FASTTARGETPRED

USER GUIDE & DATASET PREPARATION

Installation

The FastTargetPred program is accessible at the following url:
https://github.com/ludovicchaput/FastTargetPred

Step by step setup and quick overview

FastTargetPred is a ready-to-use program.

1. Download the complete Github folder.

2. In the terminal, use the command change directory (cd) and go to the directory where the file FastTargetPred.py is located.

3. To list the different parameters available run the command:

```
> python3.7 FastTargetPred.py -h
```

4. The first time users run FastTargetPred, if MayaChemTools is not in the path (e.g., for instance check with a command such as: echo $PATH or printenv...), one has to specify the folder location of the downloaded MayaChemTools (bin) directory (e.g., /mypath/mayachemtools/bin - this install location will be stored by FastTargetPred for the next run. You can download MayaChemTools here: http://www.mayachemtools.org/ and unzip or uncompress the directory anywhere on your system).

5. Then to test FastTargetPred you can use a query input SDF file, for instance available in the directory **test_set**. Just copy the file sample_1.sdf that contains one query compound in your working directory and run in the terminal window:

```
> python3.7 FastTargetPred.py sample_1.sdf
```

The SDF file must be in the standard version V2000. Each compound should start with a name or ID. At the end of the SDF file, please remove empty lines if any.

the output in the terminal window (default output) should be:

```
* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *

      TargetPred version 1.2

* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *

Checking local system ...               ok
Checking input arguments ...            ok

A total of 1 molecules has been found.

Starting maya calculation ...

Estimated time :      1s ...            ok

Time for maya calculation: 2.1895759105682373

Starting tanimoto computation on 4 cores.

Compound : >118<

               1          CHEMBL269132   1.000  10014      P21556
PTAFR_CAVPO reviewed Platelet-activating factor receptor (PAF-R) (PAFr)
PTAFR Cavia porcellus (Guinea pig) CHEMBL5136  chemotaxis [GO:0006935];
cytokine production [GO:0001816]; inflammatory response [GO:0006954];
inositol trisphosphate biosynthetic process [GO:0032959];
phosphatidylinositol-mediated signaling [GO:0048015]


Elapsed time for the all script : 4.344098091125488.
```

This command will also create the “**out**” directory containing in this specific example two files: log_ECFP4.log (general information about the ECFP4 computation with MayaChemTools) and sample_1_ECFP4.fpf (the computed fingerprint).

Remarks

The program can run on the following operating system:

- Linux
- MacOS
- Windows

NB: for MayaChemTools on Windows, users have to install Perl (for instance: http://strawberryperl.com/) if it is not already available.

Usage

Help menu

FastTargetPred includes several features. All of them are shortly described in the help menu accessible through the following command:

```
> python3.7 FastTargetPred.py -h
```

FastTargetPred requires Python version 3.7 or above.

Basic usage

Below is the example of a basic command:

```
> python3.7 FastTargetPred.py myquery.sdf
```

The first argument is the file that contains the query molecule(s), it should be in SDF[1] format. This command will compare the query compound(s) with the compounds of the curated ChEMBL25 (see the article and supplement file) and propose a list of identified “similar compounds”. The default fingerprint is ECFP4.

Advanced usage

Example 1

```
> python3.7 FastTargetPred.py myquery.sdf -fp ECPF6 -tc 0.9
```

In the above example, 2 additional arguments are provided: the type of fingerprint (fp) and the Tanimoto coefficient threshold (tc). The Tanimoto coefficient is defined between 0 and 1. Higher is the Tanimoto, higher is the molecule similarity. This value differs for the different fingerprints.

Example 2

```
> python3.7 FastTargetPred.py myquery.sdf -fp ECPF6 MACCS -sd 10
```

In the above example, 2 fingerprints are provided, it activates the consensus mode. A second argument provides the score threshold defined as a standard deviation (sd).

Once the consensus mode is activated, Tanimoto scores are normalized and centered (z-score), then the mean z-score on the n given fingerprints is calculated and used to sort and filter FastTargetPred results. The sd argument can be combined with the Tanimoto threshold (tc), if so, the output provides results that match the 2 criteria.

Example 3

```
> python3.7 FastTargetPred.py myquery.sdf -o output.csv -f csv
```

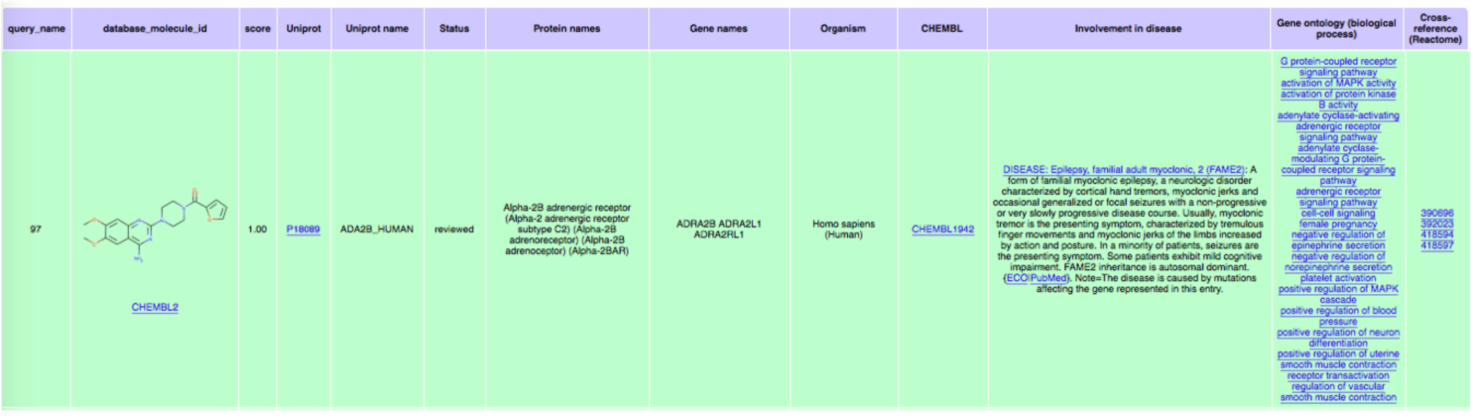
In this example, the output file name (specified with -o) is set to “output.csv” (written in the current directory). The second argument (-f) defines the file format. The output file format can be plain text or CSV file.

This output.csv file can be further processed with a utility script as follow:

```
> python3.7 FastTargetPred_workup.py -i output.csv -o test.html
```

[1] For this first step, the fingerprints are calculated with MayaChemTools, this toolkit requires input file in SDF format.

The html file will contain the following information:



query_name	database_molecule_id	score	Uniprot	Uniprot name	Status	Protein names	Gene names	Organism	CHEMBL	Involvement in disease	Gene ontology (biological process)	Cross-reference (Reactome)
97	CHEMBL2	1.00	P18089	ADA2B_HUMAN	reviewed	Alpha-2B adrenergic receptor (Alpha-2 adrenergic receptor subtype C2) (Alpha-2B adrenoreceptor) (Alpha-2B adrenoceptor) (Alpha-2BAR)	ADRA2B ADRA2L1 ADRA2RL1	Homo sapiens (Human)	CHEMBL1942	DISEASE: Epilepsy, familial adult myoclonic, 2 (FAME2): A form of familial myoclonic epilepsy, a neurologic disorder characterized by cortical hand tremors, myoclonic jerks and occasional generalized or focal seizures with a non-progressive or very slowly progressive disease course. Usually, myoclonic tremor is the presenting symptom, characterized by tremulous finger movements and myoclonic jerks of the limbs increased by action and posture. In a minority of patients, seizures are the presenting symptom. Some patients exhibit mild cognitive impairment. FAME2 inheritance is autosomal dominant. {ECO|PubMed}. Note=The disease is caused by mutations affecting the gene represented in this entry.	G protein-coupled receptor signaling pathway activation of MAPK activity activation of protein kinase B activity adenylate cyclase-activating adrenergic receptor signaling pathway adenylate cyclase-modulating G protein-coupled receptor signaling pathway adrenergic receptor signaling pathway cell-cell signaling female pregnancy negative regulation of epinephrine secretion negative regulation of norepinephrine secretion platelet activation positive regulation of MAPK cascade positive regulation of blood pressure positive regulation of neuron differentiation positive regulation of uterine smooth muscle contraction receptor transactivation regulation of vascular smooth muscle contraction	390696 392023 418594 418597

NB: If users have installed Python via Anaconda, they must have Matplotlib library available for plotting stats on the job results. One can do this using the "-p" argument and the "-cn X" argument to control the number of bin on the graph while processing output csv file.

Additional options

Several other options allowed to control the program (please see the -h option):

- Change the maximum number of targets to reports. By default, it is restricted to 50 by query compound.
- Define the number of CPUs used by FastTarget pred (-cpu).
- The possibility to output all hits for the same target (-bppt). By default, only the molecule with the highest score is reported.
- Removing contextual information about the targets.

In-house "database" preparation

In addition to the curated ChEMBL25 and the approved drugs dataset, a user can prepare its own annotated compound collection.

Required files

We provide 2 Python scripts to prepare the data:

- MayaFPcsv2bin.py
- tlt_gen.py

The user needs 2 input files:

- a SDF file (V2000) containing the annotated molecules.

- and a list of the protein-active molecule pair.

Note: the identifier of the molecule must be the same in the two files. If the target identifier is a ChEMBL identifier, FastTargetPred will add contextual information in the output file. These additional information, which come from merging Uniprot and ChEMBL data, are stored in the db/uniprot_database_ChEMBL.csv file.

Preparation

Fingerprint preparation

First, the user has to generate the fingerprints using MayaChemTools (http://www.mayachemtools.org). It can be ECFP4, ECFP6, Path Length or MACCS.

The fingerprint should be formatted in hexdecimal representation.

MayaChemTools command lines example:

```
# ECFP 4 1024 bits

ExtendedConnectivityFingerprints.pl --CompoundIDMode MolName -m
ExtendedConnectivityBits -r fp_mycompounds_ECFP4_1024Hex -o
mycompounds.sdf

# ECFP 6 1024 bits

ExtendedConnectivityFingerprints.pl --CompoundIDMode MolName -m
ExtendedConnectivityBits -n 3 -r fp_mycompounds_ECFP6_radius3_1024Hex -o
mycompounds.sdf

# MACCS 322 bits

MACCSKeysFingerprints.pl --CompoundIDMode MolName -r
fp_mycompounds_MACCS_322Hex -o mycompounds.sdf -s 322 -b
HexadecimalString

# Path length 1024 bits

PathLengthFingerprints.pl --CompoundIDMode MolName -m PathLengthBits -r
fp_mycompounds_PathLength -o mycompounds.sdf -b HexadecimalString
```

Then, the first script (mayaPFcsv2bin.py) is required to convert the MayChemTools fingerprint csv file into a binary file readable by FastTargetPred. The extension of the output file is .bfp (binary fingerprint).

Target-molecule pair dataset preparation

The second script (tlt_gen.py) formats the file containing the list of target-ligand pairs into a format readable by FastTargetPred.

The input file should be formatted as follow:

```
mol1 target1

mol1 target2

mol2 target3

mol2 target4

mol3 target5

...
```

The script converts it into the following format (1 molecule per line):

```
mol1 target1 target2

mol2 target3 target 4

mol3 target 5

...
```

The extension of the output file is .tlt (target-ligand table).

It is not necessary to sort the input file.

Dataset integration to FastTargetPred

Once the binary fingerprint file (.bfp) and the target-ligand table file (.tlt) are ready. The user can put it into the db/ folder, or any other folder. These 2 files must be in the same folder.

Then, the dataset can be screened with FastTargetPred using the following argument "-db db/mydataset" where "mydataset" is the prefix of the .bfp and .tlt files and “db/” is the relative of absolute path to access the files.

The filenames are important. The prefix is the name of the dataset. The binary fingerprint files should also have a suffix matching the fingerprint type, including an underscore (“_”) character:

"_ECFP4", "_ECFP6", "_PL" or "_MACCS".

Filename example:

- mydataset.tlt,
- mydataset_ECFP4.bfp, mydataset_ECFP6.bfp, mydataset_PL.bfp, mydataset_MACCS.bfp.