Blast2GO Command Line User Manual

Version 1.1 October 2015

BioBam Bioinformatics S.L.
Valencia, Spain

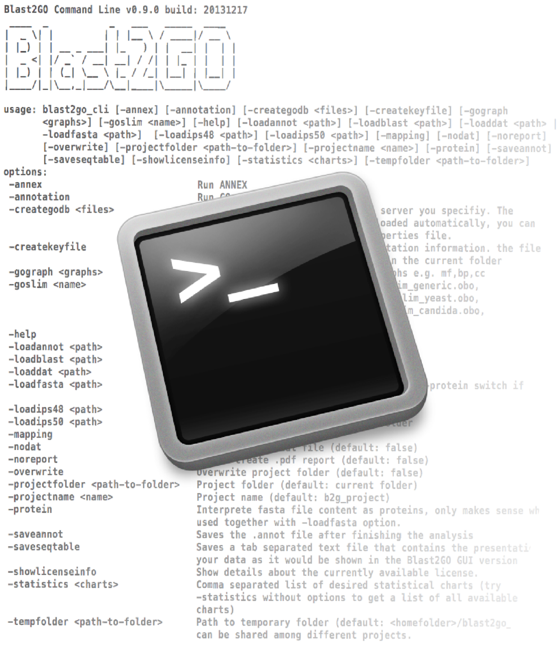
Contents

This report summarises the functional annotation process performed with the Blast2GO Command Line. The command line is based on the Blast2GO methodology, first published in 2005 (Conesa et al., 2005), for the automatic and high-throughput functional annotation and analysis of gene or protein sequences. The method uses sequence alignments (BLAST) to obtain a list of potential homologous for each input sequence. Blast2GO then maps Gene Ontology (GO) terms associated to the obtained BLAST hits and returns an evaluated functional annotation for the query sequences. Additional steps to improve the quality of the functional annotation are available. The following sections provide more detailed information about the different analysis steps as well as information about the input datasets, used parameters and the overall results.

1 Introduction

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

A Functional Annotation Pipeline: The Blast2GO Command Line is a professional solution for flexible, high-performance and automatic functional annotation tasks. This Annotation Pipeline allows your to integrate and automate your functional annotation task in a flexible way. Generate high-quality results in a reproducible way directly integrated into your data analysis workflows. The Blast2GO Command Line (CLI) version allows to combine BLAST and InterPro XML results, to perform the mapping against the Gene Ontology database and to assign the most suitable functional labels to the un-characterized sequence dataset. The command line allows to run the GO analysis on a local database which can be easily setup from the command line. Additionally the CLI allows to perform GO-Slim and Annex, to generate over 25 different summary charts and a comprehensive PDF report of each analysis.
The Command Line is based on the Blast2GO methodology, first published in 2005, [Conesa et al., 2005] for the functional annotation and analysis of gene or protein sequences. The method uses local sequence alignments (BLAST) to find similar sequences (potential homologous) for one or several input sequences. The program extracts all Gene Ontology (GO) terms associated to each of the obtained hits and returns an evaluated GO annotation for all query sequence(s). Enzyme codes are obtained by mapping to equivalent GOs and InterPro motifs can directly be added to the BLAST based annotation. A basic annotation process with Blast2GO consists of 4 steps: blasting and interpro-scan, GO mapping and functional annotation. [Götz et al., 2008]



```
Blast2GO Command Line v0.9.0 build: 20131217

usage: blast2go_cli [-annex] [-annotation] [-creategodb <files>] [-createkeyfile] [-gograph
       <graphs>] [-goslim <name>] [-help] [-loadannot <path>] [-loadblast <path>] [-loaddat <path> |
       -loadfasta <path>]  [-loadips48 <path>] [-loadips50 <path>] [-mapping] [-nodat] [-noreport]
       [-overwrite] [-projectfolder <path-to-folder>] [-projectname <name>] [-protein] [-saveannot]
       [-saveseqtable] [-showlicenseinfo] [-statistics <charts>] [-tempfolder <path-to-folder>]
options:
 -annex                             Run ANNEX
 -annotation                        Run GO
 -creategodb <files>                                            server you specifiy. The
                                                                oaded automatically, you can
                                                                perties file.
 -createkeyfile                                                 tation information. the file
                                                                n the current folder
 -gograph <graphs>                                              phs e.g. mf,bp,cc
 -goslim <name>                                                 im_generic.obo,
                                                                lim_yeast.obo,
                                                                im_candida.obo,


 -help
 -loadannot <path>
 -loadblast <path>
 -loaddat <path>
 -loadfasta <path>                                              -protein switch if

 -loadips48 <path>
 -loadips50 <path>
 -mapping                                                       lder
 -nodat                                    t file (default: false)
 -noreport                                 create .pdf report (default: false)
 -overwrite                         Overwrite project folder (default: false)
 -projectfolder <path-to-folder>    Project folder (default: current folder)
 -projectname <name>                Project name (default: b2g_project)
 -protein                           Interprete fasta file content as proteins, only makes sense wh
                                    used together with -loadfasta option.
 -saveannot                         Saves the .annot file after finishing the analysis
 -saveseqtable                      Saves a tab separated text file that contains the presentati
                                    your data as it would be shown in the Blast2GO GUI version
 -showlicenseinfo                   Show details about the currently available license.
 -statistics <charts>               Comma separated list of desired statistical charts (try
                                    -statistics without options to get a list of all available
                                    charts)
 -tempfolder <path-to-folder>       Path to temporary folder (default: <homefolder>/blast2go_
                                    can be shared among different projects.
```

1.1 Main characteristics

- **High Performance** The command-line version of Blast2GO allows you to analyse large dataset on your performant computing servers will nearly no extra effort.
- **Flexible** Easily integrate your functional annotation tasks within your custom analysis pipeline and run different analysis scenarios in parallel.
- **Automatic Data Generation** Generate all the statistics charts Blast2GO offers in an automatic fashion. This includes a summary report in PDF as well as images and text file formats.
- **Reproducability** Control the whole analysis with a simple configuration file. This allows you to set up different analysis strategies and reproduce the multiple scenario for one or various datasets.
- **Secure** Run BLAST, InterProScan and the Blast2GO annotation offline on your own servers according to your security requirements. Take 100% control of data sources and versions you use throughout the analysis.

1.2 Main features

- Perform Blast (Cloud/Local) directly from the Blast2GO Command Line
- Perform InterProScan from the Command line (online feature)
- Generate Gene Ontology Graphs now also off-line
- Improved performance local database installation/update
- Run Blast2GO on your own servers and control all analysis steps from the command line.
- Automate your functional annotation
- Reproduce your results in a consistent manner
- Handle tens of thousand of sequences
- Design advanced annotation strategies
- Integrate Blast2GO into your existing analysis pipeline
- Work online with your own databases
- Create your own local Blast2GO database
- Fast import of BLAST and InterProScan results
- Automatically generate PDF Reports
- Save all your results to specific project folders
- Work consistent and effective once you found the right settings for your analysis

1.3 Developed by

Blast2GO Command Line is developed and maintained by BioBam Bioinformatics which is internationally recognised for its expertise in functional annotation and genome analysis.

2 Setup

2.1 System Requirements

Blast2GO Command Line (CLI) is a Java application and can be run on Mac, Linux and Windows systems. It is always necessary to have Java (version 1.6 or higher preferably from Sun/Oracle) installed, at least 1GB of RAM is recommended. The Blast2GO Command Line needs a local Blast2GO database (DB) to perform the mapping step. This DB can be generated with the CLI itself; however the previous installation and configuration of a MySQL database (GPL license) is necessary.
In general this program works offline, however the following features may depend on an internet connection:

- **-gograph** drawing if Graphviz is not properly configured.
- **-cloudblast** and **-cloudblastbalance**
- **-creategodb** if the necessary files are not downloaded manually.
- **-ips** (InterProScan)

2.2 Product activation

The Blast2GO Command Line has to be activated with an individual licence key file specific for each workstation. This key has to be requested from the Blast2GO support team via email. To do so, a signature of the workstation has to be generated first. The command line parameter **-createkeyfile** will generate such a file called information.b2g, which will be used to generate the license then.

Note: On **MS Windows** all the following commands starting with ./blast2go_cli.run must be replaced with blast2go_cli.exe.

Steps to activate the Blast2GO Command Line:

1. Generate the **information.b2g** file by executing the follow command:

   ```
   ./blast2go_cli.run -createkeyfile
   ```

2. Send the resulting **information.b2g** file to our support team: clisupport@blast2go.com
3. The Blast2GO support team will create a **license.b2g** key file which has to be placed in the same folder as the blast2go_cli executable.
4. You can check the details of a licence file with the option **-showlicenseinfo**

2.3 Create Properties File

The Blast2GO Command Line needs a properties file, that contains all the information of the different paramaters that can be changed for the analysis. The properties file can be created with this command:

```
./blast2go_cli.run -createproperties cli.prop
```

Once this file has been created it is possible to edit it with a text editor.

2.4 Setting up a local Blast2GO Database

Blast2GO CLI offers the possibility to install a Blast2GO database semi-automatically with the command **-creategodb**.
Important: Before running this command please check the following requirements which will also be described more in detail below:

1. **A working MySQL database installation**
2. **Available disk space** (approx. 120GB during the installation process)
3. **Obtain all necessary and up-to-date data files**
4. **Configure the cli.prop file regarding your setup** (db-version, db-user, etc)

MySQL database installation

A local MySql server or a working client to connect to a remote server is necessary. Furthermore the credentials for a user with sufficient permissions on the database server are needed and will be prompted in the terminal window. Since the command line also creates a Blast2GO default user to access the database, it is recommended to use a user also with **grant privileges** (e.g. the root user). If not possible, a pre-created DB can be used (DB name must match the DB indicated in the properties file (see section: 2.3)).

Available disk space

During the installation the program will download and extract various files, which occupy currently (September 2015) approx. 20GB. The final database installation needs 110GB of available disk space. It is not possible for the program to ensure suficient disk space, so if the database installation seems to hang, please check your available disk space.

Obtain the necessary data files

Before starting the automatic download, please check that the file URLs provided in the properties file are up-to-date. Next, start the database installation from the Blast2GO Command Line. Obviously, for this option you need a working Internet connection. If you can not connect to the Internet from the workstation you are running the command line you can also download the necessary files directly from the indicated URLs and copy them to a temporary directory. You can specify these files in the properties file and start the database installation.
To import the Gene Ontology data into the database we need the file **go_YYYYMM-assocdb-data.gz**. To find the most recent version of this file please open the following URL in a web-browser: `http://archive.geneontology.org/latest-full/`

Configure the properties file regarding your setup

Before creating the Blast2GO database you need to configure your **cli.prop** file. The following parameters will be used during the installation process.

These lines need to be adapted to your own database configuration:

```
Dbacces.dbname=b2g_sep15
Dbacces.dbhost=192.168.0.1 // also possible with a port 192.168.0.1:3306
Dbacces.dbuser=blast2go
Dbacces.dbpasswd=blast4it
```

The information provided in these 4 lines is used to create the Blast2GO database and to use it later on for the mapping process. In the second line the mandatory host id and optional MySQL port are given. It is recommended to name the database according to the version of the **assocdb-data** file like in our example above (b2g_MMMyy).

The next section defines the different file to be imported into the database:

```
Dbacces.assocdbdata=http://archive.geneontology.org/latest-full/go_201402-assocdb-data.gz
Dbacces.geneinfo=ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene_info.gz
Dbacces.gene2accession=ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2accession.gz
Dbacces.idmapping=ftp://ftp.pir.georgetown.edu/databases/idmapping/idmapping.tb.gz
```

Here you are able to adjust the URIs for the necessary files. In general it is only necessary to change the Gene Ontology file URL (**assocdb**). The other files, even though updated frequently, maintain their URLs.
If you do not want to use online resources i.e. work offline (without URLs to third party websites) you can also point the URI to a local file like for example with:

```
Dbacces.assocdbdata=file:///path/to/file/local_b2g_db/b2g_jan14/go_201402-assocdb-data.gz
```

Note that the 3 slashes at the beginning of the URI in this example are not a typing error, but necessary.

Troubleshooting

The automatic installation is roughly divided into 4 steps, each step consists in downloading, unpacking and importing one of the 4 necessary files. If you experience problems with one of these steps (see console output), you can restart the installation process only for specific steps. For example:

1. You start the installation with

   ```
   ./blast2go_cli.run -properties cli.prop -creategodb
   ```

2. Step 1 completed successfully but the program fails in the 2nd step because of a network connection error.

3. You can now restart the installation starting with the 2nd step

   ```
   ./blast2go_cli.run -properties cli.prop -creategodb 2,3,4
   ```

Updating an existing database

An existing database can not directly be updated from the command line. We will have to create a new database with a new name as described above and the previous/old database version has to be removed manually form the server (drop database b2g_MMMyy) if desired. It should be enough to change the dbname, adjust the assocdbdata file link and to start the installation once more.

3 Command Line Parameters

This section gives a quick guide on the parameters used in Blast2GO CLI. Some command examples will be given in the end of the detailed description.

1. Load or import data commands:
   -loadannot <path> Path to .annot file
   -loadb2g <path> Path to Blast2GO .b2g file
   -loadblast <path>Path to Blast .xml file (pre 2.2.31)
   -loadblast31 <path>Path to Blast .xml/.json/.zip file (2.2.31+)
   -loaddat <path> Path to Blast2GO .dat file
   -loadfasta <path> Path to fasta file. Activate -protein option when working with amino acids.
   -loadips48 <path> Path to InterProScan 4.8 file or folder
   -loadips50 <path> Path to InterProScan 5.0 file or folder

2. Analysis commands:
   -annex Run ANNEX to complement the Gene Ontology annotation based on existing molecular functions
   -mapping Run the Gene Ontology mapping
   -annotation Run the Blast2GO annotation algorithm
   -goslim <path> Run goslim using an *.obo file. Possible subset obo files can either be downloaded from http://geneontology.org/page/go-slim-and-subset-guide, or customized by hand with OBO-Edit2.
   -cloudblast <cloublastkey> Run CloudBlast via webservice. This requires a working internet connection and a valid CloudBlast key with a positive balance.
   -cloudblastbalance <cloublastkey> Print the CloudBlast balance. This requires a working internet connection.
   -extractfasta <path> Extract features from a fasta reference to a fasta file (path). Needs configuration in the properties file.
   -ips <email> Run InterPro via webservice. This requires a working internet connection, a valid email address and that your data-set contains sequence data.
   -localblast <path> Run Blast against a local database. This requires a working internet connection in order to download the necessary Blast executable (Alternatively you can specify a binary folder manually and place the binary there). Also necessary is a correctly configured local blast database (properties file).

3. Save or export commands:
   -saveannot <path> Save the functional annotations (Gene Ontolgoy terms and Enzymes) as .annot
   -saveb2g <path> Save the project as .b2g
   -savedat <path> Save the project as .dat
   -savelog <path> Save the log in a specified file. The options -nameprefix and -workspace will be ignored.
   -savelorf <path> Convert nucleotide sequences (FASTA format) into amino acid sequences (longest Open Reading Frame, FASTA format). This function may be used to prepare a FASTA file for a local InterProScan run.
   -savereport <path> Create .pdf report
   -saveseqtable <path> Save your data as it would be shown in the Blast2GO GUI version (tab separated)
   -gograph <graphs> Provide a comma-separated list of desired graphs e.g. 'mf,bp'. Possible values are mf (Molecular function), bp (Biological Process) and cc (Cellular Component). Internet connection necessary. The option -nameprefix will be ignored. Can be used working offline if Graphviz is configured.
   -statistics <charts> Provide a comma-separated list of desired statistical charts (try -statistics without options to get a list of all available charts). '-statistics all' will try to export all statistics that are available. The option -nameprefix will be ignored.

4. Other Options:
   -createproperties <path> Path to where the default properties file should be created
   -godbstat <1,2,3,4> Print a statistic about a GO database specified in the cli.prop. Useful to get a hint if the database installation has been successful.
   -creategodb <1,2,3,4> Create a Gene Ontology database on a specified MySQL server. All necessary files (4) will automatically be downloaded if not provided by the user. The corresponding URLs (uris) can be changed in the properties file.
   -createkeyfile Create a file which contains a unique ID for your workstation. This file is nessesary to issue licence keys. The file will be placed in the current folder
   -help Display this message
   -nameprefix <name> Prefix for the output files, if you do not specify any path in particular (default: b2g_project)
   -properties <path> Path to properties file (mandatory)
   -protein Set this flag if the fasta file contains protein sequences. This option only makes sense together with the -loadfasta option.
   -showlicenseinfo Show details about the currently available license.
   -tempfolder <path> Path to temporary folder (default: System temp folder)
   -useobo <path> The obo file to use for annotation, graph drawing, some statistics and various file im- and exports.
   -workspace <path> Workspace folder, e.g. where the results will be saved if not specified (default: current folder)

   Note:
   If a path is specified for a save option (e.g. -saveannot), the options **workstation** and **nameprefix** will be ignored for this particular option (see Use Case Examples for detailed information).

4 Use Case Examples

The following examples have been executed in a Linux operating system.

Please note: If you are using MS Windows all commands must be changed accordingly: Please replace ./blast2go_cli.run with blast2go_cli.exe

1. Load a DNA fasta file, add the corresponding blast results and perform mapping and annotation. Furthermore, we want to save the .dat file and the PDF report at the current directory with the given name (example).

   ```
   ./blast2go_cli.run -properties cli.prop -useobo go_latest.obo -loadfasta
   example_data/1000_plant.fasta -loadblast example_data/1000_plant_blastResult.xml
   -mapping -annotation -savedat example -savereport
   example
   ```

2. Load nucleotide sequences, run local blast against swissprot database, mapping and annotation. Save the 3 combined graphs in a folder with nameprefix localblastSwissprot in .svg, .b2g and .txt format. In this example a blast, mapping and annotation statistics will also be stored. Finally the whole project will be saved in the example_data folder in .b2g format and also the corresponding log file. For this example the properties file have to be added accordingly.

   Properties file:

   ```
    // ** LocalBlastAlgoParameters **
    LocalBlastAlgoParameters.blastProgram=blastx-fast
    LocalBlastAlgoParameters.blastDbFile=/path_to_swissprot_database(.pal/.psq)/
    LocalBlastAlgoParameters.blastXMLResultEnable=true
    LocalBlastAlgoParameters.blastXMLResult=/path_to/example_data/LocalBlastresults_swissprot.xml
   ```

   Command Line:

   ```
   ./blast2go_cli.run -properties cli.prop -useobo go_latest.obo -loadfasta
   example_data/15_plant.fasta -workspace example_data -nameprefix localblastSwissprot
   -localblast <path_to_blastx_executables> -mapping -annotation -statistics
   bspecdis,mdbresmap,aannotscore -gograph -saveb2g -savelog
   example_data/blast2goLog
   ```

3. Load nucleotide sequences, load the new version (.json or .xml2) of blast results from a zip file, run mapping and annotation. Save the whole project and its report as example_json.b2g.

   ```
   ./blast2go_cli.run -properties cli.prop -useobo go_latest.obo -loadfasta
   example_data/15_plant.fasta -loadblast31
   example_data/json/02X9PD4T01R-Alignment.json.zip -mapping -annotation
   -savereport example_data/example_json_report -saveb2g
   example_data/example_json
   ```

4. Load example.b2g from the example_data folder, which does not contain InterProScan and run it using ips parameter. This example will save the InterProScan results in a folder and also the whole project including them.

   Properties file:

   ```
   // ** InterProScanAlgoParameters **
   InterProScanAlgoParameters.ipsXMLResult=/path_to_folder_where_to_save_ips_results/IPS
   InterProScanAlgoParameters.ipsXMLResultEnabled=true
   ```

Command Line:

```
./blast2go_cli.run -properties cli.prop -useobo go_latest.obo -loadb2g
example_data/example.b2g -ips <valid_email_address> -saveb2g
example_data/example_withIPS
```

5. Convert sequences to protein and save as fasta file.

```
./blast2go_cli.run -properties cli.prop -useobo go_latest.obo -loadfasta
example_data/15_plant.fasta -savelorf example_data/15_plant_protein
```

6. To run this example, the previous point had to be executed. Run CloudBlast, mapping, annotation on the protein sequences and save the results in .b2g and customized annotation format. For this example the properties file have to be added accordingly.

   Properties file:

```
// ** CloudBlastAlgoParameters **
CloudBlastAlgoParameters.blastProgram=blastp-fast
CloudBlastAlgoParameters.blastDB=nr_alias_viridiplantae
CloudBlastAlgoParameters.blastXMLResultEnable=true
CloudBlastAlgoParameters.blastXMLResult=/path_to/example_data/15seq_protein_plant.xml


// ** ExportAnnotParameters **
ExportAnnotParameters.format=custom
ExportAnnotParameters.desc=true
ExportAnnotParameters.go=category_and_id_and_term
ExportAnnotParameters.goseparator=tabulator
ExportAnnotParameters.column=tabulator
ExportAnnotParameters.row=sequence
```

   Here is an example output where C refers to Cellular Component as the category, followed by GO ID and Description:

```
C GO:0016021 integral component of membrane
```

   Command Line:

```
 ./blast2go_cli.run -properties cli.prop -useobo go_latest.obo
 -protein -loadfasta example_data/15_plant_protein.fasta -cloudblast
 B2G-CloudBlastKey-12435**** -mapping -annotation -saveb2g
 example_data/15_plant_protein -saveannot example_data/15_plant_annotation
```

7. Generate Combined Graphs locally without internet connection. Install Graphviz on the workstation and use it to generate the graphs. Load the example.b2g file and generate the combined graphs. In this example the parameters of the graph will be changed. A sequence filter will be applied and the graph colour is according the sequence count. The graphs will be saved in 3 different formats (.svg, .b2g and .txt) in the default "b2g_project" folder and at the current directory.

   Properties file:

```
// ** GraphvizParameters **
GraphvizParameters.dotExecutable=/usr/bin/dot
GraphvizParameters.dotExecutableEnabled=true

// ** CombinedGraphAlgoParameters **
CombinedGraphAlgoParameters.seqFilter=100
CombinedGraphAlgoParameters.graphMode=sequence_count
```

Command Line:

```
./blast2go_cli.run -properties cli.prop -useobo go_latest.obo -loadb2g
example_data/example.b2g -gograph
```

8. Load a protein fasta file, add the corresponding blast results and execute mapping and annotation. All files (.dat, .b2g, .pdf, .annot and .txt) will be saved with the nameprefix "p53" in the workspace folder "work_dir" at the current directory. In addition the data distribution pie chart and enzyme statistics will also be saved in the "work_dir" folder.

   Properties file:

   ```
   // ** EnzymeStatisticsAlgoParameters **
   EnzymeStatisticsAlgoParameters.ecFirstLevel=main
   ```

   Command Line:

   ```
   ./blast2go_cli.run -properties cli.prop -useobo go_latest.obo -loadfasta
   example_data/1000_seq_protein.fasta -protein -loadblast
   example_data/1000_plant_protein_blastResult.xml -mapping -annotation
   -workspace work_dir -nameprefix p53 -saveb2g -saveannot
   -savedat -savereport -saveseqtable -statistics gdatadispie,aecdis
   ```

9. Load a .dat file, apply plants GO Slim and save the results as .b2g, which will be saved with the default nameprefix "b2g_project" and at the current directory.

   ```
   ./blast2go_cli.run -properties cli.prop -useobo go_latest.obo -loaddat
   example.dat -goslim example_data/goslim_plant.obo -saveb2g
   ```

10. Load a fasta file, a blast result file and InterProScan 5.0 files, perform mapping, annotation and ANNEX. Then create all 3 GO graphs and all statistical charts. As a result we want to obtain the .b2g and the PDF report, which will be saved with the default nameprefix "b2g_project" and at the current directory.

    ```
    ./blast2go_cli.run -properties cli.prop -useobo go_latest.obo -loadfasta
    example_data/1000_plant.fasta -loadblast example_data/1000_plant_blastResult.xml
    -loadips50 example_data/plant_ipsr -mapping -annotation -annex -gograph all
    -statistics all -saveb2g -savereport
    ```

11. Here we spread some light on the usage of the **-nameprefix** and **-workspace** parameters which allows you to define in a very precise way where and how to generate all output files. Please be aware that the **-nameprefix** for the options **-gograph** and **-statistics** will be ignored and to save all generated output files in a particular folder the **-workspace** can be provided. Please note that the following examples are made with the -saveannot option but can be applied to any other save option (e.g. -savedat).

Command parameters	**Saved output file**
-saveannot	current_directory/b2g_project.annot
-saveannot path/abc	current_directory/path/abc.annot
-saveannot /path/abc	/path/abc.annot
-saveannot path/abc -nameprefix xyz **or** -saveannot path/abc -workspace mno **or** -saveannot path/abc -nameprefix xyz -workspace mno	current_directory/path/abc.annot
-saveannot -nameprefix xyz **or** -saveannot path/xyz	current_directory/xyz.annot
-saveannot -workspace mno **or** -saveannot mno/b2g_project	mno/b2g_project.annot
-saveannot mno/xyz	mno/xyz.annot

Bibliography

[Conesa et al., 2005] Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., and Robles, M. (2005). Blast2go: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, 21(18):3674–3676.

[Götz et al., 2008] Götz, S., Garcia-Gomez, J. M., Terol, J., Williams, T. D., Nagaraj, S. H., Nueda, M. J., Robles, M., Talon, M., Dopazo, J., and Conesa, A. (2008). High-throughput functional annotation and data mining with the blast2go suite. *Nucl. Acids Res.*, pages gkn176+.