Contents

1 Clusters of Orthologs 2
2 Orthologous Group Annotation Tool 2
3 Statistics for NOG Categories 6
4 Merge Orthologous Group GOs to Annotation 7
1 Clusters of Orthologs

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

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

2 Orthologous Group Annotation Tool

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

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

The Orthologous Group Annotation Tool is launched using the “Find Ortholog Groups (COG)” button, inside the Analysis menu (Figure 1). The annotation is performed with the “Ortholog Group Annotation” option. Blast and Mapping Blast2GO steps are needed to be completed in order to perform the Orthologous Group assignment. The Orthologous Group Annotation will have a better performance if the Blast is executed against the UniProtKb database.

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

2.1 EggNOG

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

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

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

![Figure 2: Orthologous Group Annotation Wizard](image)

- **e-Value**: Select an e-Value limit to which the sequence will be included in the analysis. The e-Value is assigned in the Blast step. Select one of the multiple options available in this widget (Default: 1E-3).

- **Filter by Similarity (%)**: Specify a minimum similarity percentage from which the input sequences are filtered out. The similarity is gathered from the Blast result, it is obtained dividing the positive matches of the alignment and the Hsp length (Default: 50%).

- **Hsp/Hit coverage filter**: Establish a Hsp/Hit coverage to filter those results that cover less hit length than the minimum specified. The coverage is a percentage, therefore it must be a number between 0-100, 0 is selected to disable this filter option (Default: 0).

2.3 Orthologous Group Annotation: Algorithm

Blast and Mapping results are required to execute the Orthologous Group Annotation. When this feature is launched, all the Blast2GO-project selected sequences are iterated. Those sequences that do not have Blast/Mapping Annotation will be skipped, while the mapping results of the annotated sequences are extracted. The program iterates over the mapping results, if the Blast parameters pass the filters set in the Orthologous Group Annotation Wizard (section 2.2), the method identifies the Orthologous Group annotation of each mapping result (if it has been described).

Orthologous Groups are assigned to the project using the EggNOG database (section 2.1). Since EggNOG does not provide an API to assign the annotation directly from their REST service, we use UniProt RESTful service API, which contains the information of the Orthologous Group...
via EggNOG. The information regarding the Orthologous Group Description, Category, or Gene Ontology (see section 2.4) is gathered from the EggNOG RESTful API. Such information is stored in the b2gFiles folder locally, and loaded to a Blast2GO Object, which is visualized in a Table Viewer.

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

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

2.4 Orthologous Group Annotation: Results

The information gathered from the section 2.3 is retained in a Blast2GO object and visualized in a Blast2GO Table Viewer. Figure 3 shows the default viewer for the Orthologous Group Blast2GO object.

Figure 3: Ortholog Group Annotation Main Table Results. Results are ordered by sequence

The Annotation Results are visualized in eight columns:

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

The results can be visualized by Sequence ID, as it is shown in Figure 3, or can also be displayed reordering the Table by NOG ID or by OG Category. When the Orthologous Group object is saved into a .b2g file, it can be opened in different formats. Figure 4 shows the menu that is expanded when a Blast2GO Orthologous Group Object is right-clicked in the Blast2GO File Manager. The option “Open With” opens a second menu, which reflects the table viewer options: open table by sequence, by NOG ID, or by OG Category.

Note: Merge option in the right-clicked menu is a characteristic of all Blast2GO objects, two selected objects can be merged into a single Table Viewer. Details Viewer, in the “Open with” menu, show all the analysis performed in the selected object.

Opening the Orthologous Group Table by NOG ID (Figure 5), would allow the user to determine which sequences are present in each Orthologous Group. A new column, which shows the number of sequences present in each Orthologous Group, is added to the Table Viewer.

When the Orthologous Group Table is ordered by Category (Figure 6), the NOG IDs are grouped in their Orthologous Group Categories. Also, sequences that correspond to the NOG IDs grouped, are combined to allow the recovery of sequences by Category. Here, we add two new columns: one corresponding to the number of sequences per Category, and another corresponding to the number of NOG IDs per category. The column corresponding to NOG Description is not present.

All the results can be exported by selecting the Table, and clicking in File>Export>Export Table (Generic). The Table will be exported in the format selected when the Export is requested.
3 Statistics for NOG Categories

Once each sequence has been annotated with their corresponding Orthologous Group, this tool can be executed to gather an insight of the NOG Categories present along the input data. The Statistics for NOG Categories feature, counts the number of occurrences of each annotated Category along the input data, and displays it in a graph to allow a visual analysis of the existing Categories.

The tool is launched via Analysis > Find Ortholog Groups (COGs) > Statistics for NOG Categories. The Orthologous Groups Annotation Table must be selected when this feature is initiated.

3.1 Statistics for NOG Categories: Results

Figure 7 and Figure 8 correspond to the visualization of the NOG Category counts in horizontal bar and pie chart format respectively. The menu on the right of figure 7, can be used to change between visualizations, as well as to change between font sizes and styles. There is also a 'Save as' option available, which allows to save the charts in SVG, PNG, PDF or TXT format.

Figure 7: Horizontal Bar chart representing the Ortholog Groups Category count. The menu on the right shows multiple visualization/save options.
The Category counts are represented as total counts. These charts can be used to show differences in Category prevalence between different samples or conditions.

4 Merge Orthologous Group GOs to Annotation

With the Merge Orthologous Group GOs to Annotation feature, the GOs assigned to each sequence in the Orthologous Group Annotation (section 2.3) can be merged to the Mapping GO annotation. Since the Blast2GO project only displays the most specific gene ontologies, this function only adds those GOs that are more specific than the ones being displayed in the project; it uses the Blast2GO DAG feature to achieve this.

This feature is executed via Analysis>Find Ortholog Groups (COGs)>Merge Orthologous Group GOs to Annotation. It needs the main Blast2GO project to be selected in order to run.

Figure 9: Merge Orthologous Group GOs Wizard to upload the Orthologous Group Annotation object
4.1 Merge Orthologous Group GOs to Annotation: Wizard

Figure 9 shows the wizard of this feature. It provides a widget to select the Orthologous Group Annotation Object file path. Such file will be iterated to extract the GO Annotation that will be merged with the Blast2GO project.

4.2 Merge Orthologous Group GOs to Annotation: Results

When the GOs are merged into the Blast2GO project, the GO annotation is changed as more specific GOs are added into the project and less specific are removed. Figure 10 shows a statistical evaluation after the GO merge has been performed. It includes the following information:

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

![Merge Orthology groups GO annotation](image)

Figure 10: Merge Orthologous Group GOs result. Shows the statistics of: number of GOs before and after the merge is performed, confirmed GOs, too general GOs and new GOs

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