A Sparse Bradley-Terry Model for Drug Sensitivity Ranking in Precision Medicine

Phillip Si

November 7, 2017

Abstract

Project Title: A Sparse Bradley-Terry Model for Drug Sensitivity Ranking in Precision Medicine With the rapid development of genomic sequencing technology, precision medicine by incorporating personalized genetic variations is considered an emerging approach for cancer treatment; however, its advantages and drawbacks are largely debated, including the cost and effort needed to sustain such a complicated progress. Among the various challenges, the highly noisy cell line measurements of high dimensionality impose a grand hurdle to overcome for data scientists. How can we identify those drugs which are highly specific to genetic mutations from the widely used ones, given the noisy data?

With the creation of the extensive drug-sensitivity database by Iorio et al. (Cell 166.3 (2016): 740-754.) using 990 cell lines and 265 drugs, we provide a new answer to this question using a statistical model, namely a sparse Bradley-Terry model. To deal with the high noise in IC50 values of sensitivity measurements, the original cell line vs. drug sensitivity data are converted into pairwise comparisons between drugs which are robust to the noise. Then a Bradley-Terry model is learned to capture both the common ranking of all drugs based on their average effect on the whole data, and also the personalized deviations based on specific genetic profiles and cancer types. In this model, we assume that the personalized deviations are sparse enforced by l_1-norm penalty to avoid overfitting and it can thus be trained by logistic regression with elastic net using R package `glmnet`. To our surprise, such a simple model achieves a 95.2% accuracy on the test data, almost as good as the best model using a complicated neural network approach with data enhancement

(95.8%), yet it also has with superb *interpretability*. Among various discoveries, as a confirmative result, it discloses that drugs sensitive to EGFR mutation show positive efficacy for the treatment of lung cancer (LUAD), but resistance to brain lower grade glioma (LGG), which is a hot direction in current research. Moreover, some widely used medications such as Breomycin discovered in 1962 show a diverse response to different gene mutations, which deserves future studies. These results show that the proposed methodology adds a valuable tool for precision medicine modeling with large scale data on cancer cell lines.

1 Introduction

Cancer is currently an important part of the biological research community. Researchers have come up with many ways of testing the efficacy of each drug upon cancer, but medicine for cancer isn't a one-glove-fits-all solution. Depending on the different types of cancer cell lines, different approaches may be needed. In fact, there may be new cancer cell lines springing up, causing many more tests to be needed to save patients. It is an ever-evolving problem which we are doing our best to solve. *Precision medicine* based on high throughput genetic profiles of patients, also known as *personalized medicine*, is one of the routes that doctors can take to help patients fight back against malignant tumors. Precision medicine could reduce the use of traditional chemotherapy by about 25%, and can help save more lives. However, precision medicine is currently too expensive for most people, as the medicine usually contains biologics, drugs made from natural sources, instead of the chemically synthesized medicine that is widely used today. In addition, development of specific medicine for each patient wastes too much resources if the targeted client population is

too small due to the vast number of trials needed to test the efficiency of the drug upon the cells. However, the ways to approach the solution is not only through biological approaches. Utilizing mathematical and statistical methods of analyzing the data, we can enhance the interpretation of cancer cell lines and drug components.

Garnett et al published an article in 2012 regarding the collection of sensitivity data and the analysis using a Bayesian sigmoid model using elastic net. They analyzed which mutations could be used to make cell lines more sensitive in general, including the BRAF, EGFR, and EML-ALK which are responsive to treatments (Garnett, 2012). With the creation of the extensive drug-sensitivity database by Iorio et al. in 2016 using 990 cell lines and 265 drugs, the pre-clinical test environment became much more concrete, and the data set landscape dramatically different from what was observed in the method comparison study by Costello in 2013. However, Costello's experiment did determine that the best performing models used nonlinear probabilistic determinations of sensitivity, with the best being MKL(multiple kernel learning) (2013). Therefore, to keep in touch with previous research, I decided to analyze this another way using pairwise comparisons and a logistic model. However, Barnett doesn't provide data on the actual viability for precision medicine, albeit having used his data to adapt some into I chose to use pairwise comparisons for this project because it provides a much simpler result as compared to regular IC50 values, just whether one drug is more effective than another one. In addition, the generation of pairwise comparisons provides a much larger data set than using simple regression; this would allow us to obtain increasingly accurate models. Using the Bradley-Terry model, we can at the very least predict which compounds to test first during trails, if not prescribing personalized medicine to cancer patients. The addition

of pairwise comparisons instead of regular data makes the threshold of analysis much simpler to do; for new data, rather than specifically measuring the IC50's of each drug-cell line test, pairwise comparisons can be easily observed, allowing for quick tests and prediction of the actual IC50 value and data by comparing them relative to other drugs. With a good model, we can help make personalized medicine more affordable to research, more affordable for patients, more accessible for everyone, and most, of all, safer for those patients who are currently struggling for their lives. In addition, I hoped to identify various ways which I can assist researchers in precision medicine to develop new drugs. Although we use various types of universal drugs today which target the cancer in a specific body part, it must be noted that there are only so many categorizations. Genes are a factor which have many different variations and combinations based on the cell. By analyzing this portion, doctors can have a vastly greater control of the killing of the cells. Pharmacogenomics currently is a rapidly developing, albeit relatively new, approach to precision medicine. However, many people claim that it is not implementable and not worth the effort. Therefore, in the process of mathematically analyzing the various components which contribute to drug sensitivity, I also hope to prove (or disprove) the viability of precision medicine in combating cancer, so that we can find out whether it is an avenue of medicine we would be able to support wholeheartedly. For this to happen, the coefficients for precision medicine on specified cell mutations must be significant enough compared to the baseline effects for the drug.

2 Methodology

2.1 Notation

In my calculations, the terms β_0 will refer to the simple coefficient of the intercept. For the first portion of the experiment without drug features, that would refer to the intercept of each of the drugs. That is, each of the medicines would have a specific effect that it has on the cells regardless of cell line. On the other hand, β, or sometimes referred to as β_1, refers to the regression coefficient that includes the cell line features as well as the medicine. i denotes the cell line number, and G or $\mathscr{G}$ indicates the response variables used, which in our experiment mostly consists of 1's and -1's. T means a matrix transposition, and η is a function of x that defines y, i, and j respectively denote whether the coefficient indicates drug i or drug j, the first or the second compound in the comparison.

2.2 Prediction Accuracy with 3M Pairwise Comparisons

In this part of the experiment, I used the 2.4M comparisons given as training data from Kaggle to predict the 0.6M comparisons. The initial data for this experiment was obtained from the Kaggle contest https://inclass.kaggle.com/c/drugsensitivity-3 and contains 3,000,000 binary comparisons between 265 drugs on 990 cell lines, each of which contain 1250 possible genetic features. In addition, these cell lines are classified into 31 cancer types. Sensitivity values are measured by IC50, and in binary comparison data of [Cell Line ID, Drug 1, Drug 2, Comparison Value], a comparison value of 1 denotes a higher sensitivity of drug 1 as compared to drug 2 on the cell line. It follows that a value of -1 signifies that drug 2 is the more sensitive of the two drugs. Alternatively, a value of 0 indicates a negligible difference between the two. Of the 3,000,000 binary comparisons, 2,400,000 is given as training data, and the purpose of the project is to predict

the 600,000 comparisons for which the results are not given. For the second part of the study, I obtained data from cancerrxgene.com, where the data from the preliminary part of my experiment came from.

Part 1: Sparse Bradley-Terry Model with Elastic-Net Logistic Regression

Due to the sparsity of the features for the cell lines, I first started with a sparse logistic model with *no intercepts* using glmnet, where all the drugs had penalty factor zero and the rest of the coefficients had penalty factor 1. Due to the number of comparison values of zero at 0.18% of the data, I decided to forgo this to make a simpler logistic model in comparison to a softmax model. In this case, the response variables are $\mathscr{G} = \{1, -1\}$ and we define $y_k = K(g_k = 1)$, creating a Bradley-Terry model of

$$\mathrm{P}(G = 1|X = x) = \frac{e^{\beta_0(i)-\beta_0(j)+(\beta_1(i)-\beta_1(j))^T x}}{1 + e^{\beta_0(i)-\beta_0(j)+(\beta_1(i)-\beta_1(j))^T x}}, \tag{1}$$

for cell line feature x, and after logistic transformation we obtain

$$\log \frac{\mathrm{P}(G = 1|X = x)}{\mathrm{P}(G = -1|X = x)} = \beta_0(i) - \beta_0(j) + (\beta_1(i) - \beta_1(j))^T x.$$

We will use the negative log-likelihood as a loss function defined as

$$\mathscr{L}(y, \eta) = -\sum_{k=1}^{N} \log P(y_k|x_k, \eta),$$

and substituting $\eta = \beta_0(i) - \beta_0(j) + x_k^T(\beta_1(i) - \beta_1(j))$ gives us

$$= -\sum_{k=1}^{N} y_k \log P(y_k = 1|x_k, \beta_0(i) - \beta_0(j) + x_k^T(\beta_1(i) - \beta_1(j)) + ...$$

$$... + (0 - y_k) \log P(y_k = -1|x_k, \beta_0(i) - \beta_0(j) + x_k^T(\beta_1(i) - \beta_1(j)))$$

then substituting equation 1 for P:

$$= -\sum_{k=1}^{N} y_k \log \frac{e^{\beta_0(i)-\beta_0(j)+x_k^T(\beta_1(i)-\beta_1(j))}}{1+e^{\beta_0(i)-\beta_0(j)+x_k^T(\beta_1(i)-\beta_1(j))}} + (0-y_k)\log \frac{1}{1+e^{\beta_0(i)-\beta_0(j)+x_k^T(\beta_1(i)-\beta_1(j))}}$$

$$= -\sum_{k=1}^{N} y_k(\beta_0(i)-\beta_0(j)+x_k^T(\beta_1(i)-\beta_1(j))) - \log(1-e^{\beta_0(i)-\beta_0(j)+x_k^T(\beta_1(i)-\beta_1(j))}). \quad (2)$$

Minimizing function and adding a regularization penalty, in addition to averaging it over N, makes this formula for the elastic net model

$$\min_{(\beta_0,\beta_1)\in\mathbb{R}^{p+1}} -\left[\frac{1}{N}\sum_{i=1}^{N} y_k(\beta_0(i)-\beta_0(j)+x_k^T(\beta_1(i)-\beta_1(j))) - \log(1-e^{\beta_0(i)-\beta_0(j)+x_k^T(\beta_1(i)-\beta_1(j))})\right]$$

$$+\lambda\left[(1-\alpha)||\beta_1||_2^2/2+\alpha||\beta_1||_1\right] \quad (3)$$

where α is the penalty factor, $\|\beta_1\|_2^2 := \sum_{i=1}^{265}\|\beta_1(i)\|_2^2$, $\|\beta_1\|_1 := \sum_{i=1}^{265}\|\beta_1(i)\|_1$, and p consists of 265 drugs, 1250*265 feature-drug combinations, and 31*265 cancer-drug combinations as predictors, for a total of 339730 predictors. To differentiate the drugs between drug 1 and drug 2, I set a value of 1 for drug 1 and -1 for drug 2 in the matrix. For the regularization's penalty factor, I ran ridge regression($\alpha = 0$) for the 265 drugs and LASSO($\alpha = 1$) for the rest of the predictors.

I then ran a 10-fold cross validation, and stayed with the 90th lambda value of 2.692592e-06. This model, when tested on partitioned data in a ratio of 4:2 train to test, achieved an accuracy 93.5%. When trained with 2,400,000 comparisons, this model predicted the 600,000 comparisons of test data with 95.011 accuracy.

Part 2: Floyd-Warshall Graph Extension

I also attempted to use the extended graphs through the Floyd-Warshall Algorithm to increase the number of comparisons in the training set to 18 million, but it didnt produce the desired effect.

Since the resultant regression fit registered at 118 MB, while the original model took up 128 MB. This indicates that the intercepts and coefficients were sparser, so we can conclude that more values were reduced to 0 in the extended model than the original, which may have caused the model to perform worse. The extra repetitive samples caused an overfit so the cases outside of the direct association had less accuracy. Another point to be made is that the original data was also not a complete set, so therefore there are sometimes less comparisons in one cell line as compared to another.

Part 3: Elastic Net Logistic Model with Intercepts

In this model, I performed an elastic net regularization method instead of the lasso L1 penalty, selecting a penalty factor of $\alpha = 0.6$. In addition, I added intercepts to the model, and used the extended graphs to confirm results, of which nearly all $(> 99.75\%)$ were correct, meaning that the errors in prediction must have come from more distant comparisons than those which could be deduced from just completing the graph. This model ended up achieving 95.198% on the test data, which, as of July 27, 2017, ranked second on the Kaggle leaderboard.

2.3 Coefficient Analysis on Full Data Set

For this stage, I gathered my data directly from a public data bank (Genomics of Drug Sensitivity in Cancer; www.cancerRxgene.org) in the form of an excel table with IC50 values for each cell-line and drug pair. The site also provided a similar 990 cell line features in addition to drug features which total 880 public chemicals (for the known 212 drugs). I converted the IC50 data

into pairwise comparisons for logistic regression.

Part 1: Coefficient Analysis

After confirming the prediction power in the experiment outlined above, I used the same model but with the complete pairwise comparison data, which I generated with the IC50 values table from cancerrxgene.org. This should make the data less skewed as compared to the pairwise comparison data I generated from the Floyd-Warshall algorithm, since it is generated from a nearly complete graph rather than linking together given data points. This leaves me with approximately 24 million pairwise comparisons, which I plugged into the model. In the initial model and the Floyd-Warshall model, the lambda.min only changed minutely, so we can assume it should be similar due to the closeness in size of the Floyd-Warshall model and this model using the full data set. I reduced the cell line features which were completely zero in the revamped model. When using the previous models, we see that 99.75% of the pairwise comparisons which can be directly extended from the current acyclic graph could already be predicted, so reducing the data appropriately into halves would not make the training set less accurate, but on the other hand would prevent overfitting. The resultant coefficients become more balanced than in the previous models due to a more complete data set. From this model, I extracted a coefficient matrix so that I can further investigate which factors tie into the sensitivity of the drugs the most.

Part 2: Full Data Set, Drug Features Although there were drugs which were tested at two different sites (and therefore had slightly different results), they had the same starting concentration and ending concentration so I just treated them as different drugs so that they would have differ-

ent baselines. In addition, I added an extra feature detailing whether the drug was repeated or not, which would be considered as a correction feature for those drugs that were tested in different places and/or under different conditions. For this model, each drug has a similar baseline, and each drug feature contains its own baseline in addition to the drug feature - cell line feature coinciding. This allows for a greater predictive power when it comes to cell line features, and it also allows for the design of personalized medicine based on the specific features of the cancer strain the patient has. I removed the cell line features for which there were no cell lines which contained them, bringing the total features for the cell lines to 1063, and adding on the extra feature for alternate conditions gives us 1064 cell line features. Each drug feature has a baseline effect that it has on the cell line, making the initial features 880 drug features for the β_0 term for the regression formula. Combining that with the cell line features and cancer types gives 880 drug features * (1064 cell line features + 31 cancer types) = 963600 total features. Adding on the baseline terms provides us with a total of 964480 possible features to analyze for each comparison. However, due to the density of the drug features, as compared to the previous sparsity of the cell line feature matrix, this venture costs multiple times more processing power than the experiments before.

3 Illustrations

Provided below in Figure 1 is the graph of the coefficients for the changing lambda values, contrasted with a graph of the full data set and its lambda values in Figure 2:

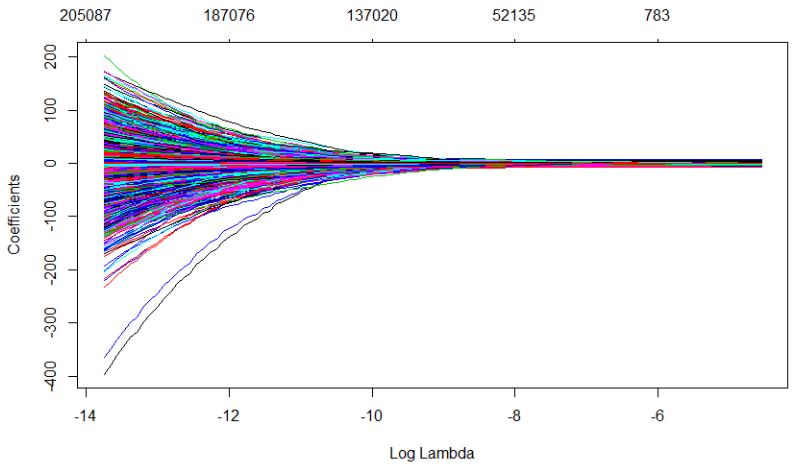

Figure 1: Coefficients (β_1) of logistic regression plotted against Lambda values with pairwise data from kaggle. Two very predictive components make up the negative portion, the CUL3 mutation with LY317615 and the RHOT1 mutation with Bleomycin, 50 M. However, in the full graph extension, the CUL3 and LY317615 pair become more reduced, while the RHOT-Bleomycin pair still exists. For Bleomycin, the correlation between the coefficients of RHOT and COAD/READ are similarly resistant, showing that RHOT has a strong correlation with pancreatic cancer

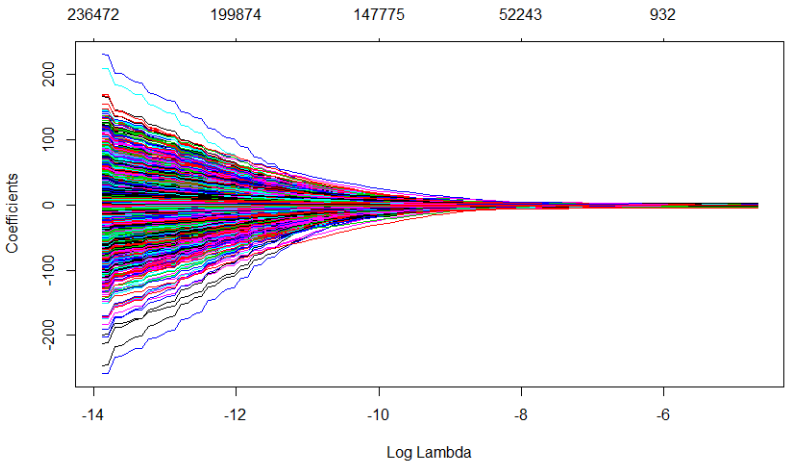

Figure 2: Coefficients plotted against Lambda values for full data set, data set is more normally distributed as compared to the one formed from the 2.4M partitioned data

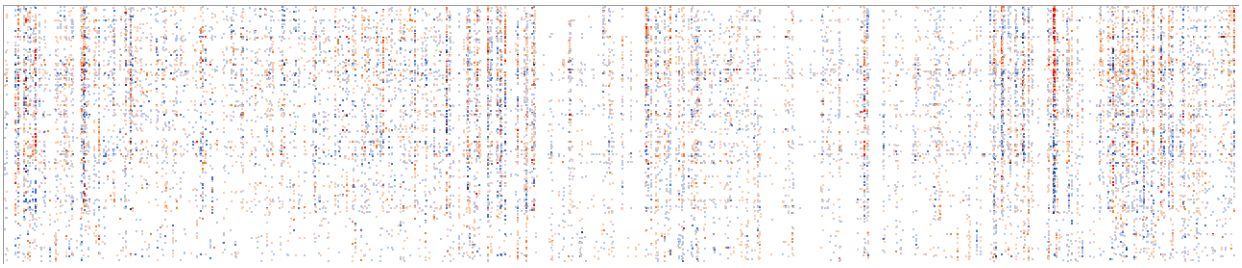

Figure 3: Unsorted colored data values of the regression coefficients in β_1, with y axis medicine and x axis mutations and cancer types. Darker shades of blue denote more sensitive mutation-drug combinations, and darker shades of red denotes more resistant combinations.

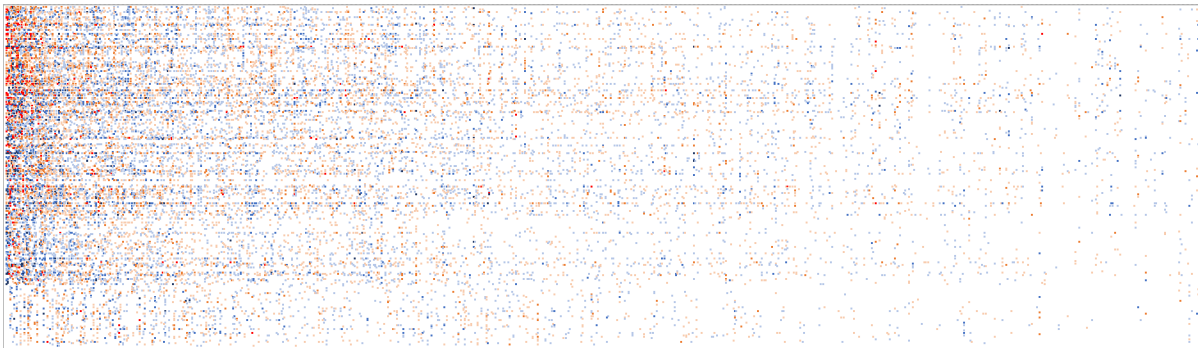

Figure 4: Sorted first by column and then by row so that the most significant coefficients (calculated by the sum of the absolute values for the columns/rows of β_1) drift towards the top left.

4 Results and Discussion

A brief overview of the visual in Figure 3 reveals the fact that mutations affect the sensitivity of the drug much more than expected. The medicine does affect the outcome, but trends in the mutations are clearly visible, as mutations which show strong sensitivity and/or resistance tend to affect the outcome more, even on different drugs. On the other hand, there seem to be many columns with many values which are more neutral due to the shrinkage enacted by the elastic net penalty.

At the same time, these trends indicate that some of the cell line features' sensitivity has a correlation with some of the drugs in either an extremely sensitive or extremely resistant association. I organized the data into the most influential components ordered by columns first then by rows, after future noting that there were trends among drugs too. Analyzing which components of the drugs affect the cell line features the most would contribute majorly to the experiment. From just this sorted model's row 1, Obatoclax Mesylate could be defined as having a major impact upon various mutations. A search to the analysis in cancerrxgene redefines this: a volcano plot for Obatoclax Mesylate contains many of the extremely resistant tendencies that we see in the data visualization, but also many extremely sensitive values appear. However, Obatoclax Mesylate is an interesting case, as even though it has more extreme resistant coefficients, the number of moderate sensitive coefficients greatly outnumber them, so that although not many of its coefficients are as definitive as bleomycin or etoposide, it still occupies the top position in the sum of the L1 norm.

Bleomycin is one of the most used chemotherapy drugs for a reason, because it has great tumor killing potential in a variety of cancer cell types. The effectiveness changes further depending on the dosage amount used. From my sorted chart in terms of overall impact in sensitivity (both positive and negative), bleomycin's two tests appear in second and fourth rows respectively. Note that the two were not only tested in different dosages of 0.195-50 uM and 0.25-64 uM, but they also were tested under different environments of Sanger Institute and Massachusetts General Hospital. The fact that they were respectively placed fourth and second in the coefficient analysis was no coincidence: the cell lines showed trends of large sensitivity coefficients in various cell lines and drugs.

Etoposide's (a commonly used chemotherapy drug) coefficients are similarly high as Bleomycin's, especially for the extreme values where the cell lines are very sensitive. Although etoposide's overall L1 norm isn't as high as CX-5461 and LY317615, those two drugs lean more to the resistant side of the drugs.

These two drugs both analyze CDC73 as a very drug sensitizing mutation, which is a very interesting point of study that has not been researched deeply into. In terms of overall impact in the high variability section, it places third. According to its coefficients, in reference to the aforementioned Bleomycin examples and the Etoposide drug, which place in the top 1% for the chart, CDC73 shows great potential in being a cancer sensitizing genomic markers, which biologists may be very interested in testing. The bleomycin of 50 uM and Etoposide also coincide in NTRK2 and NTN4 mutations, and these two drugs are known to be used together when treating testicular cancer. In fact, the NTRK2 and NTN4 mutations show great collinearity in terms of coefficients, so there is a possibility that they influence each other in turn, making for a intriguing direction to work towards.

However, these drugs seem to only target some specific cell mutations, as when we take a look at their β_0 coefficients, the coefficients on the top are quite low compared to those toward the bottom of the list (which indicates less variability with the cell line features). To confirm this, I did a simple binomial regression analysis to get a common ranking for all the drugs, for which I tested on the kaggle contest to achieve an 83% prediction accuracy. All of the 10 drugs which are analyzed to have the highest common ranking values are on the bottom of the list of variability. Bortezomib, the top of the common ranking according to our model, was the first therapeutic

proteasome inhibitor to be used in humans, demonstrating its viability and the validity of the model. Many of the rest of the top drugs are also being used and is confirmed by our model. Epothilone and Docetaxel, ranked two and third in the model respectively, are within the first tier of cancer treatments. Vinblastine and two types of Camptothecin, are also being used regularly. However, these drugs all place towards the lower ends of the spectrum when comparing the magnitude of the cell features. On the other hand, the drugs towards the higher ends of the variability spectrum are not used as much, despite vastly higher values concerning drug mutations. Why? One thing that many treatments prioritize is the repeatability of the drugs on various mutations due to the lack of precision medicine. Comparison of top and bottom total mutation variability drugs reveal that the β_0 values for the bottom values are more stably positive, and the magnitudes tend to drift higher. Although there are 1250 cell line features (and well over 50,000 in the true data, after which we concentrated it down to the 1250 most important features), there are only 31 cancer types. It is impossible to do specific targeting of a mutation just by knowing the cell line. Therefore, rather than drugs which target specific features, we choose ones which can do well over a variety of features. However, we can judge from the fact that the coefficients which are generated from the β_1 values vastly outstrip the magnitude of the values from the β_0 field that precision medicine could have massive effects upon cancer treatment, as the β_1 cell features contains quite a few values well over 20, while the latter common ranking usually ranges only from -4 to 4 at its extreme values, reflecting the large variability of cell line sensitivity due to these features.

Comparison of the β_0 values and the sum of the coefficients for those drugs. Observation of data shows that drugs which have a higher β_0 values corresponds to higher total sum values when

adding the total drug drug coefficients together, despite them having smaller values in general, meaning that the drugs on the bottom of the list have less variability with values, but they stably positive. These drugs towards the bottom act as a general solution to most cancer therapy. Those drugs near or at the top of the list are more useful for research in precision medicine, as it targets specific cell features rather than work on them as a whole. If used incorrectly, it could be very ineffective, but if used correctly, it can help vastly improve the conditions of the cancer patient. For cancer researchers testing precision medicine, it would be more effective to test the top values such as Bleomycin and Etoposide instead of the bottom values as those are more universal drugs which have a moderate effect on most cancer cell lines. The formula $\beta_0 + x^T \beta_1$ gives the precision drug ranking conditioning for cell line feature x. Although precision medicine analysis and testing could still be used on those drugs due to the fact that there are still negative values, maximizing the effect of the low variability drugs based on the cell line mutations is not as effective as perfecting the top variability drugs. However, these universal drugs could be utilized well if we use them as a baseline for testing the precision medicine.

The model also confirms current biological interpretations for mutations. A phenomenon that occurs is the fact that the EGFR mutation improves lung cancer treatment results, but does adversely for Glioma. In our data, there is a positive correlation between the EGFR mutation and the LUAD/LUSC mutation, as more sensitive values for a drug in one of them corresponds to the same in another. On the other hand, both of these have a negative correlation with LGG cancer, statistically supporting the argument.

During the course of the experiment, I first discovered that logistic regression has an astonishingly high prediction accuracy when using it for pairwise comparisons on data constructed from the original IC50 chart. A pairwise comparison model makes data more accessible and creates a simple model which people can easily interpret. For example, to the regular patient the IC50 value is a number which is hard to understand, with a plethora of explanation is needed to make them understand that the lower the values the more effect the drug has. However, what they truly care about is just a simple statement of whether a drug is effective. A confirmation that drug 1 is more effective than drug 2 would be more relevant. Accurate predictions can serve as a guide for early cancer drug development, pointing them which compounds to test first, or even prompting the creation of a new compound which can be personalized for every person.

Unfortunately, all cell lines for these databases are performed on cultivated cell lines. During the implementation of precision medicine, there may be some unpredictable implications that arise. The fact that two of the same drug, such as Selumetinib, tested at the same concentration but under different environments of the Wellcome Trust Sanger Institute and the Massachusetts General Hospital Cancer Center had radically different results shows that the sensitivity of cell lines needs repetitive testing to confirm its effect. In addition, through observation we can see that some of the cell features exist in only one cell line, and some in none at all throughout the 990 that are tested (1063 being the number of features which actually have values). Due to this, construction of a conclusion based upon only a single case study of the coefficients does not have much statistical significance, as shown by the significant results of LY317615-CUL3 pair (Figure 1), which disappears when the full data set is applied (Figure 3). One avenue of research we can undertake to

resolve this is to analyze the specific conditions under which the cell lines were tested, and gather data on the external factors which could increase or decrease the sensitivity of the drugs. However, to get field data on cancer drug tests is a very difficult process, and the risk is very high, and there may also be some unexpected adverse effects if implemented incorrectly. Hopefully the database can first be completed, as there are still some unfilled IC50 values in the chart, so that we can more effectively analyze the efficiency of the medicine so that when it comes the time to take real trials, there wouldn't be any problems. In addition, we can tell from the coefficients of bleomycin that when changing conditions and dosage, the cell lines that the drug has an impact on may change. Therefore, it would also be best to do laboratory tests in various dosages and detail the conditions the drugs and cell lines were tested on. If we manage to complete that step, then the future data would be much more helpful for analysis. I hope that my project will not only unveil details about cancer treatment, but also inspire the biology oriented researchers to collect more data to enhance the analysis we can do.

5 Conclusions and Future Work

The sparse Bradley-Terry (logistic) model is accurate and simple with interpretability on biological phenomena. It learned some coefficients (β_0, β_1) from a large set of pairwise comparisons of drugs based on their sensitivity comparisons in cell lines. Here the coefficient vector β_0 describes a common ranking of all drugs based on their overall sensitivity on the whole dataset, while the sparse coefficient matrix β_1 measures the variability of drug sensitivity over genetic features and cancer types. As a result, $\beta_0 + x^T \beta_1$ gives rise to the personalized drug ranking when a feature x is

presented, which is desired by precision medicine. Some phenomena are confirmed by this model. For example, when x represents the EGFR mutation, the top ranked drugs indicated by $\beta_0 + x^T \beta_1$ show good efficacy in the treatment of LUAD (lung cancer) but are low ranked in LGG (glioma), which is a trending research direction in our labs. Moreover, large row or column l_1-norms in matrix β_1 indicates the high diversity in sensitivity over drugs or cell line features, e.g. Bleomycin as a widely used cancer medication are shown to be extremely sensitive or resistant to different mutations which deserves further attentions in clinical studies.

On the other hand, some hypothesis derived from the data still deserves further studies due to the small number of trials against the wide number of possibilities. Some cell lines describing major effects such as the effect of CUDC-101 upon LUAD and EGFR only contain a total of 5 data points, which is not sufficient to draw a stable conclusion from. In particular, some measurements in IC50 for cell lines could vary greatly when compared to the use of the drugs in clinical trials, which indicates the big uncertainty based on cell line studies. Therefore, it is desired to build large comprehensive datasets for precision medicine in the future. Another future direction is, with the growth of data, it is desired to extend this simple model to analyze changes due to multiple cell line and/or drug features interactions. As we know, various mutations could work together to make the cell line either more resistant or more vulnerable, usually the former, to various types of medicine. Pinpointing emergent properties of the cell lines and medicine could further enhance understanding of early drug development and personalization of treatment.

References

Costello, James C., et al. "A community effort to assess and improve drug sensitivity prediction algorithms." Nature biotechnology 32.12 (2014): 1202-1212.

Garnett MJ, et al. "Systematic identification of genomic markers of drug sensitivity in cancer cells." Nature. 2012;483:570575.

Iorio, Francesco, et al. "A landscape of pharmacogenomic interactions in cancer." Cell 166.3 (2016): 740-754.

My code could be found at https://github.com/CompetitionEntrant9192017/Code