Cancer Drug Effectivity Prediction

Overview

The efficacies of cancer drug treatments vary significantly among different patients, which poses a challenge for doctors to identify the most suitable or the optimal drug in different cases. We propose that a prediction of the most effective drug based on the RNA-seq profiles of tumor cells could address this challenge. Furthermore, this approach could also facilitate drug repurposing in the drug development and reusage.

Members of the project

Zhenyu Wu, Mingming Abundo, Palak Arora, and Kwok Wai (Ken) Ma

Data Sets

1. **RNA-seq data of cell lines, from CCLE database**

   RNA (ribonucleic acid) is crucial for gene expression as it acts as a messenger molecule, carrying instructions from DNA to the protein synthesis machinery of the cell, which determines the feature of cells. Different cells will have different transcriptome patterns (the complete set of RNA molecules). RNA sequencing is a powerful molecular biology technique used to analyze the presence and quantity of RNA in a biological sample.

2. **Drug IC50 data of cell lines, from GDSC**

   The IC50 (half maximal inhibitory concentration) of a drug is the concentration required to inhibit a specific biological or biochemical function by 50%. In this study, the IC50 value is a measure of the potency of a drug. And the drug resistance of cell lines could be inferred based on the IC50 data.

Datasets link:

https://www.cancerrxgene.org/gdsc1000/GDSC1000_WebResources/Home.html
https://depmap.org/portal/download/all/

In this project, we will combine these two data sets to build different machine learning models. Specifically, the RNA transcriptome patterns (the values depend on the type of tumors) are the features in the model. The efficacies of various drugs on treating different tumors, which have different RNA transcriptome patterns, are the targets that we need to predict.

Goals of the project

1. Design, build, train, and test different machine learning (ML) models to predict the efficacies of different cancer drugs on treating a given tumor, based on the RNA transcriptome pattern of that tumor.

2. With a given threshold, build machine learning models to predict whether the drugs can be effective or not to treat a given tumor with a known RNA transcriptome pattern.

3. Based on the above results (if the ML models can be generalized after cross-validation and testing), suggest the most or certain more effective drugs to treat different tumors.

4. Report the results in a systematic and comprehensive using tools and techniques from data visualization.

Stakeholders

Oncologists, Biologists, Pharmaceutical manufactories

Key Performance Indicators (KPI)

1. Root Mean Squared Error (RMSE) between true and predicted drug efficacies

2. Receiver operating characteristic (ROC) curve, the area under curve (AUC), and F1 score in classifying effective and ineffective drugs

Schedule

1. Data collection and cleaning

2. Exploratory data analysis and basic data visualization

3. Feature selection, feature engineering, possible PCA

4. ML model design (Possibilities: Regression, Classification, Ensemble Learning, Neural Network, Pretrained model etc)

5. Model training, hyperparameter fine tuning, model optimization

6. Results visualization and presentation