Project: anti-HIV

Development process of Project:

1) Retrieve all the bioactivity data for anti-cancer is downloaded from NCI

(https://wiki.nci.nih.gov/display/NCIDTPdata/Chemical+Data)

(https://wiki.nci.nih.gov/display/NCIDTPdata/AIDS+Antiviral+Screen+Data)

./data/aids\_ec50\_may04.txt #####ec50 data

./data/chembl\_human\_hiv\_10um.tab #####ec50 data from chembl

./data/Open\_2D\_Oct2014.smi or nciopenb\_smi.zip #####SMILES string for all compounds

2) Active and inactive compound information was obtained using python script (active(<10000 nm) and inactive(>100000 nm))

python step1a\_get\_data\_active\_inactive.py

input: ./data/aids\_ec50\_may04.txt

output: ./data/aids\_ec50\_may04\_active.txt

output: ./data/aids\_ec50\_may04\_inactive.txt

python step1b\_get\_data\_smi

input: ./data/aids\_ec50\_may04\_active.txt

input: ./data/aids\_ec50\_may04\_inactive.txt

input: ./data/chembl\_HIV\_active.smi

input: ./data/Open\_2D\_Oct2014.smi

output: ./data/chembl\_NCI\_aids\_active.smi

output: ./data/chembl\_NCI\_aids\_inactive.smi

3) Remove the conflict compounds between active and inactive compounds using python script

python step3a\_get\_similarity\_remove.py

python step3b\_get\_finaldata\_remove.py

input: ./data/chembl\_NCI\_aids\_active.smi

input: ./data/chembl\_NCI\_aids\_inactive.smi

output: ./data/chembl\_NCI\_aids\_active\_final.smi

output: ./data/chembl\_NCI\_aids\_inactive\_final.smi

4) Get the balance dataset by random selection of the inactive compounds using python script

python step3\_get\_inactive\_random.py

input: ./data/chembl\_NCI\_aids\_inactive\_final.smi

output: ./data/chembl\_NCI\_aids\_inactive\_final\_ok.smi

5) Write fingerprint of active and inactive compunds for the generation of predict model

python step4\_write\_bitzp\_and\_shuffle.py

input: ./data/chembl\_NCI\_aids\_inactive\_final.smi

input: ./data/chembl\_NCI\_aids\_inactive\_final\_ok.smi

output: ./data/alldata.pkl.z

output: ./data/alldatalabel.pkl.z

6) Cross validation of random forest (search best parameters)

python step5\_RF\_model\_performance.py

output: ./data/RF\_performance\_final.txt

7) Cross validation of support vector machine (search best parameters)

python step6\_SVM\_model\_performance1.py

python step6b\_SVM\_model\_performance2.py

output: ./data/svm\_performance\_final.txt

8) Generated final model using above parameters(RF and SVM)

python step7\_generate\_model.py

output: ./data/HIV\_SVM\_model\_500\_rbf\_win.pkl

output: ./data/HIV\_RF\_model\_1000\_win.pkl.z

9) Scripts for the prediction of anti-bacterial activity of query compounds

python step8\_antiHIVpre.py

output: anti\_HIV\_prediction.txt