fCAT

One of the critical steps in a genome sequencing project is to assess the completeness of the predicted gene set. The standard workflow starts with the identification of a set of core genes for the taxonomic group, in which the target species belongs to. The fraction of missing core genes serves then as a proxy of the target gene set completeness.

fCAT is a **f**eature-aware **C**ompleteness **A**ssessment **T**ool, that helps to answer the question “How complete is my gene set?”. In particularly, fCAT checks for the presence of conserved genes (the core genes) of a specific taxonomy clade in the target gene set using feature-aware directed ortholog search (fDOG). In addition to the length criteria for classifying the found orthologs (as same as BUSCO), fCAT utilizes the domain architecture similarity FAS scores to further validate the orthologs. The later gives an alternative view on the accuracy of the target gene models, which shows how different the target orthologs in comparison to the core genes in their domain architecture.

fCAT outputs both the summary result in a tabular text file and the phylogenetic profile of the core genes, which can be visualized using the tool PhyloProfile. By analyzing the profiles of the entire orthologous groups within a specific taxonomy clade, we can further identify and ultimately correct erroneous gene annotations.

Table of Contents

How to install

fCAT tool is distributed as a python package called *fcat*. It is compatible with Python ≥ v3.7.

You can install *fcat* using `pip`:

```
python3 -m pip install fcat
```

or, in case you do not have admin rights, and don't use package systems like Anaconda to manage environments you need to use the `--user` option:

```
python3 -m pip install --user fcat
```

and then add the following line to the end of your **~/.bashrc** or **~/.bash_profile** file, restart the current terminal to apply the change (or type `source ~/.bashrc`):

```
export PATH=$HOME/.local/bin:$PATH
```

Usage

The complete process of *fCAT* can be done using one function `fcat`

```
fcat --coreDir /path/to/fcat_data --coreSet eukaryota --refspecList
"HOMSA@9606@2" --querySpecies /path/to/query.fa [--annoQuery
/path/to/query.json] [--outDir /path/to/fcat/output]
```

where **eukaryota** is name of the fCAT core set (equivalent to BUSCO set); **HOMSA@9606@2** is the reference species from that core set that will be used for the ortholog search; **query** is the name of species of interest. If `--annoQuery` not specified, *fCAT* fill do the feature annotation for the query proteins using FAS tool.

Output

You will find the output in the */path/to/fcat/output/fcatOutput/eukaryota/query/* folder, where */path/to/fcat/output/* could be your current directory if you not specified `--outDir` when running `fcat`. The following important output files/folders can be found:

- *all_summary.txt*: summary of the completeness assessment using all 4 score modes
- *all_full.txt*: the complete assessment of 4 score modes in tab delimited file
- *fdogOutput.tar.gz*: a zipped file of the ortholog search result
- *mode*1, mode*2*, *mode_3* and *mode_4*: detailed output for each score mode
- *phyloprofileOutput*: folder contains output phylogenetic profile data that can be used with PhyloProfile tool

Besides, if you have already run *fCAT* for several query taxa with the same fCAT core set, you can find the merged phylogentic profiles for all of those taxa within the corresponding core set output (e.g. */path/to/fcat/output/fcatOutput/eukaryota/*.phyloprofile*).

To learn how to interpret the phylogenetic profiles using PhyloProfile, please watch this video.

fCAT score modes

The table below explains how the *specific ortholog group cutoffs* for each fCAT core set were calculated, and which *value of the query ortholog* is used to assess its completeness, or more precisely, its functional equivalence to the ortholog group it belongs to. If the value of a query ortholog is *not less than* its ortholog group cutoff, that group will be evaluated as **similar** or **complete**. In case co-orthologs have been predicted, the assessment for the core group will be **duplicated**. Depending on the value of each single ortholog, a *duplicated* group can be seen as **duplicated (similar)** or **duplicated (dissimilar)** in the full report (e.g. *all_full.txt*).

Score mode	Cutoff	Value used for comparing
Mode 1 - Strict mode	Mean of FAS scores between all core orthologs	Mean of FAS scores between query ortholog and all core proteins
Mode 2 - Reference mode	Mean of FAS scores between refspec and all other core orthologs	Mean of FAS scores between query ortholog and refspec protein
Mode 3 - Relaxed mode	The lower bound of the confidence interval calculated by the distribution of all-vs-all FAS score in a core group	Mean of FAS scores between query ortholog and all core proteins
Mode 4 - Length mode	Mean and standard deviation of all core protein lengths	Length of query ortholog

*Note: **FAS scores** are bidirectional FAS scors; **core protein** or **core ortholog** is protein in the core ortholog groups; **query protein** or **query ortholog** is ortholog protein of query species; **refspec** is the specified reference species*

Contact

For further support or bug reports please contact: tran@bio.uni-frankfurt.de