432 Proposal: Predicting Drug Sensitivity of Cancer Cell Lines

Lam Vo - NetID: lamvo2

11/20/2018

Motivation

As sequencing technologies make it possible to collect vast amount of data about the cellular environment in complex diseases, the use of statistical learning techniques to integrate and analyze these data sources and guide personalized treatment decisions has attracted great interest from the scientific community. One exciting open challenge is how to accurately predict the potency of drugs on cancer tumors given the tumor cells' omics profiles (such as genomics, transcriptomics, or proteomics).

Different kinds of features representing both the drug and the cancer cells have been used to trained predictive models to address this challenge. While mRNA levels and mutation profiles are popular choices for representing cancer cells, I believe protein levels offer a more accurate depiction of the cellular environment than the two mentioned data sources. Drug molecules have often been represented by structural features (e.g. size and topology) and physical properties (e.g molecular weight, polarity, and lipophilicity). However, the link between these features and the drugs' effects on the cellular environment is rarely clear.

A new method for profiling small molecules' bioactivities is morphological profiling. In this method, a library of compounds are given to a standard cancer cell line, and changes in the cells' morphology (e.g. size and shape of the cell, location of the nuclei and organelles as well as their fluorescent intensities upon staining with fluorescent dyes) after the treatment are measured from microscopic images. The molecules' effects on the cellular environment have been shown to manifest in these changes in morphology, and thus I believe they could be used as effective descriptors and predictors for the molecules' bioactivities.

Analysis Task

My project aims to train a regression model that can take as input a query drug's morphological profile and a query cell line's protein levels (measured by reverse phase protein arrays) and predict the drug's potency on the cancer cell line as measured by area-under-percent-viability-curve (AUC) of that drug-cell line pair. The definition of AUC is illustrated in plot B in Figure 1 below. The horizontal axis is the drug's concentration and the vertical axis is the percent of cells from a population of the cell line killed by the drug. AUC is the red area in the plot. *A high AUC means the drug achieves high killing percentage at a low dose, or in other words, the drug is more potent on the cell line in the drug-cell line pair being investigated.*

If time allows, I will also train another model using the same drugs' structural and chemical features for comparison of predictive power between these kind of features and morphological profiles.

Datasets and Data Preprocessing

The drugs' morphological profiles were extracted from GigaDB at this link and removed missing and duplicate data points. For every row, the first column is the drug' master ID in the CTRP database, and the remaining columns contain measurements for morphological properties of U2OS (an osteosarcoma cell line) cells upon treatment with that drug. There are 80 drugs and 1532 morphological properties.

```
drugs = read.csv('drugs.csv')
head(drugs[, 1:3])
```

```
##   Master.ID Cells_AreaShape_Area Cells_AreaShape_Center_X
## 1    417262           0.10068269              -0.01068128
## 2    415688           0.06699153               0.05301241
## 3     27894          -0.22886003               0.02317770
## 4     50715          16.94456642               0.40456857
```

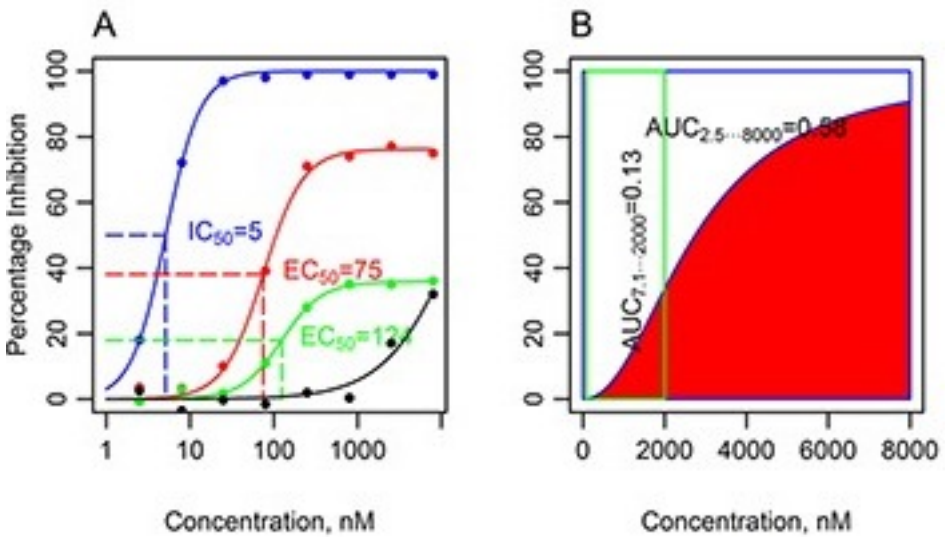

Figure 1: AUC definition

```
## 5    414479           0.22325481                0.02705412
## 6     25393           0.26196895                0.03207342
```

```
dim(drugs)
```

```
## [1]   80 1533
```

```
sum(is.na(drugs)) # check for missing data points
```

```
## [1] 0
```

```
sum(duplicated(drugs)) # check for duplicate data points.
```

```
## [1] 0
```

The cell lines' protein levels data were downloaded from the Cancer Cell Line Encyclopedia CCLE and removed missing and duplicate data points. For every row, the first column is the name of the cell line and the subsequent columns are the names of the proteins whose expression levels are being measured. There are 899 cell lines and 214 proteins.

```
cellProteins = read.csv('cellProteins.csv')
head(cellProteins[, 1:3])
```

```
##   Cell.Line X14.3.3_beta X14.3.3_epsilon_Caution
## 1     DMS53    -0.104888                0.060414
## 2    SW1116     0.358504               -0.180291
## 3  NCIH1694     0.028738                0.071902
## 4     P3HR1     0.120039               -0.066802
## 5     HUT78    -0.268997               -0.060281
## 6     UMUC3    -0.171170                0.055813
```

```
dim(cellProteins)
```

```
## [1] 899 215
```

```
sum(is.na(cellProteins)) # check for missing data points
```

```
## [1] 0
```

```r
sum(duplicated(cellProteins)) # check for duplicate data points.
```

```
## [1] 0
```

The data on drugs' potency on cell lines were downloaded from the Cancer Therapeutics Response Portal CTRP and removed missing and duplicate data points. For every row, the first column contains AUC value, the second column is the drug's master ID in the CTRP database, and the third column is the name of the cell line treated with the drug. There are 54805 observations.

```r
AUCs = read.csv('AUCs.csv')
head(AUCs)
```

```
##      AUC Master.ID Cell.Line
## 1 13.390     23256      CAS1
## 2 14.385     25036      CAS1
## 3 15.570     26870      CAS1
## 4 13.510     26914      CAS1
## 5 13.149     26972      CAS1
## 6 12.430     27894      CAS1
```

```r
dim(AUCs)
```

```
## [1] 54805     3
```

```r
sum(is.na(AUCs)) # check for missing data points
```

```
## [1] 0
```

```r
sum(duplicated(AUCs)) # check for duplicate data points.
```

```
## [1] 0
```

I have already preprocessed and cleaned the data frames after downloading them from online sources. All 3 do not contain any null values or duplicate rows (as shown aboved).

I have merged these 3 data frames into one big one. This data frame contains 49536 drug-cell line pairs, the identities of the drug and the cell line in each pair, their corresponding AUC values, and features of both the drugs and the cell lines.

```r
big = merge(AUCs, drugs, all = FALSE)
big = merge(big, cellProteins, all = FALSE)
dim(big)
```

```
## [1] 49536  1749
```

```r
head(big[, 1:5])
```

```
##   Cell.Line Master.ID    AUC Cells_AreaShape_Area Cells_AreaShape_Center_X
## 1     22RV1    438691 14.447            0.4965107             -0.013010540
## 2     22RV1    447732 14.436           -0.1705984              0.000462578
## 3     22RV1    375596 14.569            0.9539691              0.014277678
## 4     22RV1    375354 14.439            0.3862486              0.029696109
## 5     22RV1     48589 11.386            5.6386064              0.024269823
## 6     22RV1    442141 14.320            0.3456213             -0.031117445
```

```r
sum(is.na(big))
```

```
## [1] 0
```

```r
sum(duplicated(big))
```

```
## [1] 0
```

Method

I plan to employ the following regression algorithms:

- Elastic Net
- Random Forest
- Deep Neural Network

Train-test split and k-fold cross-validation will be used to compare the performance of these algorithms. I plan to use Root Mean Square Error (RMSE) and R2 score as performance metrics.

Challenges

The merged data frame is pretty big (over 1 GB in size), so computational power and time would pose a challenge. Another challenge is the large number of features and features being highly correlated. Extensive feature selection and feature engineering will be needed for best predictive performance. Dissecting the models afterward (if they do perform well enough) to extract insights into factors affecting cancer cell lines' drug sensitivity would be interesting but challenging as well.