HOME
INTRODUCTION
DESCRIPTION
ARCHITECTURE
EVALUATION
DISCUSSION

IWBBIO 2014 (2nd International Work-Conference on Bioinformatics and Biomedical Engineering)

Massive Automatic Functional Annotation (MAFA)

Nelson Pérez - nelsonp@correo.udistrital.edu.co
Cristian Rojas - carojasq@correo.udistrital.edu.co
Nelson Vera - neverap@udistrital.edu.co

IGUN – CECAD AGREEMENT

In 2012 the Institute of Genetics at the National University of Colombia joined the High performance Computing center of district University of Bogotá

INSTITUTO DE GENÉTICA

CECAD
Centro de Computación de Alto Desempeño

UNIVERSIDAD NACIONAL DE COLOMBIA

UNIVERSIDAD DISTRITAL
FRANCISCO JOSE DE CALDAS

HOME
INTRODUCTION
DESCRIPTION
ARCHITECTURE
EVALUATION
DISCUSSION

A project of this group is to develop a Bioinformatics Platform for NGS

BIOINFORMATICS PLATFORM FOR NGS

Genomics data

- PREPROCESSING
- ASSEMBLY
- MAPPING
- GENOME COMPARATION
- GENE PREDICTION
- ANNOTATION

Transcriptomics data

- PREPROCESSING
- ASSEMBLY
- ANNOTATION
- QUANTIFICATION AND DIFFERENTIAL EXPRESSION

HOME
INTRODUCTION
DESCRIPTION
ARCHITECTURE
EVALUATION
DISCUSSION

The tool presented today (MAFA) is part of the annotation module

BIOINFORMATICS PLATAFFORM FOR NGS

Genomics data

- PREPROCESSING
- ASSEMBLY
- MAPPING
- GENOME COMPARATION
- GENE PREDICTION
- ANNOTATION

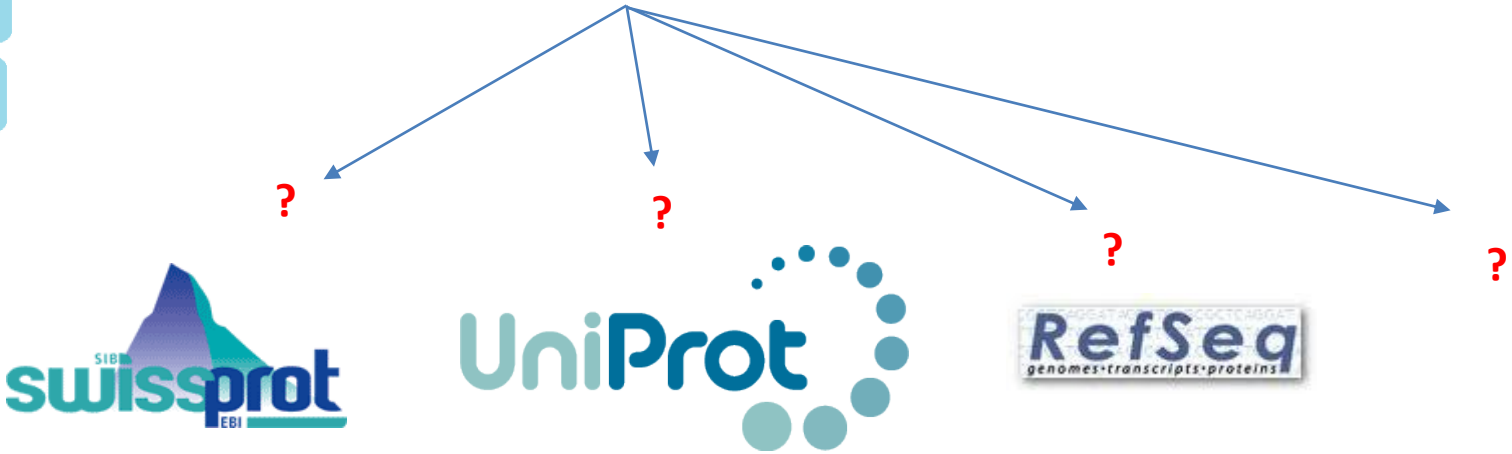
Transcriptomics data

- PREPROCESSING
- ASSEMBLY
- ANNOTATION
- QUANTIFICATION AND DIFFERENTIAL EXPRESSION

HOME
INTRODUCTION
DESCRIPTION
ARCHITECTURE
EVALUATION
DISCUSSION

ANNOTATION

```
VNNHSGEKLYECNERSKAFSCPSHLQCHKRR
---------YECNQCGKAFAQHSSLKCHYRT
         ****: .***:  * *:** *
```

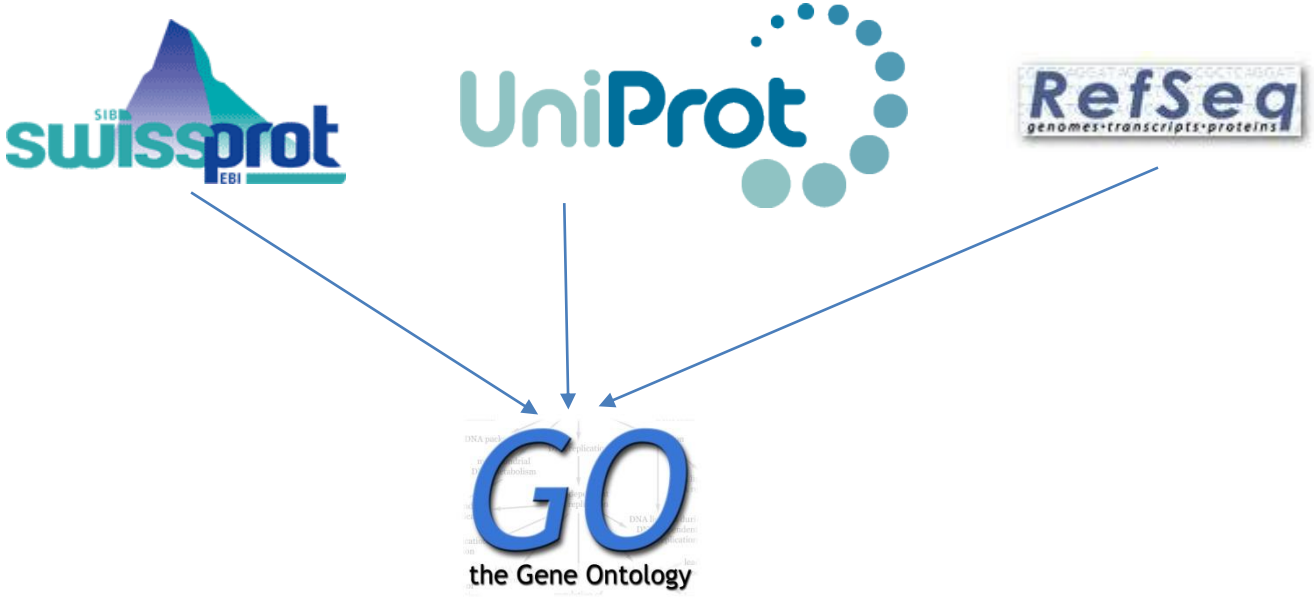


ANNOTATION: Comparison of sequences using alignments to search for similar sequences in databases of known sequences.

Association of unknown sequences with known sequences !!!

HOME
INTRODUCTION
DESCRIPTION
ARCHITECTURE
EVALUATION
DISCUSSION

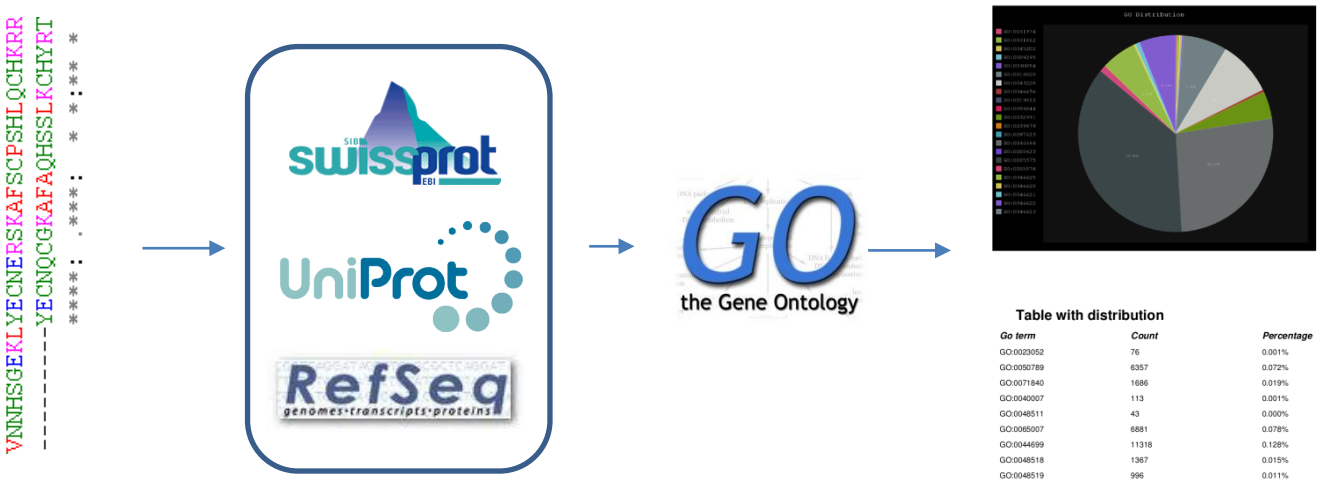
FUNCTIONAL ANNOTATION



FUNTIONAL ANNOTATION: Association of sequences (Known IDs) with functional groups (Cellular component, Biological process, Molecular function).

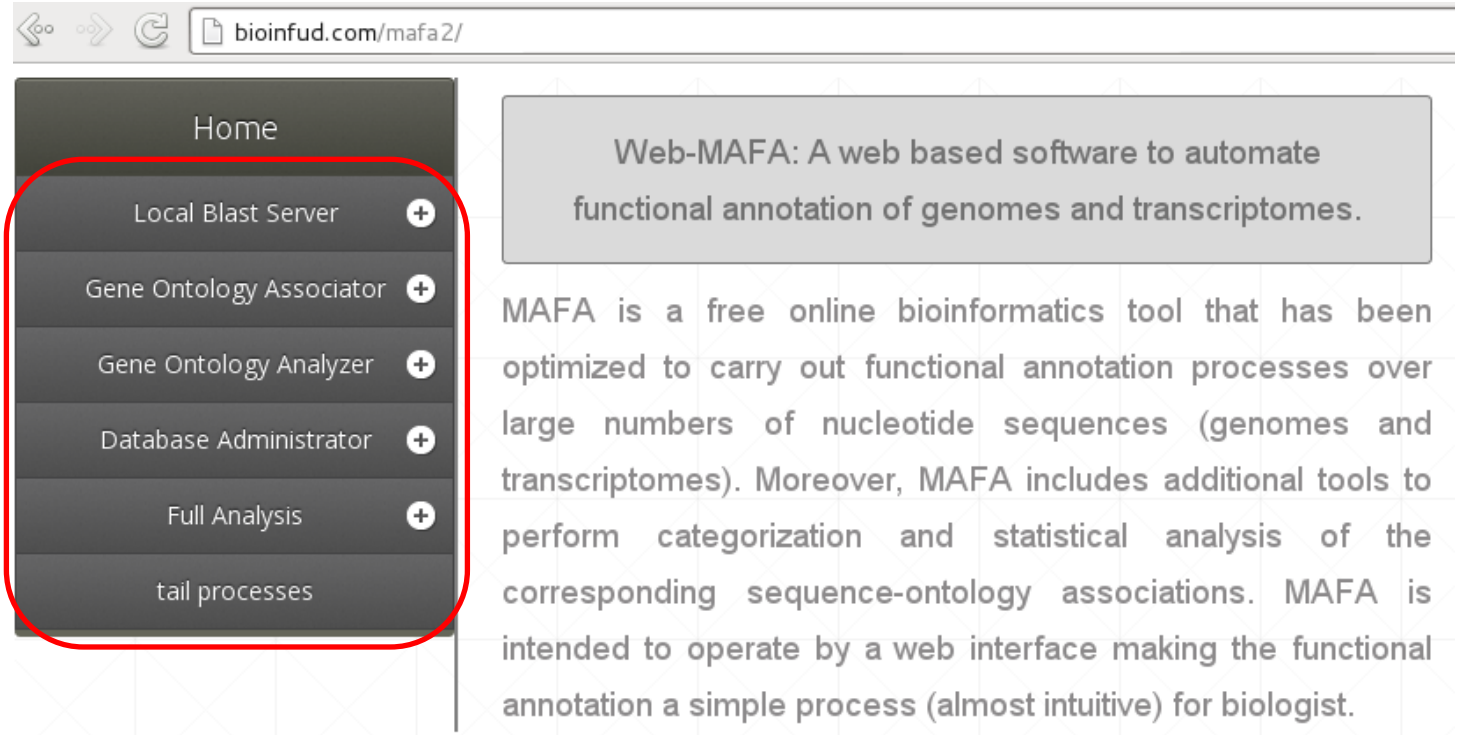
HOME
INTRODUCTION
DESCRIPTION
ARCHITECTURE
EVALUATION
DISCUSSION

MAFA



MAFA is an free bioinformatics tool that has been optimized to carry out functional annotation processes over large numbers of nucleotide sequences (genomes and transcriptomes). Moreover, MAFA includes additional tools to perform categorization and statistical analysis of the corresponding sequence-ontology associations.

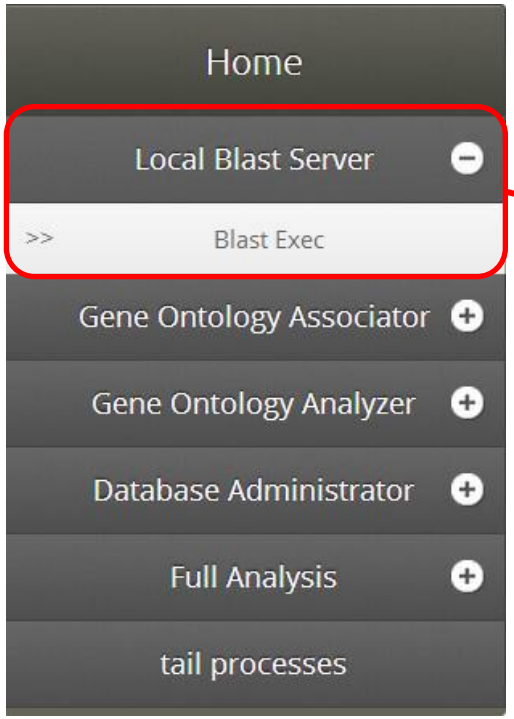
MAFA



Web-MAFA: A web based software to automate functional annotation of genomes and transcriptomes.

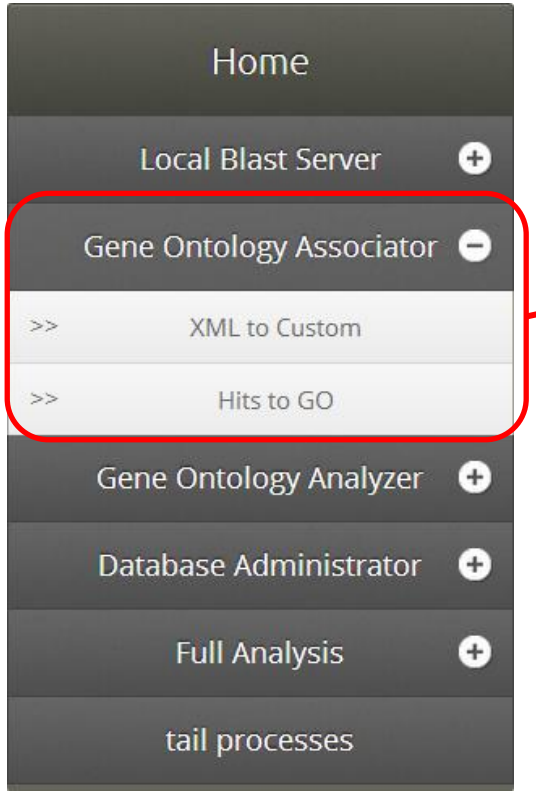
MAFA is a free online bioinformatics tool that has been optimized to carry out functional annotation processes over large numbers of nucleotide sequences (genomes and transcriptomes). Moreover, MAFA includes additional tools to perform categorization and statistical analysis of the corresponding sequence-ontology associations. MAFA is intended to operate by a web interface making the functional annotation a simple process (almost intuitive) for biologist.

MAFA



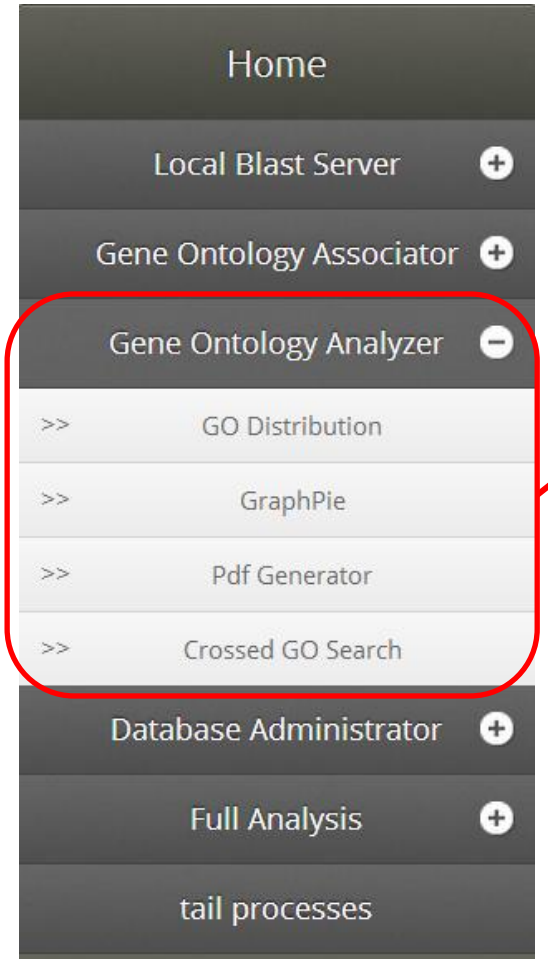
This item is in charge of running BLAST (Nucleotides vs Amino-acids) and also of storing the corresponding output using the XML format

MAFA



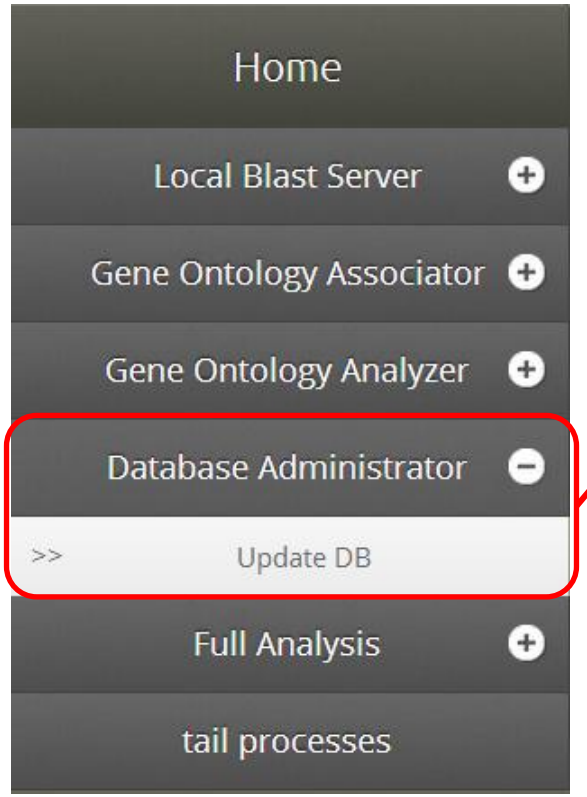
This item establishes the existing associations between the best hits, obtained from BLAST, and the terms from Gene Ontology

MAFA



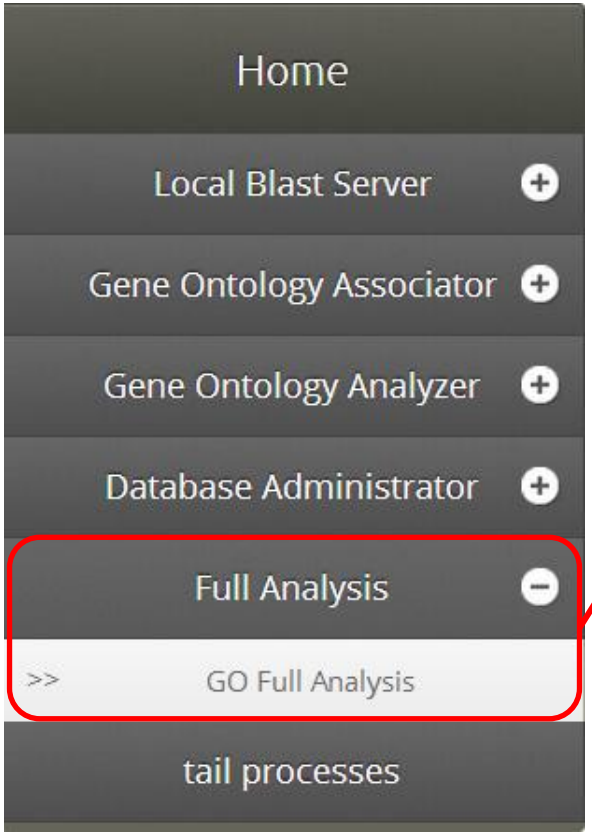
This item categorizes the GO terms according to user's interests

MAFA



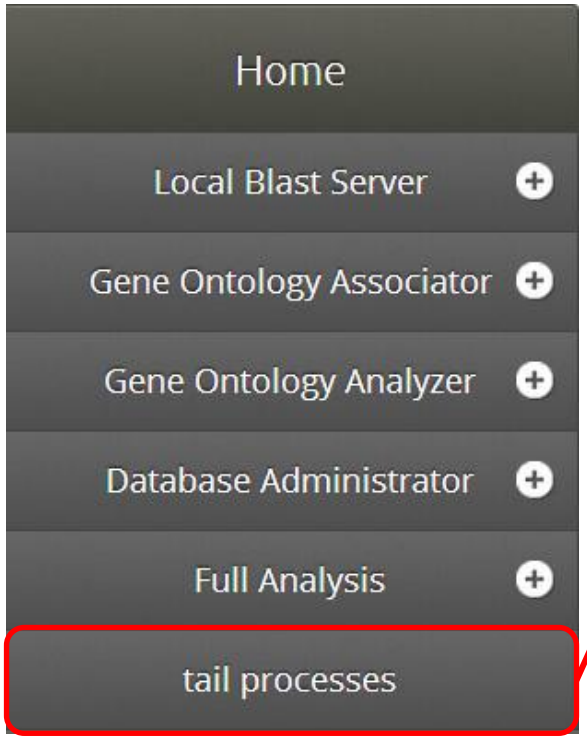
This item carries out updating tasks over the databases of both sequences and mapping so that the databases are available in the local server

MAFA



In this item the system run all the scripts in one single process.

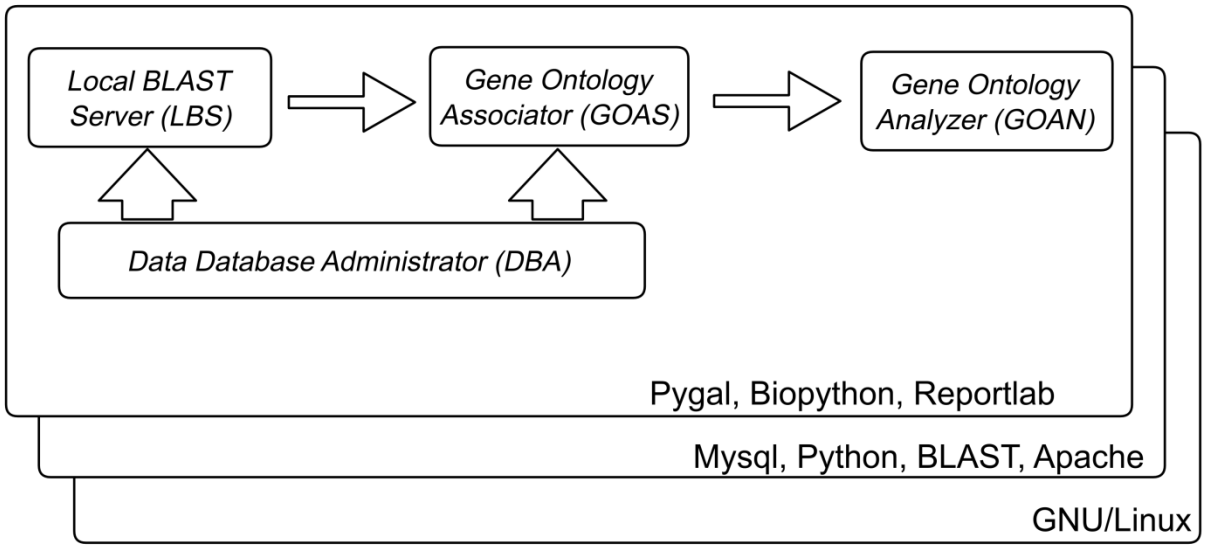
MAFA



This item indicates the states of the processes.

HOME
INTRODUCTION
DESCRIPTION
ARCHITECTURE
EVALUATION
DISCUSSION

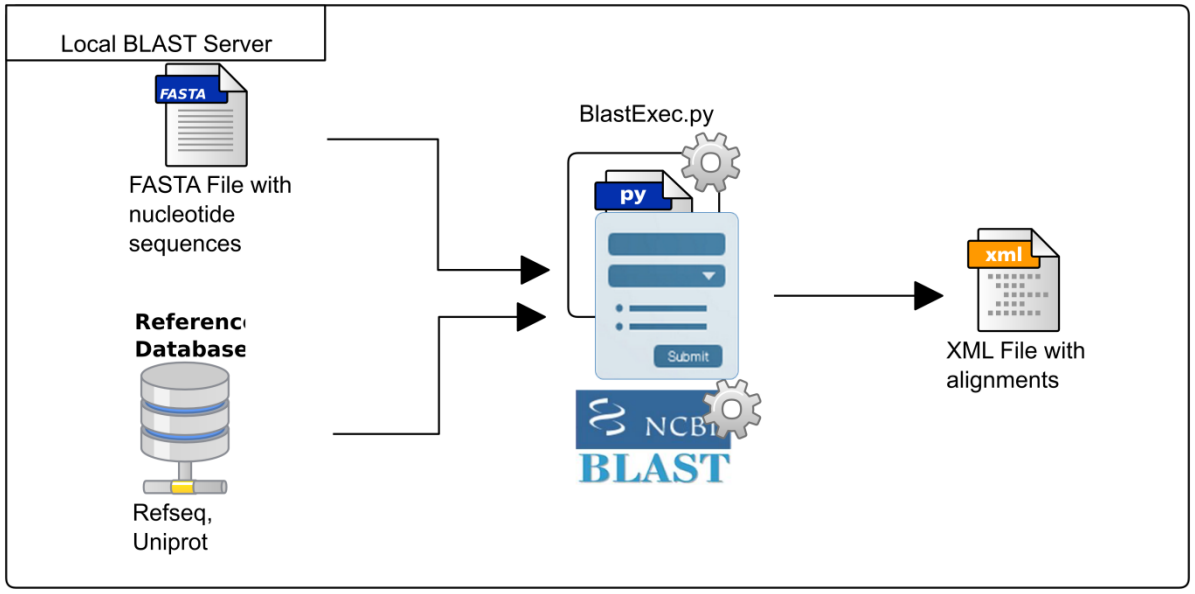
GENERAL ARCHITECTURE



MAFA consists of 4 modules that constitute a work flow. In order to run and integrate the modules, it is necessary to use additional tools that apply to all modules (cross-module applicable tools)

HOME
INTRODUCTION
DESCRIPTION
ARCHITECTURE
EVALUATION
DISCUSSION

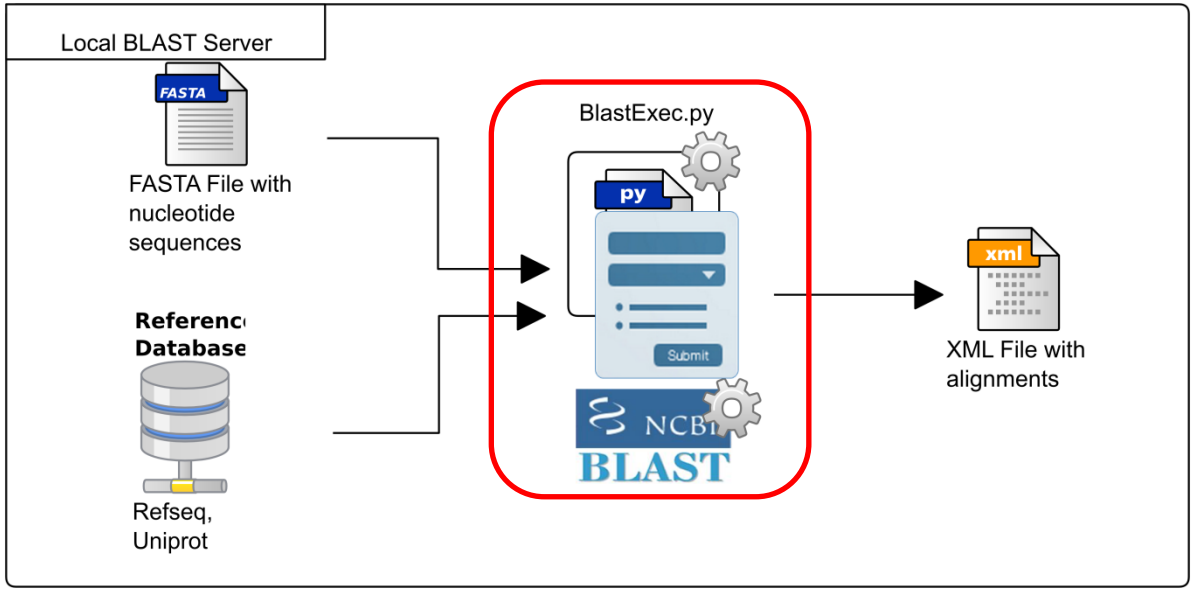
LOCAL BLAST SERVER



This module is in charge of running BLAST (Nucleotides vs Amino-acids) and also of storing the corresponding output using the XML format. Is composed by 1 scrtip.

HOME
INTRODUCTION
DESCRIPTION
ARCHITECTURE
EVALUATION
DISCUSSION

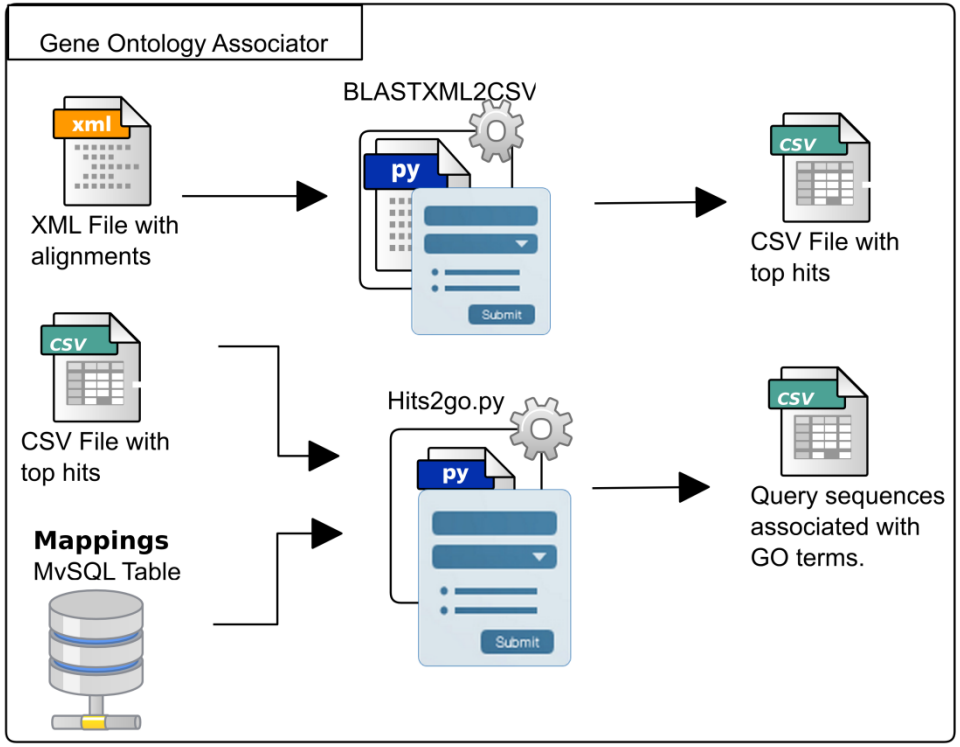
LOCAL BLAST SERVER



BlastExec.py The process orders the system to run blastx using various core

HOME
INTRODUCTION
DESCRIPTION
ARCHITECTURE
EVALUATION
DISCUSSION

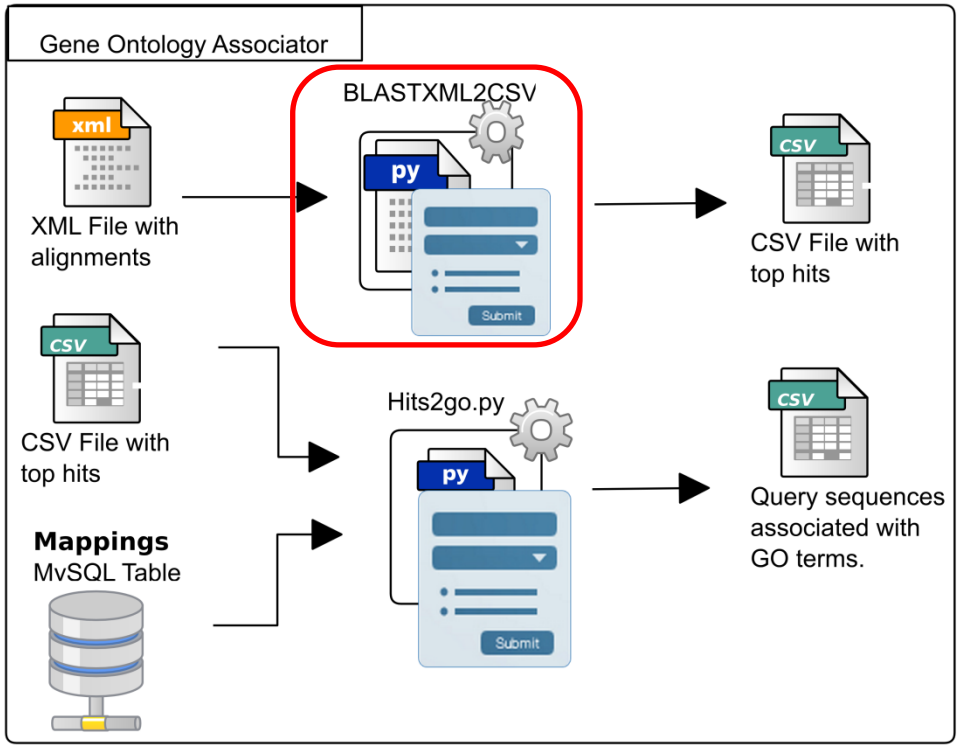
GO ASSOCIATOR



This module establishes the existing associations between the best hits, obtained from BLAST, and the terms from Gene Ontology.

HOME
INTRODUCTION
DESCRIPTION
ARCHITECTURE
EVALUATION
DISCUSSION

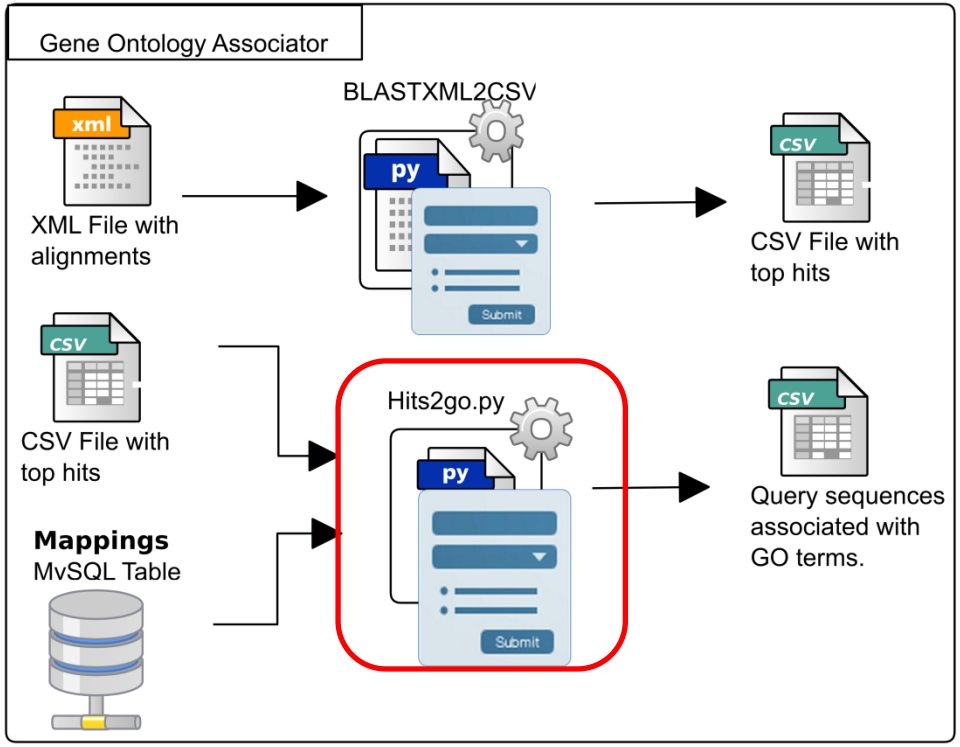
GO ASSOCIATOR



BLASTXML2CSV.py:: Selects the best alignment per sequence (top hit) and also writes the new file in CSV format

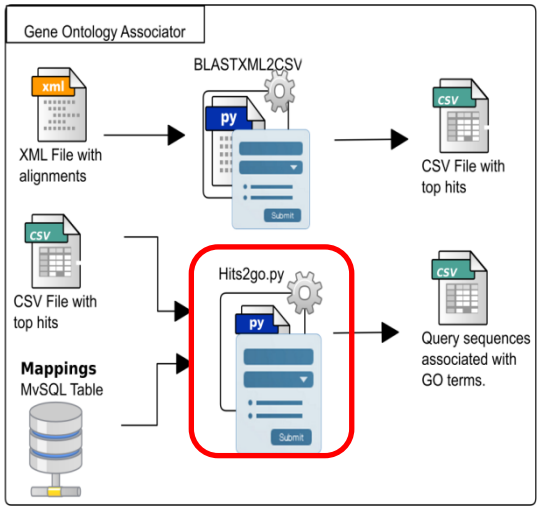
HOME
INTRODUCTION
DESCRIPTION
ARCHITECTURE
EVALUATION
DISCUSSION

GO ASSOCIATOR



Htis2go.py: The process makes an association between sequence identifiers and GO terms.

GO ASSOCIATOR



swissprot
UniProt
RefSeq
genomes·transcripts·proteins

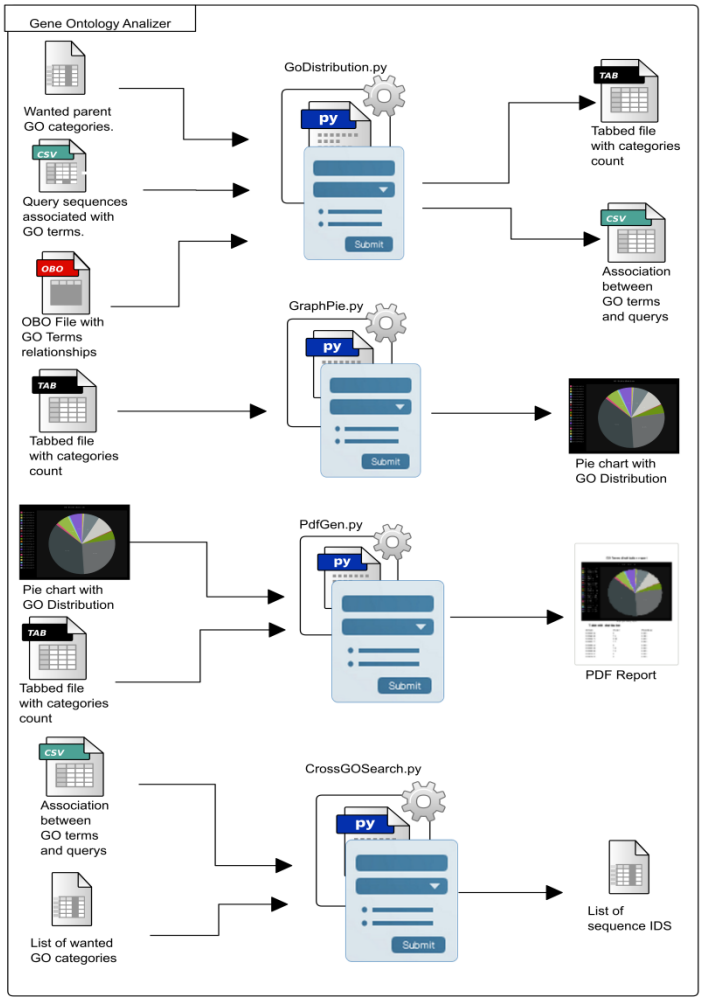
GO
the Gene Ontology

YLR229C	GO:0019317
YGR152C	GO:0042355
YOR212W	GO:0042354
YAR035W	GO:0006004
YCL066W	GO:0031424
YGL178W	

HOME
INTRODUCTION
DESCRIPTION
ARCHITECTURE
EVALUATION
DISCUSSION

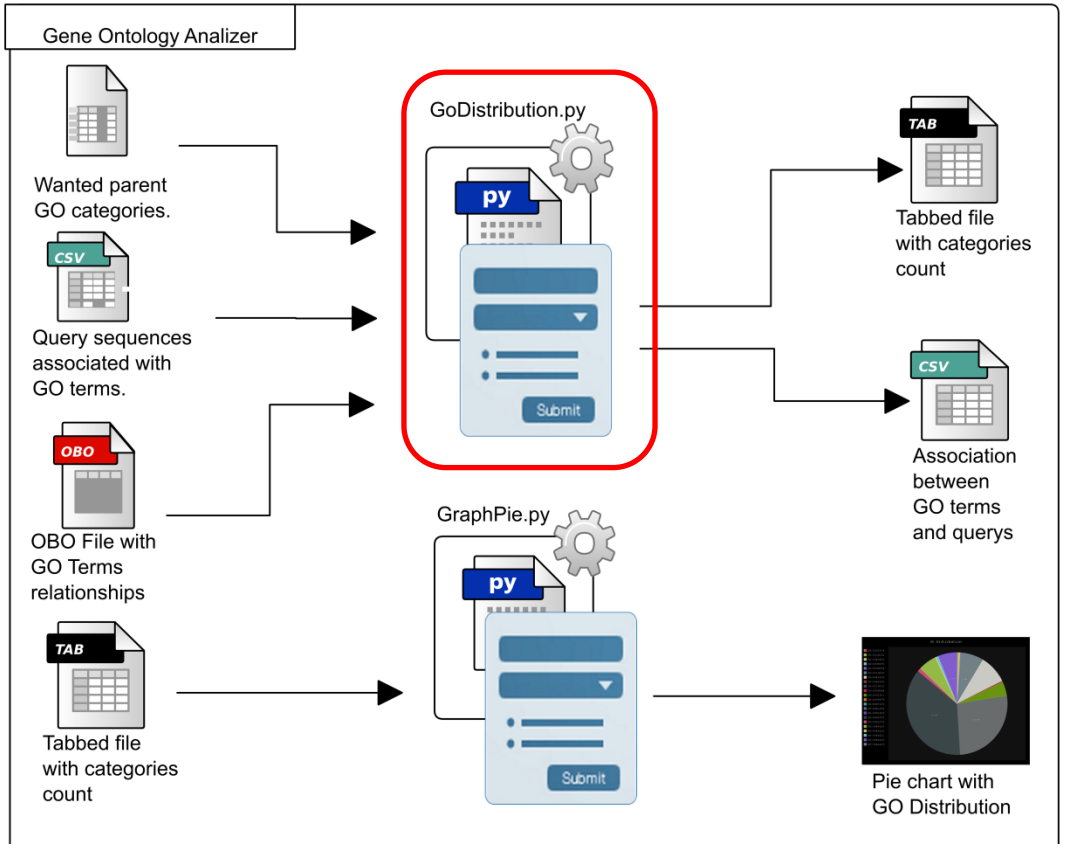
This module categorizes the GO terms according to user's interests. The module also counts how many times particular input sequences appear into the per-user categories and produces a complete report of the results

GO ANALYZER



HOME
INTRODUCTION
DESCRIPTION
ARCHITECTURE
EVALUATION
DISCUSSION

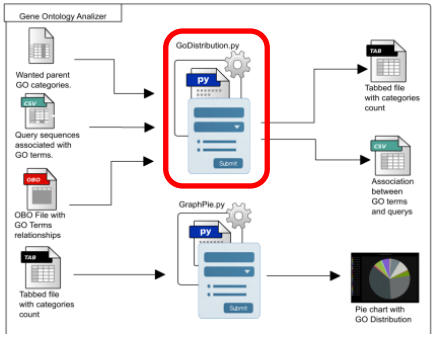
GO ANALYZER

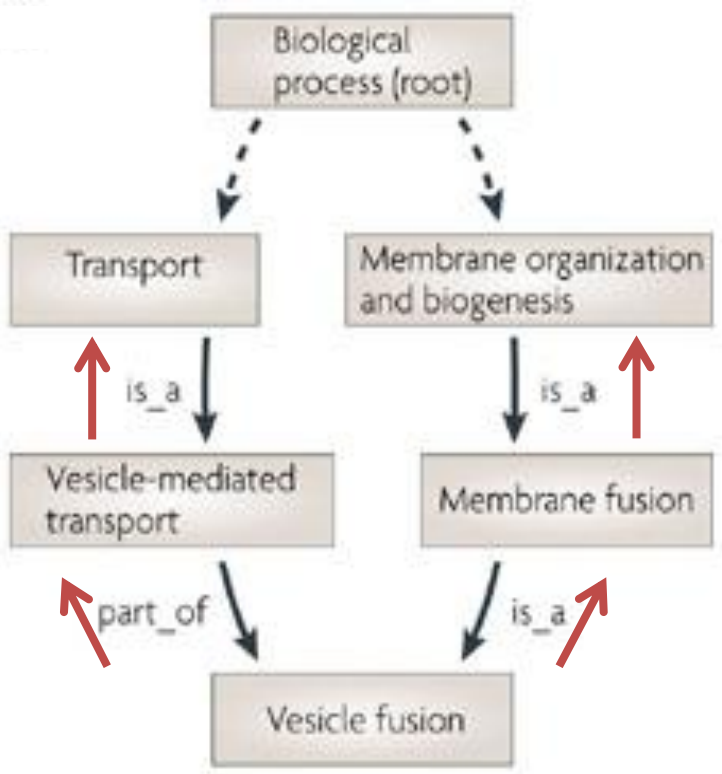


GoDistribution.py:
The process associates the desired GO categories (desired by users) to the more specific terms; it also counts how many times input sequences appear per desired GO category

HOME
INTRODUCTION
DESCRIPTION
ARCHITECTURE
EVALUATION
DISCUSSION

GO ANALYZER

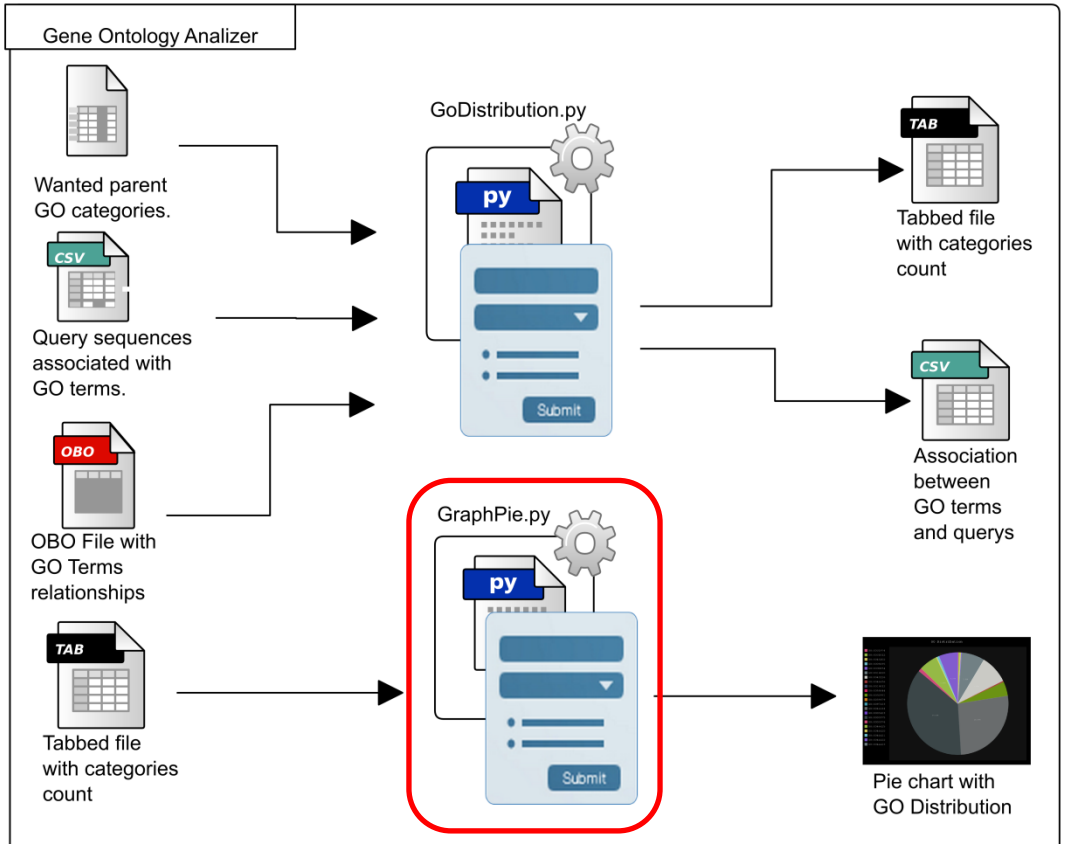




GoDistribution.py:
The process associates the desired GO categories (desired by users) to the more specific terms; it also counts how many times input sequences appear per desired GO category

Source: Nature review Genetics. 9:509-515 (2008)f

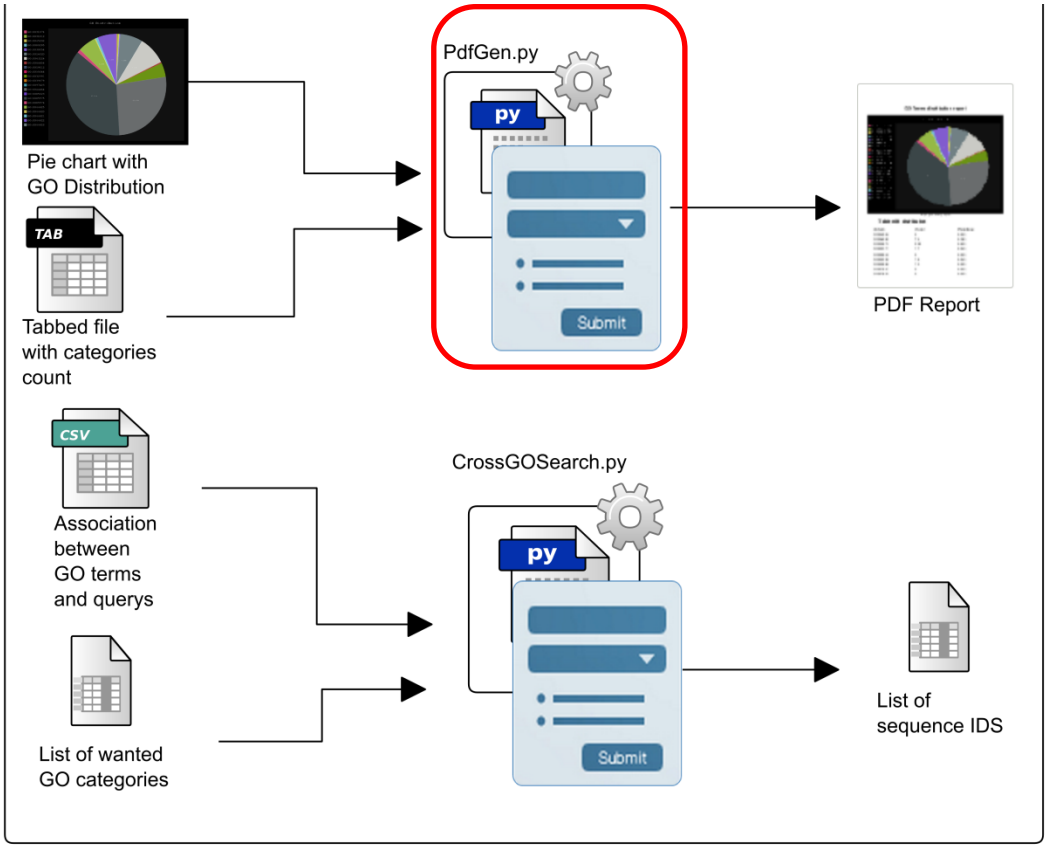
GO ANALYZER



GraphPie.py :
Produces a circular graph that illustrates the distribution of the categories.

HOME
INTRODUCTION
DESCRIPTION
ARCHITECTURE
EVALUATION
DISCUSSION

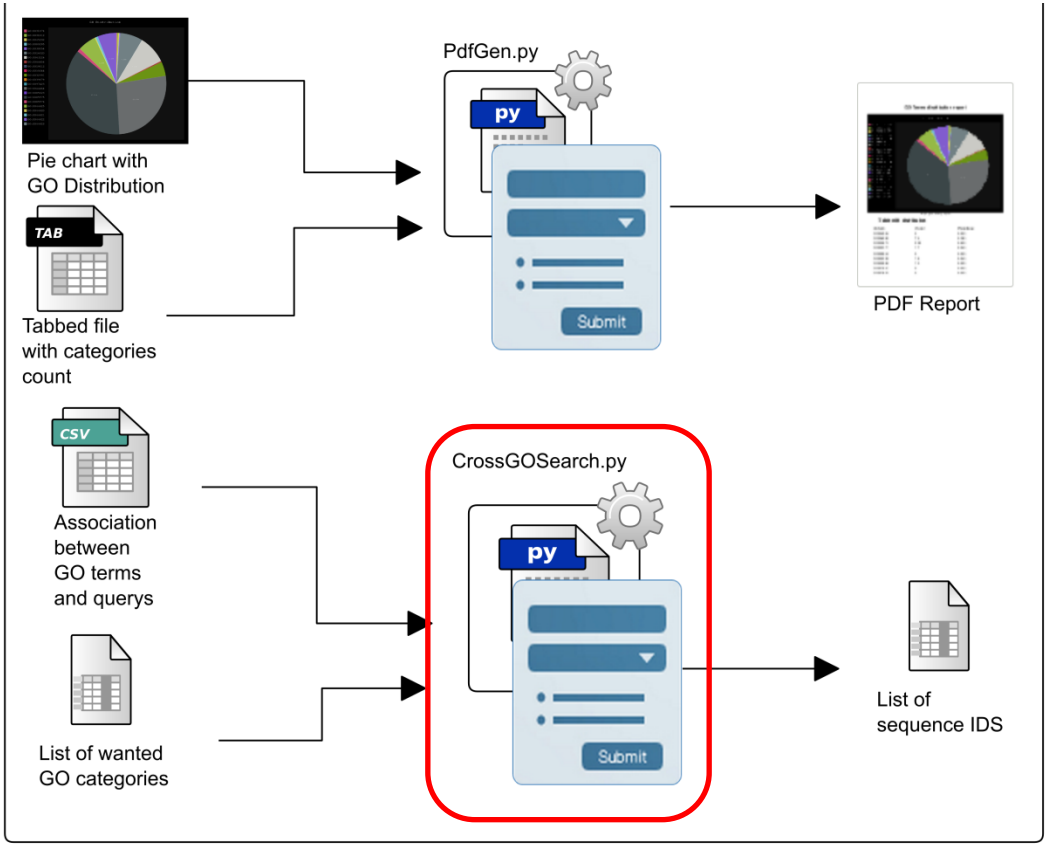
GO ANALYZER



PdfGen.py:
The process produces a PDF-format report that contains the analysis results.

HOME
INTRODUCTION
DESCRIPTION
ARCHITECTURE
EVALUATION
DISCUSSION

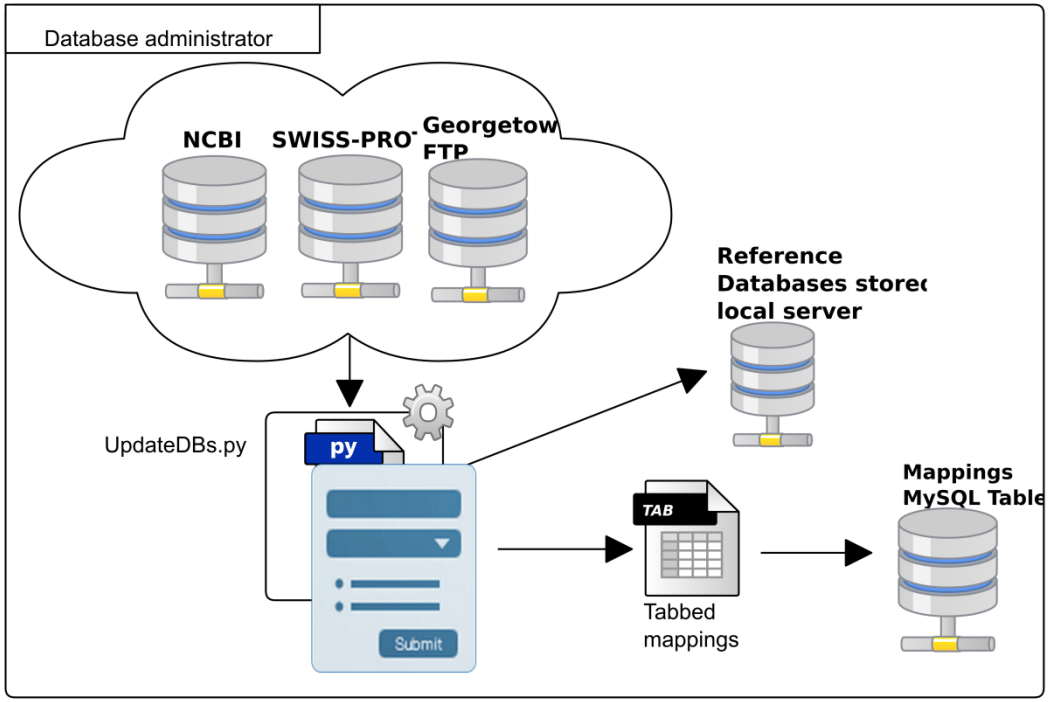
GO ANALYZER



CrossedGOSearch.py:
The process is a filter of all the sequences that appear in various GO categories at the same time.

HOME
INTRODUCTION
DESCRIPTION
ARCHITECTURE
EVALUATION
DISCUSSION

DATABASE ADMINISTRATOR



This module carries out updating tasks over the databases of both sequences and mapping so that the databases are available in the local server.

HOME
INTRODUCTION
DESCRIPTION
ARCHITECTURE
EVALUATION
DISCUSSION

DATA SET

Organism:	Diploria Strigosa.
Sequence type:	Transcriptomics.
Number of sequences:	500, 1000, 2000, 4000.
Format:	FASTA.
Database:	Uniprot, Non-Redundant

METRICS

- Processing Time
- Number of Results
- RAM Usage

HOME
INTRODUCTION
DESCRIPTION
ARCHITECTURE
EVALUATION
DISCUSSION

RESULTS

Databases	Number of sequences			Module				
				LBS	GOAS	GOAN	TOTAL	
	Original	BLAST hits	Annotated with GO	Time (S)	Time (S)	Time (S)	RAM (MB)	Time (S)
Uniprot	500	180	170	3460	0	2	1050	3462
	1000	384	367	4497	2	17	1050	4516
	2000	729	689	8678	23	67	1050	8768
	4000	1585	1513	19067	67	125	1050	19259
Refseq	500	191	176	35976	3	8	9134	35987
	1000	406	376	86780	34	23	9134	86837
	2000	759	706	190670	89	178	9134	190937
	4000	1672	1572	287808	140	201	9134	288149

HOME
INTRODUCTION
DESCRIPTION
ARCHITECTURE
EVALUATION
DISCUSSION

RESULTS

Databases	Number of sequences			Module				
				LBS	GOAS	GOAN	TOTAL	
	Original	BLAST hits	Annotated with GO	Time (S)	Time (S)	Time (S)	RAM (MB)	Time (S)
Uniprot	500	180	170	3460	0	2	1050	3462
	1000	384	367	4497	2	17	1050	4516
	2000	729	689	8678	23	67	1050	8768
	4000	1585	1513	19067	67	125	1050	19259
Refseq	500	191	176	35976	3	8	9134	35987
	1000	406	376	86780	34	23	9134	86837
	2000	759	706	190670	89	178	9134	190937
	4000	1672	1572	287808	140	201	9134	288149

LBS is the module that requires longer processing times is Local Blast Server.

HOME
INTRODUCTION
DESCRIPTION
ARCHITECTURE
EVALUATION
DISCUSSION

RESULTS

Databases	Number of sequences			Module				
				LBS	GOAS	GOAN	TOTAL	
	Original	BLAST hits	Annotated with GO	Time (S)	Time (S)	Time (S)	RAM (MB)	Time (S)
Uniprot	500	180	170	3460	0	2	1050	3462
	1000	384	367	4497	2	17	1050	4516
	2000	729	689	8678	23	67	1050	8768
	4000	1585	1513	19067	67	125	1050	19259
Refseq	500	191	176	35976	3	8	9134	35987
	1000	406	376	86780	34	23	9134	86837
	2000	759	706	190670	89	178	9134	190937
	4000	1672	1572	287808	140	201	9134	288149

The relation between processing time and the number of sequences is almost linear, reaching database-dependent rates of 4.8 seconds per processed sequences (for Uniprot) and 80.3 seconds per processed sequence (for Non-redundant).

HOME
INTRODUCTION
DESCRIPTION
ARCHITECTURE
EVALUATION
DISCUSSION

RESULTS

Databases	Number of sequences			Module				
				LBS	GOAS	GOAN	TOTAL	
	Original	BLAST hits	Annotated with GO	Time (S)	Time (S)	Time (S)	RAM (MB)	Time (S)
Uniprot	500	180	170	3460	0	2	1050	3462
	1000	384	367	4497	2	17	1050	4516
	2000	729	689	8678	23	67	1050	8768
	4000	1585	1513	19067	67	125	1050	19259
Refseq	500	191	176	35976	3	8	9134	35987
	1000	406	376	86780	34	23	9134	86837
	2000	759	706	190670	89	178	9134	190937
	4000	1672	1572	287808	140	201	9134	288149

Regarding RAM usage, there is direct dependency on the database in use.

There is no dependency on the number of sequences to be processed.

For Uniprot, RAM usage is approximately 1GB; for Non-redundant, RAM usage is 9GB.

HOME
INTRODUCTION
DESCRIPTION
ARCHITECTURE
EVALUATION
DISCUSSION

DISCUSSION (1)

MAFA is a tool that allows functional annotation and further annotation classification provided there are some given term-specific categories of Gene Ontology. MAFA's main functions include the following: the generation of structured-data outputs that advertise the amount of sequences associated to each GO term, and the establishment of relations between the target term identifiers of Gene Ontology and the identifiers of the given sequences.

HOME
INTRODUCTION
DESCRIPTION
ARCHITECTURE
EVALUATION
DISCUSSION

DISCUSSION (2)

Additionally, MAFA generates easy-to-interpret graphs for users as well as complete PDF reports containing the results from the corresponding analysis. It is also possible to conduct search processes in order to find sequences that are simultaneously associated to various categories or GO terms.

HOME
INTRODUCTION
DESCRIPTION
ARCHITECTURE
EVALUATION
DISCUSSION

DISCUSSION (3)

Regarding performance of the tool (MAFA), a linear behavior was observed when analyzing processing time and the number of sequences. In this respect, database-dependent rates (using the 8 cores of a Xeon E7450 processor and 256GB RAM) were found to be 4.8 seconds per sequence for Uniprot and 80.2 seconds per sequence for Non-redundant. Additionally, it was observed that RAM usage patterns are independent of the number of sequences to be processed and only depend on the reference database in use.

THANKS

Aviability: http://bioinfud.com/mafa2/.

Source codes: https://github.com/BioinfUD/MAFA.