blast2GO *.dat* file. This file can be opened by the application at a later moment to continue the analysis.
- Import. Apart from importing existing BLAST or Annotation files, for those sequences with available annotation in public databases Blast2GO can retrieve their GO annotation when imported in the application either as Accession numbers or Gene Symbol. In the later case, the species must also be specified. Accession import can be done by just importing a file containing a list of ACC's (separated by line breaks). Gene symbols has to follow the same list format but need to have a taxa id added to it by a horizontal line (e.g.: ABP1—4932 for gene ABP1 from *Saccharomy Cescerevisia*) since the gene symbols are not always unique.
- Export. Export Sequence Table. Export the current Main Sequence Table for the selected sequences.
- Export as FASTA. Export the selected sequences in FASTA format.
- Export Mapping Results. Will export all GO terms retrieved at the Mapping step.
- Close File. Closes current project.
- Close. Closes Blast2GO application.

BLAST Menu

- Reset BLAST Results. Delete BLAST results for the selected sequences.
- Load Grid Session. This option is in development and not available to users. In the future it will allow launching Blast2GO queries on a Grid System for speeding up processing. Aparicio et al. (2006)

Mapping Menu

- Reset Mapping. Delete Mapping results for the selected sequences.

Annotation Menu

- Make Enzyme Annotation. Finds Enzyme codes from the current GO term annotations. You may use this function when GO terms have been imported or changed manually.
- Annotation Validation. Blast2GO annotation generates lowest node annotations. This is not always guaranteed when Annotations have been imported or changed manually. This function can be run to ensure that no parent-child redundancy is present in the annotated set.

Tools

- Handling sequences. Here there are different function for selection and unselecting sequences. This allows deleting or resetting functions for a given subset of the loaded sequences.
- Select sequence by color. This function allows selecting/unselecting of sequences on the basis of their color code, i.e., the processing stage they have.
- Select sequences by Name. This is a general unction for selecting sequences by loading a file containing a list sequence IDs
- Deselect Sequence by Name. This is a general function to deselect sequences loading a file containing a list sequence IDs.
- Delete Sequence Selection. This function will delete selected sequences of the Main Sequence Table
- Start Batch. This function allows the user to run Blast2GO annotation on a set of sequence files. Files will be loaded in Batch and BLAST, Mapping and Annotation will be subsequently run with the current application parameters.
- Make Filtered BLAST-GO-DB. You can use this function to create you a costume DB containing only GO-annotated sequences. You can use this type of DB in the BLAST step in combination with an own WWW-BLAST installation.

Info At the Info Menu information can be found about the Blast2GO version, authors and also available documentation and examples.

14 Blast2GO for advanced users

1. The `blast2go.properties` file: The property file of Blast2GO can be found in the local user profile directory (e.g.: `/home/yourname/blast2go/blast2go.properties`) and is used to change specific application parameter, default settings and to store temporal configurations. Here you will find a list of the most important settings which can not be changed through the Blast2GO interface but in some special cases a change might be useful:
 - The buffer size of the application message tab (default: 10000 character)
`MainGui.consoleBuffer=10000`
 - The possible databases to run the BLAST (separated by comma).
`Blast.blastdbs=nr,swissprot`
 - The possible BLAST algorithms to use:
`Blast.programs=blastx,blastp,blastn,tblastx`
 - The servers to run a WWW-BLSAT. Put here your own www- BLAST installation path (separated by comma).
`Blast.urls=http://www.ncbi.nlm.nih.gov/blast/Blast.cgi,
http://www.yourDomain.edu/Path ToYour/wwwBlast/Blast.cgi,...`
 - The hit which matches the String given by this parameter will be filtered out. This parameter can be e.g. used to exclude an organism while blasting against NR.
`Blast.hitDescFilter=Arabidopsis`
 - This parameter defines the position of sequences descriptions within a BLAST-Result. (default: 5)
`Blast.hitDescPosition=5`
 - This parameter defines if the xml BLAST tag `<id>` is also used as hit descriptio.
`Blast.addIdtoDef=true`
 - Maximum number of nodes in a graph; prevents graphs getting to big to be displayed.
`MakeGraph.nodeLimit=700`
 - Maximum enriched terms of a Fisher's Exact Test; prevents results from getting too long.
`Fisher.maxresultnr=100`
2. Local Blast2GO database installation: If you are interested in installing a own Blast2GO database locally with the aim to not depend on the Blast2GO server, you can find a tutorial on the Blast2GO website in the download section including a step-by-step installation guide. Basically will need a MySQL server, the latest GO database dump and some additional "mapping tables" (NCBI and PIR flat-files). By following several few steps this data is imported into your database.
3. Blast2GO Pipeline: The pipeline version (Blast2GO4Pipe) runs Blast2GO without any graphical interface. Its is a Java application which can be run platform independent on any machine offline if there is local Blast2GO database installed or with a Internet connection to the Blast2GO DB. The annotation process is configured by a self-describing property file (`b2g4pipe.properties`). The output files generated by Blast2GO4Pipe are equivalent to the file generated by the desktop application. The annotation file (`.annot`) is a tab separated file containing the sequence names and gene ontology Ids. You can find the distribution and an installation description (read-me) in the Download section of the Blast2GO website.

Bibliography

- Aparicio, G., Götz, S., Conesa, A., Segrelles, D., Blanquer, I., García, J. M., Hernandez, V., Robles, M., and Talon, M. (2006). Blast2go goes grid: developing a grid-enabled prototype for functional genomics analysis. *Stud Health Technol Inform*, 120:194–204.
- Blüthgen, N., Brand, K., Cajavec, B., Swat, M., Herzel, H., and Beule, D. (2005). Biological profiling of gene groups utilizing gene ontology. *Genome Inform*, 16(1):106–115.
- Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., and Robles, M. (2005). Blast2go: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, 21(18):3674–3676.
- Götz, S., Garcia-Gomez, J. M., Terol, J., Williams, T. D., Nagaraj, S. H., Nueda, M. J., Robles, M., Talon, M., Dopazo, J., and Conesa, A. (2008). High-throughput functional annotation and data mining with the blast2go suite. *Nucl. Acids Res.*, pages gkn176+.
- Myhre, S., Tveit, H., Mollestad, T., and Laegreid, A. (2006). Additional gene ontology structure for improved biological reasoning. *Bioinformatics*, 22(16):2020–2027.
- Pietriga, E. (2005). A toolkit for addressing hci issues in visual language environments. In *Visual Languages and Human-Centric Computing, 2005 IEEE Symposium on*, pages 145–152.
- Rattei, T., Tischler, P., Arnold, R., Hamberger, F., Krebs, J., Krumsiek, J., Wachinger, B., Stümpflen, V., and Mewes, W. (2008). Simap—structuring the network of protein similarities. *Nucleic acids research*, 36(Database issue):D289–D292.
- The_gene_ontology_consortium (2008). The gene ontology project in 2008. *Nucleic acids research*, 36(Database issue).

List of Figures

1	Blast2GO (v.2)	2
2	Java(TM) Web Start 1.6.0 Interface	3
3	Blast2GO Interface	5
4	BLAST Configuration Dialog	9
5	E-Value Distribution	10
6	Similarity Distribution	10
7	Species Distribution	10
8	Evidence Code Distribution of BLAST hits	11
9	Evidence Code Distribution for sequences	11
10	Source DBs urces of Mapping	12
11	Annotation Rule	12
12	Annotation Configuration	13
13	Evidence Code weight configuration	13
14	Annotation Distribution	14
15	Single Sequence Menu	16
16	BLAST Result	17
17	Single Graph Drawing Configuration	17
18	Single Sequence Graph	17
19	Combined graph visualization	20
20	Combined Graph Drawing Configuration Dialog	20
21	Combined Graph Pie and Bar-Charts	21
22	GO Distribution by Level	22
23	Sequence Distribution/GO as Bar-Chart	22
24	Sequence Distribution/GO as Multilevel-Pie (#score or #seq cutoff)	22
25	Fisher's Exact Test Dialog	23
26	Fisher's Exact Test result table	24
27	"Enriched" Graph	24
28	Bar-Chart representation of Fisher's Exact Test results	24
29	Values of an Enriched Graph Node	24

Nomenclature

B2G	Blast2GO
BLAST	Basic Local Alignment Search Tool
EC	Evidence Code
GO	Gene Ontology
JRE	Java Runtime Environment
JWS	Java Web Start
EC	Enzyme Code
PIR	Protein Information Resource
URL	Uniform Resource Locator
XML	Extensible Markup Language