

PcGs-finder

Overview

PcGs-finder is a computational pipeline to identify Polycomb-group proteins (PcGs) homologous in eukaryotes. PcGs-finder can be run using any set of sequence data, as long as sequences are in fasta format. PcGs-finder is a python script to apply pipeline employs hmmer (jackhmmer) protein search and eggNOG mapper to find a robust protein homologous within a big protein database. Finally, the selected protein searched for the domain architecture using hmmer. All results will be reported in a tab-delimited file.

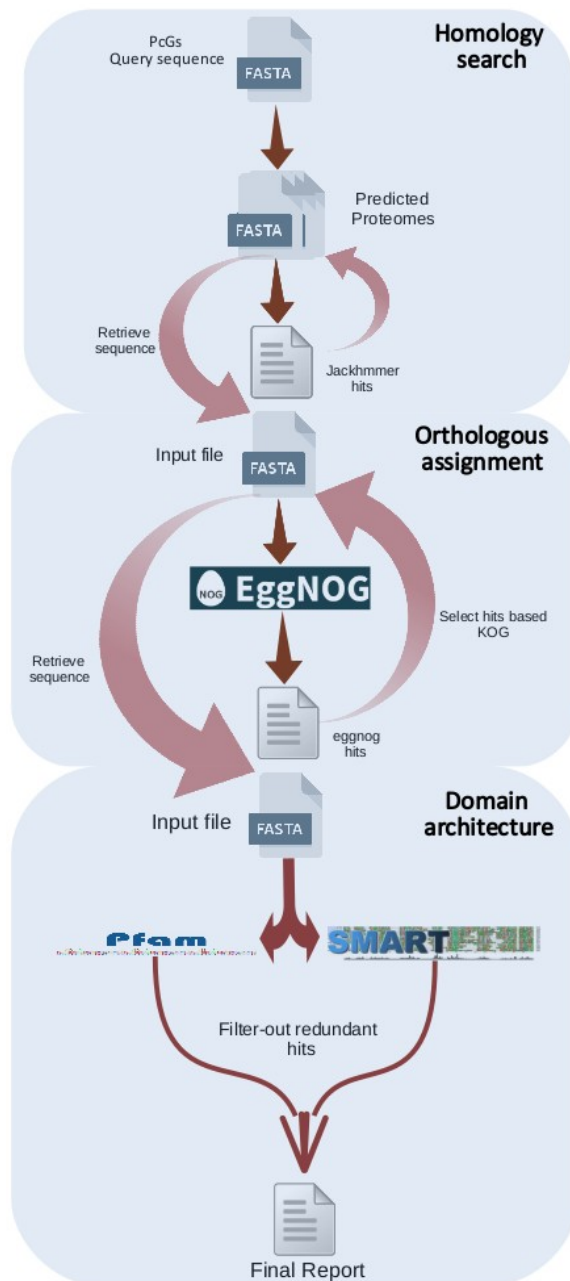
PcGs-finder can be used to process sequences downloaded directly from GenBank/NCBI local sequence data (e.g. sequences not downloaded from GenBank such as unpublished data), or a combination of both. PcGs-finder process total predicted proteins based on genomics or transcriptomics data. PcGs-finder offers the option to select a different PcGs' sub-units (ezh,suz12, esc, or p55); to allow users to find the homologous of each sub-unit separately. Also, PcGs-finder offers a multi-threading option for fast and short-time computing.

PcGs-finder is scalable and can be used to process a variety of datasets, ranging from small datasets (one species) to large datasets with thousands of proteins. PcGs-finder was intended to be objective and repeatable and provides a meaningful output at every step to help guide user decisions.

PcGs-finder is described in more detail in the following publication:

The article is still in preparation.

A visual overview of the major steps in PcGs-finder is shown below:



Installation

PcGs-finder is written in Python (compatible with 3.0 or more). The required dependencies will be downloaded and executed during the installation of PcGs-finder as a Python package or library. PcGs-finder directory must be used as a working directory, then the input file(s) must be transferred to the PcGs-finder directory but the output folder can be elsewhere. PcGs-finder requires a 40 GB free-space to download the necessary databases. There are only two external dependencies that should be installed before using PcGs-finder:

- HMMER: biosequence analysis using profile hidden Markov model.
<http://hmmerr.org>
- eggnog-mapper.

<https://github.com/eggnogetdb/eggnoget-mapper>

Helpful installation instructions for these dependencies can be found on the previous links.

PcGs-finder scripts can be run using Mac OSX (10.10+) and Linux, and can also work with Windows using a program like Cygwin. PcGs-finder can install using the following command:

```
./setup.py  
OR  
python3 setup.py
```

Usage

PcGs-finder can be run using following command:

```
./PcGs_finder.py [-h] -i INFILE -pg PCGS -o OUTFOLDER [-t THREADS] (it's 10 by default)
```

PcGs-finder can be tested for the first time through the available test data (Athaliana.fasta; the total predicted protein of *Arabidopsis thaliana*).

Examples commands:

for ezh sub-unit:

```
./PcGs_finder.py -i Athaliana.fasta -pg ezh -o TEST
```

for suz12 sub-unit:

```
./PcGs_finder.py -i Athaliana.fasta -pg suz12 -o TEST
```

for esc sub-unit:

```
./PcGs_finder.py -i Athaliana.fasta -pg esc -o TEST
```

for p55 sub-unit:

```
./PcGs_finder.py -i Athaliana.fasta -pg p55 -o TEST
```