DSE 601: Project 1

Sourav Yadav

11 October 2020

Problem

The aim of this project is to classify compounds as active or inactive based on their inhibition values.

Overview

The Coronaviridae are a family of positive single stranded encapsulated viruses. They typically cause mild respiratory diseases, but infections with the ß-coronavirus SARS-CoV, MERS and SARS-CoV-2 can lead to acute respiratory diseases and high mortality, particularly in individuals with underlying health conditions. In the last 20 years, Coronaviridae have emerged in two severe outbreaks, 2002/2003 with SARS-CoV, 2012 with MERS and one pandemic, in late 2019 with SARS-CoV-2. At the time of writing, the coronavirus SARS-Cov-2 pandemic has led to close to 39 million confirmed infections and 1 million deaths (https://coronavirus.jhu.edu).

Multiple interventional clinical trials have been initiated in the search for effective pharmacological treatments against SARS-CoV-2 infection and the related disease Covid-19. Bioinformatics analyses have proposed repurposed drugs based on the interactome between viral encoded proteins and host-cell pathways. In the absence of safe and effective vaccines against SARS-CoV-2, repurposing of existing drugs represents a first pragmatic strategy for the treatment of Covid-19 patients. We thus identify potential inhibitors of SARS-CoV-2 in-vitro cellular toxicity in human (Caco-2) cells by predicting its bioactivity using ML after training the data on a large scale drug repurposing collection[3]. Bioactivity describes the characteristic of an implant material to interact with or initiate a specific reaction of living tissue upon exposure. The biochemical systems encountered by a drug molecule (implant material) are extremely complex. The factors affecting the bioactivity[4] may

be divided into three categories:

1. Physicochemical properties such as solubility, partition coefficients, and ionization.

2. Chemical structure parameters such as resonance, inductive effect, oxidation potentials, types of bonding, and isosterism.

3. Spatial considerations such as molecular dimensions, interatomic distances, and stereochemistry

Data

To identify possible candidates for progression towards clinical studies against SARS-CoV-2, the authors of the paper[1] screened a well-defined collection of compounds. We obtained this data via the data base - ChEMBL27 SARS-CoV-2 release under the title Identification of inhibitors of SARS-CoV-2 in-vitro cellular toxicity in human (Caco-2) cells using a large scale drug repurposing collection[3].

We obtain this data in two parts as follows:

1. **inhibition.csv** – Data in this data set is used for stating the bioactivity measure which are trained against the corresponding attributes from compounds.csv to classify compounds into two categories active and inactive. A compound is called active if it's inhibition value is greater than 75% otherwise it is classified as inactive.

2. **compounds.csv** – Data in this data set is used for stating factors that affect towards bioactivities as stated previously. The factors selected from the given set of attributes are Molecular Weight, AlogP, PSA, HBA, HBD, CX ApKa, CX BpKa & CX LogD which we shall use as our descriptors.

Here is a rough overview of our descriptors and targets. Here we see the reason as to why the following descriptors are chosen by comparing their definitions to the factors that affect bioactivity of a compound.

DESCRIPTORS

1. *Molecular Weight* : Measure of the mass of a given molecule.
2. *AlogP* : Measure of lipophilicity which is a key physiochemical property that plays a crucial role in determining ADMET(adsorption, distribution, metabolism, excretion and toxicity) properties and overall suitability of drug candidates.
3. *PSA* : Measure of the polar surface area(PSA) of a molecule is defined as the surface sum over all polar atoms or molecules, primarily oxygen and nitrogen, also including their attached hydrogen atoms.
4. *HBA* : Hydrogen Bond acceptors atoms.
5. *HBD* : Hydrogen Bond donor atoms.
6. *CX ApKa, CX BpKa* : Measure of pH
7. *Log D* : Measure of distribution coefficient. It is the ratio of the sum of the concentrations of all forms of the compound(ionized plus un-ionized) in each of the two phases, one essentially always aqueous; as such, it depends on the pH of the aqueous phase, and log D = log P for non-ionizable compounds at any pH.

TAGETS

Bioactivity : This is boolean variable which is active when inhibition value of particular compound is greater than 75% otherwise the compound is said to be inactive.

We further use Pandas Profiling to see a bried visualization of our data against parameters such as Correlations, etc.

Approach

As stated in the paper, it will be key to determine whether any clinical-stage compounds or related molecules could safely achieve active concentrations at targeterd sites.

Compounds in out data sets are screened for their inhibition of viral induced cytotoxicity using the human epithelial colorectal adencarcinoma cell line Caco-2 and a SARS-CoV-2 isolate obtained from an individual originally exposed to the virus in the Wuhan region of China.

We thus have to identify inhibitors of SARS-CoV-2 in-vitro cellular toxic-

ity in human(Caco-2) cells using a large scale drug repurposing collection for progression towards clinical studies against SARS-CoV-2 by predicing the efficacy of compounds.

Strategy

We implent a SVM based classifier, that essentially classfies compounds as active or inactive based on their inhibition values. Compounds with inhibition values greater than 75% are called as active and rest are termed as inactive. The model does a good job in classfying the compounds. Here are the Cross Validation results(r^2 scores) for train and test sets using 5-folds,

Cross Validation Train set : 0.94826457 0.94891945 0.94891945 0.94891945 0.94829843

Cross Vaidatoin Train set mean : 0.9486642700158061

Cross Validation Test set : 0.95026178 0.94764398 0.94764398 0.94764398 0.95013123

Cross Validation Test set mean : 0.9486649901746574

Conclusion

We observe that the SVM model does a great job in classifying compounds with a mean r^2 score around 0.95.

References

1. https://www.researchsquare.com/article/rs-23951/v1

2. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3524573/

3. https://www.ebi.ac.uk/chembl/document_report_card/CHEMBL4303101/

4. https://www.drugtimes.org/how-drugs-act/factors-affecting-bioactivity.html

5. https://www.researchgate.net/figure/Values-MW-clogP-HBA-HBD-PSA-logBB-and-logK
tbl1_259626002