## Contents

1 **Blast2GO - Functional Genomics Made Easy** 4
   1.1 Introduction .................................................. 4
   1.2 Requirements .................................................. 4
   1.3 Activation ...................................................... 4
   1.4 The Desktop Application ..................................... 5

2 **Quick Start** 11

3 **Sequences** 13
   3.1 Load sequences from start .................................. 13
   3.2 Show Sequence .................................................. 13
   3.3 Sequence Length Statistics .................................. 14
   3.4 Add sequences to existing Blast2GO project ............. 14
   3.5 Export FASTA Sequence with Annotation Results ........ 15

4 **BLASTing** 16
   4.1 Run BLAST at the NCBI ...................................... 17
   4.2 Run BLAST using CloudBLAST (PRO Feature) .............. 19
   4.3 Run BLAST using AWS Blast (PRO Feature) ............... 21
   4.4 Run BLAST Locally ............................................. 21
   4.5 Show BLAST Results .......................................... 22
   4.6 Statistics ....................................................... 22
   4.7 Load BLAST results ........................................... 25
   4.8 Make Blast Database .......................................... 25
   4.9 Other BLAST Functions ....................................... 26

5 **InterProScan Annotation** 27
   5.1 General .......................................................... 27
   5.2 Run InterProScan ............................................... 27
   5.3 Merge InterProScan GOs to Annotation ..................... 28
   5.4 Statistics ....................................................... 29
   5.5 Load InterProScan Results .................................. 30

6 **Gene Ontology Mapping** 31
   6.1 General .......................................................... 31
   6.2 Show Individual Mapping Results ......................... 31
   6.3 Statistics ....................................................... 32
   6.4 Export Mapping Results ..................................... 33
7 Gene Ontology Annotation

7.1 Blast2GO Annotation Rule 34
7.2 Individual Annotation Results 36
7.3 Statistics 36
7.4 Augment Annotation by Annex 38
7.5 Annotate GOs from Blast Descriptions 38
7.6 Exporting Annotation 39
7.7 Load Annotation Results (.annot) 39
7.8 Other Annotation Functions 40

8 Data Distribution Charts 41

8.1 Project Statistics 41

9 Gene Ontology Graphs 43

9.1 Directed Acyclic Graphs 43
9.2 Pies and Bar Charts 43

10 Quantitative Analysis 44

10.1 Descriptive analysis. Combined Graph Function 44
10.2 Coloured GO Graphs from a text file 51
10.3 Make GO Graph 53

11 Select 54

11.1 Select Sequences (PRO Feature) 54
11.2 Select sequence by color 55
11.3 Invert Selection 55
11.4 Delete Selected Sequences 55
11.5 Create ID List From Selection 55
11.6 Sort by Selection 55
11.7 Export Selected Sequences as DAT 55
11.8 Hide unselected sequences 56

12 Analysis Menu 57

12.1 GO-Slim 57
12.2 Enzyme Code and KEGG 57
12.3 Enrichment Analysis (Fisher’s Exact Test) 59
12.4 Gene Set Enrichment Analysis (GSEA) 61
12.5 RFAM 63

13 Differential Expression Analysis 65

13.1 Pairwise Differential Expression Analysis 65
13.2 Time Course Expression Analysis 73
1 Blast2GO - Functional Genomics Made Easy

1.1 Introduction

Blast2GO (Conesa et al., 2005) is a comprehensive bioinformatics tool for the functional annotation and analysis of genome scale sequence datasets. The software was originally developed to provide a user-friendly interface for Gene Ontology (The gene ontology consortium, 2008) annotation (Götz et al., 2008). Over the last years many improvements have considerably increased the functionality of Blast2GO and many different functional genomics tools are now available. Additionally the application offers a wide array of graphical and analytical tools for data manipulation and mining.

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

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 basic use case of Blast2GO 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 Logo (v.3.0)

1.2 Requirements

Blast2GO is an operating system independent Java Application and works on Mac, Windows and Linux systems. We recommend at least 1GB of RAM, to open Blast2GO with 4GB or more you will need to have a 64 bit architecture. A working Internet connection is required to use most of the features of the application. Application specific proxy settings are available. The folder Blast2GO is located in has to have write permissions for automatic updates to work.

1.3 Activation

On startup and if not yet provided, Blast2GO asks for activation. It is possible to use Blast2GO in 3 different mode:

- **DEMO:** This mode allows to use all features of Blast2GO but only with an example dataset of 1000 predefined sequences.
- **BASIC:** To use Blast2GO in Basic mode an activation key is needed. This key can be requested online. In Basic mode it is possible to run a basic functional annotation analysis with publicly available resources. More advanced features are not available in this mode. A detailed up-to-date list of features available in Basic and PRO mode is available online.
- **PRO:** To activate Blast2GO in PRO mode the corresponding subscription has to be obtained via the Blast2GO website. A PRO subscription provides access to all features, cloud
services, updates and support. More details about all the advantages of a PRO subscription can be found online. A free trial account can be requested via the Blast2GO website.

To activate Blast2GO the software requires access to the Internet. If necessary, proxy settings can be directly adjusted directly from the activation dialog (see 18.5).

![Activation dialog of Blast2GO: Demo, Basic and Pro Mode](image)

**Figure 2: Activation dialog of Blast2GO: Demo, Basic and Pro Mode**

### 1.4 The Desktop Application

This section describes briefly the main parts of the Blast2GO applications.

#### 1.4.1 Main User Interface

The general components of the Blast2GO main user interface are (Figure: [Figure]:

1. **Menu Bar**: The main application menu contains 5 items:
   - **File**: This menu groups all functions for opening, saving and closing Blast2GO files as well as to load or export data in many different formats.
   - **Analysis**: This menu hosts different options for the analysis of your datasets like e.g. computing GoSlim view, Enzyme Code annotation with KEGG maps and performing statistical analysis of GO distributions for groups of sequences.
   - **Tools**: This menu contains rather technical features for data-manipulations like batch-renaming, ID mapping, data redundancy checks etc.
   - **View**: This menu allows to open different utility taps like the file manager, the application messages or the Java memory monitor.
   - **Help**: The help menu provides access to various support features. You can review your subscription details, the CloudBlast history and the App-Manager. The App-Manager allows to install and update additional tools for Blast2GO.
2. Main Analysis Icons: These will execute the whole analysis.

- **start**: Quick start to load a Blast2GO Project or recent projects or fasta sequences or annotation file.
- **genefind**: Allows to run eukaryotic and prokaryotic gene predictions.
- **blast**: Contains functions for performing BLAST searches and resetting results.
- **interpro**: It will run the InterProScan, reset the results or merge the GO results to the existing Annotation.
- **mapping**: This function fetches GO terms associated to hit sequences obtained by BLAST.
- **annot**: Includes different functions to obtain and modulate GO
- **charts**: The Charts icons collects a number of chats 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.
- **graphs**: Allows to make Combined and Colored Graphs as well as Single Graphs by providing GO IDs or GO Terms.
- **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.
- **diff-expr**: Allows to perform pairwise and time-course expression analysis for un-normalized read counts.

3. Application Tabs:

- **Progress tab**: Shows running, queued and finished jobs. Allows to stop jobs and open a separate log-viewer. Each finished job shows the time it took to execute. All finished jobs can be removed by clicking on the small triangle in the top left corner.
- **Application Messages tab**: Displays information on the progress of the analysis.
- **File Manager**: The File Manager allows to view, open and organize different Blast2GO related files.
- **Java Memory Monitor**: Provides information about the memory consumption of the application.
- **Log tabs**: Most of the analysis steps provide additional logging. These logs can be viewed via the Progress tab.

4. Data Tabs: Each object type can be opened and saved in a separate tab. In general, table viewer tabs like e.g. the Blast2GO sequence table will be opened by default in the upper-wide area while result tabs will be opened in the lower-right area (e.g. charts, graphs, etc.).
Figure 3: The Blast2GO File Manager allows to view, organize and open different Blast2GO related files. The “Merge” option allows to combine various file into one. The merge function may not be available for all types of datatypes and can only be used to object of the same type.

Figure 4: Blast2GO Main User Interface

1.4.2 Table Viewers

From within the FileManager under the option “Open” and “Open With” different data viewers can be chosen depending on the object type. Not all objects can be opened with a table viewer (e.g. bar charts, graphs, etc.).

Available Viewers:

- Blast2GO Table: The traditional colored Blast2GO tables (see section: 1.4.2.1).
- Generic Table: A spreadsheet style table for multiple data objects (.b2g files) (see section: 1.4.2.2).
• RFAM Table: Shows the RFAM results obtained from the EBI
• Fisher’s Exact Test Result: Shows the enriched terms with tags: over and under
• Mapping Results Table: Shows the GO mapping results of a particular sequence.

1.4.2.1 The Blast2GO Table  The Blast2GO sequence table shows the details and progression of the analysis for the loaded sequence dataset. This table is the one that is opened when loading sequences to Blast2GO. The rows are colored according to the analysis progress. The table allows to hide or show single columns via a checkbox menu by right-clicking on the column header.

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 ID list associated to the sequence.
• GO Names list
• Enzyme Code (EC) associated to the sequence.
• InterProScan results.
1.4.2.2 The Generic Table  The Generic Table allows to open different types of objects (.b2g files) in spreadsheet style. Additionally to the columns representing the dataset it contains a TAG column which represents the different status of each line according to the object type (Figure: 5). The table allows to hide or show single columns via a checkbox menu by right-clicking on the column header. In case, multiple objects are viewed (e.g. Blast2GO plus Rfam results), column headers are colored differently.

![Generic Table Example](image)

Figure 5: In this example tags show if a functionally enriched GO term is over or under represented.

The context menu allows to create ID lists of a selected column, extract subsets of entries as well as to copy part of the content into the clipboard (Figure: 6). The values of a given column can also be visualized as a category or distribution plot in various formats (bar-chart, pie-chart, etc.).

![Context Menu Example](image)

Figure 6: Context menu of the generic table which allows to extract, copy or convert entries.

- Copy Selected Entries in Tabular Format
- Create Category Statistics (ID, Value) from Selected Entries
- Create Distribution Statistics (#occurrences) from Selected Entries
- Extract Selected Entries

1.4.2.2.1 Generate ID lists from a generic table:  The generic table allows to create ID lists based on a selected column. Once the new ID list is created it can be saved as ID object. These objects can then be used for selections (Toolbar: Select Sequences) as well as to determine the test and reference set of a enrichment analysis (Figure: 7).

To generate an ID list, a saved project has to be opened with the generic table (see section: 1.4.2.2). We now select the desired sequences and create a new ID list via right click on “Create ID list from selected Entries (SeqName column)” (see Figure: 8). A new tab is opened containing the ID list of the selected sequence names. This list can now be saved for further use (e.g. enrichment analysis, selections, etc.). It is also possible to generate sequence ID lists from Graphs (10.1) and from the select arrow (see section 11).
1.4.3 File Types

From version 3.1 upwards the .b2g file type replaces the previous .dat file. The .dat files will still be supported for opening and export. All Blast2GO project and results (enrichment results, charts, graphs, etc.) will be saved with the new file type .b2g. This files can be viewed and opened directly within the FileManager tab. All other file types can be open form the FileManager via the systems default application.

1. .b2g: File type for all Blast2GO objects (project, results, id lists, etc.). Some Blast2GO objects can be opened with different viewers like for example the the Blast2GO Table or the Generic Table.

2. .dat: Legacy format for previous Blast2GO projects. This format had been replaced in 2015 by the more flexible and performance .b2g format. The .dat format is still maintained for compatibility with older datasets and application versions as well as the Blast2GO plugins.
2 Quick Start

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

1. Load data: Go to File → Load → Load Sequences → Load Fasta File and select your .fasta file containing the set of sequences in FASTA format. Alternatively you can load the example sequences into Blast2GO choosing File → Load → Load Example Sequences. Several example files can also be found online at https://www.blast2go.com/support.

2. BLAST: Click on the orange blast toolbar icon. In the BLAST Configuration Dialog (Figure 13) select the type of BLAST mode which is appropriate for your sequence type (blastx for nucleotide and blastp for protein data), click Next for the advanced settings and to choose where to save the Blast results and click Run to start the BLAST search against NCBI’s non redundant NR database. Alternatively you can choose to use different BLAST such as CloudBlast, AWS Blast and Local Blast on the arrow next to the blast (toolbar). See Section 4 for further details on this.
   • Once your BLAST analysis is finished visualize your results at charts (toolbar) → BLAST Statistics.
   • On the Main Sequence Table, right-click on a sequence to open the Single Sequence Menu (Figure 16). Select Show BLAST Result to open the BLAST Browser for that sequence.

3. By clicking on the interpro icon the corresponding Wizard will be shown. Please provide a valid email address. Since this search is executed via a web service provided by the EBI. To run IPS is highly recommended in order to improve the quality of the annotations. Once InterProScan results are retrieved use Merge InterProScan GOs to Annotation to add GO terms obtained through motifs/domains to the current annotations. InterProScan can be run in parallel with BLAST.


5. Annotation: Click on annot (toolbar) to open the Annotation Configuration Window. Click Next to change the evidence codes and finally click Run to start the annotation. Annotated sequences will turn blue.
   • Once the annotation is completed you are able to visualize your results with charts (toolbar) → Statistics → Annotation Statistics.
   • On the Main Sequence Table, right-click on a sequence to open the Single Sequence Menu. Select Draw Graph of GO-Mapping with Annotation Score to visualize the annotation on the GO DAG for that sequence.
   • If desired, modify the annotation by clicking with the right mouse button and select Change Annotation and Description or change the extent of annotation by adding implicit terms Annotation (arrow) → Run ANNEX or reducing to a GO-Slim representation (Analysis → GO-Slim → Run GO-Slim (online)).
   • During the annotation process, Enzyme Codes (EC) will be also given when a GO-term/EC number equivalence is available.

6. Enrichment Analysis: Blast2GO provides tools for the statistical Analysis of GO term frequency differences between two sets of sequences. Go to Analysis → Enrichment Analysis (Fisher’s Exact Test). A new Dialog window will open (Figure 72). 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 website. Select a second set of sequences as reference/background set if desired. If no reference set is provided the all annotations of the corresponding project will be used as reference. Click Run to start the analysis. A table containing the results of this analysis will be displayed in a new tab.
- Click on Make Enriched Graph icon to visualize the results of the Fisher’s Exact Test on the GO DAG.
- Click on Show Bar Chart to obtain a bar chart representation of GO term frequencies.
- The results can be reduced to more specific GO terms in the corresponding icon and saved as text format (Save as Text).

7. 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 select (toolbar) → Select Sequence. Now select Select by Features and Select by Name or ID. 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. Now go to graphs (toolbar) → Make Combined Graph. Now click Run.

8. Save results: File → Save saves the current Blast2GO project as .b2g file.

9. Export results:
   - File → Export allows to export the generated data in many different formats.
   - File → Export Annotations exports the actual annotation results as .annot file or generate own formatted annotation file as .txt file.
   - The enrichment analysis results can be exported in various formats from the Fisher Exact Test Result Viewer. “Save as Text” exports the results as a tabulator separated text file.
   - To export GO graphs use the sidebar of the corresponding graph viewer. Graphs can be saved/exported in .pdf, .png, .svg and .txt.
3 Sequences

To start a new Blast2GO project you just have to load your sequence data from a file into Blast2GO.

3.1 Load sequences from start

At the “File” menu, go to Load → Load Sequences → 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 or FASTQ format (see Figure ). These files must have the extension .fasta, .fnn, .faa, .fna, .ffn, .txt, .fq or .fastq 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
MCAAQRSAAALAAAAPRTVYAFSARPLAGGEFPNFLSSLRGKVLILIVNASLUGTTVRDYTMND
LQRPLPGRLVVLGFPCNQFGQHQAENAKNEELNCLKYVRPGGF

3.2 Show Sequence

Once the sequences have been loaded to Blast2GO, it is possible to see them by right-clicking on the Sequence Table. The “Single Sequence Menu” (context menu) will appear (Figure 9). This menu provides some functions for sequences individually, i.e. will apply to the sequence at that position of the Table.

![Figure 9: Show Sequence](image-url)
3.3 Sequence Length Statistics

Blast2GO allows you to visualise the length distribution of your sequences in the arrow next to the “Chart” icon.

3.4 Add sequences to existing Blast2GO project

Use the File Manager context menu to merge to project. Select two or more project, open the context menu with a right click on one of the files and select Merge.

In case one of your existing files is in .dat format just open it by selecting the file-type in File → Load Project (.dat Legacy Format). If the loaded project file has only Blast results and no sequence information it is still possible to add the corresponding sequences to Blast2GO project by clicking on the arrow next to the “Start” icon and select “Load Sequences”. Now two options will be displayed “Create new project” see Figure 11 and “Add to the existing project”. The “Add to the existing project” option should be selected and in the next page you can browse for the fasta file and “Overwrite” option should be selected.
3.5 Export FASTA Sequence with Annotation Results

After executing the whole functional annotation of the sequences in Blast2GO (BLAST, Mapping and Annotation) it is possible to export the sequence in FASTA format with the corresponding sequence description and GO ID or GO term.

An example of the Exported FASTA format with Sequence Description and GO IDs:

```
>C04018C10|mitogen-activated protein kinase 3|GO:0005634;GO:0004707;GO:0005515
acaaacgagacgcgtagaaaattattagagagaagagagagagtaaatgggtgagctgcacggtcagctacagtataaattttgga
acgtggttgaatcgcagaaagcgctagctcgtctcgcttggtcagttttaattatggttgctcaatggctgattttcctgcggtaccgacgcacggcggtcagtttatacagtacaatatattttggaaaacttgtttgaaatcacggccaagttaaatcttgagaattttcaacattttctgcggaactcatggacactgacctttaccaaatattcgttcaaatcaaatgtagttttacgagacgcgtagctcgtctcgaactcatctgactacactgacactacactacagt
```
4 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](http://www.ncbi.nlm.nih.gov/BLAST) for details on the BLAST function. Figure 13 shows the BLAST Configuration Dialog Window that controls the BLAST step.

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

1. CloudBlast. This 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. Please see section 4.2 for more information.

2. 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.

3. Remote BLAST. Blast2GO will download the latest BLAST+ executable form NCBI and will use it to query NR or other databases remotely.

4. 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.

5. AWS Blast. The NCBI provides via Amazon Web Services (AWS) a preconfigured machine image (AMI) which contains the latest BLAST+ release. One can access to the AMI’s account through Blast2GO. The AMI’s URL has to be provided to Blast2GO and the BLAST searches will run in the Amazon Cloud.

6. Local BLAST against own database. It is possible to use BLAST+ executable to query a local/own database. At [https://www.blast2go.com/make-own-database-and-blast](https://www.blast2go.com/make-own-database-and-blast) one can see how to prepare and blast locally an own fasta database.

QBlast at NCBI is the only feature available for Blast2GO Basic users.

The next figure shows the menu manner to select between NCBI-, remote- or local- BLAST as well as CloudBlast, AWS Blast or blasting against an own database.
4.1 Run BLAST at the NCBI

Here, the user can specify the following parameters, which are divided in three different sections: Blast Configuration in Figure 13(a), Advanced in Figure 13(b) and Save Results Page 13(c).

4.1.1 Blast Configuration Page

- Your e-mail address in case you are using the NCBI BLAST web service.
- BLAST program: The algorithm you want to use:
  - blastp - Compares an amino acid query sequence against a protein sequence database.
  - blastn (-task 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.
  - blastx-fast
  - blastp-fast
  - blastp-short
  - blastn (-task megablast)
  - blastn (-task dc-megablast)
  - blastn-short
  - tblastn-fast
- 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://data.biobam.com/ncbi_blast_dbs_protein.pdf
• 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.

• Number of BLAST hits: The number of alignments you want to achieve (0-100).

• BLAST Description Annotator: The BDA finds the best possible description for a new sequence based on a given BLAST result.

4.1.2 Advanced Page (PRO Feature)

• Blast Parameters:
  ◦ Word size: One of the important parameters governing the sensitivity of BLAST searches is the length of the initial words. The word size is adjustable in blastn and can be reduced from the default value to increase sensitivity. This word size can also be increased to increase the search speed and limit the number of database hits.
  ◦ 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.

• Filter Options:
  ◦ 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.
  ◦ Filter by description: Filter-out Blast hits by description
  ◦ Filter by taxonomy: Search for Blast results only in the selected taxonomy.

4.1.3 Save Results Page

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 Save Results Page. If the chosen file already exists, upcoming results will be appended. Choose a format type to additionally save your BLAST results.

• XML2: This is a new BLAST results provided by NCBI and can also be loaded into Blast2GO.

• XML: It is recommended to save your BLAST results as XML as this format is supported by the Blast2GO Load BLAST Results function.

• TXT: It saves the blast results of each sequence in text file format.

• HTML: For each sequence a file in html format will be saved.
4.2 Run BLAST using CloudBLAST (PRO Feature)

CloudBlast offers a highly optimized, self-sustained HPC solution to address a very specific need of the Blast2GO PRO community.

CloudBlast is a BLAST service totally independent from the NCBI servers to provide fast and reliable sequence alignments. It consists of a high performance computing cluster dedicated exclusively to Blast searches.

All Blast2GO PRO subscriptions include "ComputationUnits" to make use of this resource and allows you to perform 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 and the blast algorithm (blastx, blastp) and parameters used. The smaller the database you blast against the more sequences you can analyse with 6,000,000 ComputationUnits (see section 17.2 to know how to monitor the ComputationUnits). This means that e.g. if you blast against the vertebrate NR-subset you would be able to blast approx. one million (1,000,000) sequences. If you decide to blast against the NR database, the largest protein database available, it should allow you to blast approx. 80,000 sequences (with an average length of 800nt per sequence).

For the advanced and save parameters page please see section 4.1.2 and 4.1.3 for detailed information.

Figure 14: CloudBlast Configuration Page
4.3 Run BLAST using AWS Blast (PRO Feature)

The AWS Blast option allows to run stand-alone searches with the BLAST+ applications at Amazon Web Services (AWS). The NCBI is providing a BLAST machine image (AMI) hosted at Amazon Web Services (AWS). Once you setup one of these AMIs you can submit searches directly from Blast2GO via the NCBI-BLAST URL API. The AMI includes a FUSE client that downloads the desired BLAST databases during the first search. Please find more information on how to setup an AMI here: https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=CloudBlast

To start blasting against your AMI via Blast2GO you simply have to provide the URL to your Blast.cgi of your AMI in the “Blast using AWS” wizard page. To obtain a public URL for your AMI please follow these instructions: http://ncbi.github.io/blast-cloud/doc/running-web-blast.html. The URL you will need to provide will look similar to this (ec2-54-82-43-97.compute-1.amazonaws.com/cgi-bin/blast.cgi). Make sure that your AWS security group allow access through HTTP.

4.4 Run BLAST Locally

With Local BLAST you can blast the sequences against own database. Blast2GO allows to create a Blast database from a FASTA file with the option “Make Blast Database” (see section: 4.8). Download and format your database and choose the corresponding folder see Figure 15. Databases have to be formatted for NCBI Blast+.

The main parameters in the Local BLAST Configuration page are very similar to the ones in NCBI and CloudBlast. The main difference is when choosing the database as Blast2GO is expecting a “.pal” file. On the Advanced Page at the “Run Parameters” it is possible to select the number of threads to be used. This field has not to be set up as Blast2GO detects the number of threads in the computer. The section 4.1.2 provides detailed description of each parameter. As in CloudBlast the BLAST results will be saved in XML file format.

![Figure 15: Local Blast Configuration Page](image-url)
4.5 Show BLAST Results

As the BLAST search progresses, sequences with successful BLAST results change their color on the Main Sequence Table from white to orange 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.

With a mouse the right click on a sequence, the Single Sequence Menu will be displayed and it is possible to see the BLAST results for each sequence individually. Show BLAST Results (Figure 16) will generate a tab in the Results containing information on the results of the similarity search of 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 hsps mapped
- GO-Terms with its evidence code
- UniProt codes of the hit sequences.

![Show BLAST Results](image1)

**Figure 16: Show BLAST Results**

4.6 Statistics

Different BLAST statistics charts (Figure 19(a), 19(b) and 19(c)) 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. To generate the BLAST Statistics charts just go to the arrow next to the “Chart” icon and select the statistics to be displayed (see Figure 18).

- **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 E-Value cutoff in the annotation step.

![Individual BLAST Results](image2)

**Figure 17: Individual BLAST Results**
• 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-Hit Species Distribution: Bar chart showing the species distribution of all Top-Blast hits.

• Hit Distribution: This chart shows a distribution of the number of hits for the blasted sequences in a data-set.

• 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.

Figure 18: Blast Statistics
Figure 19: Blast Statistics Graphs
4.7 Load BLAST results

If a BLAST result is already available in XML format, it can be directly loaded into Blast2GO by using Load → Load Blast Results in the File menu. You can choose here to import the Blast results as XML file or the new XML2/JSON format. These new formats can be loaded as Zip file.

In the Load Blast Results dialog a whole directory containing a collection of BLAST XML files or a single XML file can be selected Figure 20. The BLAST results will be added to your current Blast2GO session.

Blast2GO PRO also allows the input of TimeLogic DeCypher Blast results.

![Load Blast Results](image)

Figure 20: Load Blast Results

4.8 Make Blast Database

This option allows to create a BLAST database from the sequence of any /blast2go project or from a FASTA file (Figure 21). This option can be found in the arrow next to the blast icon.

- Current project: Blast2GO will use the loaded sequences to create the Blast database. Note: If the resulting database will be used for further GO mapping a proper ID and description line with “GO mappable” information is needed.
- FASTA file: This option allows to choose own FASTA file. The FASTA file have to be correctly formatted for NCBI Blast+.
- Output Folder: Select the directory where to save the created Blast database.
- Blast Database Name: Provide a name for the Blast database
- Taxonomy Options:
  - Taxonomy ID: Introduce the NCBI species ID.
  - Mapping file: If the sequences come from different species, it is possible to generate a text file with the sequence names and its species id to map to the corresponding sequence in the FASTA file.

Example:
4.9 Other BLAST Functions

- Remove Blast Results: This option will remove the BLAST results from the selected sequences.

- Run Blast-Descriptor-Annotator (BDA): This will run the BDA algorithm. For further details, please see section 4.1.1.

- Recover original Best-Blast-Hit Description: When this option is executed the sequence description column on the Main Sequence Table will contain the top blast hit description and not the one from the BDA.
5 InterProScan Annotation

5.1 General

The functionality of InterPro annotations in Blast2GO allows to retrieve domain/motif information in a sequence-wise manner. Corresponding GO terms are then transferred to the sequences and merged with already existing GO terms. InterProScan results can be viewed through the Single Sequence Menu (Figure 24(a)) and saved in TXT and XML format (Figure 23(b)). The sequences will turn violet if no other analysis has been executed before.

![InterProScan options](image)

Figure 22: InterProScan options

- Run InterProScan. Start sending sequences to the EBI.
- Merge InterProScan GOs to Annotation. Add GO terms obtained through motifs/domains to the current annotations.
- Remove InterProScan. Delete InterProScan results for the selected sequences.

5.2 Run InterProScan

InterProScan can only be performed if the sequences shown in the sequence table contain the actual sequence information (loaded via fasta file). This can be the case e.g. if you created a project via a blast XML file or if you loaded an .annot file. To add the sequences to the current Blast2GO project see section 3.4.

You can save the InterProScan results in different file formats, in tab separated values (TVS), XML, which is the default output, GFF3 and the input (query) sequence itself (Figure 23(b)).

If you are working with nucleotide sequences, Blast2GO translates it to the longest open reading frame and sends it to InterProScan. For this particular case when exporting the input sequence Blast2GO will save the protein sequence itself and not the nucleotide one.

Once the InterProScan has finished it is possible to view the results of each sequence via the context menu (Figure 24(a)).
5.3 Merge InterProScan GOs to Annotation

The InterProScan GOs results can now be added to the already existing annotations based on the BLAST results. This option is available from the IPS sub menu (little arrow).

Once the merge has finished a distribution chart is displayed in the Results menu showing the number of GOs that have been added to (or confirmed) the current annotation results (Figure 23(b)).
5.4 Statistics

On the sub menu of the “Charts” icon it is possible to select InterProScan statistics to see how many sequences still do or do not have IPS results and how many sequences have GOs resulting from InterProScan.

- InterProScan Results: This chart reflects the effect of adding the GO-terms retrieved through the InterProScan results (Figure 27).
• InterProScan Families Distribution: Bar chart representing the number of sequences that belong to a particular IPS family.

• InterProScan Domains Distribution: Bar chart showing the number of sequences that belong to a particular IPS domain.

• InterProScan Repeats Distribution: Bar chart reflecting the number of sequences that belong to a particular IPS repeat.

• InterProScan Sites Distribution: Bar chart representing the number of sequences that belong to a particular IPS sites.

• InterProScan IDs Distribution: Bar chart showing the number of sequences that have been annotated with that InterProScan IDs.

• InterProScan IDs by Database: Pie chart reflecting the number of sequences of the InterProScan IDs for a particular InterProScan Database. In figure 26 the Pfam database is selected.

5.5 Load InterProScan Results

The InterProScan results saved in XML format can be loaded in the current Blast2GO project (File → Load → Load InterProScan Results).
6 Gene Ontology Mapping

6.1 General

Mapping is the process of retrieving GO terms associated to the Hits obtained by the BLAST search. Blast2GO performs four different mappings steps:

1. BLAST result accessions are used to retrieve gene names or symbols making use of two mapping files provided by the 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. GeneBank identifiers (gi), the primary blast Hit ids, 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.

![Run Mapping and Remove Mapping](image)

Figure 28: Mapping options

- Run Mapping. Mapping will start.
- Remove Mapping. Delete Mapping results for the selected sequences.

6.2 Show Individual Mapping Results

For each sequence it is possible to see the mapping results individually.

![Mapping Results for sequence C02006A02](image)

Figure 29: Show Mapping Results

Figure 30: Mapping Results for sequence C02006A02
1. Show Mapping Results. A new table will be displayed (see Figure 30). The resulting table shows the GO mapping results for a particular sequence. See Generic Table section 1.4.2.2 to manipulate/extract the results from this table.

2. Show GO Descriptions. 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. Annotate Sequence. This function allows changing annotation parameters for the selected sequence and re-running automatic annotation.

4. 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 (see Figure 37) is available for marking sequences with manual annotation. The sequence will get the pink label on the Main Sequence Table.

5. Make Graph of GO-Mapping-Results with Annotation Score. Displays a DAG with all GO terms related to one sequence. Shows all the GOs from the mapping step as well as final annotations (highlighted). The wizard (Figure 31 allows to filter the hits which will be taken into account (see Section 9 for more details about visualization in Blast2GO)

(a) 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.

(b) HSP-Hit Coverage CutOff: Includes only those hits which are overage with the HSP for a given percentage.

![Figure 31: Single Graph Drawing Configuration](image)

6.3 Statistics

If a BLAST result is successfully mapped to one or several GO terms, these will be shown in the GOs column of the Main Sequence Table and this sequence row will turn light-green. Assigned GOs can be reviewed in the BLAST results Table (see Section 4.5 and Figure 17).

Three different charts are available to summarise the mapping step:

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

- EC Distribution for Blast Hits (Figure 32(a)): Evidence Codes associated to the obtained GO pool
• EC Distribution for Sequences (Figure 32(b)): 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 Resources of Mapping (Figure 32(c)): This chart gives the distribution of the number of annotations (GO-terms) retrieved from the different source databases e.g. UniProt, PDB, TAIR etc.

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.

(a) Evidence Code Distribution of BLAST hits  
(b) Evidence Code Distribution for sequences  
(c) DB Resources of Mapping

Figure 32: Mapping Statistics Graphs

6.4 Export Mapping Results

A tab separator text file can be exported with the corresponding mapping results (File → Export → Export Mapping Results”).

Copyright 2016 - BioBam Bioinformatics S.L.
7 Gene Ontology Annotation

7.1 Blast2GO Annotation Rule

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 33.

\[
DT = \max(\text{similarity} \times EC_{\text{weight}})
\]
\[
AT = (\theta \times GO - 1) \times \text{GO_{weight}}
\]
\[
AR: \text{lowest node}(AS(DT + AT)) \geq \text{threshold}
\]

Figure 33: 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 34) and in the Evidence Code Weight Configuration Dialog (Figure 35).

1. Annotation Cut-Off (threshold). The annotation rule selects the lowest term per branch that lies over this threshold (default=55).
2. GO-Weight. This is the weight given to the contribution of mapped children terms to the annotation of a parent term (default=5).
3. Filter GO by taxonomy: The filter will remove the Gene Ontology terms known not to be in the given taxonomy using the restrictions defined by Gene Ontology. You can select one of the given options or simply write a taxonomy id.
4. 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).
5. **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.

6. **Hit Filter.** This option allows you to consider only the first N hits during annotation. This option is correlative with “Only hits with GOs” feature.

7. **Only hits with GOs.** This option together with the “Hit Filter” option allows to apply it only on hits that have a GO term candidate.

8. **EC-Weight.** EC code weights can be modified at the following pages of the Run Annotation dialog by clicking Next. 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.
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.

### 7.2 Individual Annotation Results

Annotation results for each sequence can also be visualized on the GO DAG by selecting “Draw Graph of GO-Mapping with Annotation Score” at the context menu. Additionally, the “Change Annotation and Description” options of this menu offer also the possibility to adjust annotations specifically for a single sequence.

This function edits the annotation of the selected and allows typing and deleting of annotation or sequence description. A manual annotation check-box (see Figure 37) is available for marking sequences with manual annotation. The sequence will get the **pink** label on the Main Sequence Table.

### 7.3 Statistics

An overview of the extent and intensity of the annotation can be obtained from the Annotation Distribution Chart (Figure 38), which shows the number of sequences annotated at different amounts of GO-terms.
There are different statistic distribution graphs that can be generated:

- **Annotation Distribution**: This chart informs about the number of GO terms assigned per sequence.
- **GO Annotation Level Distribution**: A bar chart which shows all GO terms for all 3 categories for a given GO level taking into account the GO hierarchy (parent-child relationships).
- **Annotation Score Distribution**: A chart that shows the number of sequences per annotation score.
- **Annotated Seqs/Seq-Length**: Shows the relation between amount of annotated sequences and sequence lengths.
- **Number of GOs/Seq-Length**: Shows the relation between sequence length and number of GOs.
- **Go Distribution by 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**:
  - Molecular Function: 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.
  - Biological Process: Same as above but for Biological Process.
  - Cellular Component: Same as above but for Cellular Component.

In order to display the Annotation Statistics Wizard, Figure 39, go to the “charts” icon and select Annotation Statistics.
7.4 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 find this option in the “Annotation” menu. For more details visit the Annex Project at [http://www.goat.no](http://www.goat.no).

7.5 Annotate GOs from Blast Descriptions

This tool looks at every significant alignment (Right-Click → Show Blast Result on a sequence) for each sequence and searches their description lines for GO ids. These GOs are now directly annotated to the sequence, if the alignments similarity passes the desired minimum. Validation can also be applied and is recommended, it will remove intermediate GO terms.
7.6 Exporting Annotation

The annotation results can be exported in a variety of formats. This function is available under File → Export → Export Annotation.

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. **Genespring format.** One single row is given by sequence where 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.

3. **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)

4. **WEGO format (native).** One single row is given by sequence, including those without annotated GOs. Belonging GOs are added to each sequence separated by tabs. The format corresponds to the “WEGO native format”, shown in this example: [http://wego.genomics.org.cn/docs/input01.lst](http://wego.genomics.org.cn/docs/input01.lst)

5. Custom: It is possible to customize the exportation of the annotation file according to the information desired or the column separator see the next figure.

Blast2GO allows to export two additional annotation file formats.

1. Export Annotations in GO Annotation File Format (GAF v.2), which is the primary format currently used by the GO Consortium [http://geneontology.org/page/go-annotation-file-formats](http://geneontology.org/page/go-annotation-file-formats).

2. Export Annotation Descriptions.

7.7 Load Annotation Results (.annot)

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):
Figure 42: Export Annotations Custom Configuration

Blast2GO Annotation File (.annot)

Seq1   GO:0001234, glycolipid transfer protein-like  
Seq1   GO:0001264, GO:0004567, ...
Seq1   GO:0034567
Seq1   EC:2.1.2.10
Seq2   GO:0001234, sorbitol transporter
Seq2   GO:0001244
Seq3   GO:0001234, GO:0004567, GO:0009123
Seq3   EC:1.2.4.1, EC:3.1

7.8 Other Annotation Functions

There are still other annotation functions available in the sub menu:

- Remove Annotation. Delete Annotation results for the selected sequences.
- Filter Annotation by GO Taxa
- Validate Annotations. 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.
- Remove 1. Level Annotations
8  Data Distribution Charts

8.1  Project Statistics

Once finished any step or at the beginning we can obtain a general character chart in which it shows how is the data distribution depending on the state of each sequence. We will be able to know the amount of sequences that belong to a concrete state.

The data distribution can be visualized in two different charts, one a bar chart (Figure 43) and the other as a Pie chart (Figure 44).

Figure 43: Data Distribution Bar chart

Figure 44: Data Distribution Pie Chart

These are the different states we are going to find in the chart:

1. Total: The total amount of in the project (only in bar chart)
2. Without Analysis: Sequences without processing or have been reset in the BLAST menu.
3. With Only InterProScan Sequences that only have InterProScan and nothing else
4. Without Blast Hits: It has not been found any hit for these sequences after BLAST step.
5. With Blast: Successful sequences after BLAST step or they have been reset in the Mapping menu.

6. With Mapping: Successful sequences after Mapping step or they

7. With GO Annotated: Successful sequences after Annotation step.

8. With Manual Annotation: Manually annotated sequences before or after executing the annotation step

9. With GO-Slim Annotation: Sequences with GO-Slim Annotation

Each state will have assigned the specific colour as the Main Table Legend (Figure 113).

It is also possible to see the progress of the analysis. From the following chart we understand that the from the 1000 sequences 700 have blast results, but still need to be processed.

![Analysis Progress Chart](image)

Figure 45: Analysis Progress Chart
9 Gene Ontology Graphs

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

9.1 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.

9.1.1 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.

9.2 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.
10 Quantitative Analysis

As a Data Mining tool, Blast2GO provides various ways for the joint analysis of groups of annotated sequences.

10.1 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 a 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 “Graph” icon. To understand the different types of shapes please see section 17.4 where you see the corresponding legend.

![Combined graph visualization](image)

**Figure 46: Combined graph visualization**

![Combined Graph Drawing Configuration Dialog](image)

**Figure 47: Combined Graph Drawing Configuration Dialog allows to provide a graph title header and to choose between the different GO categories**

**Figure 47** shows the Combined Graph Drawing Configuration Dialog, where the following parameters are available:

- **Graph Title:** The user can enter a title for the graph result tab.
- **GO Categories:** Select which type of the Gene Ontology category you want to visualize. If you check all boxes, the three graphs will be visualized in three different tabs.

For each Gene Ontology category a graph will be displayed. Blast2GO allows to extract information from the graph nodes such as tool tip (Figure 19), create a subgraph from that specific
GO, create an Id list of the sequences that have been annotated with that particular GO (Figure: 50). The generated Id list can then be used within Blast2GO in the select by sequences feature (Section: 11).

Figure 48: Molecular Function Combined Graph

Figure 49: Graph Node Tooltip

Figure 50: Extract Node Information

10.1.1 Graph Side Panel

The generated combined graph is interactive and its parameters can be modified from the side panel.

- **View.** This section controls the graph visualization within its area.
  - Zoom
  - Collapse All: The nodes will collapse and only the root will be visualized.
Figure 51: Combined Graph Side Panel
• Expand All: The nodes will expand to the original graph visualization.
• Re-Layout: The whole graph will be re-scaled to adjust to the visualization area.

• **Search.** Allows to search for GO IDs/ Terms/ Description in the Combined Graph.

• **Node Info.** This parameter controls the information shown at a node. Possible values are:
  - GO ID: If checked the GO ID will be included in the node.
  - GO Name: The GO Names are shown in the node.
  - GO Description: When checked the GO Description will be included in the node.
  - Nodescore: The node score will be shown in the node.
  - Sequence Names: The names of the sequences annotated at each GO are included in the node. The limit number of names names to be displayed is 15.
  - Sequences: The number of sequences annotated with that particular GO will be displayed in the node.

• **Layout.**
  - Edge Labels: When checked the labels on the edges will be shown.
  - Expand/Collapse Icon: If checked the ions that represent expand/collapse on the node are displayed.
  - Only “is a” Relations: Only the is a relations between nodes will be displayed if the box is checked.
  - Color
    - Ontology: All nodes will be colored according to the ontology category. Biological Process - green; Molecular Function - blue; Cellular Component - yellow.
    - White: The nodes will turn white.
    - By Nodescore: A Score is computed at each node according to the formula: 
      \[ \text{score} = \sum_{\text{GOs}} \text{seq} \times \alpha^{\text{dist}} \] (1)
      where \( \text{seq} \) is the number of different sequences annotated at a child GO term and \( \text{dist} \) the distance to the node of the child. GO term Coloring by Score will highlight areas of high annotation density.
    - By Sequence Count: Node color intensity will be proportional to the number of contributing sequences at the node.

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

• **Charts.** (see section 10.1.2)

• **Save as.** The information present in a Combined Graph can be saved as image (.png) or in table format. This will generate a .txt file where all information related to each node of the plotted Graph is provided in different columns.

---

Copyright 2016 - BioBam Bioinformatics S.L.
• **Overview.** Provides a radar-like view of the graph, which allows to adjust the visible window.

• **Open With.** Open the graph information as TreeMap or WordCloud (see sections 10.1.4 and 10.1.3 respectively).

### 10.1.2 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 52).

![Combined Graph Pie and Bar-Charts](image)

**Figure 52: Combined Graph Pie and Bar-Charts**

Four possibilities are available:

1. Sequence distribution by GO level (Pie-Chart): This pie chart represents the number of sequences for each Gene Ontology term for a given level. See Figure 53.

2. Sequences per GO terms (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 54.

3. Top 50 GO terms (Bar-Chart): A bar chart representing the GO terms according to the number of annotated sequences. See Figure 55.

4. Sequence distribution by GO level (Bar-Chart): This bar chart represents the number of sequences for each Gene Ontology term for a given level. See Figure 56.

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

![Graph Level 2 Pie Chart](image)

**Figure 53: Sequence distribution by GO level**
Figure 54: Sequence Distribution/GO as Multilevel-Pie (#score or #seq cutoff)

Figure 55: Top 50 GO terms

Figure 56: Sequence distribution by GO level
10.1.3 WordCloud

A WordCloud is a visual representation for list of labels. The importance of words, here GO terms, are represented by its font size. The font size depend either the sequence count or the NodeScore of each GO term. The list of words can be limited to a specific Gene Ontology category (BP, CC or MF). The coloring is random. Several options to change the graphical appearance are available like the number of words, the orientation and shape of the cloud as well as the color scheme. See Figure 57.

Figure 57: Word Cloud

10.1.4 TreeMap

The TreeMap viewer allows to visualize graphs (hierarchical, tree-structured data in general) as a set of nested rectangles. Each branch of the tree is given a rectangle, which is then tiled with smaller rectangles representing sub-branches. The size of each rectangle represents the number of sequences associated to a given GO term or a GO’s NodeScore. See Figure 58.

Figure 58: A TreeMap representing a Gene Ontology Graph. The size of the rectangles represents the number of sequences or the NodeScore of each GO term.

Figure 59: Tree Map
10.2 Coloured GO Graphs from a text file

We can generate a GO graph from a text (.txt) file which contains a list of GOs and the desired colour for each of them. It is also possible to label groups of GOs with the same name. Figure 61 shows an example that was created introducing the following text file:

<table>
<thead>
<tr>
<th>GO</th>
<th>Value</th>
<th>Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0000003</td>
<td>6</td>
<td>Group A</td>
</tr>
<tr>
<td>GO:0040007</td>
<td>8</td>
<td>Group B</td>
</tr>
<tr>
<td>GO:0050896</td>
<td>1</td>
<td>Group B</td>
</tr>
</tbody>
</table>

The text file has to follow a simple structure, to be processed correctly. It may contain from 2 to 3 columns in each line. The first column has to contain a GO, the second a number (0.0 to ∞) and the optional third column contains a text that will be written into the octagon of the corresponding GO. The columns must be separated with a tabulator character.

Figure 60: Colour Configuration Window

Figure 61: Coloured GO Graph by Group
According to the example above Group B has two GO IDs that contain a different values. It is also possible to differentiate these GO IDs by colouring according to their values. In order to colour the octagon according to the value you should select the gradient colour in the next page on the colour graph configuration window (see Figure 62).

Figure 62: Select Colour to differentiate values within the same group.

Figure 63: Coloured GO Graph by Group value
10.3 Make GO Graph

The “Make GO Graph” function allows to visualize any set of GO terms/Ids.

Figure 64: Make GO ID Graph

Figure 65: Make GO Graph
11  Select

There are different functions for selecting and deselecting sequences. Most functions in Blast2GO are only applied to selected elements. Selections allow to create subset or apply certain functions to parts of a given dataset.

11.1  Select Sequences (PRO Feature)

The Select Sequences feature can be applied to Blast2GO Projects only and allows to select sequences for many different criteria. Selections can be added to existing ones, subtracted or created from scratch.

- Sequence Name. This is a general function for (de)selecting sequences by loading a file containing a list sequence IDs.
- Sequence Description. Blast2GO allows to (de)select sequences according to Blast result description.
- Species.
- Function (GO-Terms or GO-IDs). This is a general function for (de)selecting sequences by loading a file containing a list of GO-Terms or GO-IDs.
- InterProScan IDs.
- Enzyme IDs.

![Figure 66: Start New Selection](image)
11.2 Select sequence by color

This function allows (de)selection of sequences on the basis of their color code i.e. the processing stage they have (see Section 17.3).

![Select by Color](image)

Figure 67: Selection by Color

11.3 Invert Selection

This function will invert the current selection. Those sequence that are not selected will now be selected and vice versa.

11.4 Delete Selected Sequences

This function will delete selected sequences from the Main Sequence Table

11.5 Create ID List From Selection

A Blast2GO ID list object with the sequence names will be generated.

11.6 Sort by Selection

The selected sequences will be on the top of the Main Sequence Table.

11.7 Export Selected Sequences as DAT

Once you have the desired sequences selected you can export these instead of saving the whole dataset. In order to do so go to File → Export → Export Selected Sequences as DAT and browse for the folder you want to save.
11.8 Hide unselected sequences

This function will hide the sequences that are not selected from the Main Sequence Table. Only the selected sequences are shown.
12 Analysis Menu

In this menu contains various data-analysis features:

1. GO-Slim
2. Enzyme Code and KEGG
3. Enrichment Analysis (Fisher’s Exact Test)
4. Gene Set Enrichment Analysis (GSEA)
5. Run Rfam

Figure 68: Analysis Menu Options

These functions will be discussed in detail in the following section.

12.1 GO-Slim

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.

12.2 Enzyme Code and KEGG

Blast2GO provides EC annotation through the direct GO ->EC mapping file available at the GO website. 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 Pathway-Maps from KEGG (online) Figure: 69. 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 a new results tab. 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.

It is possible to load the KEGG pathway map of only one sequence on the context menu. It will show the KEGG maps associated to that particular sequence.

12.2.1 Statistics

To see the main Enzyme classes in the dataset it is possible to generate a distribution Enzyme Code chart on “charts” icon Figure: 70.
• 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.

Figure 69: Load Kegg Pathway Map

Figure 70: Enzyme Code Statistics

**Enzyme Code Distribution of Tomato_analysis**

<table>
<thead>
<tr>
<th>EC Classes</th>
<th>#Seqs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.-Oxidoreductases</td>
<td><img src="image" alt="Bar Chart" /></td>
</tr>
<tr>
<td>2.-Transferases</td>
<td><img src="image" alt="Bar Chart" /></td>
</tr>
<tr>
<td>3.-Hydrolases</td>
<td><img src="image" alt="Bar Chart" /></td>
</tr>
<tr>
<td>4.-Lyases</td>
<td><img src="image" alt="Bar Chart" /></td>
</tr>
<tr>
<td>5.-Isomerases</td>
<td><img src="image" alt="Bar Chart" /></td>
</tr>
<tr>
<td>6.-Ligases</td>
<td><img src="image" alt="Bar Chart" /></td>
</tr>
</tbody>
</table>

Figure 71: Enzyme Code Distribution
12.3 Enrichment Analysis (Fisher’s Exact Test)

Blast2GO has integrated the FatiGO 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 7 of this tutorial. This functionality can be found under Analysis → Enrichment Analysis (Fisher’s 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 72).

![Fisher’s Exact Test Dialog](image)

Figure 72: Fisher’s Exact Test Dialog

When there is no reference set selected, the whole dataset present in the project will be taken as reference. Click on the run button to start the analysis. Once competed the results will be shown in the Statistics Tab where the adjusted p-values of each GO term above a given threshold will be shown.

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

For further details please refer to the FatiGO publication Al-Shahrour et al. 2004.

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 74): 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 9 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 75).
Figure 73: Fisher’s Exact Test result table

Figure 74: “Enriched” Graph

Figure 75: Bar-Chart representation of Fisher’s Exact Test results
12.4 Gene Set Enrichment Analysis (GSEA)

Blast2GO includes the GSEA computational method that determines whether an a priori defined set of genes shows statistically significant, concordant differences between two biological states. GSEA considers experiments with genome wide expression profiles from samples belonging to two classes, labeled 1 or 2. Genes are ranked based on the correlation between their expression and the class distinction by using any suitable metric. Given an a priori defined set of genes S (e.g., genes encoding products in a metabolic pathway, located in the same cytogenetic band, or sharing the same GO category), the goal of GSEA is to determine whether the members of S are randomly distributed throughout L or primarily found at the top or bottom.

For further details please refer to the GSEA publication (Subramanian et al., 2005).

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 7 of this tutorial. This functionality can be found under Analysis ➔ Gene Set Enrichment Analysis (GSEA). A dialog screen appears. Ranked list of genes can be selected by uploading text files containing the lists of sequence IDs and a statistical value for each one or selecting a previous ID-Value-List created in Blast2GO (Figure 76).

![Figure 76: GSEA Dialog](image)

Click on the run button to start the analysis. Once completed the results will be shown in the bottom part where the p-values of each GO term above a given threshold will be shown (Figure 77).

The GSEA results viewer has two options for the visual display of the results:

1. Make Enriched Graph: Click here to generate a representation on the GO DAG: Nodes are color highlighted proportionally to their significance value. In the graph viewer the user can choose which type of calculated p-value to use for highlighting and the threshold for filtering out nodes.

2. Show Global Statistics: This option generates two charts: a plot of p-values versus normalized enrichment scores, which provides a quick, visual way to grasp the number of enriched gene sets that are significant and an histogram of enrichment scores across gene sets, which provides a quick, visual way to grasp the number of enriched gene sets. (Figure 78).
Figure 77: GSEA result table

Figure 78: Global Statistics of GSEA results
12.5 RFAM

The Rfam database is a collection of RNA families, each represented by multiple sequence alignments, consensus secondary structures and covariance models (CMs). The families in Rfam break down into three broad functional classes: non-coding RNA genes, structured cis-regulatory elements and self-splicing RNAs. Typically these functional RNAs often have a conserved secondary structure which may be better preserved than the RNA sequence. The CMs used to describe each family are a slightly more complicated relative of the profile hidden Markov models (HMMs) used by Pfam. CMs can simultaneously model RNA sequence and the structure in an elegant and accurate fashion (Rfam description from: [http://rfam.xfam.org/](http://rfam.xfam.org/)). Please cite: [Nawrocki et al., 2014].

---

12.5.1 Rfam Results and Statistics

Once Rfam analysis has finished a table with the corresponding results will be displayed. The results can be visualized in different charts.

---

Copyright 2016 - BioBam Bioinformatics S.L.
1. Hit Distribution: This chart shows a distribution of the number of hits for the blasted sequences in the Rfam analysis.

2. Biotypes Pie Chart: This pie chart shows the distribution of the Rfam families of blasted sequences.

3. Biotypes Distribution: The same as the former but bar-style.

4. E-Value Distribution: This chart plots the distribution of E-values for the Rfam blastn hits.

5. TreeMap: This visualisation allows to see the Rfam families (hierarchical, tree-structured data in general) as a set of nested rectangles.

Figure 81: Rfam Statistics Graphs and Visualization
13 Differential Expression Analysis

13.1 Pairwise Differential Expression Analysis

13.1.1 Introduction

This tool is designed to perform differential expression analysis of count data arising from RNA-seq technology. This application, based on the edgeR program \cite{Robinson2010}, allows identification of differentially expressed genomic features (e.g. genes) in a pairwise comparison of two different experimental conditions. The software package "edgeR" (empirical analysis of DGE in R), which belongs to the Bioconductor project, implements quantitative statistical methods to evaluate significance of individual genes between two experimental conditions. If you need more information about edgeR please visit: \url{https://bioconductor.org/packages/release/bioc/html/edgeR.html}.

13.1.1.1 General Workflow

The workflow to be followed to perform an analysis of differential expression is described in Figure 83.

13.1.1.2 Load Data

Go to File \rightarrow Load \rightarrow Load Count Table and select your .txt file containing the count table in tab-delimited format. Rows should correspond to genes and columns to independent libraries or samples (Figure 84). Counts represent the total number of reads aligning to each gene (or any genomic locus). The first column must contain gene identifiers.

Note:

- This application only accepts raw counts without any type of normalization.
- Replicates for each experimental condition are necessary.

Current count table can be saved as Count Table object (File \rightarrow Save).
13.1.2 Run Pairwise Differential Expression Analysis

Go to `diff-expr` (toolbar) → Run Differential Expression Analysis and choose the “Pairwise Differential Analysis Option”. Here you can specify the following parameters, which are divided in three different sections: Preprocessing Data (Figure 86(a)), Experimental Design (Figure 86(b)) and Comparison and Test (Figure 86(c)).

13.1.2.1 Preprocessing Data Page:

- **Filter low count genes:**
  - CPM Filter: Establish a filter to exclude genes with low counts across libraries, as those genes may interfere with the subsequent statistical approximations. Filtering is performed on a count-per-million (CPM) basis to account for differences in library size between samples (e.g. a CPM of 1 corresponds to a count of 6 in a sample with 6 million reads).
  - Samples reaching CPM Filter: Set a minimum number of samples in which the gene’s CPM is above the filter level (is expressed). If this value is set to e.g. five, at least 5 of the samples have to be above the given CPM. The number of samples of the smallest group is usually used (e.g. in an experiment that has two replicates for each condition (or group), a gene should be expressed in at least two samples). Set value to 0 if no filter is desired.

- **Calculate normalization factors to scale the raw library sizes:**
  - Normalization Method: Here the normalization takes the form of scaling factors for library sizes that enter into the statistical model. These correctional factors are used
to compute the effective library sizes. For further details please refer to the edgeR User’s Guide. You can select the normalization method to be used:

- TMM: Weighted trimmed mean of M-values. In this method weights are obtained from the delta method on Binomial Data (this method is recommended).
- RLE: Relative log expression. Scale factors are the median ratio of each sample to the median library (geometric mean of all samples).
- Upper-quartile: 75% quantile for the counts for each library is used to calculate the scale factors.
- None: Normalization factors are set to 1.

### 13.1.2.2 Experimental Design Page:

- **Experimental design file:** Select your .txt file containing your experimental factors with the experimental conditions associated to each sample in tab-delimited format. As demonstrated by Figure 85, rows correspond to samples and columns to experimental factors. Make sure that the names in the first column of the experimental design table are exactly the same as the sample names in the count table header. If your experimental design file has less samples than in the count table, only the samples contained in this file will be analyzed.
13.1.2.3 Comparison and Test Page:

- **Design Type:** The design type you want to adjust your analysis to.
  - Simple design: Makes a pairwise comparison between samples belonging to two experimental conditions. You only have to select the experimental factor of interest and establish the comparison selecting the reference and contrast condition of samples in “Primary Target”.
  - Paired design: Makes a pairwise comparison between samples belonging to two experimental conditions, adjusting for baseline differences of other experimental factors (e.g. “Individual” in Figure 86(b)). In this design, you have to establish the conditions for the comparison in “Primary Target” and the experimental factor for baseline difference in “Secondary Target”. This design type is appropriate for paired or blocking design, or experiments with batch effects.
  - Multifactorial Design: Makes a pairwise comparison between samples belonging to two experimental conditions with two experimental factors. For this design, you have to select the two experimental factors of interest and establish the reference and contrast group for each in “Primary Target” and “Secondary Target”. This design type is appropriate if you want to analyze effects of combined experimental conditions on gene expression.

- **Statistical Test:**
  - Select a Statistical Test:
    - Exact Test: Based on the quantile-adjusted conditional maximum likelihood (qCML) methods (similar to Fisher’s exact test). It is only applicable to datasets with a single factor design (simple design).
    - GLM (Likelihood Ratio Test): Based on fitting negative binomial Generalized Linear Models (GLMs) with the Cox-Reid dispersion estimates. Is a good choice for inferences with GLMs.
    - GLM (Quasi Likelihood F-Test): The empirical Bayes quasi-likelihood F-test is an alternative to Likelihood Ratio Test and provides more robust and reliable error rate control when the number of replicates is small.
  - Robust: Estimation is strengthened against potential outlier genes.
Figure 86: Run Pairwise Expression Analysis Dialogs
13.1.3 Results

Once the input counts have been processed and analyzed via the “Pairwise Differential Expression Analysis” tool, a new tab is opened containing results (Figure 87):

- logFC: A measure that describes how much the expression changes between conditions (log2-fold-changes are shown).
- logCPM: The average log2-counts-per-millions.
- LR: Likelihood ratio statistic for the GLM (Likelihood Ratio Test).
- F: Quasi-likelihood F-statistic for the GLM (Quasi Likelihood F-test).
- FDR: False Discovery Rate calculated by the Benjamini-Hochberg method (multiple hypothesis testing correction).
- Tags: Indicate whether a gene is upregulated (FDR < 0.05, logFC > 0) or downregulated (FDR < 0.05, logFC < 0).

Genes that have not passed the filtering step are not shown in the new tab.

![Figure 87: Generic Table Viewer](image)

Results can be saved as a Pairwise Results object. Note that it is not possible to perform the analysis on this object. For this purpose you have to open the Count Table object. If you want to see both count table and results, go to the File Manager and open the two .b2g files together.
A result page will show a summary of the pairwise differential expression analysis results (Figure 88).

![Figure 88: Results Summary](image)

If you want to filter for differential expression based on other FDR and/or logFC cutoffs, you can go to `diff-expr (toolbar) → View Results: Pairwise → Set Up/Down Tags` and establish new values for both cutoffs. Tags will be updated, and the result section of the Result Summary will change according to the new cutoffs. To view the updated summary results go to `diff-expr (toolbar) → View Results: Pairwise Analysis → Open Result Summary`. 
13.1.4 Charts and Statistics

Different statistics charts can be generated for a global visualization of the results. Go to `diff-exp` (toolbar) → View Results: Pairwise → Statistics and select the chart which you want to view.

- **Library Size per Sample**: Bar chart showing the number of read counts contained in each sample (Figure 89(a)).

- **MDS Plot**: Generates a two-dimensional scatterplot in which the distances represent the typical log2 fold changes between samples. You can select an experimental factor by which you want to color the MDS graphic (Figure 89(b)).

- **Volcano Plot**: Scatter chart that is constructed by plotting the negative log of the adjusted p-values (FDR) on the y-axis versus the log of the fold changes on the x-axis (Figure 89(c)). Upregulated and downregulated genes are shown in green and red respectively (FDR < 0.05, logFC > 1 or logFC < -1).

- **MA Plot**: A smear plot showing the log of the fold changes on the y-axis versus the average of the log of the CPM on the x-axis. DE genes (FDR < 0.05) are marked in red (Figure 89(d)).

![Figure 89: Charts and Statistics](image-url)
13.2 Time Course Expression Analysis

13.2.1 Introduction

This tool is designed to perform time course expression analysis of count data arising from RNA-seq technology. Based on the maSigPro program [Nueda et al. 2014], this application allows the detection of genomic features (e.g. genes) with significant temporal expression changes and significant differences between experimental groups. The software package ‘maSigPro’, which belongs to the Bioconductor project, implements a two steps regression strategy to find genes for which there are significant expression profile differences in time course RNA-seq experiments. If you need further information about maSigPro please visit: https://www.bioconductor.org/packages/release/bioc/html/maSigPro.html.

Figure 90: Differential Expression Analysis Interface

13.2.1.1 General Workflow

The workflow to be followed to perform a time course expression analysis is described in Figure [91].

13.2.2 Load Data

Go to File → Load → Load Count Table and select your .txt file containing the count table in tab-delimited format. Rows should correspond to genes and columns to independent libraries or samples (Figure [92]). Counts represent the total number of reads aligning to each gene (or any genomic locus). The first column must contain gene identifiers.

Note:

- This application only accepts raw counts without any type of normalization.
- Replicates for each experimental condition are required.

Current count table can be saved as Count Table object (File → Save).
Figure 91: General Workflow

Input data
Count table (txt)

Time Course Expression Analysis
Counts
Experimental design
Parameters

Output Results
Statistical Values
Result Summary

Charts
Venn Diagram
Expression Profile by Gene
Experiment-wide Expression Profiles
Summary Expression Profiles

Figure 92: Count Table File
13.2.3 Run Time Course Expression Analysis

Go to diff-expr (toolbar) → Run Differential Expression Analysis and choose the “Time Course Expression Analysis” option. Here you can specify the following parameters, which are divided in three different sections: Preprocessing Data (Figure 94(a)), Experimental Design (Figure 94(b)) and Analysis Options (Figure 94(c)).

13.2.3.1 Preprocessing Data Page:

- **Filter low count genes:**
  - CPM Filter: Establish a filter to exclude genes with low counts across libraries, as those genes may interfere with the subsequent statistical approximations. Filtering is performed on a count-per-million (CPM) basis to account for differences in library size between samples (e.g. a CPM of 1 corresponds to a count of 6 in a sample with 6 million reads).
  - Samples reaching CPM Filter: Set a minimum number of samples in which the gene’s CPM is above the filter level (is expressed). If this value is set to e.g. five, at least 5 of the samples have to be above the given CPM. The number of samples of the smallest group is usually taken (e.g. in an experiment that has two replicates for each condition (or group), a gene should be expressed in at least two samples). Set value to 0 if no filter is desired.

- **Normalization procedure:**
  - Normalization Method: Normalization is an important step to make the samples comparable and to remove possible biases (as sequencing depth bias) in count data. You can select the normalization method to be used:
    - TMM: Weighted trimmed mean of M-values. In this method weights are obtained from the delta method on Binomial Data (this method is recommended).
    - RPKM: Reads Per Kilobase per Million mapped reads. This method correct for gene length and the number of sequencing reads (gene length is required).
    - Upper-quartile: 75% quantile for the counts for each library is used to calculate the scale factors for normalization.
    - None: It is not applied any normalization procedure.
  - Feature Length File: For RPKM normalization load a tab-delimited file (or ID-Value object) with two columns containing name and length of each gene or genomic feature.

13.2.3.2 Experimental Design Page:

- **Experimental design file:** Select your .txt file containing your experiment descriptors associated to each sample in tab-delimited format. As demonstrated by Figure 93 rows correspond to samples and columns to experimental descriptors. A column must contain the associated time points for each sample, and another column should show the assignment of samples to experimental groups. Make sure that the names in the first column of the experimental design table are exactly the same as the sample names in the count table header. If your experimental design file has less samples than count table, only the samples contained in this file will be analyzed.
13.2.3.3 Analysis Options:

- **Design Type:** The design type you want to adjust your analysis to.
  - Single Series Time Course: Detects genes that show significant expression changes over time. You only have to select the time factor of your experimental design in “Targets”.
  - Multiple Series Time Course: Find genes with significant temporal expression changes and significant differences between experimental groups. You have to establish the time and experimental factors, and select the control condition of your experimental design in “Targets”.

- **Statistical Settings:**
  - Significance Level (Alfa): The level of FDR control used for variable selection in the stepwise regression.
  - R-squared Cutoff: Cutoff value for the R-squared of the regression model.

- **Visualization of Results:**
  - Number of Clusters: Establish a number of clusters to group genes by similar expression profiles.
  - Clustering Method: Choose a clustering method for data partitioning.
    - Hierarchical Clustering: Performs a hierarchical cluster analysis using a set of dissimilarities for the features being clustered.
    - K-Means Clustering: Is intended to divide the points into K clusters such that the sum of squares of the points to the centers of the clusters assigned is minimized.
    - Fuzzy c-means: Groups features based on the Euclidean distance and the c-means objective function which is a weighted square error function.
Figure 94: Run Pairwise Expression Analysis Dialogs
13.2.4 Results

Once the input counts have been processed and analyzed via the “Time Course Expression Analysis” tool, a new tab is opened containing statistical results obtained by the stepwise regression statistical test (Figure 95):

- P-value of the regression ANOVA.
- R-squared of the model.
- P-value of the regression coefficients of the selected variables.
- Tags: Indicate the list/s of significant genes in which the feature appears (R-squared > R-squared Cutoff).
  - Red tags: Lists of significant genes for each experimental group (only available in “Multiple Series Time Course”).
  - Blue tags: List of significant genes for each variable of the regression model.

Only the genes that have passed the established Significance Level (13.2.3.3) are shown in the new tab. For further details please refer to the maSigPro User’s Guide.

![Figure 95: Generic Table Viewer](image)

Results can be saved as a TC Results object. Note that is not possible to perform the analysis on this object. For this purpose you have to open the Count Table object. If you want to see both count table and results, go to the File Manager and open the two .b2g files together.
A result page will show a summary of the time course expression analysis results, including the cluster of features with similar expression profiles (Figure 96). Go to `diff-expr (toolbar)` → View Results: Time Course Analysis → Open Result Summary in order to visualize the result summary.
13.2.5 Charts and Statistics

Different statistics charts can be generated for a global visualization of the results. Go to diff-expr (toolbar) → View Results: Time Course Analysis → Statistics and select the chart which you want to view.

- **Venn Diagram**: Diagram showing all possible logical relations between a finite collection of different feature sets. (Figure 97(a)). You can choose between two types of Venn Diagram (“Pairwise” or “Triple”), and select the sets of significant genes to display.

- **Expression Profile by Gene**: Graph of gene expression profiles over time for a particular gene (Figure 97(b)). It is possible to see them by right-clicking on the chosen gene, and selecting the “Show Expression Profile” option.

- **Experiment-wide Expression Profiles**: Plot showing the expression level levels across samples for each cluster of genes (Figure 97(c)).

- **Summary Expression Profiles**: Plot showing the median level expression of each cluster of genes across time (Figure 97(d)).

Figure 97: Charts and Statistics
14 Gene Finding

Thanks to the advances of next-generation sequencing methods, complete genomic sequences are becoming more and more abundant. Most frequently, after a first genome assembly, the consensus sequence is used for structural annotation. This type of annotation, among others, provides information about gene locations. Gene-Finding is an essential step preceding the functional characterisation of genomics elements and fits perfectly in the Blast2GO workflow.

There are basically two types of gene finding, ‘ab initio’ and ‘hint based’. For the ‘ab initio’, we only need the DNAseq data and, using HMMM, we are able to build models to mathematically and probabilistically predict the positions of genes.

Note: In all ‘ab initio’ gene prediction approaches, the number of genes is overestimated, i.e ‘ab initio’ methods raise the number of false positives in order to minimize the false negatives.

14.1 Gene Finding Tool

With this tool we intend to provide an easy and fast way to locate the genes using the ‘ab initio’ methodology on your prokaryotic or eukaryotic query genome, without need of RNA-seq data, and directly obtaining fully a exportable Blast2GO project and GFF files.

This tool integrates the Augustus and the Glimmer algorithm within Blast2GO, executing them in the service cloud. This allows to obtain the results in any OS without excessive memory consumption on the client.

14.2 Eukaryotic Gene Finding by Augustus

Augustus (Hoff and Stanke [2013]) is a program that predicts genes in eukaryotic genomic sequences; it is one of the most accurate programs for the species it is trained for. In the human ENCODE project, it proved to be the most accurate gene finder among the tested ‘ab initio’ programs. In the more recent nGASP (worm) project, it was again among the best in the ‘ab initio’ and transcript-based categories.

The accuracy of Augustus lies on his precomputed models which facilitate fast and accurate gene prediction.

14.2.1 Eukaryotic Gene Finding: Wizard

In order to speed up the gene finding process, the fasta will be split by sequence, i.e, each fasta entry will be sent to a different node for parallel execution.

14.2.2 Eukaryotic Gene Finding: Page 1

- **FASTA file**: The query file contains the DNA input sequence which must be in decompressed (multiple or single) FASTA format. Every letter other than a,c,g,t,A,C,G and T is interpreted as an unknown base. Digits and white spaces are ignored. The number of characters per line is not restricted.

  Note: The differences that make the fasta identifiers unique must be within the first 30 characters to be recognized by Augustus.
• **Closest species**: This list allows for the selection of the closest related organism to your query, in order to obtain the most accurate prediction.

• **Strand**: Here you can choose the sense of the gene search, obtaining the predicted genes on the forward strand, the backward strand or on both strands.

• **Type of gene**: With this option you can select the gene model.
  
  ◦ partial: allows prediction of incomplete genes at the sequence boundaries (default)
  ◦ intronless: predicts only single-exon genes like in prokaryotes and some eukaryotes
  ◦ complete: predicts only complete genes

• **Output type**: Specify whether the output sequences will be extracted as nucleotides or amino acids.

• **Protein length threshold**: Set a minimum length of the predicted proteins.

• **Allow in frame stops**: Activating this checkbox will allow the detection of genes containing a stop codon in its reading frame, detecting fragment genes with some undetected zones; normally it’s the most suitable option for an ‘ab initio’ search.
The eukaryotic gene finding can be executed 'ab initio', using only DNAsq data, or using 'hints' obtained from the RNAsq alignment in order to increase the truthfulness of the predicted genes.

- **RNAsq alignment file**: The file containing the alignments in BAM format. This file is the output of every RNAsq aligner program as TopHat, BWA or STAR. To be able to locate hints in the alignment file, it must not be filtered by any parameter, that means that it must be the same file that you obtain from the aligner. For this reason, the alignment files from Ensembl are not suitable for retrieving hints as they are filtered and processed.

- **Qmap threshold**: This parameter allows to filter the aligned reads that will be used to create the intron 'hints'. The Qmap corresponds to the mapping quality in a range from 0 to 60 and it is calculated as:

  \[ Q = -10 \times \log_{10}(Wrong\ mapping\ position\ probability) \]

  Meaning that a Qmap of 50, corresponds to a mapping error of \(5 \times 10^{-5}\). Default: 50

- **Minimum read alignment**: Specify the minimum length of the read that must map to the reference genome at the beginning of the intron. If this value is too small, it can lead the program to detect an intron derived from a miss-alignment. Note: This value has 0 as minimum and the maximum depends on your reads length. Default: 11

- **Minimum intron length**: Sets the minimum intron length. Default: 32
14.3 Prokaryotic Gene Finding by Glimmer

Glimmer (Gene Locator and Interpolated Markov ModelER) (Delcher et al., 1999) is a system for finding genes in microbial DNA, especially the genomes of bacteria, archaea and viruses. Glimmer uses Interpolated Markov Models (IMMs) to identify the coding regions and to distinguish them from non-coding DNA. Glimmer was the primary microbial gene finder used at The Institute for Genomic Research (TIGR), where it was first developed, and since then has been used to annotate the genomes of hundreds of bacterial and archaea species from TIGR and other labs.

The precision of Glimmer lies in its Interpolated Context Models (ICM), which are built for every query genome, by calculating and adapting the algorithm parameters to the GC content, the start and stop codons, etc.

14.3.1 Prokaryotic Gene Finding: Wizard

To create the most accurate model for your genome, this tool joins all the input fasta files, and builds the model with it. Once the model is built, it performs the gene finding for each entry in the files. This methodology allows you to save the model created with all your sequences (belonging to the same organism), and use it to find the genes on a short sequence without loading the entire genome. If you are running this tool on small genomic fragments, the genome of the closest available evolutionary relative of the target organism can be used to provide a training set of genes, if no genome is available for your organism.

14.3.2 Prokaryotic Gene Finding: Page 1

This page groups the main settings regarding your query genome.

- **FASTA file**: The query file contains the DNA input sequence and must be in decompressed (multiple or single) FASTA format. You can select a folder or multiple fasta files. *Note: Be sure to select only the fasta files containing the sequences of your query organism.*

- **Genetic code**: Here you can choose the genetic code for your genome. Only the 1st, 2nd, 11th, corresponding to the General genetic code, the Mycoplasma/Spiroplasma Code and the Bacterial and Archeal code are available.

- **Minimum gene length**: Allows to set the length threshold for the found genes in nucleotides.

- **Maximum overlap**: Here you can choose the maximum overlap length. Unlike eukaryotic genes, prokaryotic genes often have their genes overlapped.

- **Minimum gene score**: Every ORF found has an assigned score depending on his length, start and stop codons. Here you can modify the limit of the score necessary to consider an ORF a gene. Lowering these values will increase the number of genes found, but will also increase prediction errors.

- **Genome Shape**: Here you can specify the genome shape.
14.3.3 Prokaryotic Gene Finding: Page 2

The second wizard page is dedicated to the Interpolated Context Model (ICM) creation parameters. The ICMs are a further extension of Interpolated Markov Models (IMMs) used to identify the coding regions and distinguish them from non-coding DNA. This step is one of the most sensitive points of the process, as it will determine the accuracy of all the following gene predictions.

First, you can choose to create a new ICM, or to use one created previously. If you choose to create a new ICM, you can create one with the default parameters or modify the parameters by selecting the advanced parameters checkbox:

- **Allow in-frame stops**: ORFs with in-frame stop codons are omitted in the building of the model *Default: off*
- **ICM depth**: The maximum number of positions in the context window that will be used to determine the probability of the predicted positions. *Default: 7*
- **ICM Width**: Set the width of the ICM to the specified number. The width includes the predicted position. *Default: 12*
- **ICM Period**: The period is the number of different submodels for different positions in the text in a cyclic pattern, i.e., if the period is 3, the first submodel will determine positions 1, 4, 7, . . . ; the second submodel will determine positions 2, 5, 8, . . . ; and the third submodel will determine positions 3, 6, 9, . . .. For a non-periodic model, use a value of 1. *Default: 3*
- **Gene entropy cutoff**: If this cutoff is raised, more sequences will be identified as coding, resulting in more candidate genes.
Only genes with an entropy distance score smaller than x will be considered. This parameter is inspired by the observation that the coding sequences can be translated to an amino acid sequence capable of folding into a protein, whereas the non-coding sequences do not have this function. The class of amino acid sequences capable of folding to a protein has a global organizational order in contrast to those pseudo-amino-acid sequences generated from non-coding (or completely random) DNA sequences. Looking at the amino acid composition (or abundance) of a sequence we can determine the entropy of the resulting protein which allows us to cluster two kinds of sequences (coding and non-coding). Default: 1.15

If you choose to create a new ICM, you can save it by checking the option and selecting the output folder. The ICM file (.icm) can be used in posterior runs to saving computation time.

14.3.4 Prokaryotic Gene Finding: Page 3

This page groups the settings which pertain to the gene finding process. All of these settings are made in pairs. The first member of each pair is a checkbox allowing transition from the automatic value to the manually set value. Note: if the value is set as ‘Automatic’, these values will be calculated automatically.

- **GC content**: Allow the percentage of the content of G+C to be set.
- **Start codons**: Allow the start codons to be set as a comma separated list. Note: If you want to use only one start codon, it’s suitable to set the three start codons, and to change the weight of the desired start codon to 1 in the ‘start codons weight’ parameter.
- **Start codons weight**: Specify the probability of different start codons (same number and
order as in the ‘Start codons’ parameter). If the start codons have been specified without weights, then each start codon will be assigned equal weights (which is very unusual).

- **Stop codons**: Allow the stop codons to be set as a comma separated list.
14.4 Gene Finding: Results

Once the gene finding tools have finished, two objects will automatically be opened:

- **Sequence table**: Here you can see the traditional Blast2GO table showing the sequence name corresponding to the fasta ID line plus a gene identification, and the sequence length. *Note: this sequence can be on nucleotides or in amino acids, depending on the wizard selection.*

- **GFF3 table**: Here you can see the results as a gff file with:
  - **Sequence**: The name of the source sequence that belongs to this feature.
  - **Source**: The name of the program that has predicted this feature, in this case ‘Augustus’.
  - **Type**: The type of the feature, that can be ‘gene’, ‘mRNA’, ‘CDS’, ‘gene’, ‘Start’, ‘stop’, ‘gene’
  - **Start**: The coordinate of the start codon.
  - **End**: The coordinate of the stop codon.
  - **Score**: The score assigned to the feature, except the exons.
  - **Strand**: The strand of the feature, where a ‘+’ means that the feature is forward oriented and ‘-’ backwards.
  - **Phase**: The correct frame to translate this feature, the values can be ‘0’, ‘1’ or ‘2’. A gene ‘set’ of features can have variant phase values, due to a frame shift in an intron.
  - **Attributes**: Here we can see all the attributes assigned to each feature.

A Result Viewer is also opened to display some job statistics:

- **Eukaryotes**: The number and name of sequences per spitted file, the average number of exons, the minimum, maximum and average gene length, and the number of genes per strand.

- **Prokaryotes**: The name of each sequence present in the fasta file, the number of genes per sequence, the minimum and maximum gene length, and the strand position of the genes found.
Figure 105: Augustus Gene Finding Summary Results

Figure 106: Glimmer Gene Finding Summary Results
15 Tools (PRO Feature)

1. Set-to-Sense (Based on Best-Blast-Hit): Convert all selected sequences with a negative reading frame Best-Blast-Hit to anti-sense i.e. query-sequences will be translated to its reverse compliment (e.g.: ATTG -> CAAT). The tag “antisense” will be added to the end of the sequence names. Use the batch rename function to undo the name change.

2. Translate Longest ORF: Convert all selected sequences to its longest ORF protein sequence. The tag “ORF” will be added to the sequence names. Use the batch rename function to undo the name change. The user may select the reading frame, the genetic code depending to the species that will be considered to the prediction.

3. Search Loaded Annotations in Another Annotation Set: Compare a set of annotations for a given group of sequences against the annotations already loaded in Blast2GO.

4. Find Duplicated Sequences: Mark as selected or directly remove all sequences in the dataset which have the exact same sequence string.

5. Find Similar Sequences: Detect, Select and/or remove similar sequences within one project.

6. Batch Rename: Perform a batch rename of all selected sequences by converting, replacing or adding text to the actual sequence name.

7. Retrieve Blast Top-Hit: See 15.3

8. Retrieve Blat Top-Hit: See 15.4

15.1 Find Similar Sequences

This function allows to search for similar sequences within a dataset. The search for similar sequences is done via BLAT ([Kent, 2002]) alignments. The function searches a list of sequences against itself and reports all alignments above a certain similarity percentage. It is possible to remove similar sequences from the project or to extract a less redundant result dataset into a new project.

15.2 Find Duplicated Sequences

This function allows to quickly identify and remove redundant sequences (exactly the same sequences) within a dataset.

15.3 Retrieve Blast Top-Hit

This feature allows to retrieve the sequence information of Top Blast Hits in a Blast2GO project. Data can be obtained from the the NCBI, Ensembl or Uniprot webservices and stored in a new project or replace the existing IDs/sequences (see figure 107). A possible use case scenario would be a so called “Double-Blast”: The blast results of a first run are used to replace the sequence data for a second run against a different set of query sequences. Imagine an RNA-seq data-set with a high percentage of sequences without any alignments against a protein database (e.g. blastx against NR). This feature could be used to select and extract the sequences without hits (red ones) into a new project. These sequences could be basted first against a set of EST sequences. The initial unaligned sequences are now replace with the ESTs. Now the initial blastx search is repeated again the protein. For each Top-Hit (first significant alignment from an already performed BLAST), apply the filters (bottom part of the dialog) and search them in the corresponding database (online). It is possible to either replace the sequence from your data-set or to extract them into a new data-set (Action option). You can also decide whether you want to keep the original sequence names or if you want to rename them to the downloaded sequences names. The latter will add a small note to the sequence description, telling you the original name. The last remaining option allows you to decide whether you want to replace your sequences with the downloaded ones or if you just want to retrieve their name. This option is activated by default.
15.4 Retrieve Blat Top-Hit

This tool is very similar to “Retrieve Blast Top-Hit” explained above (see 15.3), but it employs BLAT ([Kent 2002]) instead (See figure 108). The dialog is therefore quite similar and the first 3 options are identical. BLAT needs a reference FASTA file which it uses to search for similar sequences. The last 2 options allow you to filter by similarity and if BLAT should consider the reverse strand.

15.5 Data Import and Export (PRO Feature)

Under the File and Tools menu there are several useful PRO features that can be used to manipulate sequence data.

15.6 Load

1. Extract and import sequences from a FASTA and a GFF/GTF file (Figure 109).
2. Load Accession List: Load Gene Ontology annotations via an Accession list.
4. Load GI-List: load Gene Ontology annotations via a GenInfo Identifier (gi) list. Please consider the identifier to be between vertical bar e.g. gi|356569257|.
5. Load Data from BioMart: Load Gene Ontology annotations from BioMart.
Figure 108: Retrieve Blat Top-Hit Dialog.

Figure 109: Extract and import sequences from a FASTA and a GFF/GTF file.
15.6.1 Export

1. Generic Export: This option allows you to export all the desired information to a text file.

2. Export Selected Sequences as Project: Only the selected sequences can be exported and saved in .dat file.

3. Export Sequence Table: Export the current Main Sequence Table for the selected sequences.

4. Export TopBlast data: It will export the best-blast-hit for each sequence, this is the hit with the lowest e-value.

5. Export GO Propagation: Exports the GO parents up to the root for the annotated sequences.

6. Export Sequences per GO (Gene Set):

7. Export GFF2/GTF:
16 View Menu

- Application Messages
- Welcome Messages
- Progress
- Java Memory Monitor

17 Help Menu

At the Help Menu you can find this Manual, Blast2GO papers and information of the authors. In case of a bug or a malfunction of Blast2GO you can save the log file and send it to support@blast2go.com or via the priority support (PRO Feature).

- App Manager: This option allows you to install/ uninstall Apps [17.1] available on Blast2GO website [https://www.blast2go.com/blast2go-pro/apps] and once installed the app will be displayed in the Analysis Menu (see section 12).
- Send Support Mail: Send an email to support with the log file already attached.
- Save Log to File.
- Startup Announcement
- Online Resources
- Table Legend (see section 17.3)
- Graph Legend (see section 17.4)
- CloudBlast Activity history (see section 17.2)
- User Account Information: Provides the information of the user account (PRO, Basic, Activation Key).
- About Blast2GO: Provides information of Blast2GO, Java and the computer where Blast2GO is installed.

17.1 App Manager

The App Manager allows you to install, update and uninstall Blast2GO apps.

![App Manager](image)

(a) Show featured and other Apps
(b) Update and uninstall installed Apps

Figure 110: App Manager
17.2 Cloud Activity History (PRO Feature)

The user can open the CloudBlast Activity History in order to monitor the number of consumed ComputationUnits or processed sequences and success jobs.

![CloudBlast Activity History](image1)

Figure 111: CloudBlast Activity History

17.3 Table Legend

The colour 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. Orange: 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. Pink: Manually annotated sequence
7. Violet: With only InterProScan
8. Yellow: GO-Slim view

![Main Table Legend](image2)

Figure 112: Main Table Legend that shows all colour categories
17.4 Graph Legend

The GO Graphs are displayed in different shapes. In order to understand what does each shape mean the Graph Legend (Figure...) will provide you with this information.

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

Figure 113: Graph Legend that shows the graph shapes
18 General Application Functions

18.1 Auto-Save (PRO Feature)

Blast2GO PRO allows to automatically and continuously save Blast2GO results after a certain amount of time.

![Figure 114: Blast2GO Auto-Save function](image)

18.2 General Preferences

In the General Preferences it is possible to enter a valid email address, which will be used in the QBLAST and also in the InterProScan searches. Furthermore, the path to the Blast2GO workspace can be provided and all results such as BLAST, InterProScan, charts and Blast2GO projects will be saved here.

![Figure 115: Blast2GO General preferences](image)
18.3 Update

Blast2GO allows automatic software updates during the application startup. These updates contain improvements, new features or bug fixes. It is possible to choose if you want to be notified of new updates or if you want to install software updates automatically (recommended).

It is also possible to specify the update behaviour of installed Apps. We differentiate between “Featured” and normal Apps. New “featured” Apps can be installed and updated automatically. Normal, non-featured Apps have to be installed manually but can be updated automatically.

Figure 116: Wizard to configure the Blast2GO update behavior

18.4 Data Access

To execute the Mapping step one of the following options have to be selected.

(a) Public Server Connection
(b) PRO Server Connection

Figure 117: DataAccess Configuration Dialog

18.4.1 Public Blast2GO Database

The Public Blast2GO Database contains the Gene Ontology database and all the information necessary to perform the mapping step i.e. to be able to link the different protein IDs to the functional information of the Gene Ontology database (see Section 6).

To be able to connect to the public Blast2GO port 80 has to be opened at your institution for out-going connections and you can see if you have the right database name (host-name or ip-address) and DB-Name (see Launch B2G) selected (in the Blast2GO menu under File → Preferences → Data Access (see Figure 117(a)).
18.4.2 Blast2GO Server (PRO Feature)

This tab allows you to choose the closest, fastest and most up-to-date Blast2GO server. Simply choose a location and a database version and click OK to connect Figure 117(b).

18.4.3 Local Blast2GO database (PRO Feature)

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.

18.5 Proxy

Proxy Settings. If a proxy server or a firewall is used to access the internet here you can define the proxy settings. A HTTP or a Socks proxy can be configured. In this window you can configure the proxy settings only for Blast2GO and this will overcome the system wide settings. If the Use Direct Connection checkbox is selected, the application will try to connect directly to the internet skipping any system settings. To use your defined proxy settings select the HTTP or Socks Proxy checkbox and complete the required fields.

![Proxy settings dialog](image)

Figure 118: Proxy settings dialog

Copyright 2016 - BioBam Bioinformatics S.L.
19  Troubleshooting

19.1  How to Change the Tempfolder Location of Blast2GO

Blast2GO can handle big data, but it needs plenty of free disk space in the systems temporary files folder. If you run into troubles with shortage of disk space, follow this guide to manually change the location of the folder for Blast2GO temporary files to e.g. another partition. To change the location of the tempfolder, add the `-Djava.io.tmpdir` parameter to the Blast2GO configuration file (Blast2GO.ini):

(...)
-XX:+UseConcMarkSweepGC
-XX:+DisableExplicitGC
-XX:SoftRefLRUPolicyMSPerMB=5
-Djava.io.tmpdir=D:\temp\orary_folder

See [19.3] on where to find and how to modify Blast2GO.ini.

◦ Please make sure that the folder you selected exists and that you have read/write permission.
◦ The path to the desired temp folder has to be absolute.
◦ The new tempfolder should offer enough free space.
◦ Any change to the Blast2GO.ini only takes effect after restarting Blast2GO!

19.2  How to Set Maximum Memory for Blast2GO

Blast2GO automatically reserves up to 80% of the available system memory (i.e. RAM), which is fine in most cases. However, sometimes and especially on Linux, this can lead to excessive swapping and will slow down Blast2GO. Follow this guide to manually change the maximum amount of memory Blast2GO will be allowed to use. Add the `-Xmx` parameter followed by the desired amount to the Blast2GO configuration file (Blast2GO.ini):

(...)
-XX:+UseConcMarkSweepGC
-XX:+DisableExplicitGC
-XX:SoftRefLRUPolicyMSPerMB=5
-Xmx2000m

See [19.3] on where to find and how to modify Blast2GO.ini.

◦ Use at least 1000M.
◦ Do not assign more RAM than your computer has.
◦ Only change this value if you know what you are doing and if it is necessary.
◦ Any change to the Blast2GO.ini only takes effect after restarting Blast2GO!

Examples:
◦ `-Xmx1000m` → 1000 Megabyte
◦ `-Xmx1g` → 1 Gigabyte
◦ `-Xmx6g` → 6 Gigabyte

On Ubuntu Linux you may also consider lowering `vm.swappiness` to 1, read this tutorial for more information:

https://help.ubuntu.com/community/SwapFaq#What_is_swappiness_and_how_do_I_change_it

Copyright 2016 - BioHam Bioinformatics S.L.
19.3 Where to find Blast2GO.ini

1. Locate your Blast2GO.ini (default: C:\Program Files\Blast2GO\Blast2GO.ini) and open it with administrator rights in order to be able to save it after modifying. Press the WINDOWS key and paste the following: notepad C:\Program Files\Blast2GO\Blast2GO.ini

2. Press CTRL + SHIFT + ENTER and click yes in the upcoming dialog to open notepad with administrator rights.

Locate your Blast2GO.ini (default: /home/user/Blast2GO/Blast2GO.ini) and open it with your favourite text-editor.

Go to the place where you installed Blast2GO (e.g. Applications/Blast2GO). Right click on the Blast2GO app, in the menu select Show package content. This will show you the files inside the app. Now go to Contents/MacOS/launcher/, inside the app. There you will find the Blast2GO.ini file which you can open in your favourite editor. (The default absolute path for the ini file is /Applications/Blast2GO/Blast2GO.app/Contents/MacOS/launcher/Blast2GO.ini).
Bibliography


<table>
<thead>
<tr>
<th>Nomenclature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACC</td>
<td>Accession</td>
</tr>
<tr>
<td>B2G</td>
<td>Blast2GO</td>
</tr>
<tr>
<td>BLAST</td>
<td>Basic Local Alignment Search Tool</td>
</tr>
<tr>
<td>EBI</td>
<td>European Bioinformatics Institute</td>
</tr>
<tr>
<td>EC</td>
<td>Evidence Code</td>
</tr>
<tr>
<td>GO</td>
<td>Gene Ontology</td>
</tr>
<tr>
<td>JRE</td>
<td>Java Runtime Environment</td>
</tr>
<tr>
<td>JWS</td>
<td>Java Web Start</td>
</tr>
<tr>
<td>EC</td>
<td>Enzyme Code</td>
</tr>
<tr>
<td>PIR</td>
<td>Protein Information Resource</td>
</tr>
<tr>
<td>URL</td>
<td>Uniform Resource Locator</td>
</tr>
<tr>
<td>XML</td>
<td>Extensible Markup Language</td>
</tr>
</tbody>
</table>