

PanFunPro: set-up and run instructions

Oksana Lukjancenko*

1 What is PanFunPro?

PanFunPro stands for PAN-genome analysis based on FUNctional PROfiles. PanFunPro is a tool for pan-genome analysis. It has several following functionalities (1) homology detection and genome functional characterization by three HMM-collections, (2) pan-/core genome calculation within the set of proteomes, (3) pairwise pan-/core-genome analysis, (4) specific genome calculation for different subsets of genomes as well as pairwise analysis of specific proteomes, (5) basic statistics for the output genes from the pan-/core-/specific-genome calculation, (6) analysis of available pathway information for the output genes from the pan-/core-/specific-genome calculation.

PanFunPro is provided in two parts: PanFunPro2profiles.pl and PanFunPro2apply.pl

PanFunPro2profiles.pl is homology detection and functional characterization process (1), which is performed in four-step:

1. Collect functional domain information for each protein in the analysed proteome. Three databases are used for functional domain hunting: PfamA, Superfamily, and TIGRFAM.
2. Functional profile formation based on functional domain content in the protein. Databases are considered in the following order: PfamA, TIGRFAM, Superfamily; meaning that first PfamA domain information is encountered. Whether protein found no hits in PfamA database possible TIGRFAM domains are considered. And further, if protein didnt find any matches in neither PfamA nor TIGRFAM databases Superfamily is considered.
3. All protein sequences with no matches to any of three HMM-collections are clustered using CD-HIT. Each cluster is considered to be a non-characterized profile.
4. HMM-profiles and Clustering-profiles for each genome are joined together to form Genome-profiles.

PanFunPro2apply.pl is estimation and visualization process (2-6).

2 Software requirements

- Linux/UNIX
- Perl
- BioPerl

*e-mail: oksana@cbs.dtu.dk

- GO:Parser package
- InterProScan package (RC6 and higher). InterProScan might have additional requirements, which can be described on the developer web-page: <https://code.google.com/p/interproscan/wiki/InterProScan5InstallationRequirements>
- HMMER3 package <http://hmmer.janelia.org/software>
- The Oracle /Sun Java 6 JVM
- R
- CD-HIT <http://weizhong-lab.ucsd.edu/cd-hit/download.php>

The source code is primarily developed in Perl.

3 Installing and setting up InterProScan

It is important to follow the basic InterProScan installation recommendations. Additionally, some simple tips are provided:

- Install local lookup service. This will make it run faster and GO term information is not accessible without it
- Open interproscan.properties file
- Make sure Binary file locations are provided with valid path.
- Make sure precalculated.match.lookup.service.url is provided correctly.

You might also want to adjust other parameters in this file to perform better in your specific environment.

4 How to run PanFunPro2profiles.pl

Once you obtained all the necessary software, you can run it directly from command line.

```
perl PanFunPro2profiles.pl [options]
```

Option explanation can be obtained using help:

```
PanFunPro2profiles.pl h
```

5 How to run PanFunPro2apply.pl

Once profiles for each proteome are created, you can PanFunPro2apply.pl using profile information.

```
perl PanFunPro2apply.pl [options]
```

Option explanation can be obtained using help:

```
PanFunPro2apply.pl h
```