BeatAML Challenge - predict quantitative ex vivo drug sensitivity

Mingyuan Chen, Gek Chuah Kenneth Goh, Tong Wu

Abstract*— *The objective of the project is to participate in the Beat AML challenge to predict quantitative ex vivo drug sensitivity based on genomic variants, gene expression and clinical data. The project mostly focuses on using RNA sequence data. Ridge regression, linear regression and Lasso regression were implemented respectively. Top most varying data was selected as the feature selection model. Spearman score and Pearson score were calculated for evaluating the models. The highest score was from ridge regression model.

1 INTRODUCTION

1.1 Introduction of Acute Myeloid Leukemia

Acute myeloid leukemia (AML) is a type of cancer of the blood and bone marrow. It's also called acute myelogenous leukemia, acute myeloblastic leukemia, acute granulocytic leukemia, and acute nonlymphocytic leukemia (1). AML is the most common acute leukemia in adults. Over 20,000 patients are diagnosed with AML in the United States each year and over 10,000 patients die because of AML annually in the US (2). Figure 1 shows the diagnosis rate and the death rate of AML in the United States from 1992 to 2016 (3). Most of the patients with AML receive chemotherapy treatments which will kill cancer cells rapidly. The standard therapies include a combination of anthracyclines (daunorubicin) and nucleoside analogs(cytarabine). Some of the elderly patients who are not fit for the combination treatment would be treated with alternative medicines such as hypomethylating agents (HMAS) or low-dose cytarabine which are less-toxic. However, according to the National Cancer Institute (NCI), the present five-year survival rate for AML is around 28.3% (3), which means that for the people who are living with AML, only 28.3% of them are still living five years after their diagnosis. Unfortunately, there are very few improvements to the treatments for AML in the past decades. According to NCI, AML is a disease that is difficult to treat and the majority of the patients do not have good outcomes.

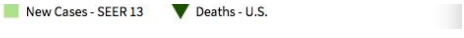

Figure 1. Diagnosis rate and death rate of AML in the United States from 1992 to 2016 (3)

There are a variety of other myeloid malignancies that are related to AML. For example, myelodysplastic syndromes (MDS) and myeloproliferative neoplasms (MPN) which are distinct diagnosis categories from AML, have overlapping mutations signatures including epigenetic regulatory, splicing pathways and signaling pathways with AML (4). It's been known that healthy individuals who do not have evidence of malignancies but have clonal hematopoiesis would suffer much higher risk for the development of AML and other myeloid malignancies (5). In addition, individuals who have MDS, MPN or the overlap disorders would also have higher risk since those often can transfer to AML (6). Understanding the clinical manifestations of AML, the mechanisms contributing to its development and the role of other overlapping disorders toward AML evolution would be helpful in improving treatment methods in the future.

Some categories of diagnosing AML that have been developed are clinical history diagnosis, cytogenetic diagnosis, and diagnosis based on the blast differentiation states of neoplastic cell arranging. For example, around 10% of the AML patients occur during the prior cancer therapy, such as whole-body radiation, etc. 10 - 20% of the patients are secondary to MDS/MPN disorders (7).

1.2 Introduction of the Beat AML DREAM Challenge

The objective of the project is to participate in the Beat AML Dialogue for Reverse Engineering Assessments and Methods (DREAM) challenge, which aims to find new and better computational models and make the models available to all. The DREAM challenge has been hosted as an open science, crowdsourcing challenge since 2006. The Beat AML program collected about 900 patient samples from AML patients on which the genomics was performed. The goal of the Beat AML program is to understand associations between genetic or transcriptomic events drug sensitivity and to nominate some of the most promising associations into clinical trials.

There are two sub-challenges. Sub-challenge 1 is to predict quantitative ex vivo drug sensitivity (represented by the area under the dose-response curve values (AUC)) based on genomic variants and/or gene expression. Sub-challenge 2 is to predict the clinical response (represented as days of survival after inclusion in the study) based on ex vivo drug sensitivity data, genomic variants, gene expression, and clinical data. There is a training set data with 213 samples for training the algorithm, a leaderboard set data with 80 samples for testing the algorithm, and a validation set data with 65 samples for finally scoring the algorithm. The whole dataset includes 5 numerical clinical features, 20 categorical clinical features, 122 drug sensitivities, 399 high-confidence-somatic mutations, and 63,677 genes in RNASeq.

Section 2 describes the method that was used for

developing the prediction algorithm. Section 3 presents the result of the algorithm and the discussion of the result. Section 4 is the conclusion of the report and the project.

2 METHOD

For training the model, me mainly focus on using the RNA sequence data. We implemented the ridge regression, linear regression and Lasso regression to the data. We selected a certain number of most varying genes as the feature selection method.

Ridge Regression is a statistical method for analyzing multiple regression data that suffer from multicollinearity. When multicollinearity occurs, least squares estimates are unbiased. However, the large variance of them would contribute to differentiate from the true value. Ridge regression can reduce the standard errors by adding a degree of bias to the regression estimate (8). In statistics, linear regression is a linear approach to model the relationship between the independent value and the dependent value by fitting a linear equation to observed data. A linear regression model has an formula of the form $\boldsymbol{Y = a + bX}$, where $\boldsymbol{X}$ is the explanatory variable and $\boldsymbol{Y}$ is the dependent variable. The slope of the line is $\boldsymbol{b}$, and $\boldsymbol{a}$ is the intercept (the value of $\boldsymbol{y}$ when $\boldsymbol{x} = 0$) (9). Lasso regression is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, for example, the mean value of the data. Lasso regression tends to procedures simple, sparse models. This particular type of regression is well-suited for models showing high levels of multicollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination (10).

Support vector machine (SVM) is a supervised machine learning algorithm which can be used for both classification and regression problems. The form of equation defining the decision surface separating the classes is a hyperplane of the form: $w^T x + b = 0$ where w is a weight vector, x is input vector and b is bias. For hyper-planes H1 and H2, we have that: H1: $w.x_i + b = +1$ and H2: $w.x_i + b = -1$. The optimization problem is to: minimize $||w||$, s.t. discrimination boundary is obeyed, i.e., min f(x) s.t. g(x)=0, which we can rewrite as: min f: $\frac{1}{2} ||w||^2$ s.t.. g: $y_i (w.x_i) - b = 1$ or $[y_i (w.x_i) - b] - 1 = 0$.

Docker was used for submitting the models because of the reproducibility of the result. Docker container was used to store the model, which is a standard unit of software that packages up code and all its dependencies. Docker image was built and used to push the model to Synapse, which is a package of software that includes everything needed to run the model. Docker file was written to call and run the codes directly from the terminal.

After models were submitted to Synapse, they were evaluated by the organizers. The performance of the models were quantified by calculating their Spearman score and Pearson Score compared with the ground truth. Both of the scores are methods of calculating correlations and strengths between two variables.

First, we ran the given sample of ridge regression on the listed number of 1,000 genes and found that the Spearman and Pearson correlation scores were 0.2616 and 0.2667 respectively. Then, we varied the number of genes to be used as features using the already provided ridge regression sample and found that the Spearman and Pearson correlation scores for top 2,000 gene variances were 0.2654 and 0.2680 respectively. We then tried increasing the number of the most varying genes to 20,000, 25,000 and 30,000. We also tried applying linear regression on the top 1000 and 2,000 and more genes with variance. For top 1,000 gene variances, the Spearman and Pearson correlation scores were 0.08367 and 0.09496 respectively. For top 2,000 gene variances, the Spearman and Pearson correlation scores were 0.07581 and 0.09352. What's more, lasso regression for top 1,000 gene variances gave Spearman and Pearson correlation scores of 0.2450 and 0.2408 respectively.

We successfully ran the SVM on the listed number of 1,000 genes locally in a Jupyter notebook environment. But subsequent attempts at submitting to the Synapse Challenge in a docker environment was not successful.

3 RESULT

Table 1: Results of submission to Synapse for Ridge Regression

SynapseID	Feature Selection	Spearman Score	Pearson Score
wertyuio	Top 1,000 most varying genes	0.2616181277755627	0.2616181277755627
syn21679510	Top 2,000 most varying genes	0.2654508990707644	0.26809511210274
syn21761387	Top 20,000 most varying genes	0.2935859683763973	0.279657885407815
syn21763430	Top 25,000 most varying genes	0.295495243075551	0.27687348442222
syn21761459	Top 30,000 most varying genes	0.292085341971344	0.270727938553658

As shown in the table 1, the highest Spearman score 0.2955 from ridge regression method came with the top 25,000 most varying genes as the feature selection. The highest Pearson score was associated with the top

20,000 most varying genes, which was as high as 0.2797.

Table 2: Results of submission to Synapse for Linear Regression

SynapseID	Feature Selection	Spearman Score	Pearson Score
syn21679510	Top 1,000 most varying genes	0.0836794110502165 5	0.094962656846804 66
syn21679510	Top 2,000 most varying genes	0.0758110554335891 4	0.093526867938361 47
syn21763548	Top 20,000 most varying genes	0.1073751184347128	0.107744969350279
syn21763881	Top 30,000 most varying genes	0.089205072301867	0.101076873910849

It was found that in Table 2, both of the cores increased when the number of the top most varying genes increased to 20,000, decreased when the number increased to 30,000. The highest scores from linear regression are 0.1074 and 0.1077 respectively.

Table 3: Results of submission to Synapse for Lasso Regression

SynapseID	Feature Selection	Spearman Score	Pearson Score
syn21679510	Top 1,000 most varying genes	0.2450536959558377	0.240821624675251
syn21765988	Top 2,000 most varying genes	0.241023260112439	0.23845239898865
syn21764324	Top 10,000 most varying genes	0.234432532220273	0.232176589158474
syn21765661	Top 20,000 most varying genes	0.222429106547529	0.216526315909
syn21765897	Top 30,000 most varying genes	0.216041208854558	0.207498224937317

Table 3 shows the result of submissions for modeling with Lasso regression. It's shown that, as the number of the top varying genes increased, both of the Spearman score and Pearson score decreased. The highest scores for Lasso regression were 0.2451 and 0.2408 respectively when there were 1,000 mots varying genes applied.

Overall, the ridge regression model with feature selection as top 25,000 most varying genes achieved the highest Spearman score, the ridge regression model with feature selection as top 20,000 most varying genes achieved the highest Pearson score.

4 CONCLUSION

It appears that linear regression is not a good method to predict drug sensitivity response for each patient under each inhibitor since the scores were not as high as with the other regression models.

Ridge regression technique seems to be more promising since increasing the number of genes variances to be tested increase both the Spearman and Pearson correlation scores. Ridge regression is particularly useful to mitigate the problem of multicollinearity in linear regression which commonly occurs in models with large numbers of parameters.

REFERENCES

1. National Cancer Institute, General Information About Adult Myeloid Leukemia, retrieved from https://www.cancer.gov/types/leukemia/patient/adult-aml-treatment-pdq , 2019.
2. Kouchkovsky I., Abdul-Hay M., Acute myeloid leukemia: a comprehensive review and 2016 update, *Blood Cancer Journal*, 6, e441, 2016.
3. National Cancer Institute, retrieved from https://seer.cancer.gov/statfacts/html/amyl.html, 2020.
4. Pati H., Veetil K., Myelodysplastic Syndrome/Myeloproliferative Neoplasm (MDS/MPN) Overlap Syndromes: Molecular Pathogenetic Mechanisms and Their Implications, *Indian J Hematol Blood Transfus,* 35(1): 3-11, 2019
5. Bowman R., Busque L., Levine R., Clonal Hematopoiesis and Evolution to Hematopoietic Malignancies, *Cell Stem Cell*, 22(2)L 157-170, 2018
6. Mangan J.K., and Luger S.M., A Paraneoplastic Syndrome Characterized by Extremity Swelling with Associated Inflammatory Infiltrate Heralds Aggressive Transformation of Myelodysplastic Syndromes/Myeloproliferative Neoplasms to Acute Myeloid Leukemia: A Case Series, *Case Rep Hematol,* 2012.
7. Greenberg P., etc, Myelodysplastic Syndromes, *J Natl Compr Canc Netw,* 9(1):30-56, 2011
8. NCSS Statistical Software, Chapter 35 - Ridge Regression.
9. Introductory business statistics with interactive spreadsheet, Chapter 8 - regression basics, retrieved from https://opentextbc.ca/introductorybusinessstatistics/chapter/regression-basics-2/
10. Tibshirani R., The Lasso: a brief review and a new significance test, Stanford University, 2014, retrieved from http://statweb.stanford.edu/~tibs/ftp/ubctalk.pdf
11. R. Berwick, An Idiot's guide to Support vector machines (SVMs), retrieved from http://web.mit.edu/6.034/wwwbob/svm-notes-long-08.pdf