Predicting Drug Therapy Response in Cancer Patients

Project Deliverable 3: Final Report

Rebecca Sarto Basso, Emma Besier, India Bergeland
Claire Dubin, Oskar Radermecker, Nikhil Yerasi

Data-X
Spring 2019

Contents

1. Introduction

1.1 Objective

The oncology pharmaceutics and genome sequencing industries are booming, and a significant challenge in modern medicine is analyzing the influx of data from these sources. Over one thousand cancer drugs are currently in development, and patient genomic data is becoming readily accessible to physicians. There is little room for trial and error when treating cancer, and it is imperative that physicians make informed decisions when prescribing drugs. By combining drug response data and genomic information, physicians can make informed treatment decisions based on the unique biology of each patient — a practice known as personalized medicine.

In this project, multiple methods of feature selection (PCA, Lasso, RFE, RF, Netphlix) were used to identify important markers in mutation data that affect drug therapy response. Machine learning methods (SVM, Logistic Regression, MLP, Random Forest) were then applied to train classifiers used to predict the sensitivity of drugs in cancerous cell lines.

Our final deliverable is a web application that makes this tool available to physicians. Clinicians can upload their patient's mutation data directly into the application, and will be provided with a list of drugs that the patient will likely be sensitive to. This is of critical importance in a world where more and more anti-cancer drugs are being developed, and medical doctors face the ever-increasing challenge of prescribing the right drug that will work best for an individual patient.

2. Materials and Methods

2.1 Dataset Description and Data Cleaning

This work involves the use of two main datasets, one that includes drug therapy responses and the other that describes the genetic mutations in a number of cell lines.

2.1.1 GDSC

The Genomics of Drug Sensitivity in Cancer (GDSC) (Benes et al., 2012) provides the drug response of 1065 cell lines and 251 drugs.

Cell lines are identified by a unique "Cosmic ID" and drug response is quantified using the **AUC** or area under the dose-response curve. It is obtained by fitting the dose-response curve which plots the activity of the cell population as a function of the drug concentration. The response of the cells can be assessed by the metabolic activity of the cells. The smaller the AUC, the less responsive and the more resistant that cell line is to the drug. Thus, small AUC values indicate that a higher concentration of the drug is needed to kill the cell line.

2.1.2 CCLE

The other dataset comes from the Cancer Cell Line Encyclopedia (CCLE). It contains the mutations of 1457 cell lines for 84,434 different genes. After basic preprocessing, we have a dataframe where each row corresponds to a cell line and a 0-1 vector which represents the presence or lack of a mutation in that cell population. Cell lines are identified by a unique "CCLE Name".

2.1.3 Merging the Data

Merging the data was one of the biggest challenges of this project. Most of the cell lines are common between the two datasets, but their names are different. Finding a conversion matrix which would efficiently cover the overlap proved to be more difficult than expected. In order to combat the initial conversion loss of 60% of all cell lines, we decided to merge the information contained in three different subsets of the GDSC dataset. The first one was an older version of the GDSC dataset on which an efficient conversion had been performed. The second one was an updated version of the GDSC dataset, but for which conversion tables were limited. The last one was obtained from the Depmap Project [source: Aviad Tsherniak et al., Defining a Cancer Dependency Map. Cell July 27, 2017. DOI: j.cell.2017.06.010]. The Depmap Project aims to map cell lines for different data types, including mutations and drug response data, to enable scientists to run joint analyses. By merging all of that data together, we managed to reduce the cell loss to 30%.

We then looked at the distribution of AUC values (Figure 2.1) and categorized the drug sensitivity into three classes: "resistant", "medium" and "sensitive".

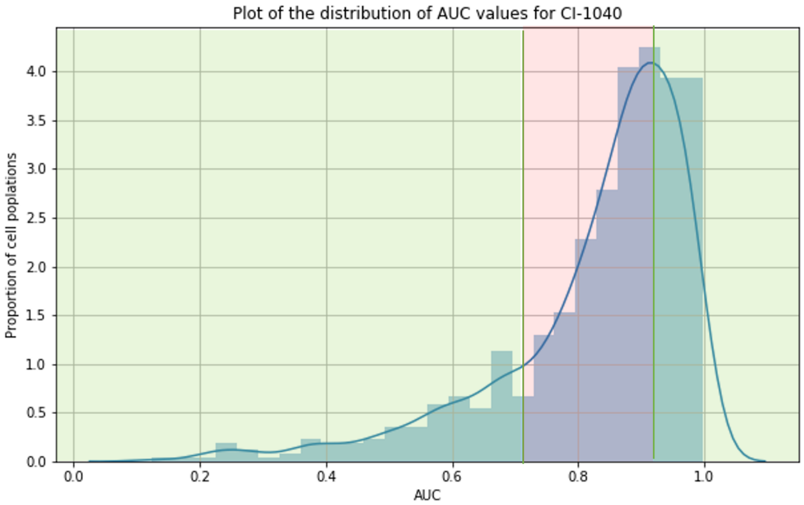
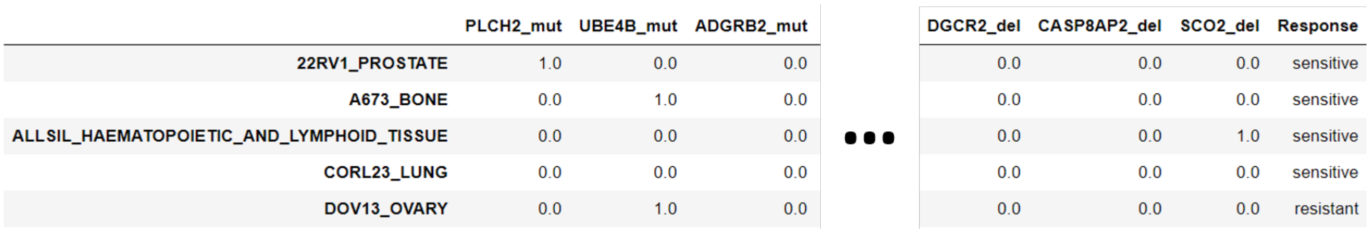


Figure 2.1: Distribution of AUC values and thresholds for drug CI-1040. The cell lines located in the red zone were discarded; the ones in the green zones were classified as either "resistant" or "sensitive".

Similar to results found in the literature Jang (2014), the lower and upper quartiles were used as thresholds. The cells in the lower quartile were labeled as resistant, whereas cell lines in the upper quartiles were labeled as sensitive. Everything in the middle was discarded. Depending on individual drugs, we were then left with around 450 cell lines. This dataframe was then use as the input to feature reduction methods and machine learning models. A screenshot of the final dataframe is displayed in Figure 2.2.

All missing values were discarded due to the biological aspect which does not allow us to make up data from a statistical standpoint.



	PLCH2_mut	UBE4B_mut	ADGRB2_mut		DGCR2_del	CASP8AP2_del	SCO2_del	Response
22RV1_PROSTATE	1.0	0.0	0.0		0.0	0.0	0.0	sensitive
A673_BONE	0.0	1.0	0.0		0.0	0.0	0.0	sensitive
ALLSIL_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE	0.0	0.0	0.0	•••	0.0	0.0	1.0	sensitive
CORL23_LUNG	0.0	0.0	0.0		0.0	0.0	0.0	sensitive
DOV13_OVARY	0.0	1.0	0.0		0.0	0.0	0.0	resistant

Figure 2.2: Screenshot of the final matrix.

Because of the way the two classes were subdivided, the final baseline accuracy is 50%.

2.2 Feature Reduction

As described above, our data set had a large amount of features as compared to the number of samples. We implemented and investigated several feature reduction techniques to decrease the number of features. To avoid data leakage, we computed all supervised selection methods on the training set first. The same transformations were then blindly applied on the test set.

First, a rough thresholding on the variance for each feature was performed. All the features with a variance lower than 0.1 were discarded. This allowed us to simplify our training and test sets by discarding genetic mutations which were only present in a small subset of our population. This method decreased the number of features from 64,000+ to ∼450.

Second, we implemented five more complex statistical-based feature reduction methods:

1. Principal Component Analysis (PCA)
2. Random Forest-Based Feature Selection
3. LASSO
4. Recursive Feature Elimination (RFE)

Because of the lack of high accuracies, another novel feature selection algorithm was used. Netphix, contrary to the other feature selection methods, takes advantage of both statistical and biological aspects in order to select the most informative features (Kim et al. (2019)). By combining the final matrix with a protein to protein interaction network (PPI), Netphix applies integer linear programming to select mutations that are on the same cellular pathway and associated with a good or bad drug response. As for PCA, this method was applied on the original training set, without the initial variance-based feature reduction.

Using the Netphlix algoritm to identify important features in a machine learning problem is an approach that has never been done in the literature before. Thus, promising results could potentially lead to a paper and an advance in the field.

2.3 Selected Models

After reducing the number of features using each of the techniques described above, four statistical methods were applied on the data to predict the sensitivity to a given drug: logistic regression, random forest, support vector machine and multi-layer perceptron.

3. Model Performances & Discussion

3.1 Performance Evaluation

The above feature selection and machine learning techniques were applied to a total of six drugs, chosen because of their known association between patient outcome and genetic mutations (Kim et al., 2019). Overwhelmingly, all drugs performed better when features were selected based on biological significance.

Amongst all model combinations, the best performing drug was CI-1040 using Netphix for feature selection. There was no significant difference in the performance between statistical machine learning models (LR, MLP, RF, SVM). The results for C1-1040 are shown below in Figure 3.1.

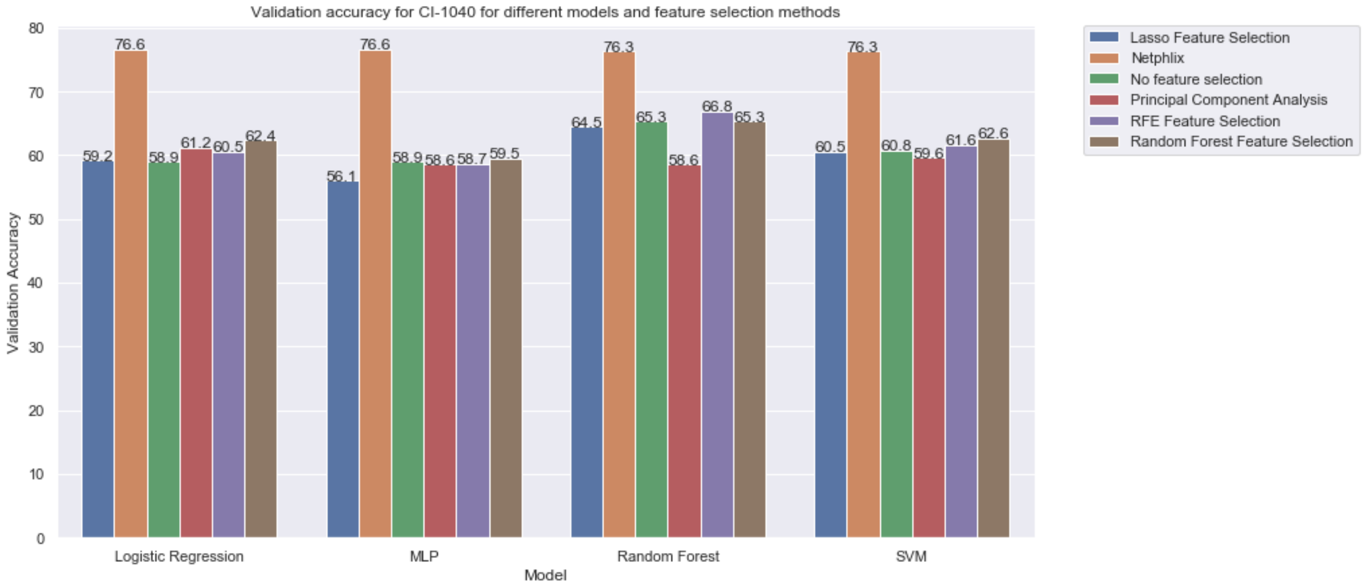
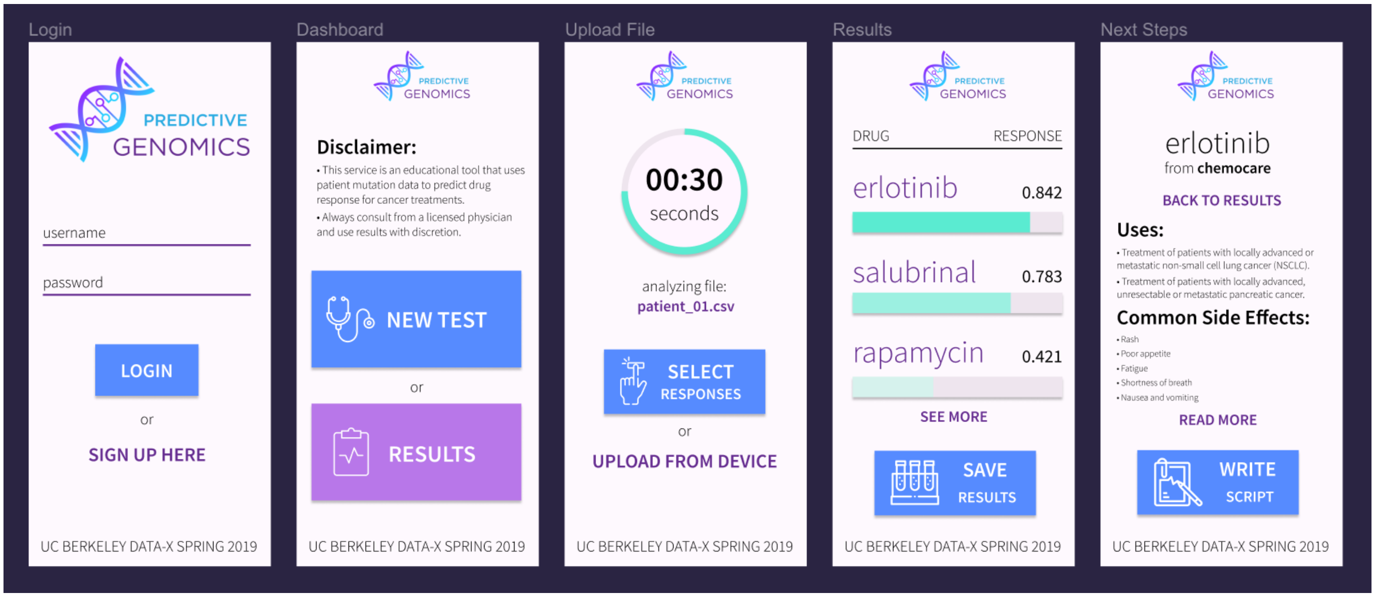


Figure 3.1: Final validation accuracies for four different machine learning algorithms and five different feature selection methods. Models were run with a 5-fold cross validation.

These results make it clear that the method of feature selection is imperative in model performance. After incorporating the Netphlix algorithm in our approach to solving this problem, we were able to achieve test accuracies close to or on-par with the "industry standard".

4. User Interface

After developing the machine learning model, our group built out the user interface. The application requirements include reading in new patient data, using our trained models to predict drug responses, displaying top drug matches succinctly and clearly, and providing next steps for the user.



Disclaimer:

- This service is an educational tool that uses patient mutation data to predict drug response for cancer treatments.
- Always consult from a licensed physician and use results with discretion.

analyzing file: patient_01.csv

DRUG	RESPONSE
erlotinib	0.842
salubrinal	0.783
rapamycin	0.421

erlotinib
from **chemocare**

Uses:

- Treatment of patients with locally advanced or metastatic non-small cell lung cancer (NSCLC).
- Treatment of patients with locally advanced, unresectable or metastatic pancreatic cancer.

Common Side Effects:

- Rash
- Poor appetite
- Fatigue
- Shortness of breath
- Nausea and vomiting

Figure 4.1: Screenshot of Wireframe in Figma

The UI/UX is simple and clear (see Figure 4.1), providing actionable insights for treatment methods and a consistent design. The application is geared towards healthcare providers and assumes a basic understanding of the background context by the physician.

Users begin by authenticating or signing up for an account. Our Firebase backend handles both this and storage of patient data and model results. Users can then choose between predicting drug responses for new patient data or viewing results from prior sessions, organized by patient. Once authenticated, users upload a .csv file containing their patient's mutation data for analysis. Automatic uploads prevent the risk of human error and allows for input of patient data with high dimensionality. The .csv is parsed into a DataFrame and processed by our pre-trained model, which selects features to analyze from the new data and returns recommended drugs.

In the future, we plan to include information about the drugs chosen for the patient as well, much of which we will scrape from reputable and complete sources like Chemocare.

5. Conclusion

In this project, multiple feature selection and machine learning methods were used to predict drug sensitivity in cancer patients. Our results show that methods of feature selection have a much higher impact on the accuracy of the results than the specific machine learning algorithm used.

We were able to harness a recently published feature selection method ("Netphlix") to achieve test accuracies up to 76.6%. This is a notable result, given that this algorithm has not been used in scientific literature of this nature before. Our highest performing models were then integrated into a user interface that allows clinicians to predict drug response for new patients using our trained models. All code is available on GitHub.

Bibliography

Benes, C., D. A. Haber, D. Beare, E. J. Edelman, H. Lightfoot, I. R. Thompson, J. A. Smith, J. Soares, M. R. Stratton, N. Bindal, P. A. Futreal, P. Greninger, S. Forbes, S. Ramaswamy, W. Yang, U. McDermott, and M. J. Garnett (2012, 11). Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Research 41*(D1), D955–D961.

Jang, Sock, e. a. (2014). Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data. *Pacific Symposium on Biocomputing*.

Kim, Y.-A., R. Sarto Basso, D. Wojtowicz, D. S. Hochbaum, F. Vandin, and T. M. Przytycka (2019). Identifying drug sensitivity subnetworks with netphlix. *bioRxiv*.