MEDICATION FOR CANCER CELL LINES

[1]S.Shobana, [2] M.Ramya, [3] N.C.Swathi, [4]Ms.Meena
[1, 2, 3]Student, [4]Professor
[1, 2, 3,4]Computer Science and Engineering,
[1, 2, 3, 4] Prathyusha Engineering College, India

Abstract- There is several hundred panels of human cancer cell lines and number of anticancer drugs available. But predicting the most effective anticancer drug for specific cancer cell line is still remains challenge. Analyzing each anticancer drug response with cancer cell lines is manually difficult. It takes more time to test each anticancer drug with cancer affected cell lines to predict the effective anticancer drug. So, our project helps to identify the most suitable drug for affected cell lines. The cancer cell lines are characterized with genomic and pharmacological data. An analyses is done with these pharmacogenomics data and the most effective anticancer drug response for cancer cell lines were predicted by identifying the IC50 (Concentration of Inhibitor) and AUC (Concentration of Drug) for the cell lines. From these values the suitable anti-cancer drug can be selected and that can be prescribed in future.

Keywords- **Data collection, Cell line, anticancer drug, pharmacoGX, Data analysis, IC50, AUC**

INTRODUCTION

Cancer genome analysis and drug prediction is a concept of data analysis to predict effective anticancer drug response for cancer affected cell lines. Analyzing millions of cancer cell lines with anticancer drug and predicting effective anticancer drug are not so easy. And manually it takes more effort, so we made an analysis with data available for cancer cell lines and anticancer drug response using data mining algorithms to predict effective anticancer drug response. Using these algorithms we can analyze those data in pharmacoGX framework and predict the anticancer drug.

EXISTING SYSTEM

The system uses NextGen sequencing approach and it is used to study the data in micro level. Genetic level variation is compared for various cancer types using the ATGC sequence. The DNA sequence is analyzed for the prediction of GC content variation that is to reveal the cases of horizontal transfer or reveal biases in mutation and the Rho value calculation to measure how over-represented or under-represented, a particular DNA word is, is done for different types of cancer. And the observation is made which leads to the next step of the study which includes the applying of IC50 values on the drugs used for treating several types of cancer that are read from the GDSC-CCLE database. Then a graph is plotted for each cancer type from the IC50 values of different cancer drugs which may assist in prescribing drugs for the individuals.

DISADVANTAGES OF EXISTING SYSTEM

- Infectious cause of other cancers could not be identified because infection status was recorded for some cancer types.
- And therefore in this system, there are limited numbers of genome sequences and fewer amounts of data sets to be compared.

PROPOSED SYSTEM

In Proposed system, we use pharmacoGX library function to access and retrieve large publicly available datasets for all cell types. This involves analyzing large set of pharmacogenomics objects to identify drug sensitivity for each cell type. Then gene expression is estimated to find the similar cell lines. This estimation is done to find drug sensitivity on cell lines because as we know similar cells have similar drug sensitivity. Then IC50 and AUC values were computed to find the suitable drug with an IC50 of value less than 1. Finally gene-drug are associated and most effective drug for the cell line is estimated.

ADVANTAGES OF PROPOSED SYSTEM

- Effective genetic diagnosis for all tumour types.
- Large amount of pharmacogenomics data are downloaded easily by using pharmacoGX framework available in R tool.
- The graphical representation makes better understanding of the result analysis.

METHODOLOGY USED

There are certain important modules used for these analyses they are as follows:

a) Data Collection
b) Estimation of drug sensitivity
c) Estimation of Gene Expression
d) Gene- drug association modeling

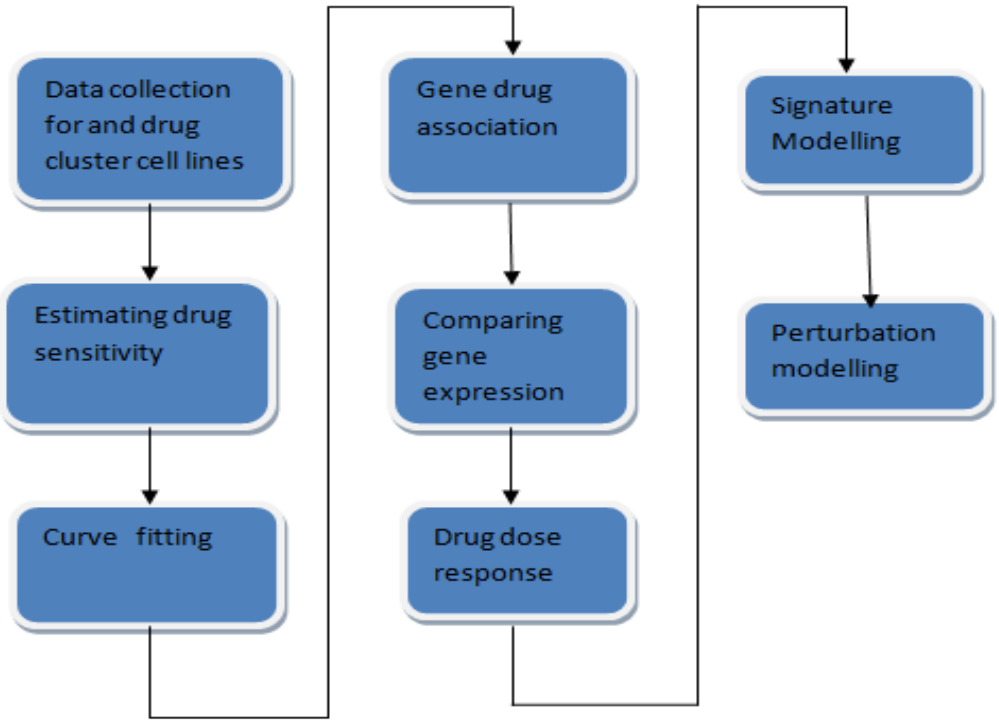

Figure.1: The flow of the project

Data Collection

The PharmacoSet objects of the publicly available datasets can be downloaded using functions. A table of available PharmacoSet objects can be obtained by using the availablePSets function. Any of the PharmacoSets can then be downloaded by calling downloadPSet, which saves the datasets into a user directory, and returns the data into the R session.

a. Installation and settings

PharmacoGX requires that several packages to be installed.

> Source ('http: //bioconductor.org/biocLite.R')

> BiocLite ('PharmacoGX')

> Library (PharmacoGX)

b. Downloading

> availablePSets ()

> GDSC <- downloadPSet ("GDSC")

ESTIMATING DRUG SENSITIVITY

In datasets, pharmacoGX includes function to Analyses estimated AUC (Area Under drug response Curve) and Ic50 values from drug dose response to measure cell viability after varying concentration of drugs.

a. Curve fitting:

In curve fitting, Ic50 value is identifying by fitting the curve to data using logLogisticRegression and drugDoseResponse function.

b. Drug response:

In drug response, a lower value of Ic50 indicates a better sensitivity of a cell lines to a given drug. Hence a curve is generated for the cell line with its IC50 and AUC values.

COMPARING GENE EXPRESSION

Here, to find the similar cell lines between two datasets the consistency of the data must be identified. The common intersection between the datasets can be found using intersectPSet. We create a summary of the gene expression drug sensitivity measures for both datasets, so we have one gene expression profile and one sensitivity profile per cell line within each dataset. Then we compare the gene expression and sensitivity measures between the datasets using a standard correlation coefficient.

GENE-DRUG ASSOCIATION MODELLING

PharmacoGX provides methods to model the association between drugs and molecular data such as genomics and proteomics (study of protein and their function). Sensitivity studies allows the discovery of molecular features that improve or inhibit the sensitivity of cell lines to various compounds, by looking at the association between the expression of the feature and the response towards each compound.

a. PERTURBATION MODELLING

1. PERTURBATION

Perturbation studies on the other hand look at the molecular profiles of a cell before and after application of a drug, and therefore can characterize the action of a drug on the molecular level. It is hypothesized that drugs which act to down-regulate expression of known disease biomarkers would be effective in reversing the cell from a diseased to healthy state. The function drugPerturbationSig models the molecular profiles of drugs tested in a perturbation dataset.

2. PERTURBATION MODELLING

The molecular response of a given drug is modeled as a linear regression model by adjusting experimental batch effects, cell specific differences, and duration of experiment to isolate the effect of the concentration of the drug applied:

i. $G = \beta 0 + \beta iCi + \beta tT + \beta dD + \beta bB$

Where G denotes the molecular feature expression (gene), Ci denotes the concentration of the compound applied, T the cell line identity, D denotes the duration of the experiment, B denotes the experimental batch, and βs are the regression coefficients, βi represents strength of feature response.

b. SENSITIVITY MODELLING

The association between molecular features and response to a given drug is modeled using a linear regression model adjusted for tissue source:

$$Y = \beta 0 + \beta iGi + \beta tT + \beta bB$$

Where Y denotes the drug sensitivity variable, Gi, T and B denotes the expression of gene, the tissue source and the experimental batch respectively, and βs are the regression coefficients, βi strength of gene-drug association. The variables Y and G are scaled (standard deviation equals to 1) to estimate standardized coefficients from the linear model.

MODULE IMPLEMENTATION

Dataset collection using downloadPset () function

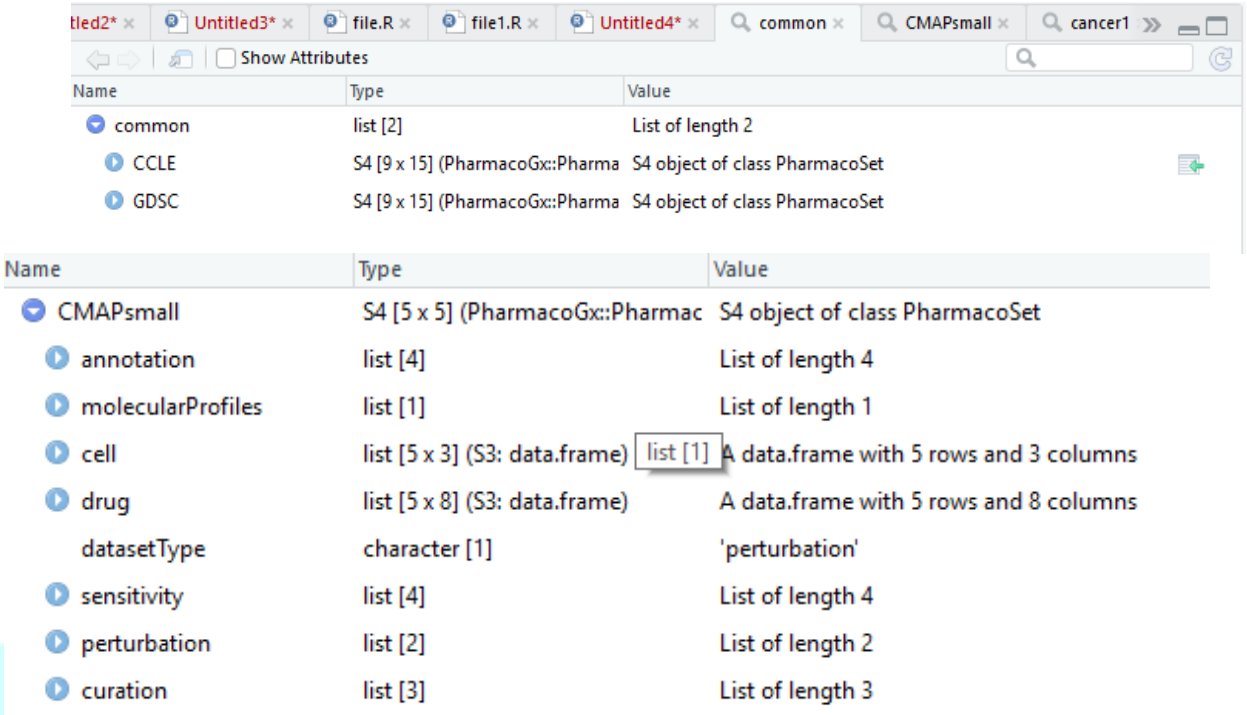

Figure.2: Screenshot of the datasets collected in R tool

Estimating drug sensitivity by calculating IC50 and AUC values

	22RV1	23132-87	5637	639-V	647-V	697	769-P	786-0	8-MG-BA
PD-0325901	4.707269	NA	15.02965125	0.01045464	NA	2.091583e+00	1.1957565	0.4173959	2.53238108
17-AAG	3.491684	NA	0.05740293	0.19312821	NA	1.570437e+00	0.2127999	0.1538108	0.09472408
Nilotinib	155.269917	NA	219.93454954	92.17712520	NA	3.063552e+00	19.6335139	137.0668825	NA
PHA-665752	NA	NA	NA	NA	NA	2.876326e+01	NA	NA	193.21006177
lapatinib	NA	NA	NA	NA	NA	1.321992e+02	NA	NA	104.77487492
Nutlin-3	12.805900	NA	609.62438182	43.79602238	NA	2.021191e+00	66.6517965	100.9576377	387.46053560
AZD0530	NA	NA	NA	NA	NA	1.327007e+01	NA	NA	80.05644913
Crizotinib	NA	NA	NA	NA	NA	9.243291e+00	NA	NA	166.33353176
Sorafenib	NA	NA	NA	NA	NA	7.059894e+00	NA	NA	136.71958453
PD-0332991	1.678425	NA	360.36953821	21.21541415	NA	5.352074e-02	4.8592641	2.1543908	NA
paclitaxel	NA	NA	NA	NA	NA	1.495096e-03	NA	NA	0.00795841
AZD6244	63.866949	NA	153.90261004	53.64262529	NA	2.608449e+02	3.7425979	10.4166033	80.15778462
PLX4720	810.598610	NA	20.04785168	912.80657351	NA	4.574032e+00	104.6663844	NA	6.18129008
TAE684	NA	NA	NA	NA	NA	4.321481e-01	NA	NA	2.65414349
Erlotinib	NA	NA	NA	NA	NA	2.653053e+00	NA	NA	79.60031101

Figure 3: Screenshot depicting the value of IC50

Untitled4* × | common × | GDSCexpression × | CMAPsmall × | cancer1 × | CCLE.auc × | CCLE.ic5

Filter

	22RV1	23132-87	5637	639-V	647-V	697	769-P	786-0	8-MG-BA
17-AAG	0.3724600	NA	0.4828250	0.4877500	NA	0.3420875	0.40937500	0.4416250	0.4426500
PD-0325901	0.3850000	NA	0.1708125	0.3088750	NA	0.2291000	0.23250000	0.1386250	0.0758750
AZD6244	0.3551250	NA	0.0647625	0.1862000	NA	0.2122500	0.15602250	0.0839625	0.0000000
Nilotinib	0.0000000	NA	0.0072625	0.0710125	NA	0.1573375	0.00000000	0.0750125	NA
Nutlin-3	0.0769250	NA	0.0000000	0.0666625	NA	0.2588125	0.08598571	0.0000000	0.0655125
PD-0332991	0.0416250	NA	0.0000000	0.0849250	NA	0.2056000	0.12012500	0.0000000	NA
Crizotinib	NA	NA	NA	NA	NA	0.1758375	NA	NA	0.0000000
Sorafenib	NA	NA	NA	NA	NA	0.1008500	NA	NA	0.0000000
lapatinib	NA	NA	NA	NA	NA	0.0453250	NA	NA	0.0547625
Erlotinib	NA	NA	NA	NA	NA	0.0826500	NA	NA	0.0803750
PLX4720	0.1263625	NA	0.1212000	0.0000000	NA	0.1128875	0.03248750	NA	0.1074875
TAE684	NA	NA	NA	NA	NA	0.3243750	NA	NA	0.1091250
AZD0530	NA	NA	NA	NA	NA	0.1011375	NA	NA	0.2001375
PHA-665752	NA	NA	NA	NA	NA	0.1609750	NA	NA	0.3962500
paclitaxel	NA	NA	NA	NA	NA	0.9725000	NA	NA	0.8787500

Figure 4: Screenshot depicting the value of AUC

Depicting those values as a curve

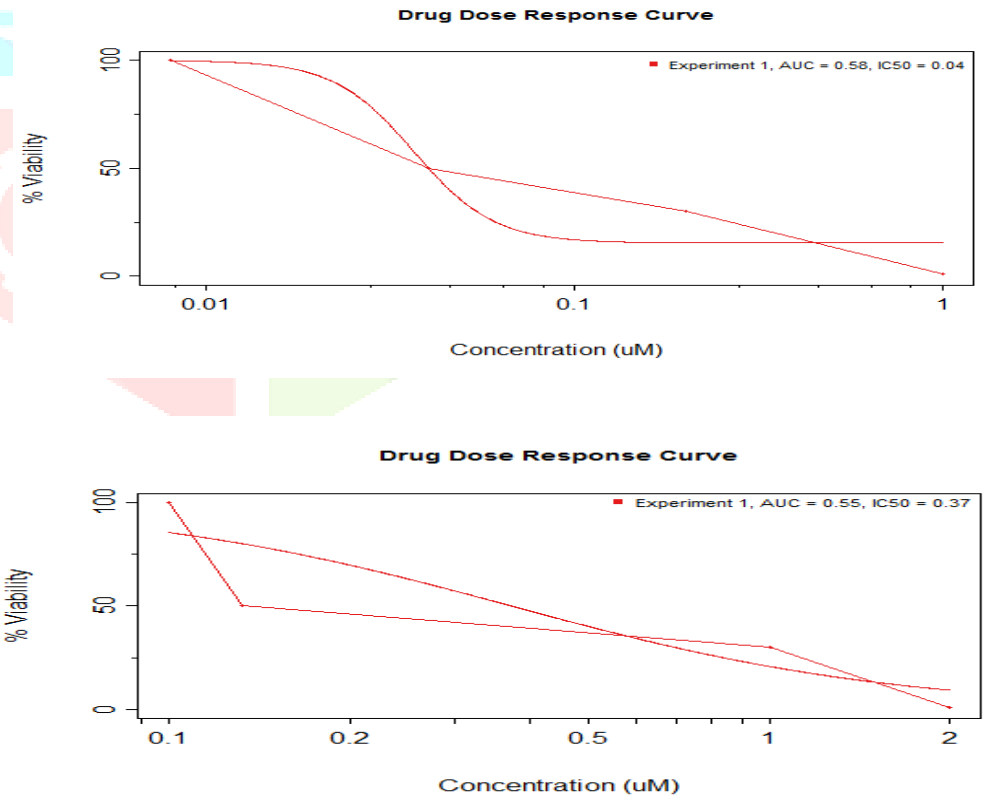

Figure 5: Graph drawn for IC50 and AUC

Finding similarity between datasets is done by comparing gene expression, IC50 and AUC values for all cell types.

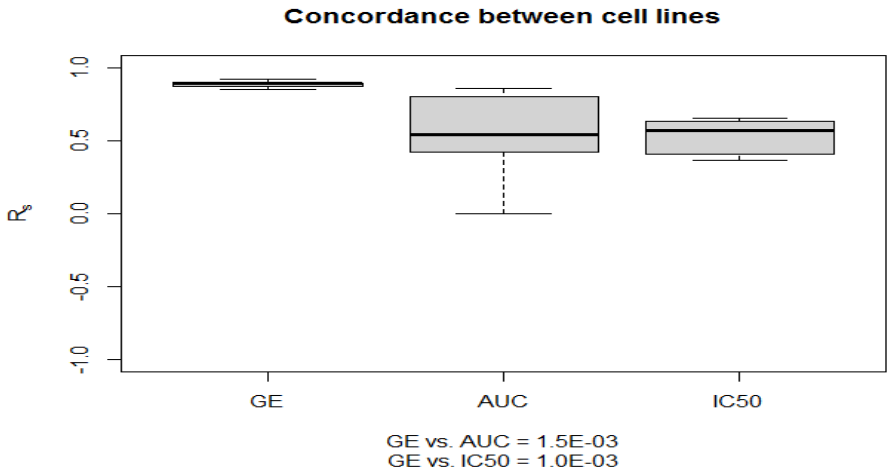

Figure 6: Cell Similarity

CONCLUSION

The pharmacoGX package enables easy and efficient analysis of the increasingly available compendium of pharmacogenomics data. This package is the first to integrate multiple pharmacogenomics datasets using structured objects incorporating standardization of cell line and drug identifier. The study on cancer cells and suggestion of drugs for the patients is a huge and complex process since it also comprises best out of bioinformatics and big data methodologies. And so whether the drug is effective or an inhibitor can be identified using AUC and IC50 values. Therefore, by this analysis, we can easily find out the affected cell lines and also we can find the efficient drug for such kinds of cancer by comparing gene expression.

ACKNOWLEDGEMENT

References

1. Wang L, Zhang L, Gao Q“Improved anticancer drug response prediction in cell lines using matrix factorization with similarity regularization”BMC Cancer are provided here courtesy of BioMed Central, February 2017.

2. Dong Z, Zhang N, Li C, Wang H “Anticancer drug sensitivity prediction in cell lines from baseline gene expression through recursive feature selection” BMC Cancer are provided here courtesy of BioMed Central, June 30, 2015.

3. Smirnov P, Safikhani Z “PharmacoGX: an R package for analysis of large pharmacogenomics datasets” Oxford University press, December 2015.

4. Menden MP, Iorio F, Garnett M, McDermott U “Machine Learning Prediction of Cancer Cell Sensitivity to Drugs Based on Genomic and Chemical Properties” PLoS ONEare provided here courtesy of Public Library of Science, April 30, 2013.

5.Barretina J, Caponigro G, Stransky N “The Cancer Cell Line Encyclopedia enables predictive modeling of anticancer drug sensitivity” IEEE Rev Biomed Eng, March 29, 2012.