

TOA (Tree-oriented Annotation)

v0.55

A software package for automated functional annotation in non-model plant species

GI Sistemas Naturales e Historia Forestal
(formerly known as GI Genética, Fisiología e Historia Forestal)
Dpto. Sistemas y Recursos Naturales
ETSI Montes, Forestal y del Medio Natural
Universidad Politécnica de Madrid

<https://github.com/ggfhf/>

Table of contents

Disclaimer	1
Introduction	2
Installation.....	7
TOA installation	7
Additional infrastructure software installation.....	8
Ubuntu Linux 18.04	8
Mac OS X 13.10	8
First steps	10
Starting TOA	10
Configuring the TOA environment	13
Installing bioinformatic software	14
Consulting submitted processes and troubleshooting	18
A step by step example	19
Load genomic data into TOA database	19
Basic data	19
Gymno PLAZA 1.0	20
Dicots PLAZA 4.0.....	20
Monocots PLAZA 4.0	20
NCBI RefSeq Plant	20
NCBI BLAST database NR.....	20
NCBI Protein GenInfo viridiplantae identifier list.....	21
NCBI Gene	21
InterPro	21
Gene Ontology	21
Create and run an annotation pipeline (amino acid).....	21
Create the config file	21
Edit the config file	24
Run a pipeline.....	24
Consult statistics data and show plots	26
Alignment - # HITs per HSPs	26
Annotation datasets - Frequency distribution	28
Species - Frequency distribution	31
Gene Ontology - Frequency distribution per term.....	34

Gene Ontology - Frequency distribution per namespace	36
Gene Ontology - # sequences per #terms.....	39
How to cite	42

Disclaimer

The software package TOA (Tree-oriented Annotation) is available for free download from the GitHub software repository (<https://github.com/GGFHF/TOA>) under GNU General Public License v3.0.

Introduction

Functional annotation is an essential stage of bioinformatic analysis of next-generation sequencing experiments. The purpose of this analysis is to assess the biochemical and biological functions of the sequences yielded after an assembly, the stage where reads are aligned and merged in longer fragments: contigs or scaffolds. In order to perform the functional annotation, it is necessary to access functional information deposited in genomic databases. Functional annotation is usually a slow task because it involves manual querying to the databases and the subsequent processing of the information obtained.

TOA aims to establish workflows geared towards woody plant species that automate the extraction of information from genomic databases and the annotation of sequences. TOA uses the following databases: Dicots PLAZA 4.0, Monocots PLAZA 4.0, Gymno PLAZA 1.0, NCBI RefSeq Plant and NCBI Nucleotide Database (NT) and NCBI Non-Redundant Protein Sequence Database (NR). Although TOA was primarily designed to work with woody plant species, it can also be used in the analysis of experiments on any type of plant organism. Additionally, NCBI Gene, InterPro and Gene Ontology databases are also used to complete the information.

TOA has a user-friendly front-end where researchers can perform the following tasks simply with a few clicks (Figure 1): i) download biochemical and biological data from genomic databases; ii) design the annotation pipeline that implements the most appropriate workflow for the experiment they are analyzing; iii) view statistics and graphics of results; and iv) track the status of processes during execution.

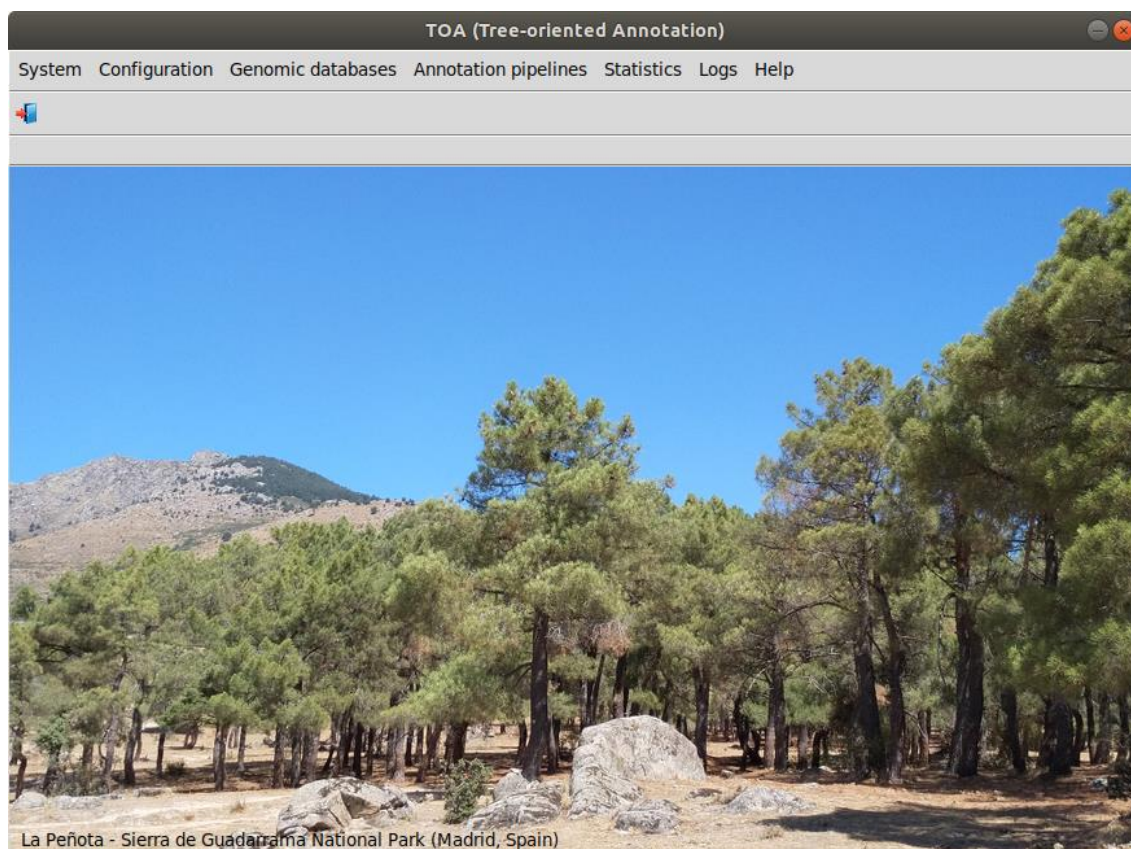


Figure. 1. Aspect of the program interface at startup.

We are going to provide an example to understand the functioning of TOA: We have carried out an RNA-seq experiment of Canary Island pine (*Pinus canariensis*) samples and we want to perform the functional annotation of the transcript sequences obtained in the *de novo* transcriptome assembly. First, we have to determine the genomic databases that we are going to use and the order in which we will carry out the processes. TOA is designed to explore the information in the databases sequentially. In this example, we will first explore the phylogenetically closest databases to the focal organism (in this case Plaza Gymno database), to end with the less specific database (NT). The database order is Gymno, Dicots, Monocots, RefSeq Plant and NT (selecting plant sequences). However, the order of the databases can be defined by the user.

Each time a specific database is queried is considered an iteration. An iteration has conceptually three steps (Figure 2):

- (1) extraction of the information from the external genomic database and loading in the TOA database.
- (2) performing an initial iteration of the full FASTA file with the nucleotide sequences against the sequences stored in the database selected in the first place.
- (3) alignment of the sequences that were not annotated in the previous iteration against the existing sequences in the remaining database.
- (4) annotation of the sequences aligned with the biochemical and biological information obtaining four files:

- (i) the alignments.
- (ii) the annotation description.

(iii) the annotated sequences in FASTA format; and (iv) the non-annotated sequences in FASTA format. The file of non-annotated sequences will be the input for the alignment of the next iteration

If we consider using predicted peptides instead of nucleotides, it is necessary a previous step to extract ORFs and predict coding regions of the transcript sequences. In this case, we will use the NR database instead of the NT database.

Once the iterations have been performed for the selected databases, the output files generated in the different iterations are merged to have files with unified information. Summary statistics and plots are generated with these files. In nucleotide pipelines, a FASTA file with annotated transcripts is output as well.

The TOA output files can be used with other related software, e.g. the unified alignment file by Blast2Go or the frequency distribution per GO term file by REVIGO.

Then, we will perform a new iteration with the non-annotated sequences using the plant databases to detect potentially contaminated sequences; in our example: fungi, bacteria, etc. To do so, we will process the non-annotated sequences with the complete NT or NR database (Figure 3). As a result, we will have a set of annotated contaminated sequences and other

sequences that have not been annotated. This second set can be due to chimeric sequences generated during the assembly stage, but it can also contain sequences that correspond to novel genes that lack functional annotation.

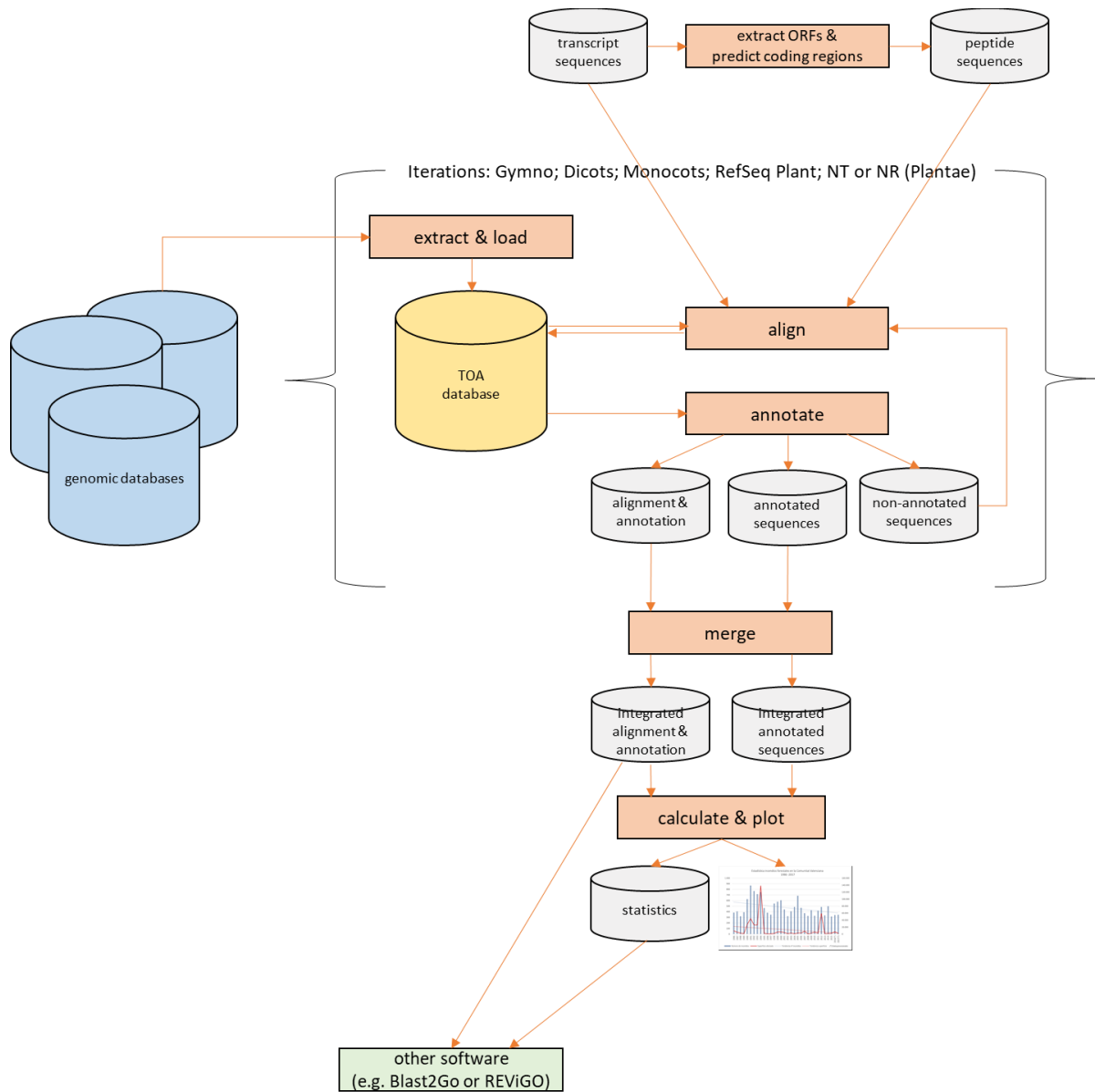


Figure. 2. Flowchart of sequence file annotation subjected to plant databases in TOA.

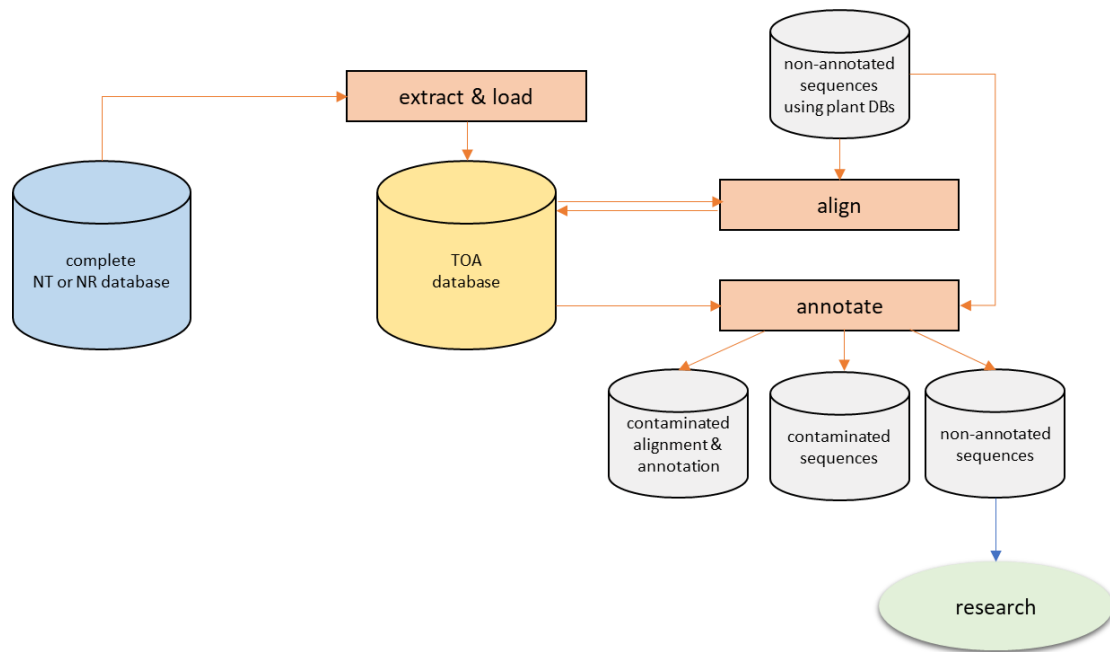


Figure. 3. Flowchart of the potentially contaminated sequence identification protocol in TOA. Those sequences previously not annotated using plant databases are subjected to the complete NT/NR databases.

Installation

TOA installation

TOA was programmed in Python 3 and it generates dynamic Bash scripts to perform annotation pipelines. TOA runs in any computer with Linux or Mac OS X.

TOA is available from the GitHub software repository of the GI Sistemas Naturales e Historia Forestal (<https://github.com/GGFHF/TOA/>), and it is distributed under GNU General Public License Version 3.

To download TOA (see Figure 4), click in *Clone or download* and then in *Download ZIP*.

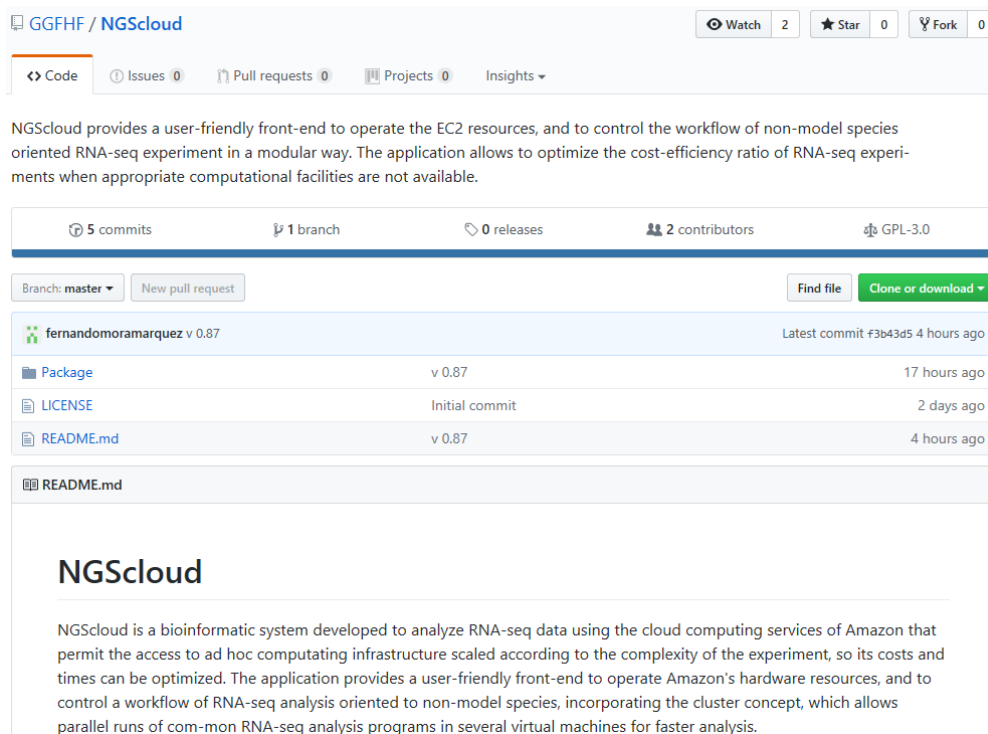


Figure. 4. TOA home at GitHub software repository.

To install TOA on Linux and Mac OS X, simply decompress the TOA-master.zip into a directory, typing the following command in a terminal window:

```
$ unzip TOA-master.zip
```

Then, the execution permissions of the programs must be set by using this command:

```
$ chmod u+x *.py *.sh
```

Additional infrastructure software installation

Python 3, version 3.6 or higher, is necessary for a correct functioning of TOA. If Python 3 is not installed in your computer, you can download it from the official website (<https://www.python.org/>), or use one of the several distributions that include Python along with other software packages for standard bioinformatic analysis such as Anaconda (<https://www.continuum.io/>).

To work properly, TOA needs the following Python modules:

- Tkinter (<https://docs.python.org/3.6/library/tkinter.html>), the standard Python interface to the Tk GUI toolkit.
- Requests (<https://requests.kennethreitz.org/en/master/>), an HTTP library.
- Plotnine (<https://plotnine.readthedocs.io/en/stable/>), an implementation of a grammar of graphics in Python based on ggplot2

Next, we present how to install this additional software in two example environments: a) an Ubuntu Linux 18.04 (Bionic Beaver) where Python3 is pre-installed in the OS; and b) a Mac OS X 10.13 (High Sierra) where Python is installed using Anaconda.

Ubuntu Linux 18.04

First, open a terminal window and type the following commands to install the Python 3 modules Tk, PIL and PIL.ImageTk and pip3, if necessary:

```
$ sudo apt-get install python3-tk python3-pil python3-pil.imagetk
```

```
$ sudo apt-get install python3-pip
```

Then install the Requests library, typing the following command in the terminal window, if necessary:

```
$ sudo pip3 install requests
```

And finally, install the Plotine library, typing the following command in the terminal window, if necessary:

```
$ sudo pip3 install plotnine
```

Mac OS X 13.10

First, install Homebrew and wget command, if necessary, typing the following commands in the terminal window:

```
$ /usr/bin/ruby -e "$(curl -fsSL https://raw.githubusercontent.com/Homebrew/install/master/install)"
```

```
$ brew install wget
```

Now download the Anaconda software file, e.g. the version 2019.09, typing the command:

```
$ wget https://repo.anaconda.com/archive/Anaconda3-2019.07-MacOSX-x86_64.sh
```

And provide execution permission to this file and run it typing the commands:

```
$ chmod u+x Anaconda3-2019.07-MacOSX-x86_64.sh
```

```
$ ./Anaconda3-2019.07-MacOSX-x86_64.sh
```

During the Anaconda installation, read the Anaconda End User License Agreement, accept the license terms and indicate the location where Anaconda will be installed. Later, review that the PATH variable incorporates the directory where Python is. So, if Anaconda installation directory is *Anaconda3_path*, edit *.bash_profile* file in the user's home directory and check the directory *Anaconda3_path/bin*. is added in the PATH variable through a sentence like this:

```
export PATH=Anaconda3_path/bin:$PATH
```

The appearance of the file *.bash_profile* must be similar to the file shown in the Figure 5.

```
# added by Anaconda3 2.3.0 installer
export PATH="/Users/jmoramarquez/bioinfo/anaconda3/bin:$PATH"

# >>> conda initialize >>>
# !! Contents within this block are managed by 'conda init' !!
__conda_setup="$( '/Users/jmoramarquez/bioinfo/anaconda3/bin/conda' 'shell.bash' 'hook' 2> /dev/null )"
if [ $? -eq 0 ]; then
    eval "$__conda_setup"
else
    if [ -f "/Users/jmoramarquez/bioinfo/anaconda3/etc/profile.d/conda.sh" ]; then
        . "/Users/jmoramarquez/bioinfo/anaconda3/etc/profile.d/conda.sh"
    else
        export PATH="/Users/jmoramarquez/bioinfo/anaconda3/bin:$PATH"
    fi
fi
unset __conda_setup
# <<< conda initialize <<<
```

Figure. 5. An example of the file *.bash_profile*.

Finally, install Plotnine, because other modules are by default when Anaconda is installed. To install it, type the following command:

```
$ conda install --channel conda-forge plotnine
```

First steps

Starting TOA

TOA runs in graphical mode using the graphical user interface (GUI), but it can also be run in console mode on server machines without GUI installed.

Start TOA in GUI mode typing the following command in a terminal window in the directory where the package of TOA is downloaded:

```
$ ./TOA.py
```

Alternatively, you can type also:

```
$ ./TOA.py --mode=gui
```

To run TOA in console mode:

```
$ ./TOA.py --mode=console
```

Initial appearance of TOA at application startup in GUI mode is shown in Figure 1.

TOA is structured in several menus:

System

Just to exit the application.

Configuration

This menu contains all the items related to:

- Recreate TOA config file
- View TOA config file
- Recreate TOA database
- Rebuild TOA database
- Install Bioinfo software

Genomic databases

Here, all options related to Genomic databases are shown:

- Basic data
 - Recreate genomic dataset file
 - Edit genomic dataset file
 - Recreate species file
 - Edit species file

- Download other basic data
 - Load data into TOA database
- Gymno PLAZA 1.0
 - Build proteome
 - Download functional annotations from PLAZA server
 - Load data into TOA database
- Dicots PLAZA 1.0
 - Build proteome
 - Download functional annotations from PLAZA server
 - Load data into TOA database
- Monocots PLAZA 1.0
 - Build proteome
 - Download functional annotations from PLAZA server
 - Load data into TOA database
- NCBI RefSeq Plant
 - Build proteome
- NCBI Plant database NT
 - Build BLAST database
- NCBI Nucleotide GenInfo Viridiplantae identifier list
 - Build identifier list using NCBI server
- NCBI Plant database NR
 - Build BLAST database
- NCBI Protein GenInfo Viridiplantae identifier list
 - Build identifier list using NCBI server
- NCBI Gene
 - Download functional annotations from NCBI server
 - Load data into TOA database
- InterPro
 - Download functional annotations from InterPro server
 - Load data into TOA database
- Gene Ontology
 - Download functional annotations from Gene Ontology server
 - Load data into TOA database

Annotation pipelines menu

The options included here allow to build and run scripts to perform the annotation process:

- TOA nucleotide pipeline
 - Recreate config file
 - Edit config file
 - Run pipeline
 - Restart pipeline
- TOA aminoacid pipeline
 - Recreate config file
 - Edit config file
 - Run pipeline

- Restart pipeline

Statistics

This menu contains all the items related to statistics corresponding to the annotation pipelines results and their plots:

- Alignment
 - # HITs per # HSPs data
 - # HITs per # HSPs plot
- Annotation datasets
 - Frequency distribution data
 - Frequency distribution plot
- Species
 - Frequency distribution data
 - Frequency distribution plot
- Family
 - Frequency distribution data
 - Frequency distribution plot
- Phylum
 - Frequency distribution data
 - Frequency distribution plot
- EC
 - Frequency distribution data
 - Frequency distribution plot
 - # sequences per # ids data
 - # sequences per # ids plot
- Gene Ontology
 - Frequency distribution per term data
 - Frequency distribution per term plot
 - Frequency distribution per namespace data
 - Frequency distribution per namespace plot
 - # sequences per # terms data
 - # sequences per # terms plot
- InterPro
 - Frequency distribution data
 - Frequency distribution plot
 - # sequences per # ids data
 - # sequences per # ids plot
- KEGG
 - Frequency distribution data
 - Frequency distribution plot
 - # sequences per # ids data
 - # sequences per # ids plot
- MapMan
 - Frequency distribution data
 - Frequency distribution plot

- # sequences per # ids data
 - # sequences per # ids plot
- MetaCyc
 - Frequency distribution data
 - Frequency distribution plot
 - # sequences per # ids data
 - # sequences per # ids plot

Logs menu

This menu allows the access to the application logs:

- View submission logs
- View result logs

Help menu

It contains the documentation of the application.

Configuring the TOA environment

When TOA starts for the first time it is required to configure the TOA environment. To do so, we select the menu item with the following path:

Main menu > Configuration > Recreate TOA config file

The Figure 5 shows the window corresponding to this menu item. Default values are presented for *Miniconda directory*, *Database directory* and *Result directory*. If necessary, modify them and press the button *[Execute]*.

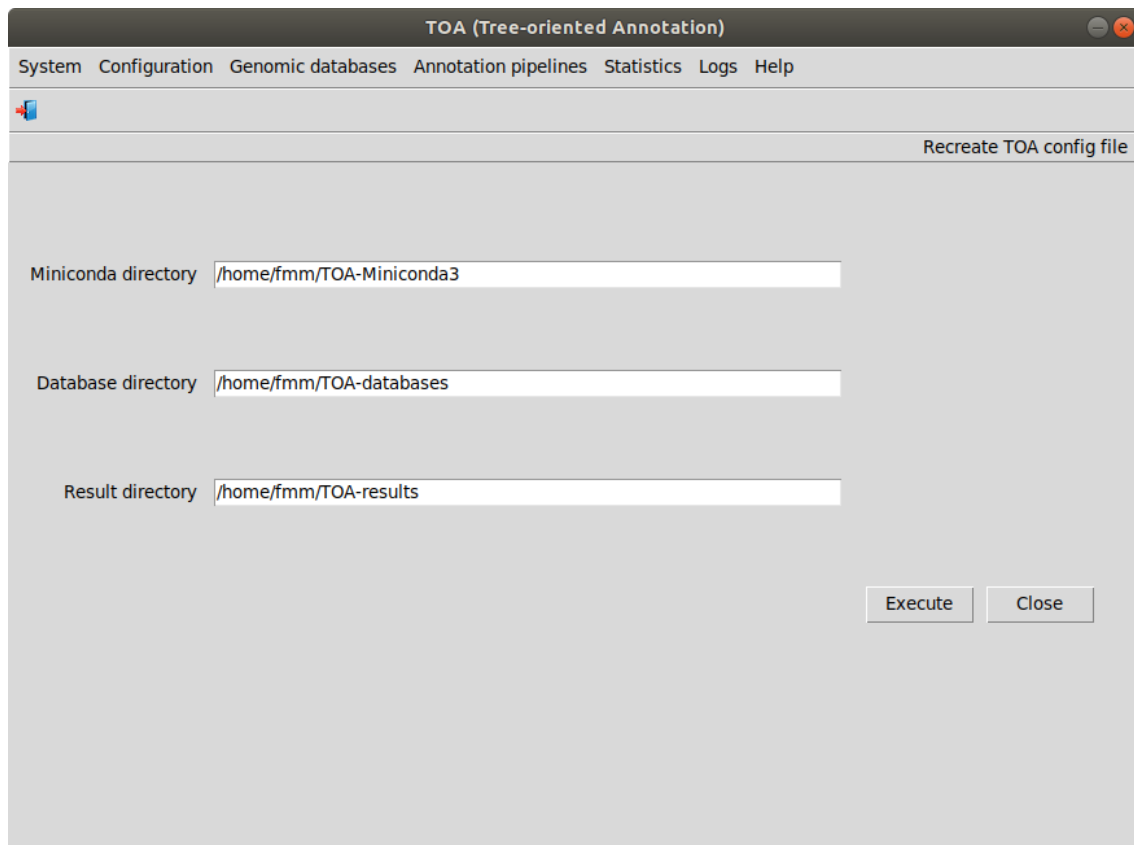


Figure 5. TOA window corresponding to the menu item *Recreate TOA config file*.

Installing bioinformatic software

TOA needs the following bioinformatic software:

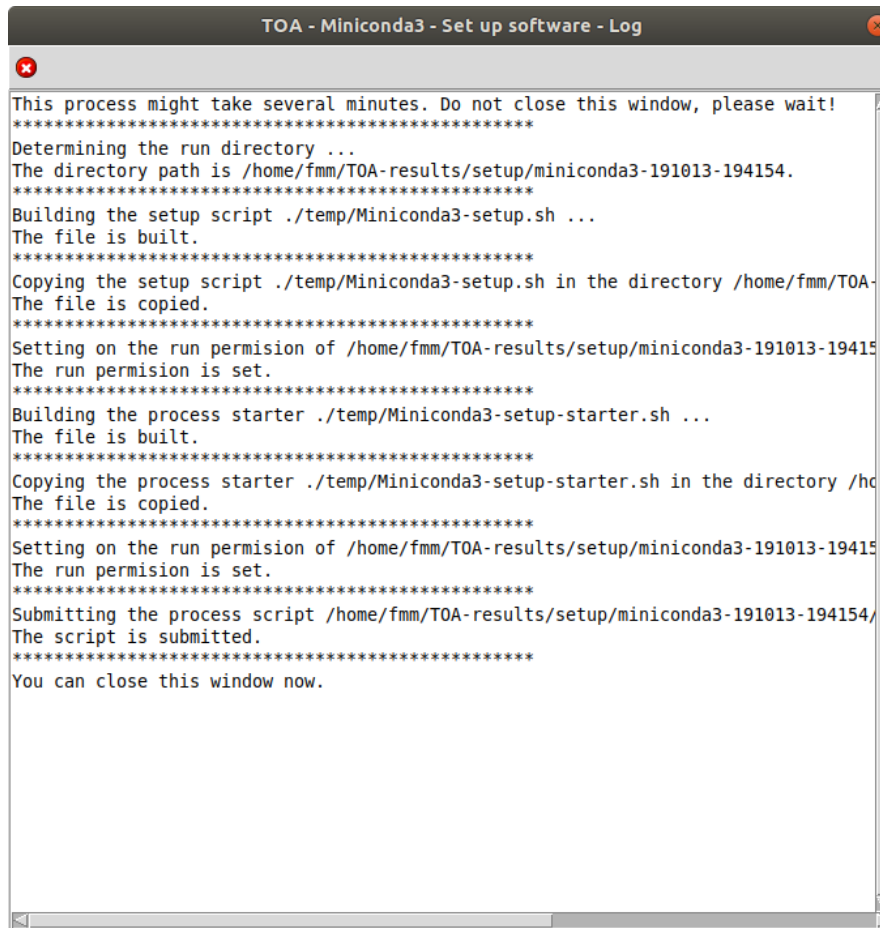
- TransDecoder (<https://github.com/TransDecoder/TransDecoder>). It is used to extract ORFs and predict coding regions within transcripts.
- BLAST+ (<https://blast.ncbi.nlm.nih.gov/>). It is used to find local similarity between transcripts or predicted peptides and sequences of genomic databases.
- Entrez Direct (<https://www.ncbi.nlm.nih.gov/books/NBK179288/>). It is a software package to download NCBI sequence identifiers of nucleotides or proteins.

The installation of these packages is automatic by using Bioconda (<https://bioconda.github.io/>), a channel of the Conda (<https://conda.io/>) package manager, which holds a large number of bioinformatics software packages.

First, setup Miniconda (Bioconda infrastructure) selecting the menu item with this path:

Main menu > Configuration > Bioinfo software setup > Miniconda3 [Execute]

Press the bottom *[Execute]*. A pop-up window will display the submission log (Figure 6)



```

TOA - Miniconda3 - Set up software - Log
This process might take several minutes. Do not close this window, please wait!
*****
Determining the run directory ...
The directory path is /home/fmm/TOA-results/setup/miniconda3-191013-194154.
*****
Building the setup script ./temp/Miniconda3-setup.sh ...
The file is built.
*****
Copying the setup script ./temp/Miniconda3-setup.sh in the directory /home/fmm/TOA-
The file is copied.
*****
Setting on the run permission of /home/fmm/TOA-results/setup/miniconda3-191013-19415
The run permission is set.
*****
Building the process starter ./temp/Miniconda3-setup-starter.sh ...
The file is built.
*****
Copying the process starter ./temp/Miniconda3-setup-starter.sh in the directory /ho
The file is copied.
*****
Setting on the run permission of /home/fmm/TOA-results/setup/miniconda3-191013-19415
The run permission is set.
*****
Submitting the process script /home/fmm/TOA-results/setup/miniconda3-191013-194154/
The script is submitted.
*****
You can close this window now.

```

Figure 6. Submission log of a Miniconda3 setup process.

To view the process log during and after the run, select the menu item with this path:

Main menu > Logs > View result logs

In the new window (Figure 7), select **setup** in the *Experiment/process* combo-box, and then press the button *[Execute]*.

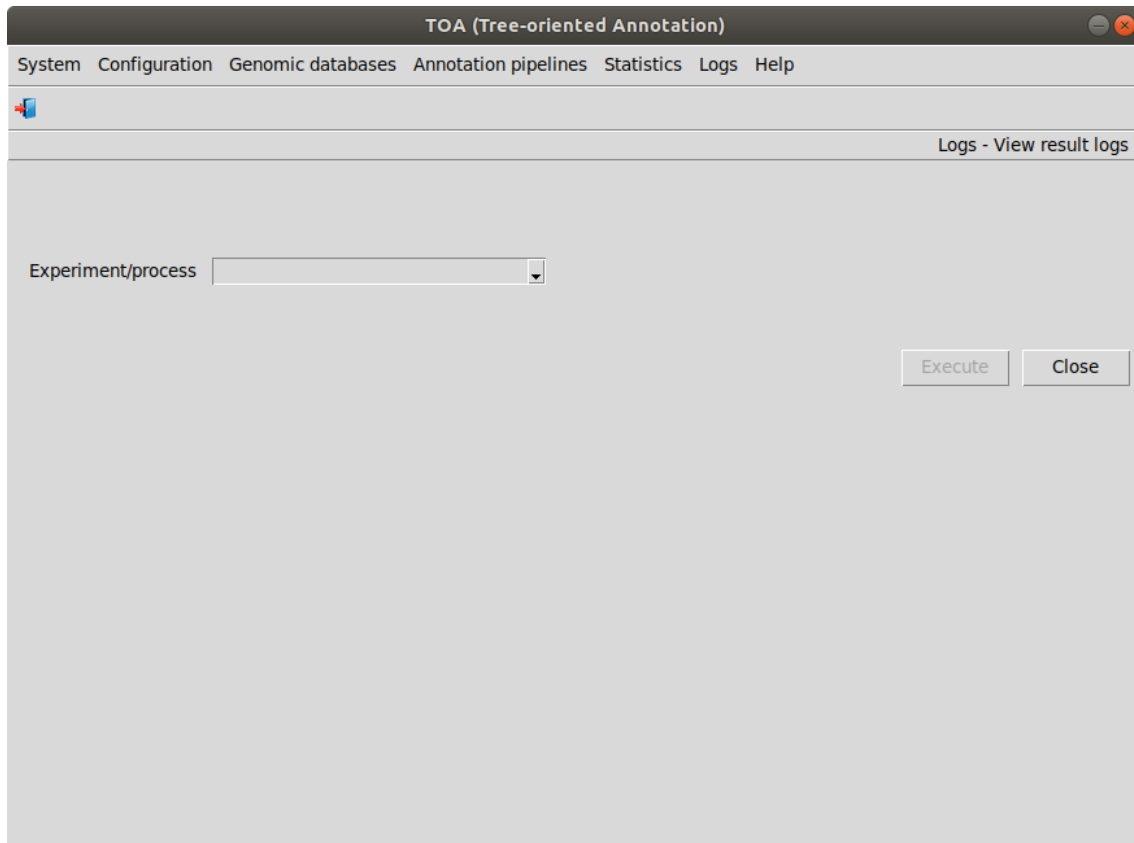


Figure 7. TOA window corresponding to the menu item *Logs - View result logs*.

Now, a pop-up window with all setup processes is shown (Figure 8).

Experiment id./Process	Bioinfo app / Utility	Result dataset	Date	Time	Status
setup	Miniconda3	miniconda3-191013-194154	2019-10-13	19:41:54	OK

Figure 8. List of setup process.

So far, we have only performed a single setup process: the process **miniconda3-191013-194154** corresponding to the last (and unique) Miniconda3 setup. Clicking on it, a new pop-up window appears with its corresponding log (Figure 9).

```

TOA - View - /home/fmm/TOA-results/setup/miniconda3-191013-194154/log.txt
#####
Script started at 2019-10-13 19:41:55.
#####
Removing Miniconda3 directory ...
The directory is removed.
#####
Downloading the Miniconda3 installation package ...

The package is downloaded.

TOA - View - /home/fmm/TOA-results/setup/miniconda3-191013-194154/log.txt
#####
Executing transaction: ...working... done
The package is installed.
#####
Installing package requests in Python 3 environment ...
Collecting package metadata (current_repodata.json): ...working... done
Solving environment: ...working... done

## Package Plan ##

  environment location: /home/fmm/TOA-Miniconda3

  added / updated specs:
    - requests

The following packages will be downloaded:

  package | build | size | channel
  -----|-----|-----|-----
  requests-2.22.0 | py37_1 | 84 KB | conda-forge
  -----|-----|-----|-----
                        Total: | 84 KB |

The following packages will be UPDATED:

  requests | pkgs/main::requests-2.22.0-py37_0 --> conda-forge::requests-2.22.0-py37_1

Preparing transaction: ...working... done
Verifying transaction: ...working... done
Executing transaction: ...working... done
The package is installed.
#####
Script ended OK at 2019-10-13 19:43:34 with a run duration of 99 s (000:01:39).
#####

```

Figure. 8. Begin and end of a Miniconda3 setup process.

In the toolbar, there is a button to refresh the run status. Clicking it, the log will be updated.

All the process logs have:

- A header with the time when it started.
- At the bottom, a summary with the status (OK, if all the programs have ended without errors; WRONG, otherwise), the end time, and the duration of the script run.

When Miniconda3 is setup, we setup BLAST+, Entrez Direct and TransDecoder selecting the menu items with these paths:

Main menu > Configuration > Bioinfo software setup> BLAST+ [Execute]

Main menu > Configuration > Bioinfo software setup> Entrez Direct [Execute]

Main menu > Configuration > Bioinfo software setup> TransDecoder [Execute]

The setup steps of each software are like for Miniconda3.

Consulting submitted processes and troubleshooting

The correct operability of the submitted processes is controlled by logs similar to the Miniconda3 setup (see Figure 8). So, the user can monitor the performance of the process at any time and detect problems. Please, confirm that each process is ended before submitting other one.

A step by step example

We have sequencing data corresponding to an RNA-seq Illumina library of an experiment about the process of cicatrization after wounding the xylem of the stem of the Canary Island pine (*Pinus canariensis*) and in this example, we are going to annotate the candidate peptides extracted from the transcriptome build in the assembly step using the following databases Gymno PLAZA 1.0, Dicots PLAZA 4.0, Monocots PLAZA 4.0, NCBI RefSeq Plant and NCBI Non-Redundant Protein Sequence Database (NR), in this order. In addition to these databases, we have to complete the biological data with other databases: NCBI Gene, InterPro and Gene Ontology.

The steps that we are going to do are:

- a) Load genomic data into TOA database of:
 - Basic data
 - Gymno PLAZA 1.0
 - Dicots PLAZA 4.0
 - Monocots PLAZA 4.0
 - NCBI RefSeq Plant
 - NCBI BLAST database NR
 - Protein GenInfo viridiplantae identifier list
 - NCBI Gene
 - InterPro
 - Gene Ontology
- b) Create and run an annotation pipeline (amino acid)
- c) Consult statistics data and show plots

The data loading of a genomic database into TOA database has to be performed once, and we only have to repeat the loading of a database when it is necessary to update its data.

The submit logs of genomic data loading can be reviewed selecting **TOA-databases** in the combo-box *Experiment/process* in the window *Logs - View result logs* (see Figure 7). And the logs of annotation pipelines can be consulted selection **TOA-pipelines** in this combo-box.

Load genomic data into TOA database

Basic data

The basic data loading requires performing several tasks. First, we create the genomic dataset file selecting the menu item with this path:

Main menu > Genomic databases > basic data > Recreate genomic dataset file [Execute]

Then, we create the species file:

Main menu > Genomic databases > basic data > Recreate species file [Execute]

Now, we download other basic data:

Main menu > Genomic databases > basic data > Download other basic data [Execute]

Finally, we load the dataset file, species file and other basic data into TOA database:

Main menu > Genomic databases > basic data > Load basic data into TOA database [Execute]

Gymno PLAZA 1.0

To Gymno PLAZA 1.0, first, we build the proteome selecting the menu item with this path:

Main menu > Genomic databases > Gymno PLAZA 1.0 > Build proteome [Execute]

Then, we download the functional annotations:

Main menu > Genomic databases > Gymno PLAZA 1.0 > Download functional annotation [Execute]

Finally, we load into TOA database:

Main menu > Genomic databases > Gymno PLAZA 1.0 > Load data into TOA database [Execute]

Dicots PLAZA 4.0

For Dicots PLAZA 4.0, how to proceed is similar to Gymno PLAZA 1.0.

Monocots PLAZA 4.0

For Monocots PLAZA 4.0, how to proceed is similar to Gymno PLAZA 1.0.

NCBI RefSeq Plant

In the case of this database, a task is only necessary, build its proteome:

Main menu > Genomic databases > Gymno PLAZA 1.0 > Build proteome [Execute]

NCBI BLAST database NR

To NCBI BLAST database NR, we only carry out the task of building the BLAST database:

Main menu > Genomic databases > NCBI BLAST database NR > Build BLAST database [Execute]

NCBI Protein GenInfo viridiplantae identifier list

Identifier list of Protein GenInfo viridiplantae is built from NCBI:

Main menu > Genomic databases > NCBI Protein GenInfo viridiplantae identifier list > Build identifier list [Execute]

NCBI Gene

To NCBI Gene, first, we download the functional annotations:

Main menu > Genomic databases > NCBI Gene > Download functional annotation [Execute]

And then, we load into TOA database:

Main menu > Genomic databases > NCBI Gene > Load data into TOA database [Execute]

InterPro

For InterPro, the procedure is similar to the case of NCBI Gene.

Gene Ontology

For Gene Ontology, the procedure is similar to the case of NCBI Gene.

Create and run an annotation pipeline (amino acid)

Create the config file

The first task to perform an annotation process with TOA is to build a config file that contains data about the transcriptome directory and file, selected databases and their order, and BLAST parameters. Taking into account these data, TOA will be able to build dynamically a Bash script corresponding to the annotation pipeline.

We create a new config file selecting the menu item with this path:

Main menu > Annotation pipelines > TOA aminoacid pipeline > Recreate config file [Execute]

Then, TOA presents a window (see Figure 9) where we indicate the *Transcriptome directory* selecting it with the corresponding button. Once this is done, we select the file in the combo-box *Transcriptome file*.

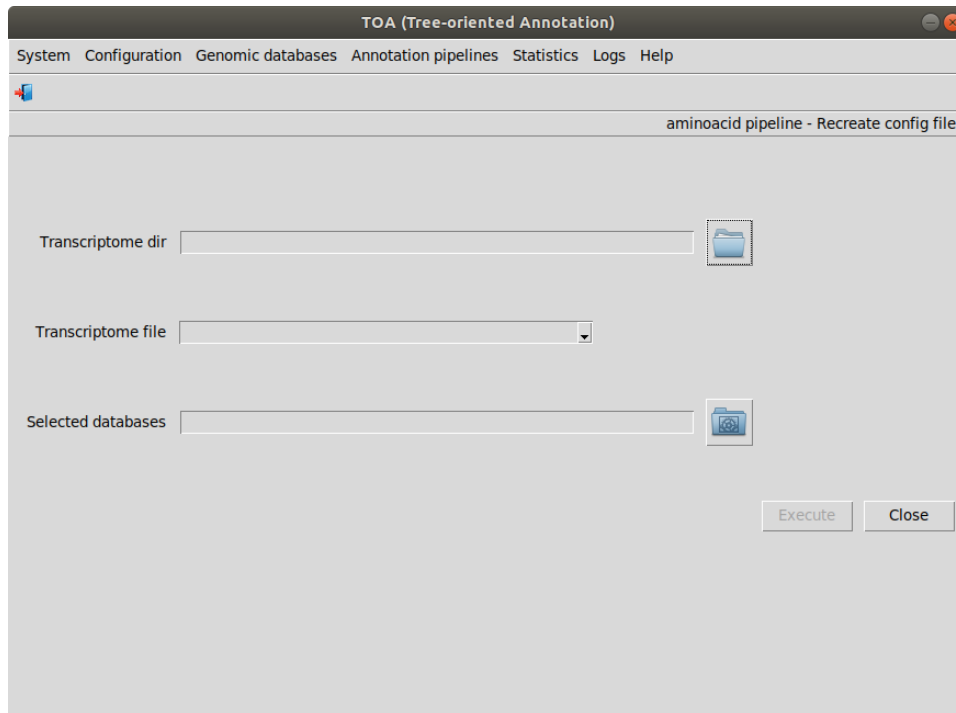


Figure. 9. TOA window corresponding to the menu item *amino acid pipeline - Recreate config file*.

Last, we have to select the databases and their order in the annotation pipeline. For it, we click the corresponding button and a window is raised (see Figure 10). In this window, we click on the first database to annotate the transcriptome, in our example, Gymno PLAZA 01. A new window is raised (see Figure 11). In the *Annotation order* entry field, we type 1 (0 means not selected; 1, the first database; 2, the second one; etc.) and press button [OK].

Database	Annotation Order	Admitted Values
gymno_01	0	integer between 0 and 6
dicots_04	0	integer between 0 and 6
monocots_04	0	integer between 0 and 6
refseq_plant	0	integer between 0 and 6
nr_viridiplantae	0	integer between 0 and 6
nr_complete	0	integer between 0 and 6

Figure. 10. TOA window where the databases the bases involved in the annotation are selected

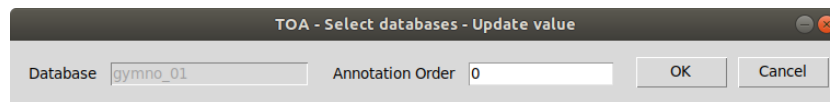


Figure. 11. TOA window where the annotation order of a selected database is set (0: not selected)

We will proceed with the remaining databases in a similar way as with Gymno PLAZA 01. When we are done with all the databases involved in the annotation, we will close the window displayed in Figure 10. Now, the appearance of the window *amino acid pipeline - Recreate config file* will be similar to the one in the Figure 11.

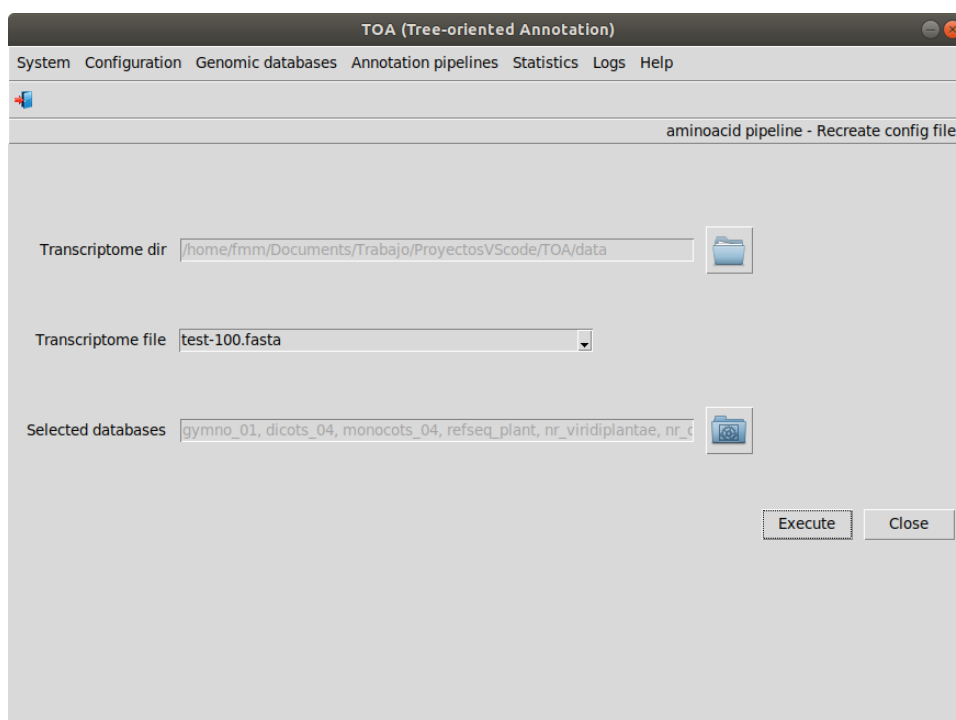


Figure. 12. TOA window corresponding to the menu item *amino acid pipeline - Recreate config file* once the data has been filled.

Then we press the button [Execute]. In the raised window (see Figure 13), we can examine the config file. In this example, it has three sections: *identification*, with the directory and file of the transcriptome; *database parameters*, where the order of the databases is set; and *BLAST parameters* where we can modify the thread number, *e_value*, the maximum number of aligned sequences, the maximum number of HSPs and the minimum coverage of alignments.

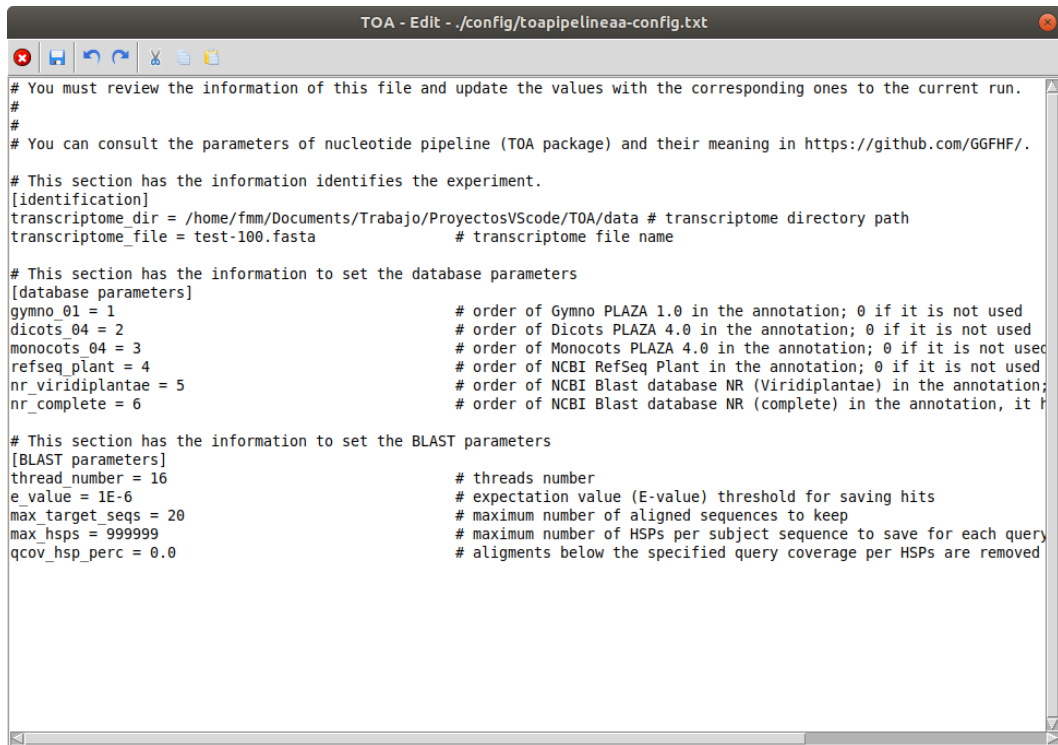


Figure. 13. Editable TOA window where the config file of a amino acid pipeline can be modified.

Edit the config file

We can review and modify the current config file selecting:

Main menu > Annotation pipelines > TOA aminoacid pipeline > Edit config file [Execute]

Run a pipeline

We run the pipeline corresponding to the current config file selecting:

Main menu > Annotation pipelines > TOA aminoacid pipeline > Run pipeline [Execute]

Then, a Bash script is dynamically built and run.

When the process end, we could go the subdirectory of TOA-results\TOA-pipelines corresponding to the script run and consult the generated files (see Figure 14). There are 3 files per database. E.g. gymno_01-alignment.xml (the BLAST alignment with the Gymno PLAZA 01 proteome), gymno_01-annotation.csv (the annotation file of aligned candidate peptides) and gymno_01-nonann-peptides.fasta (a FASTA file with the non-annotated candidate peptides). Also, there are other two files that join files of plant alignment, plant-alignment.xml, and files of plant alignment, plant-annotation.csv. In the subdirectory stats, there are the CSV files of statistics of alignment, annotation datasets, phylogenetic information and function.







































Name
 stats
 status
 transdecoder
 transdecoder.__checkpoints
 transdecoder.__checkpoints_longorfs
 dicots_04-alignment.xml
 dicots_04-annotation.csv
 dicots_04-nonann-peptides.fasta
 gymno_01-alignment.xml
 gymno_01-annotation.csv
 gymno_01-nonann-peptides.fasta
 log.txt
 monocots_04-alignment.xml
 monocots_04-annotation.csv
 monocots_04-nonann-peptides.fasta
 nr-remainder-alignment.xml
 nr-remainder-annotation.csv
 nr-remainder-nonann-peptides.fasta
 nr-viridiplantae-alignment.xml
 nr-viridiplantae-annotation.csv
 nr-viridiplantae-nonann-peptides.fasta
 pipeliner.8002.cmds
 pipeliner.8021.cmds
 pipeliner.8040.cmds
 plant-alignment.xml
 plant-annotation.csv
 plant-annotation-01.csv
 refseq_plant-alignment.xml
 refseq_plant-annotation.csv
 refseq_plant-nonann-peptides.fasta
 reidentified-peptides.fasta
 relationships.csv
 test-100.fasta.transdecoder.bed
 test-100.fasta.transdecoder.cds
 test-100.fasta.transdecoder.gff3
 test-100.fasta.transdecoder.pep
 toapipelineaa-process.sh
 toapipelineaa-process-starter.sh

Figure. 14. List of directories and files generated by the example annotation pipeline.

Consult statistics data and show plots

TOA pipelines generate statistics of alignment, annotation datasets, phylogenetic information and function. In this tutorial, we are only going to consult some of these statistics.

Alignment - # HITs per HSPs

We consult # HITs per # HSPs data of the alignment selecting the menu item with this path:

Main menu > Statistics > Alignment > # HITs per # HSPs data [Execute]

In the displayed window (see Figure 15), we select **TOA-pipeline** in the *Experiment/process* combo-box and the example pipeline in the *Pipeline dataset* combo-box. Then we press the button *[Execute]*.

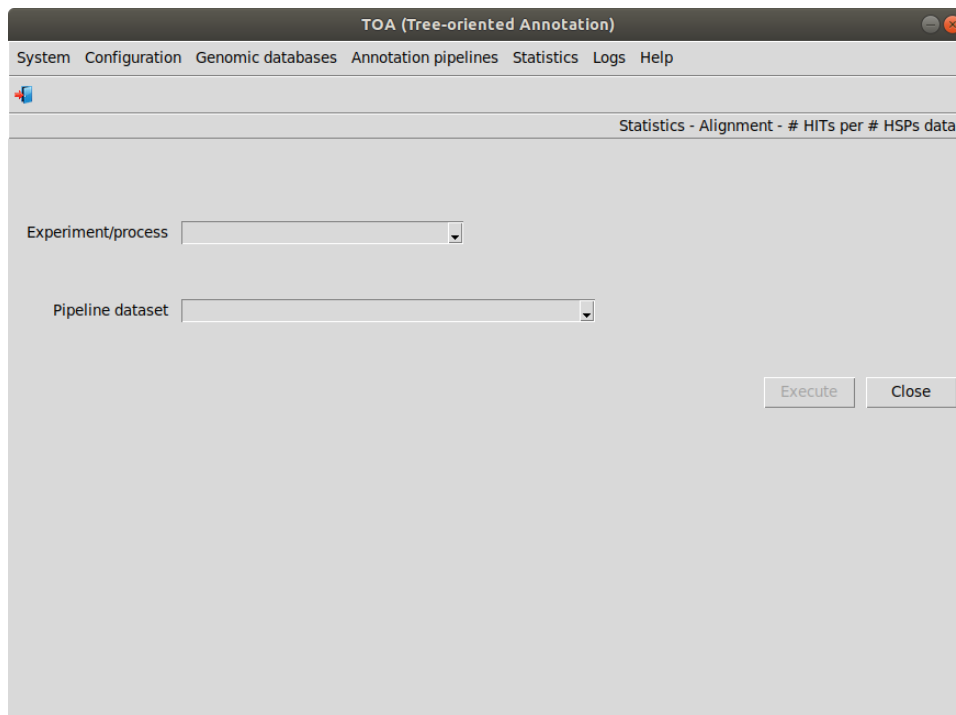
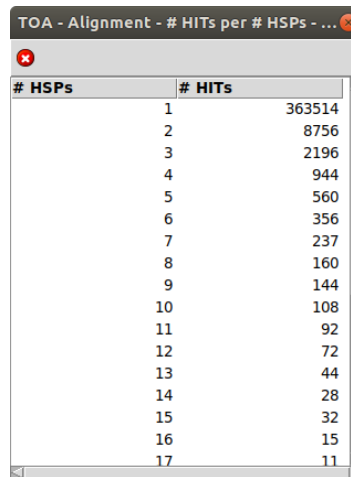


Figure. 15. TOA window corresponding to the menu item *Statistics - Alignment - # HITs per # HSPs data*.

Data are listed in the raised window (see Figure 16).



# HSPs	# HITS
1	363514
2	8756
3	2196
4	944
5	560
6	356
7	237
8	160
9	144
10	108
11	92
12	72
13	44
14	28
15	32
16	15
17	11

Figure. 16. # HITS per # HSPs data of the alignment generated by the example annotation pipeline.

We view the graphic corresponding to # HITS per # HSPs data of the alignment selecting:

Main menu > Statistics > Alignment > # HITS per # HSPs plot [Execute]

In the displayed window (see Figure 17), we select **TOA-pipeline** in the *Experiment/process* combo-box and the example pipeline in the *Pipeline dataset* combo-box. The default value of the remaining fields (*Image format*, *DPI*, *Image dir* and *Image name*) can be modified if necessary. Then we press the button *[Execute]*.

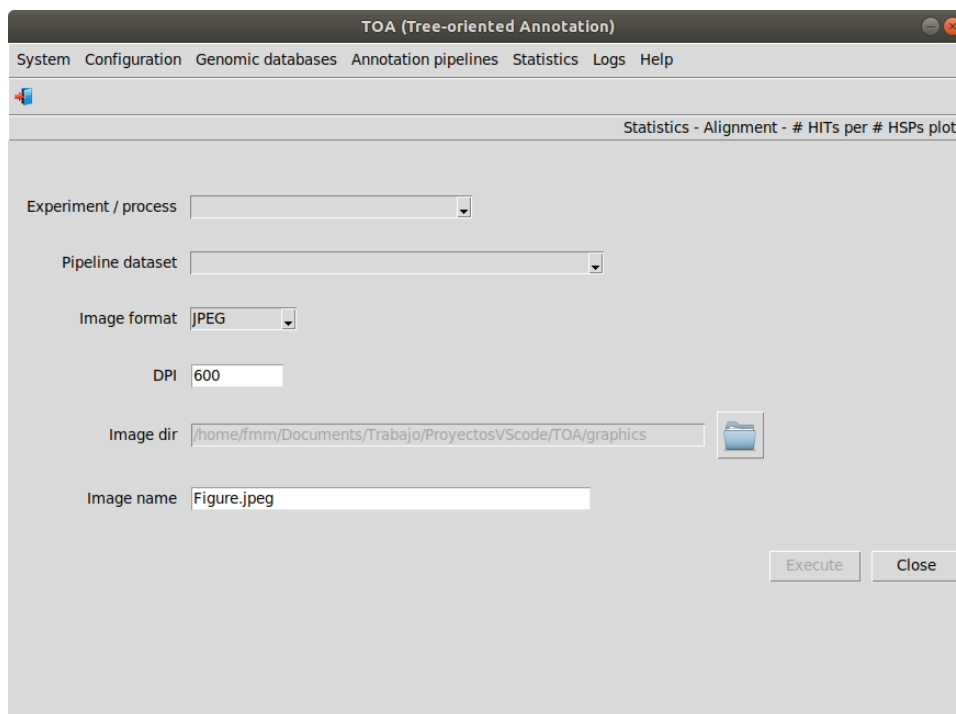


Figure. 17. TOA window corresponding to the menu item *Statistics - Alignment - # HITS per # HSPs plot*.

The plot is saved in the selected directory in *Image dir* and is displayed in a pop-up window (see Figure 18).

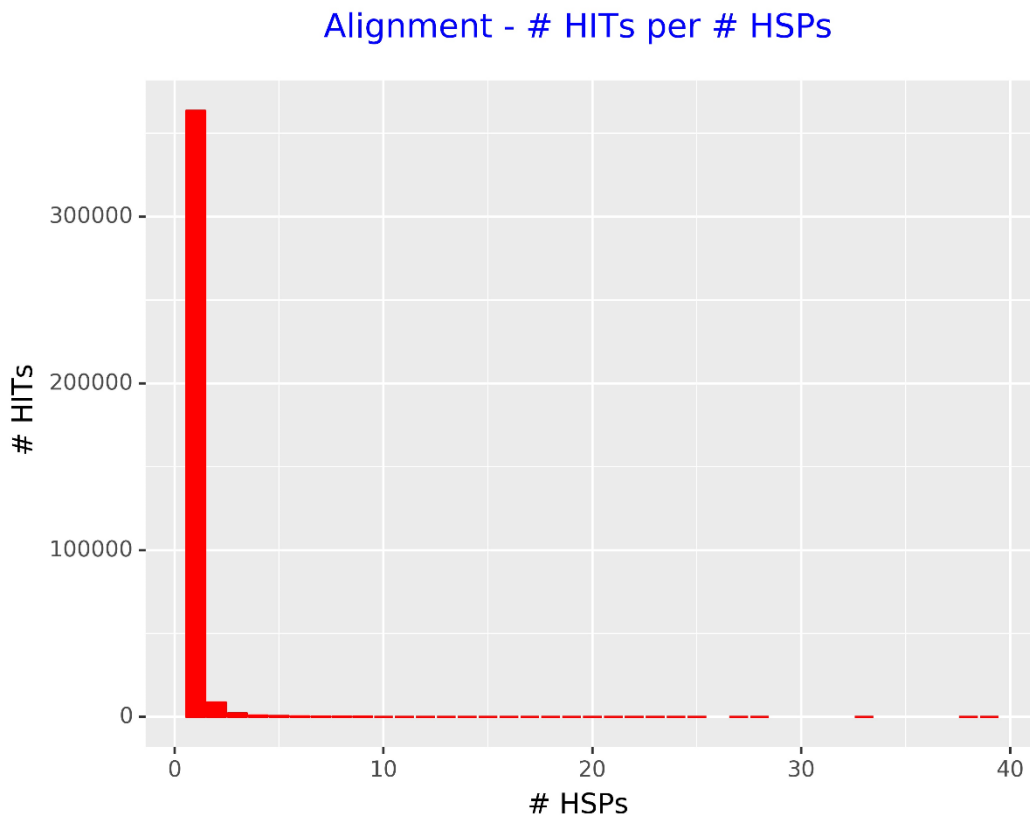


Figure. 18. Plot of # HITs per # HSPs data of the alignment generated by the example annotation pipeline.

Annotation datasets - Frequency distribution

We can consult the frequency distribution of annotation datasets selecting the menu item with this path:

Main menu > Statistics > Annotation datasets > Frequency distribution data [Execute]

In the displayed window (see Figure 19), we select **TOA-pipeline** in the *Experiment/process* combo-box and the example pipeline in the *Pipeline dataset* combo-box. Then, we press the button *[Execute]*.

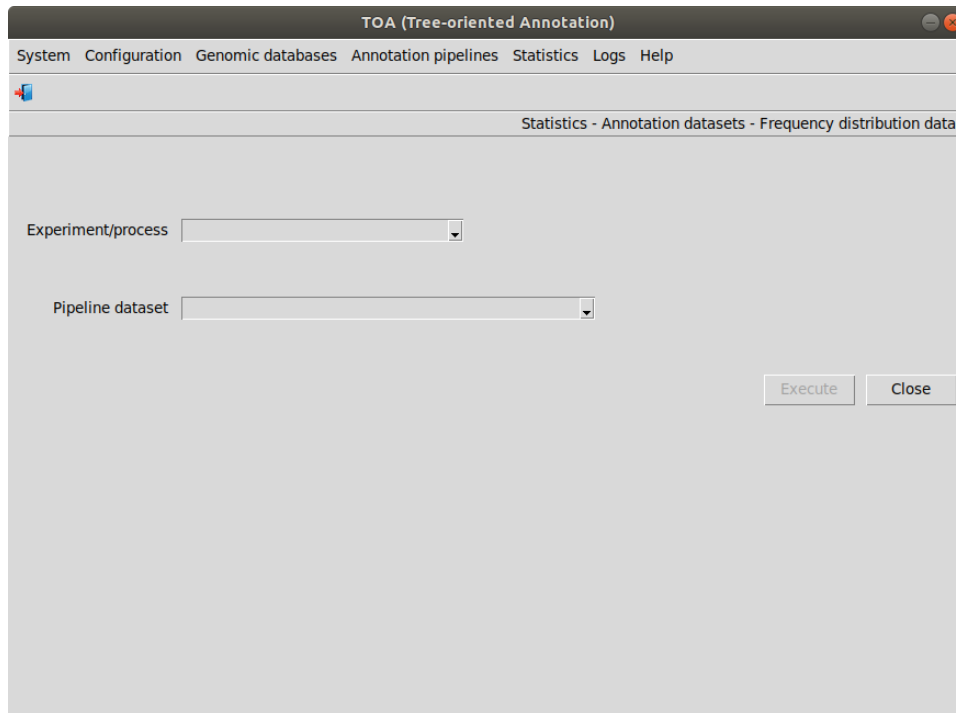


Figure. 19. TOA window corresponding to the menu item *Statistics - Annotation datasets - Frequency distribution data*.

Data are listed in the raised window (see Figure 20).

Dataset	Annotated seqs	Remained seqs
Transcriptome		44849
Predicted peptides		24674
Gymno PLAZA 1.0	22235	2439
Dicots PLAZA 4.0	324	2115
Monocots PLAZA 4.0	47	2068
RefSeq Plant	130	1938
nr Viridiplantae	358	1580
nr remainder	54	1526

Figure. 20. Frequency distribution of annotation datasets generated by the example annotation pipeline.

We view the graphic corresponding to the frequency distribution of annotation datasets selecting:

Main menu > Statistics > Annotation datasets > Frequency distribution plot [Execute]

In the displayed window (see Figure 21), we select **TOA-pipeline** in the *Experiment/process* combo-box and the example pipeline in the *Pipeline dataset* combo-box. The default value of

the remaining fields (*Image format*, *DPI*, *Image dir* and *Image name*) can be modified if necessary. Then, we press the button *[Execute]*.

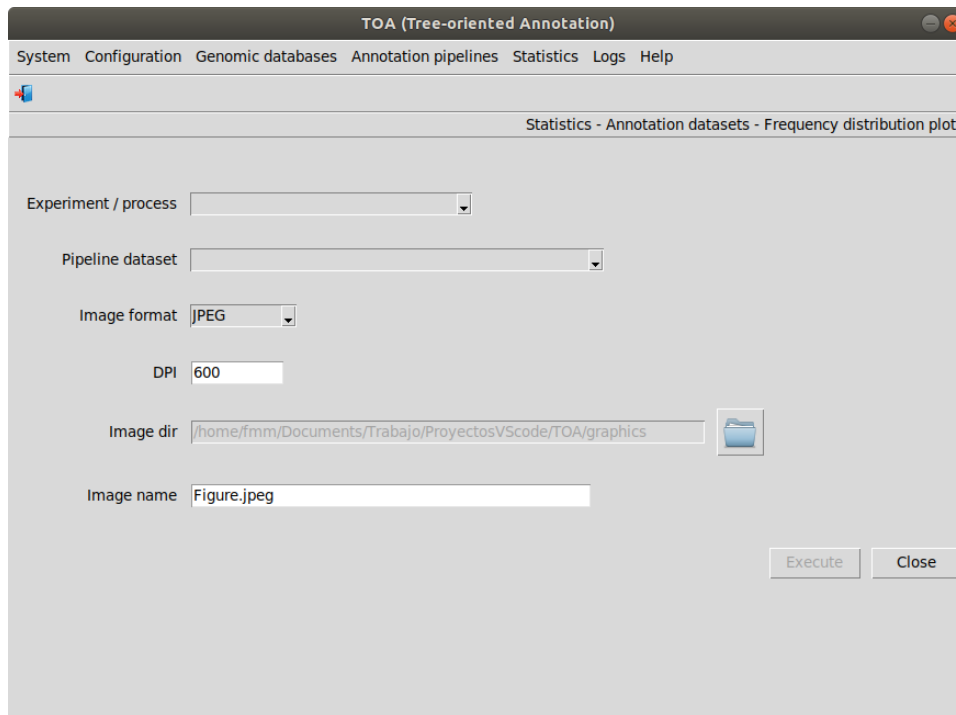


Figure. 21. TOA window corresponding to the menu item *Statistics - Annotation datasets - Frequency distribution plot*.

The plot is saved in the directory selected in *Image dir* and is displayed in a pop-up window (see Figure 22).

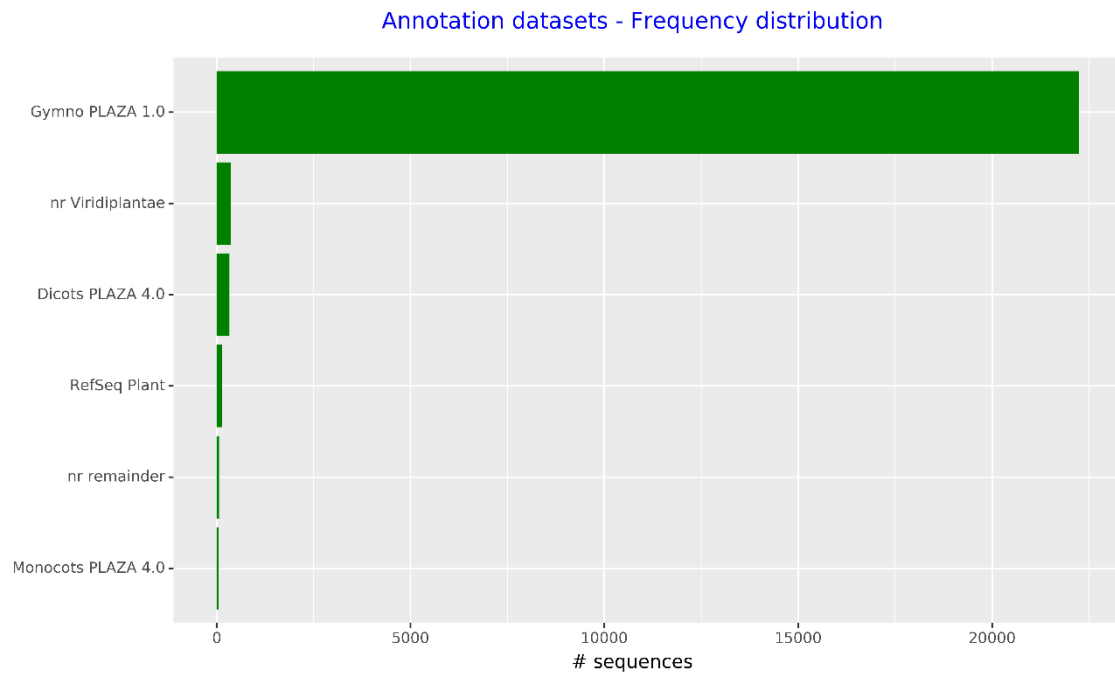


Figure. 22. Plot of frequency distribution of annotation datasets generated by the example annotation pipeline.

Species - Frequency distribution

We consult the frequency distribution of species selecting the menu item with this path:

Main menu > Statistics > Species > Frequency distribution data [Execute]

In the shown window (see Figure 23), we select **TOA-pipeline** in the *Experiment/process* combo-box and the example pipeline in the *Pipeline dataset* combo-box. Then, we press the button *[Execute]*.

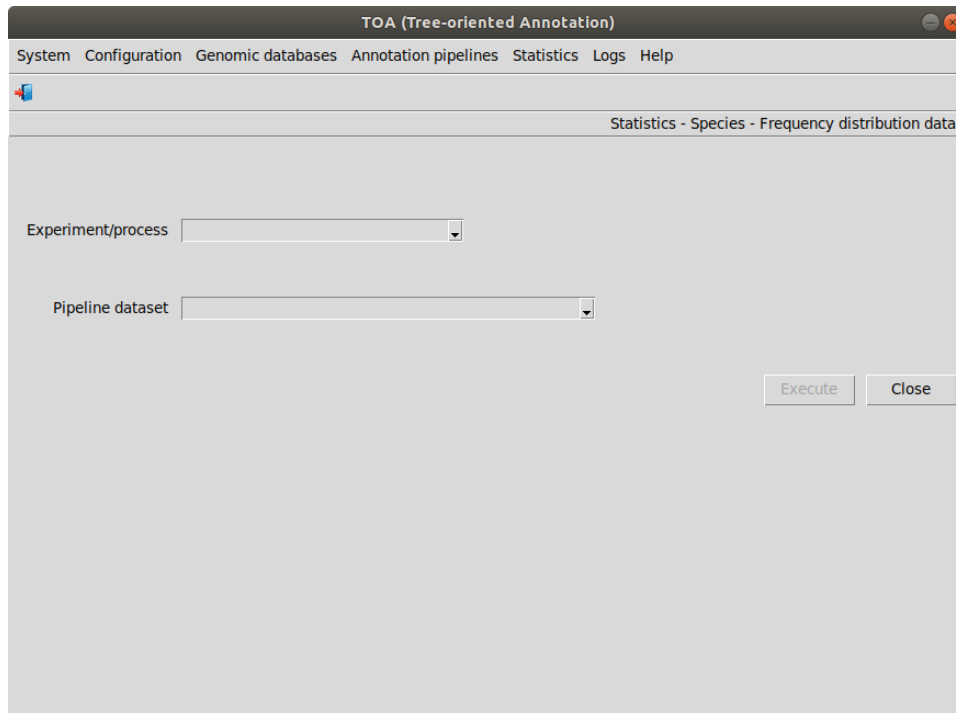


Figure. 23. TOA window corresponding to the menu item *Statistics - Species - Frequency distribution data*.

Data are listed in the raised window (see Figure 24).

TOA - Species - Frequency distribution - Table			
Species	All count	First HSP count	Min e-value count
Actinidia chinensis	31	31	1
Actuca sativa]	9	9	0
Ajanus cajan]	3	3	0
Alus domestica]	19	17	0
Amborella trichopoda	14814	13859	616
Amelina sativa]	2	2	0
Amellia sinensis]	38	38	3
Ananas comosus	8	8	4
Anicum hallii]	5	5	0
Anihot esculenta]	22	20	2
Apaver somniferum]	18	18	1
Aphanus sativus]	4	4	0
Apostasia shenzhenica	1	1	1
Apsella rubella]	2	2	0
Apsicum annum]	11	11	0
Aquilegia coerulea	1	1	1
Arabidopsis lyrata	17	17	0

Figure. 24. Frequency distribution of species generated by the example annotation pipeline.

We view the graphic corresponding to the frequency distribution of species selecting:

Main menu > Statistics > Species > Frequency distribution plot [Execute]

In the displayed window (see Figure 25), we select **TOA-pipeline** in the *Experiment/process* combo-box and the example pipeline in the *Pipeline dataset* combo-box. The default value of the remaining fields (*Alignment count level*, *Image format*, *DPI*, *Image dir* and *Image name*) can be modified if necessary. The possible values of *Alignment count level* are **all count** (all

alignments are considered), **first HSP count** (alignments with HSP 1 are considered) and **minimum e-value count** (alignments with minimum e-value per sequence are considered). Then we press the button *[Execute]*.

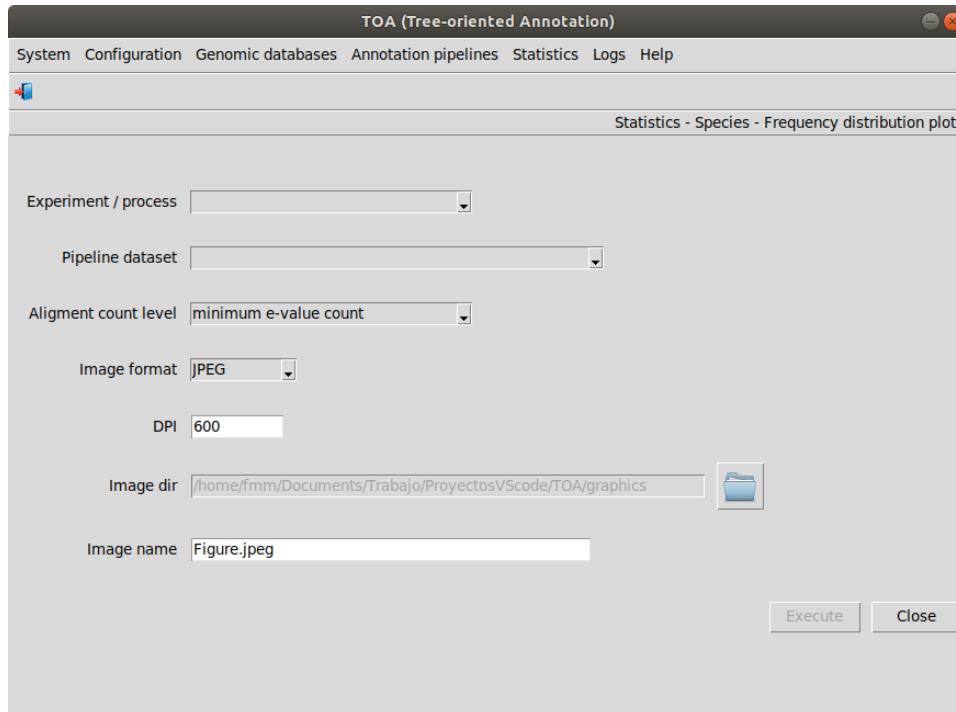


Figure. 25. TOA window corresponding to the menu item *Statistics - Species - Frequency distribution plot*.

The plot is saved in the directory selected in *Image dir* and is displayed in a pop-up window (see Figure 26).

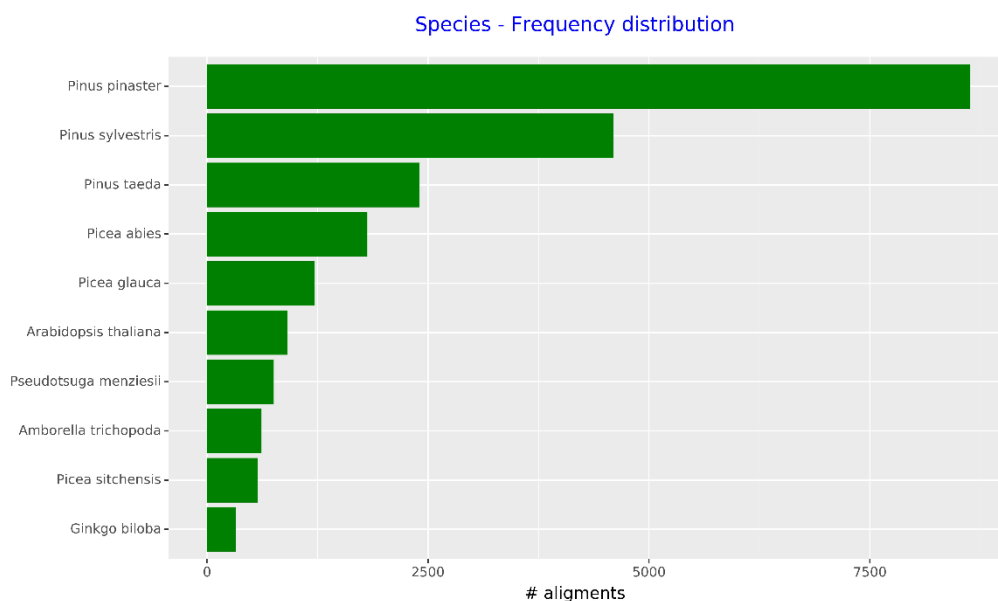


Figure. 26. Plot of frequency distribution of species generated by the example annotation pipeline.

Gene Ontology - Frequency distribution per term

We consult the frequency distribution per Gene Ontology term selecting the menu item with this path:

Main menu > Statistics > Gene Ontology > Frequency distribution per term data[Execute]

In the displayed window (see Figure 27), we select **TOA-pipeline** in the *Experiment/process* combo-box and the example pipeline in the *Pipeline dataset* combo-box. Then we press the button *[Execute]*.

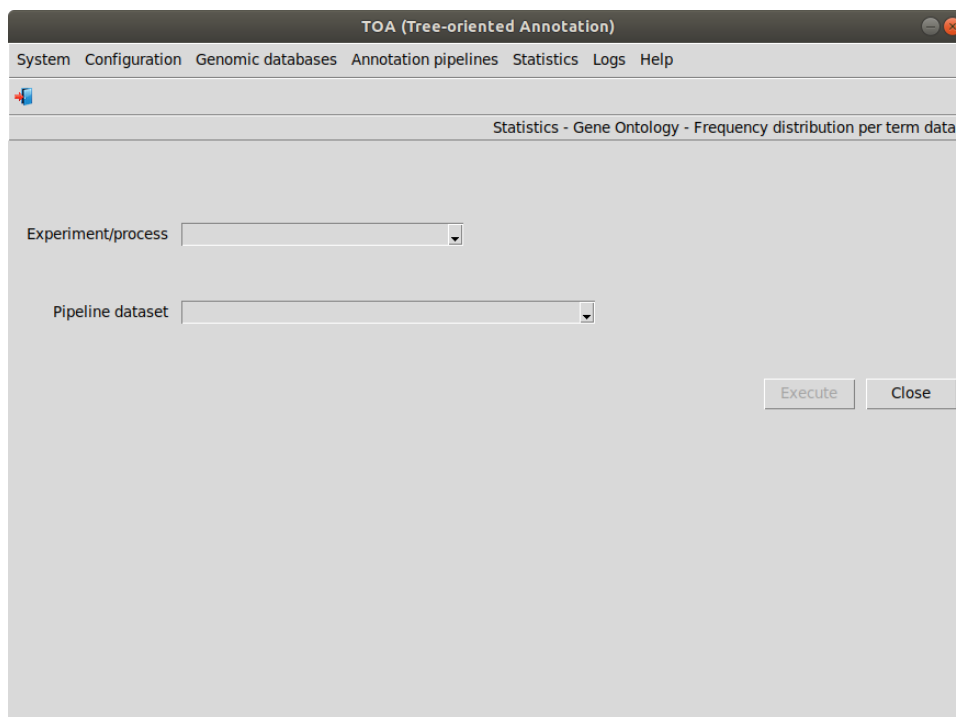


Figure. 27. TOA window corresponding to the menu item *Statistics - Gene Ontology - Frequency distribution per term data*.

Data are listed in the raised window (see Figure 28).

GO term	Description	Namespace	All count	First HSP count	Min e-value count
GO:0000002	mitochondrial genome maintenance	biological process	18	12	0
GO:0000003	reproduction	biological process	60	60	4
GO:0000012	single strand break repair	biological process	6	6	0
GO:0000014	single-stranded DNA endodeoxyribonuclease activity	molecular function	10	10	0
GO:0000015	phosphopyruvate hydratase complex	cellular component	201	201	10
GO:0000023	maltose metabolic process	biological process	177	170	19
GO:0000024	maltose biosynthetic process	biological process	9	9	0
GO:0000025	maltose catabolic process	biological process	17	12	0
GO:0000026	alpha-1,2-mannosyltransferase activity	molecular function	16	16	0
GO:0000027	ribosomal large subunit assembly	biological process	13	13	1
GO:0000028	ribosomal small subunit assembly	biological process	3	3	0
GO:0000030	mannosyltransferase activity	molecular function	3	3	0
GO:0000033	alpha-1,3-mannosyltransferase activity	molecular function	4	4	0
GO:0000035	acyl binding	molecular function	21	20	0
GO:0000036	acyl carrier activity	molecular function	13	13	0
GO:0000038	very long-chain fatty acid metabolic process	biological process	70	70	3
GO:0000041	transition metal ion transport	biological process	62	61	2

Figure. 28. Frequency distribution per Gene Ontology term generated by the example annotation pipeline.

We view the graphic corresponding to the frequency distribution per Gene Ontology term selecting:

Main menu > Statistics > Gene Ontology > Frequency distribution per term plot [Execute]

In the displayed window (see Figure 29), we select **TOA-pipeline** in the *Experiment/process* combo-box and the example pipeline in the *Pipeline dataset* combo-box. The default value of the remaining fields (*Namespace*, *Alignment count level*, *Image format*, *DPI*, *Image dir* and *Image name*) can be modified if necessary. The *Namespace* combo-box has the values **all**, **biological process**, **cellular component** and **molecular function**. The possible values of *Alignment count level* are **all count** (all alignments are taken into account), **first HSP count** (alignments with HSP 1 are taken into account) and **minimum e-value count** (alignments with minimum e-value per sequence are taken into account). Then we press the button *[Execute]*.

Figure. 29. TOA window corresponding to the menu item *Statistics - Gene Ontology - Frequency distribution per term plot*.

The plot is saved in the directory selected in *Image dir* and is displayed in a pop-up window (see Figure 30).

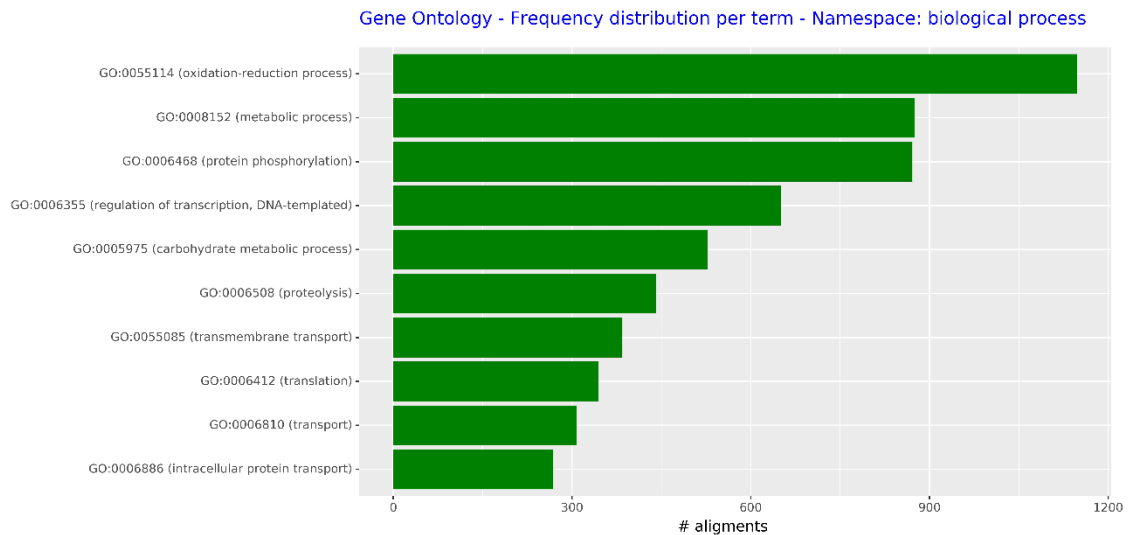


Figure. 30. Plot of frequency distribution per Gene Ontology term generated by the example annotation pipeline.

Gene Ontology - Frequency distribution per namespace

We consult the frequency distribution per Gene Ontology namespace selecting the menu item with this path:

Main menu > Statistics > Gene Ontology > Frequency distribution per namespace data[Execute]

In the displayed window (see Figure 31), we select **TOA-pipeline** in the *Experiment/process* combo-box and the example pipeline in the *Pipeline dataset* combo-box. Then we press the button *[Execute]*.

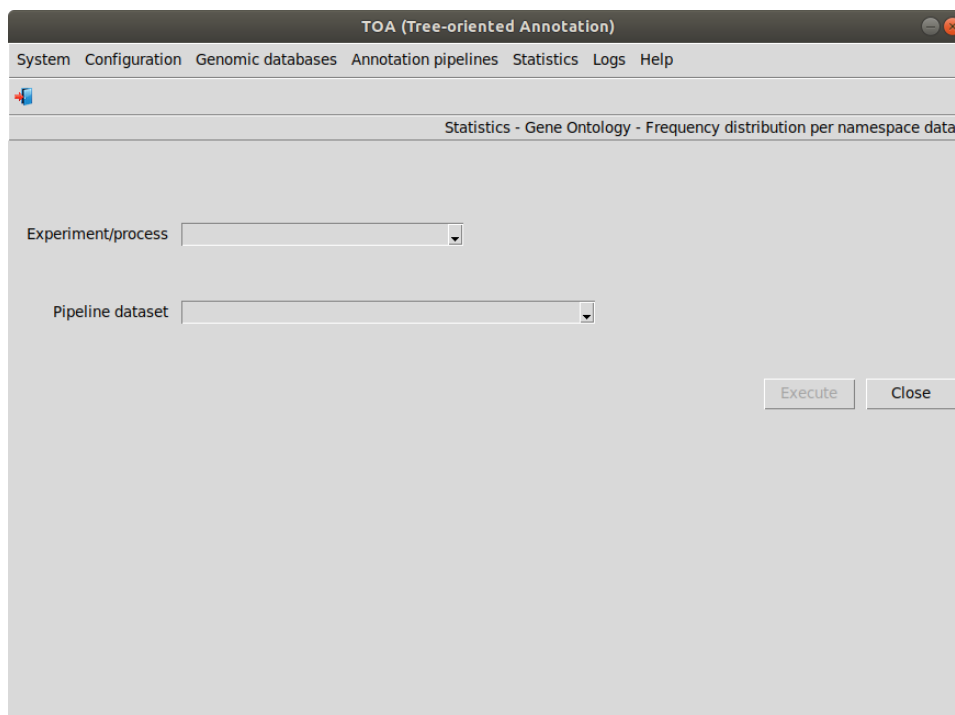


Figure. 31. TOA window corresponding to the menu item *Statistics - Gene Ontology - Frequency distribution per namespace data*.

Data are listed in the raised window (see Figure 32).

Namespace	All count	First HSP count	Min e-value count
N/A	28295	27120	819
biological process	573489	538943	23884
cellular component	272365	265210	11230
molecular function	744456	673474	31194

Figure. 32. Frequency distribution per Gene Ontology namespace generated by the example annotation pipeline.

We view the graphic corresponding to the frequency distribution per Gene Ontology namespace selecting:

Main menu > Statistics > Gene Ontology > Frequency distribution per namespace plot [Execute]

In the displayed window (see Figure 33), we select **TOA-pipeline** in the *Experiment/process* combo-box and the example pipeline in the *Pipeline dataset* combo-box. The default value of

the remaining fields (*Alignment count level*, *Image format*, *DPI*, *Image dir* and *Image name*) can be modified if necessary. The possible values of *Alignment count level* are: **all count** (all alignments are taken into account), **first HSP count** (alignments with HSP 1 are taken into account) and **minimum e-value count** (alignments with minimum e-value per sequence are taken into account). Then we press the button *[Execute]*.

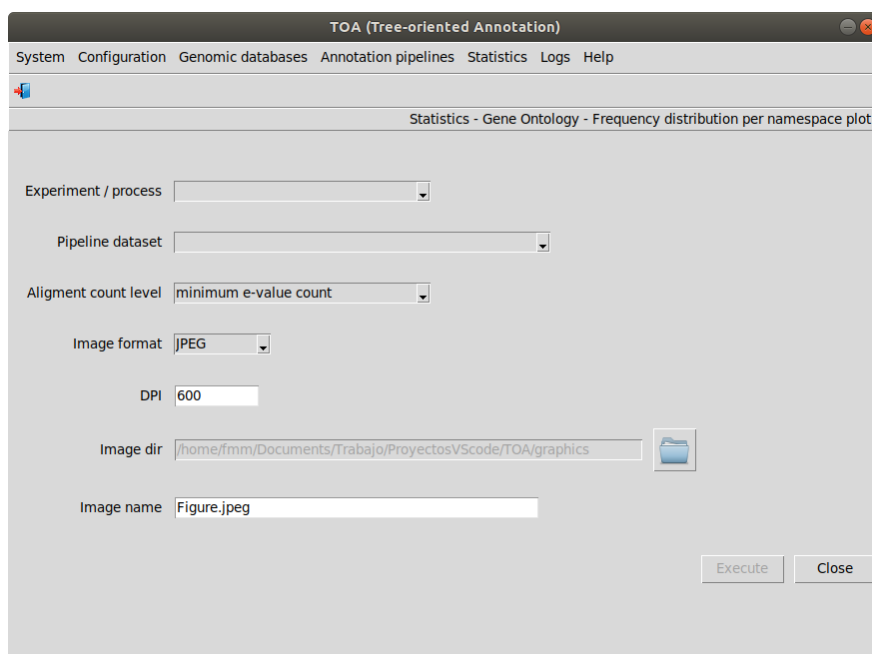


Figure. 33. TOA window corresponding to the menu item *Statistics - Gene Ontology - Frequency distribution per namespace plot*.

The plot is saved in the directory selected in *Image dir* and is displayed in a pop-up window (see Figure 34).

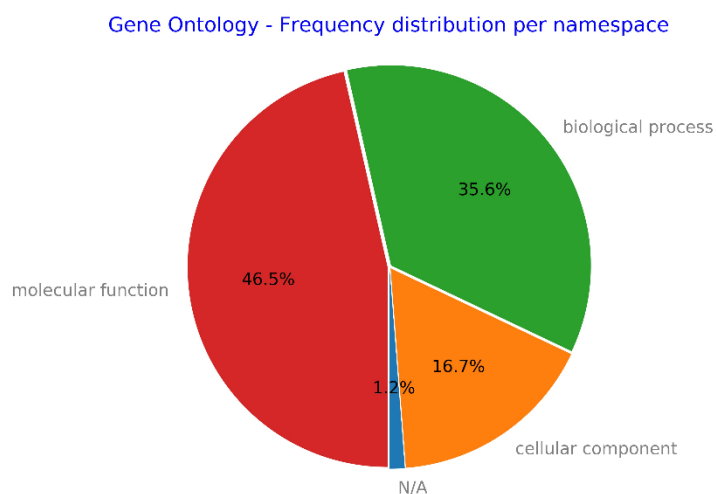


Figure. 34. Plot of frequency distribution per Gene Ontology namespace generated by the example annotation pipeline.

Gene Ontology - # sequences per #terms

We consult #sequences per # Gene Ontology terms the menu item with this path:

Main menu > Statistics > Species > # sequences per #terms data [Execute]

In the displayed window (see Figure 35), we select **TOA-pipeline** in the *Experiment/process* combo-box and the example pipeline in the *Pipeline dataset* combo-box. Then we press the button *[Execute]*.

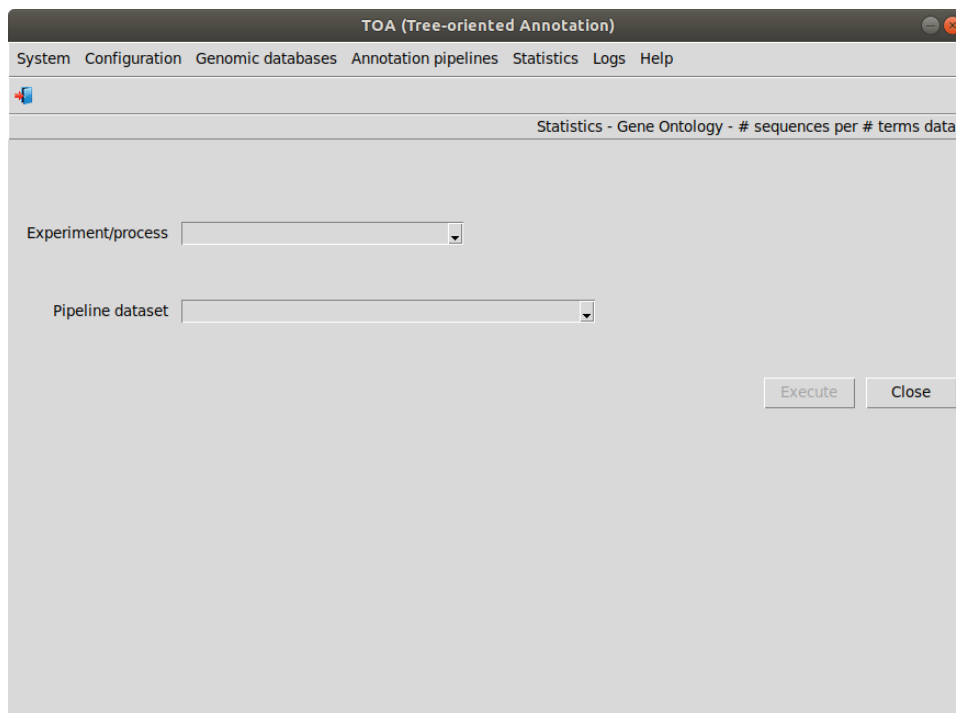
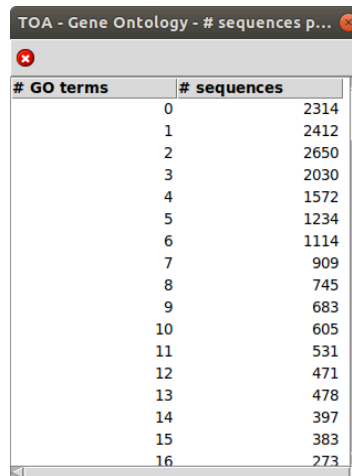


Figure. 35. TOA window corresponding to the menu item *Statistics - Gene Ontology - # sequences per #terms data*.

Data are listed in the raised window (see Figure 36).



# GO terms	# sequences
0	2314
1	2412
2	2650
3	2030
4	1572
5	1234
6	1114
7	909
8	745
9	683
10	605
11	531
12	471
13	478
14	397
15	383
16	273

Figure. 36. #sequences per # Gene Ontology terms generated by the example annotation pipeline.

We view the graphic corresponding to #sequences per # Gene Ontology terms selecting:

Main menu > Statistics > Species > # sequences per #terms plot [Execute]

In the displayed window (see Figure 37), we select **TOA-pipeline** in the *Experiment/process* combo-box and the example pipeline in the *Pipeline dataset* combo-box. The default value of the remaining fields (*Image format*, *DPI*, *Image dir* and *Image name*) can be modified if necessary. Then we press the button *[Execute]*.

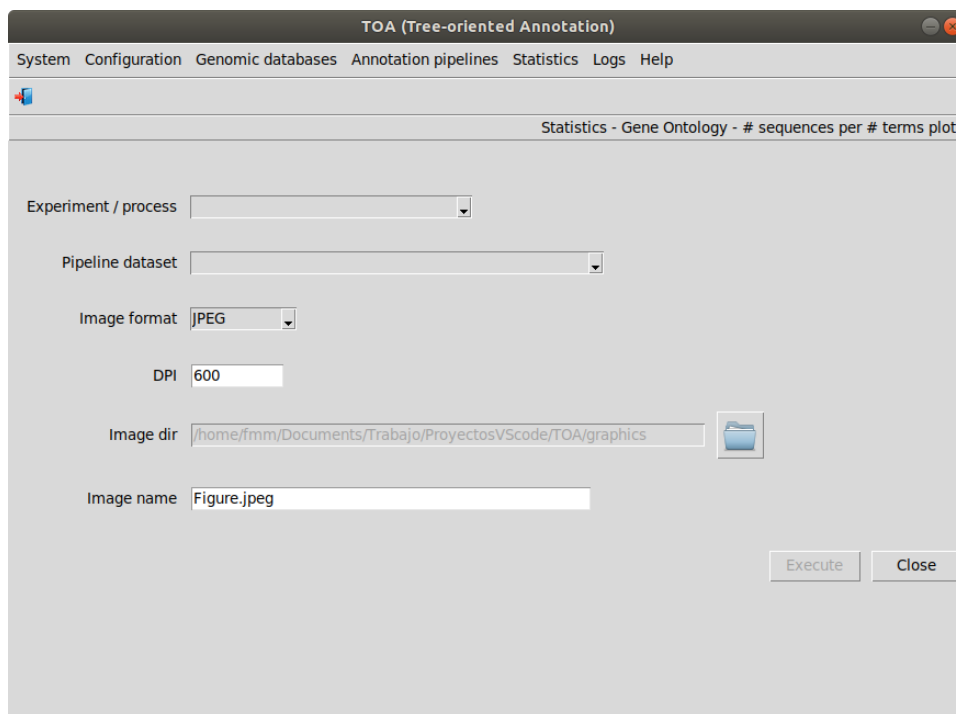


Figure. 37. TOA window corresponding to the menu item *Statistics - Gene Ontology - # sequences per #terms plot*.

The plot is saved in the directory selected in *Image dir* and is displayed in a pop-up window (see Figure 38).

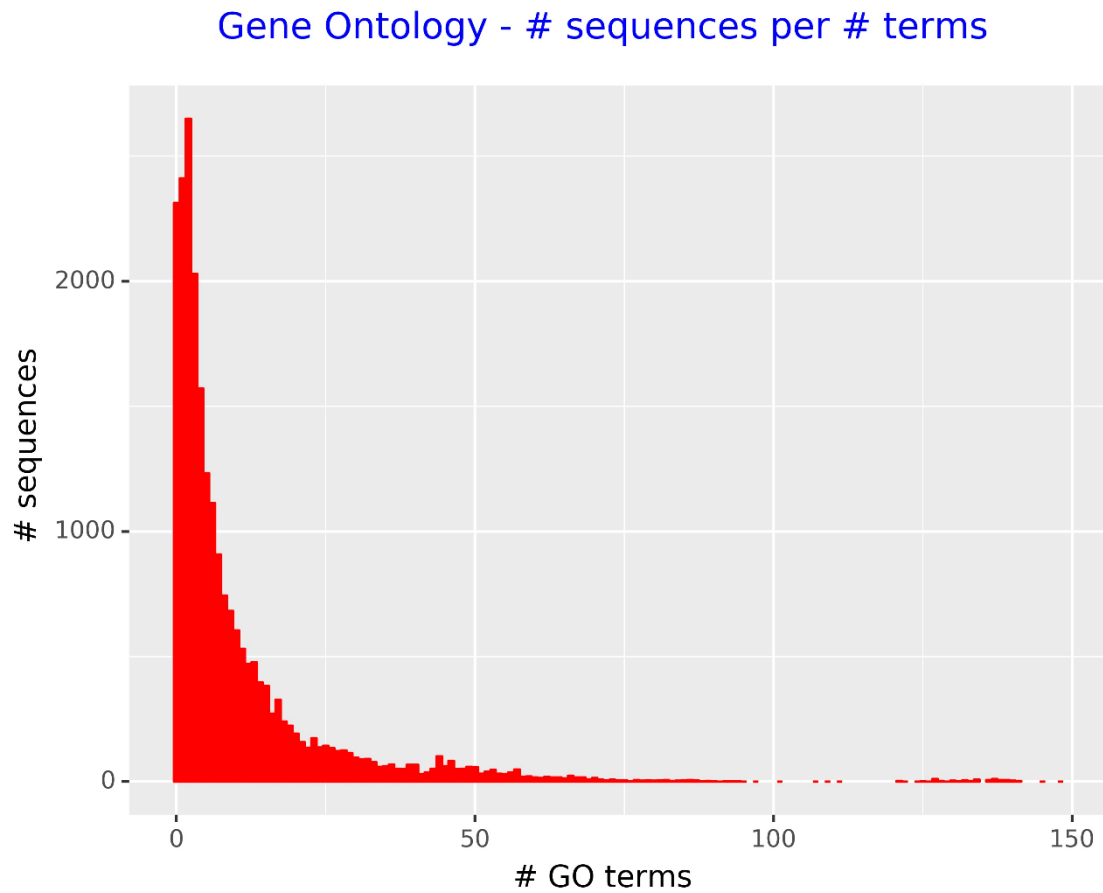


Figure. 38. Plot of #sequences per # Gene Ontology terms generated by the example annotation pipeline.

How to cite