PUBH 7475 Final Report
Predicting Drug Sensitivity Using Gene Expression Data

By: Yuting Shan, Eunice Songthangtham, Zhaoliang Zhou

Spring 2022

Introduction

Patients' responses to chemotherapeutic agents are highly variable (Mishra et al., 2010). Even though the same medication is given to patients with the same diagnosis, some may benefit from the medication with minimum side effects, while others may not benefit from the medication and experience some lethal side effects. Therefore, it is of great clinical importance to pre-classify patients into potential responders or non-responders before they physically receive the chemotherapeutics.

Recent research has shown that the predictive performance is better when considering the cumulative effect of many biomarkers in predicting complex traits like drug responses. In this project, we aimed to use the expression level of thousands of genes to predict the clinical chemotherapeutic responses. The idea of this project was motivated by a previous publication by Paul Geeleher et al (Geeleher et al., 2013), where the authors used the genomic and pharmacologic profiles of cancer cell lines as the training data to build the model and tested it using patient data. We thought the idea of using cell lines to build the model is brilliant since it is extremely difficult to acquire drug sensitivity data from real patients.

The objective of this project is to use different machine learning methods (ridge, lasso, elastic net, SVM, neural network, and PCR) to fit the model and compare their performance in predicting drug response in patient data. Moreover, we wanted to compare the computational efficiency of different models. Lastly, we aimed to use lasso regression for variable selection in order to identify the potential biomarkers in predicting in vivo drug response.

Materials and Methods

Data.

The genomic and pharmacologic profiles of cancer cell lines were obtained from the Cancer Genome project (CGP). The data contains almost 700 cancer cell lines. The expression of over 12,000 genes was measured using microarray. A total of 138 drugs were screened over all the cell lines to acquire the drug sensitivity which was quantified using IC_{50} (concentration of the drug needed to inhibit cell growth by 50%). The genomic profiles and response status of patients were acquired from two clinical trials from the Gene Expression Omnibus (GEO). Both trials contain tumor gene expression data acquired before drug treatment and drug response phenotype (responder or non-responder). The first trial examined docetaxel in breast cancer patients. A total of 24 biopsies were collected with 10 samples being sensitive to docetaxel and 14 samples being resistant to docetaxel. The expression of 8,399 genes was measured in this trial. The second trial examined bortezomib in myeloma patients. A total of 169 bone marrow biopsies were collected, with 85 samples being sensitive to bortezomib and 84 samples being resistant to bortezomib. The expression of 22,645 genes were measured in this trial.

Data preprocessing.

We followed the same data manipulation and preprocessing as described in the paper published by Paul Geeleher et al (Geeleher et al., 2013) and the detailed process can be found in Figure 1 from the paper. Therefore our results should be comparable with the results from the published paper. Here, we provided a brief summary of the data manipulation process. More details can be found in the paper (Geeleher et al., 2013), especially from Figure 1.

The response variable IC_{50} for both docetaxel and bortezomib can be found directly from the CGP website. The raw gene expression microarray can be found on ArrayExpress website. We preprocessed the data using a robust multi-array average algorithm which can be done using the *rma()* function from R library affy. This function performs background correction, quantile normalization, and media-polish summarization. To obtain the bortezomib data, we used the function *getGEO()* from the R library Geoquery to access the data directly.

We used the cell lines data as the training data, and we used the clinical trial data as the test data. First, we mapped both datasets to the official gene symbols. Since we obtained both datasets from different platforms, we then only subset the genes that are present in both datasets. Subsequently, we homogenized the two datasets using the Combat function from the sva library in R. Lastly, we removed about 20% of the genes which have the lowest variability io expression since they will not improve prediction performance.

Statistical methods.

For our study in particular, several methods are used to predict drug sensitivity including the ridge regression, LASSO regression, elastic net, support vector machine (SVM), and artificial neural network (ANN), and principal component regression (PCR). For Ridge, LASSO, and elastic net, we used the default 10-fold CV from the function *cv.glmnet()* to select the tuning parameters whereas Paul Geeleher et al. (2013) used an automatically selected lambda from the *ridge* package. For SVM and ANN, we used the default method from the function *tune()* which uses grid search to select the tuning parameters. For PCR, we select the number of components that minimizes the root mean square error of prediction (RMSEP). For ANN, to decrease the computing time, we only used the set of variables selected after LASSO for both outcomes. When using all the covariates for ANN, the computation time could take up to 6 hours. However, if we only use those variables that are selected by LASSO, we could decrease the computation time to around 30 minutes. The area under curve (AUC) values are used to compare the prediction performance of the models which is consistent with the published paper (Geeleher et al., 2013).

Results

Table 1 from the appendix section shows the AUC results for each of the models we fitted and for each of the outcomes. First, for docetaxel, we observed that LASSO had the highest AUC compared to other methods. The AUCs for the Ridge regression fitted by the paper using *LinearRidge()* are identical to the ones we fitted using *glmnet()*. In fact, the lambda values selected by two functions are very similar. Next, elastic net and SVM have similar AUC values. Figure 1 from the appendix shows the ROC curves for each of the methods for the docetaxel outcome. Although LASSO had the best performance in terms of AUC, other methods such as Ridge, elastic net, SVM all performed considerably well and did not differ from LASSO by much. Thus, the application of these models would be defined by their clinical relevance, feasibility, and interpretability. Therefore, with these considerations, we would prefer LASSO as it automatically performs variable selection and dimension reduction which is preferable in real-world clinical settings. For this data, the final model contains around 100 variables whereas the original dataset has around 400 variables.

Table 1 also includes the AUCs for different methods for the outcome bortezomib. For the bortezomib, Ridge regression had considerably higher AUC compared to other methods. This result is consistent with the paper (Geeleher et al., 2013). However, as addressed by the paper, the training data only contained bortezomib IC_{50} of 1 multiple myeloma cell line, and the performance for each method could potentially improve with more cell line data. Figure 2 from the appendix shows the ROC curves for each of the methods for the bortezomib outcome.

Discussion

One of the goals of this study is to preliminarily narrow down important genes that have important roles in drug response. Using the selected features from the lasso models, we found that 103 genes were selected for docetaxel, while only 28 genes were selected for bortezomib. Interestingly, there is only one gene, GLI3, that overlaps between both drugs. GLI3 is a gene associated with the Hedgehog pathway during the cell developmental stage (Matissek & Elsawa, 2020). It is believed that the gene itself may play a role in tumorigenesis by influencing immune cell development, hence, impact cancer cell growth and tumor progression. To validate whether the selected features from the lasso models are meaningful, a brief search on the roles of the selected genes were conducted. We found that not all selected genes were interpretable; however, the genes with the largest coefficients seem to serve a meaningful purpose in both tumorigenesis and drug response. The two most influential genes for docetaxel sensitivity are ABCB1 and

BCAT2. ATP Binding Cassette Subfamily B Member 1(ABCB1) is also known as multidrug resistance protein 1. It plays a major role in pumping foreign substrates out of the cells. Interestingly, docetaxel is a well-established substrate of ABCB1 (Shan et al., 2021). Therefore, a higher expression of this gene may result in an accelerated elimination of docetaxel from the cell. The two most important genes for Bortezomib, RBM12B and ARID5B, are more vague in terms of their role in drug sensitivity. Although a clear explanation has not been established, it is suspected that these genes may be involved in tumor development.

Despite the meaningful findings, the study has some limitations. Originally, another chemotherapeutic drug, cisplatin, was also chosen as a drug of interest. However, the models for cisplatin did not produce any interpretable results. It was presented that the issue may be due to a small number of patients being administered this cancer drug; therefore, the variability in the response was not enough for statistical significance to be achieved (Geeleher et al., 2013). Another limitation is related to the data source itself. The data used in this study are microarray data, which is less popular and reliable for gene expression analysis. Microarray data may be obtained from different platforms and can introduce bias into the data due to calibration differences between machines. A potential solution is to use RNA sequencing data, which is considered a much more popular and reliable method. Another limitation is the structure of the outcome variable IC_{50}. Because the outcome is continuous, classifying patients into responders or non-responders requires a cut-off value. The cut-off value is an arbitrary value with no specific standards in how the value should be chosen. Therefore, this value could change depending on the data as well as the investigator of the study.

Future Directions

According to the authors, eliminating a gene completely through feature selection may lead to loss of information. However, this can be advantageous. Patients can be screened for specific genes instead of scanning through the whole genome. In the field of personalized medicine, the genetic profile of the patient is used to make treatment decisions. This is particularly important in the field of cancer. Drugs that are used to treat patients are usually toxic when used long term or in high dosage. The ability to learn beforehand whether the patient will respond to the drug or is resistant to the drug is extremely beneficial.

In addition to cutting down costs and time of treatment, the prevention of adverse drug reaction (ADR) can also be prioritized through personalized medicine. Personalized medicine can be extremely helpful in a specific type of ADR called Idiosyncratic ADR where reactions to drugs are not directly tied to dosage. However, it is hypothesized that the effects are due to genetic variation of the genes for enzymes specific for drug metabolism (Adams, 2008). Aside from physical danger of adverse drug reactions, the costs associated with ADRs can be extremely high. By applying personalized medicine and considering the patient's genetic information before a treatment administration, both safety and costs can be taken into consideration at the same time. In conclusion, much more research is needed for this topic, especially research on the associations between specific genes and drugs and how to better use gene expression data to make personalized responses to treatments.

References

1. Mishra A, Verma M: Cancer biomarkers: are we ready for the prime time? Cancers (Basel) 2010, 2:190–208
2. Geeleher, Paul et al. “Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines.” Genome Biology 15 (2013): R47 - R47.
3. U.S. National Library of Medicine. (2020, August 18). Gli3 gene: Medlineplus genetics. MedlinePlus. Retrieved April 25, 2022, from https://medlineplus.gov/genetics/gene/gli3/#conditions
4. Matissek, S. J., & Elsawa, S. F. (2020). GLI3: a mediator of genetic diseases, development and cancer. Cell Communication and Signaling, 18(1), 1-20.
5. Knut & Alice Wallenberg foundation. (2021). The human protein atlas. The Human Protein Atlas. Retrieved May 3, 2022, from https://www.proteinatlas.org/
6. Cule, E., & De Iorio, M. (2012). A semi-automatic method to guide the choice of ridge parameter in ridge regression. arXiv preprint arXiv:1205.0686.
7. Adams, J. U. (2008). Pharmacogenomics and personalized medicine. Nature Education, 1(1), 194.
8. Matissek, S. J., & Elsawa, S. F. (2020). GLI3: a mediator of genetic diseases, development and cancer. Cell Communication and Signaling, 18(1), 1-20.
9. Shan, Y., Huang, Y., Lee, A. M., Mentzer, J., Ling, A., & Huang, R. S. (2021). A Long Noncoding RNA, GAS5 Can Be a Biomarker for Docetaxel Response in Castration Resistant Prostate Cancer. Frontiers in oncology, 11, 675215.
10. U.S. National Library of Medicine. (n.d.). Home - Geo - NCBI. National Center for Biotechnology Information. Retrieved April 2022, from https://www.ncbi.nlm.nih.gov/geo/
11. Sanger Institute. (n.d.). Home. Catalogue of Somatic Mutations in Cancer. Retrieved April 2022, from https://www.sanger.ac.uk/

Appendix:

Table 1: AUC values for different methods for both outcomes

Methods/Outcome	**Docetaxel**	**Bortezomib**
Ridge (paper)	0.814	0.711
Ridge (glmnet)	0.814	0.714
LASSO	0.850	0.581
Elastic net	0.836	0.581
SVM	0.836	0.677
ANN	0.821	0.579
PCR	0.75	0.591

Figure 1: ROC curves for different methods for outcome docetaxel

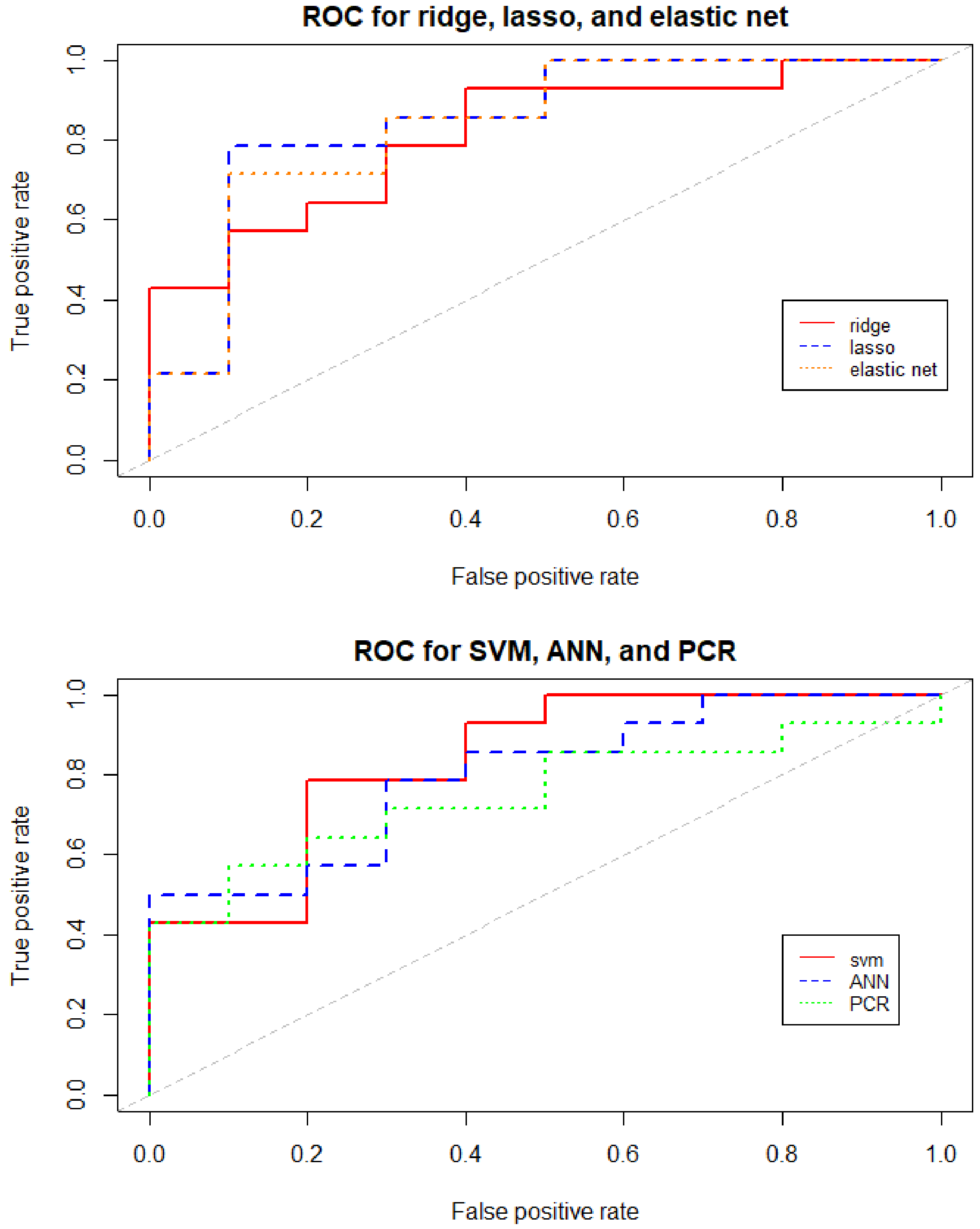
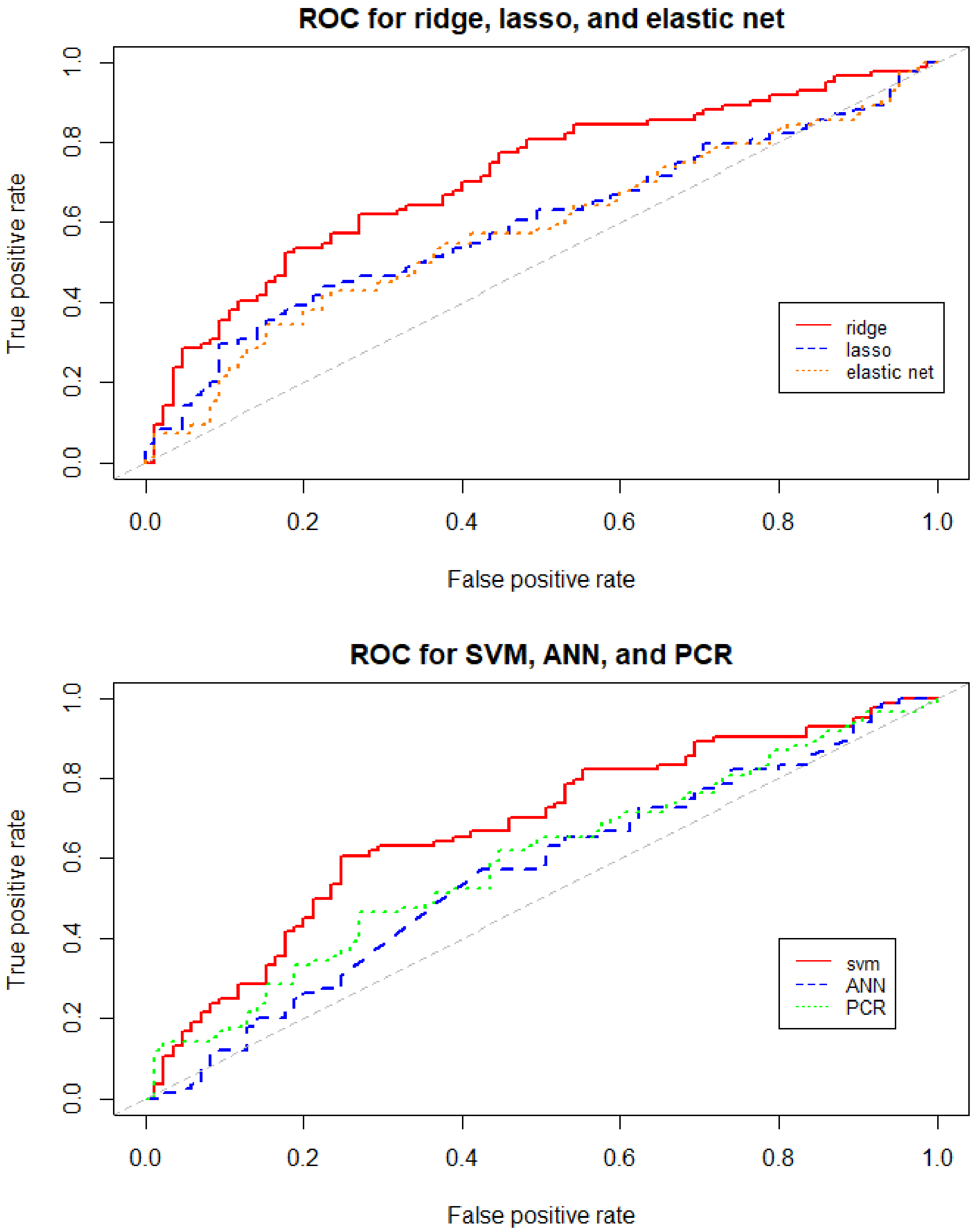


Figure 2: ROC curves for different methods for outcome bortezomib


ROC for ridge, lasso, and elastic net
True positive rate
False positive rate
ridge
lasso
elastic net
ROC for SVM, ANN, and PCR
True positive rate
False positive rate
svm
ANN
PCR