predDTI

=======

The predDTI package is developed to predict the interaction of protein and drugs. This standalone package will take a protein (from a certain protein category i.e. Enzyme, Ion Channel, GPCR and NR) and scan across multiple drugs and will generate an output stating whether the given protein interact with the drug or not.

1. System requirements

======================

The predDTI package is tested on Ubuntu system. predDTI is tested to work with Python 2.7. User is requested to install scikit-learn and XGBoost.

2. Important Links

==================

scikit-learn: https://scikit-learn.org/stable/

XGBoost: <https://xgboost.readthedocs.io/en/latest/get_started.html>)

protr: <https://cran.r-project.org/web/packages/protr/index.html>

ChemDes: <http://www.scbdd.com/chemdes>

3. Drug Web source

==================

For e.g. To search DrugID, D00348 (Isotretinoin), one can use the following web resources.

KEGG:

https://www.kegg.jp/dbget-bin/www\_bget?dr:(Append the drug ID here). Go to Other DBs section on this link. This section allows the user to navigate to various web resources to download drug files.

LigandBox:

http://ligandbox.protein.osaka-u.ac.jp/ligandbox/cgi-bin/liginf.cgi?id=D00348&source=KEGG\_DRUG

DrugBank:

https://www.drugbank.ca/drugs/DB00982

4. Input/Output formats

=======================

Input format

============

1. The input file should be a valid FASTA format that consists of a single initial line beginning with a greater-than symbol (“>”) in the first column, followed by lines of sequence data.

For e.g.,

>hsa:5916

MATNKERLFAAGALGPGSGYPGAGFPFAFPGALRGSPPFEMLSPSFRGLGQPDLPKEMASLSVETQSTSSEEMVPSSPSPPPPPRVYKPCFVCNDKSSGYHYGVSSCEGCKGFFRRSIQKNMVYTCHRDKNCIINKVTRNRCQYCRLQKCFEVGMSKEAVRNDRNKKKKEVKEEGSPDSYELSPQLEELITKVSKAHQETFPSLCQLGKYTTNSSADHRVQLDLGLWDKFSELATKCIIKIVEFAKRLPGFTGLSIADQITLLKAACLDILMLRICTRYTPEQDTMTFSDGLTLNRTQMHNAGFGPLTDLVFAFAGQLLPLEMDDTETGLLSAICLICGDRMDLEEPEKVDKLQEPLLEALRLYARRRRPSQPYMFPRMLMKITDLRGISTKGAERAITLKMEIPGPMPPLIREMLENPEMFEDDSSQPGPHPNASSEDEVPGGQGKGGLKSPA

1. The fasta file should contain ONLY naturally occurring amino acids or else the input will be ignored.
2. User is requested to provide a file containing only protein sequences.
3. Each protein sequence should be separated by a new line.
4. There is no restriction for the number of protein sequences in the input file, provided it contains valid naturally occurring amino acids and is in the fasta format.
5. Execution

==============

User need to execute the corresponding file (protein category-wise) in command line environment.

1. Protein category

=====================

Enzyme

GPCR: G-Protein Coupled Receptors

iChannel: Ion Channel

NR: Nuclear Receptors

1. Package Details

====================

Each protein category contains (Description provided w.r.t NR category):

1. An input file in fasta format that we used for prediction purpose. (Only as an example purpose for the user ,NR\_Target.fasta.)
2. A drug .csv file (Drug\_selFV.csv) that needs to be merged with the protein feature vector in order to form the orthogonal sum.
3. A python program (predDTI\_NR.py) that needs to be executed for prediction purpose.
4. Best folds (folds used during 5\*10-fold cv) that provided the best AGm values (For example, Fold2.csv, Fold3.csv, Fold5.csv, Fold6.csv and Fold 9.csv).
5. Predicted output (predDTI\_NR.out): Shows which protein interacts with which potential drugs from the analyzed set of drugs.
6. Usage

===========

Download the package from (web link http://) and extract it to a directory, for example, “~/usr”.

To execute the **predDTI** in command line environment, navigate to the “~/usr/predDTI/NR” directory and you will find a python script, predDTI\_NR.py”. The “predDTI\_NR.py” is used for protein-drug prediction. The protein sequences will be scanned across the analyzed set of drugs and produce the required output.

For example,

To execute NR protein sequences interacting with the drugs, provide the input in following manner in the command line:

$ python predDTI\_NR.py NR\_Target.fasta Drug\_selFV.csv

1. Output Format

==================

The output will be generated in the following form:

> predDTI\_NR.out

Follow the same steps for the other protein categories.

10.To predict with new set of Drugs

===================================

1. Download the drug files from the drug web source as discussed in Section 3.
2. Provide the drugs to the ChemDes webserver for drug feature vector generation.
3. Eliminate the features (columns) which contain entire rows to be zero and the features that contain missing values.
4. Normalise these feature vector using [0,1] normalisation using Weka 3.8.
5. Select only the drug features provided in the package for drugs (navigate to respective protein category folder) and form a Drug\_selFV.csv file.
6. In command line, execute the python program

$ python predDTI\_NR.py NR\_Target.fasta Drug\_selFV.csv

1. Generate the predicted output.
2. Follow the same steps for the other protein categories.