EN.601.448/648 Computational genomics: Final Project

Elysia Chou (echou4)
BME

Keefer Chern (kchern1)
BME & CS

Alexander Chang (achang56)
BME & CS

June 3, 2019

Due: 5/1/2019

1 Problem

Cancer cell lines are often used in cancer biology as a method to study in vitro characteristics of a disease that affects millions every year. However, there is much work to be done on studying the relationship between the (epi)genomic profiles of cancer cell lines and their respective drug sensitivity. Such relationships, if found, could provide insights into future directions of in vitro cancer pharmaceutics development[1]. The Cancer Cell Line Encyclopedia (CCLE) provides extensive data to study this relationship.

In this project, our aims were two-fold: we used either histone modification profiles or gene expression of around 500 cancer cell lines to predictively model drug sensitivity to the drug PD-0325901, which is a MEK inhibitor. The basis of this project comes from the the paper "The Cancer Cell Line Encyclopedia enables predictive modeling of anticancer drug sensitivity," where they use gene expression, mutation, and copy number data from the CCLE to model drug sensitivity of the roughly 500 cancer cell lines using an elastic net regression[2]. One of the limitations of this paper is that the authors selected features based on the whole dataset rather than just the training set when predictively modeling the drug sensitivity of drug PD-0325901[2].

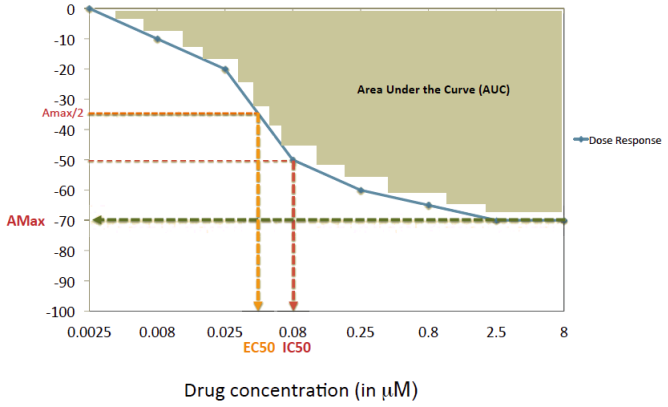
Therefore, our approach is meant to provide a possible improvement upon their results for the case of predicting PD-0325901 drug sensitivity. One of the ways we propose to improve upon their results is to properly hold out a testing set and train on a subset of the selected datasets. Using an elastic net regression model with epigenomic data and comparing it with a similar model using gene expression data would prove an interesting comparison to explore the possibilities of using the unique and promising global chromatin profiling dataset[3,4]. Previous work has suggested that epigenomic data, and more specifically, the global chromatin profiling dataset, can lead to discoveries of new therapeutic targets and drug associations[4,5]. As such, we will build an elastic net regression model that will predict drug response (activity area, A_{max}, EC50, or IC50) based off of epigenomic or genomic features and compare the two. We will also compare our gene expression model's top features with those generated by Barretina et. al's elastic net model to examine whether their results could be validated in a more rigorous way.

2 Data

The data used for our project is the global chromatin profiling (GCP) data as well as gene-centric RMA-normalized mRNA expression data (with Entrez Gene IDs) from CCLE (`https://portals.broadinstitute.org/ccle/data`). The GCP data contains information about levels of 42 different post-translational histone modifications obtained from mass spectrometry of the histone modifications in question over the whole genome (hence the term "global")[5].

Additionally, the CCLE also provides a dataset that has labeled drug response data for each cancer cell line, including IC50, EC50, A_{max}, and activity area/AUC (see figure 1[6] for a visual explanation, where the y axis is relative growth inhibition(%)).

Figure 1: Drug response curve



3 Methods

3.1 Pre-processing

To start off, we extracted our data files which include GCP features, gene expression features, and labels. There are 504 cell lines that have drug response data for PD-0325901, of which there are $n_{GCP} = 458$ cell lines overlapping with the GCP data and $n_{expr} = 491$ cell lines overlapping with the gene expression data. Our labels data is an $n \times 4$ matrix, that has n cell lines as the rows and 4 different types of labels, namely EC50, IC50, Amax, and ActArea (activity area). Our GCP feature data is a 458×42 matrix, which similarily has cell lines as rows and 42 different histone modifications as the columns. Our gene expression feature data is a 458×18988 matrix that has expression for 18,988 genes.

The next step was to process the extracted data. For both GCP and gene expression labels, the EC50 column had a large proportion of values that were not recorded (null values); we decided to drop EC50 as a label that we would use in this project. For our GCP feature data, it contained a total of 392 null values. Of those null values, 364 were contained over a span of three columns. Therefore, we decided to remove those columns from further analysis in this project. For the the rest of the null values, we replace them with the mean of the values in their respective columns.

For our gene expression feature data, we did not have any missing values. Looking at the shape of the label data, we decided to log transform activity area so it took on a more normal shape.

3.2 GCP and drug sensitivity modeling

After processing the data and splitting it into the 80% training and 20% test set, we followed Barretina et al.'s method by first performing the hyperparameter tuning for an elastic net regression. The hyperparameters that we tuned for were α, the mixing parameter, and λ, the regularization term (a λ value of 0 is just regular linear regression).

We performed a 10-fold cross-validation on the training set to determine the optimal α and λ. We used 17 values of α from $[0.2, 1.0]$ and 10 values of $\lambda \in \{0.0, 0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10\}$. We performed regression by regressing one label at a time. We obtained the average mean squared error (MSE) of the validation and training set and the average R^2

value of the cross-validation training set to determine the optimal λ and α for each label. Due to the bad MSE and R^2 obtained from the training data, we hypothesized that there may be no correlation between the GCP data and PD-0325901. Thus, we decided to look more at the data itself and see if we can find relations with another method.

First, we performed linear regression on each histone modification independently on each of the 3 labels. We performed Benjamini Hochberg multiple hypothesis correction to see which histone modification sites were correlated with the three labels. In addition, we computed the Spearman correlation between features and performed PCA to find the top 10 principle components to compute the pairwise Pearson correlation coefficient between each of the principal components and the labels.

3.3 Gene expression and drug sensitivity modeling

We performed a similar protocol with the gene expression data: after processing the data and splitting it into the 80% training and 20% test set, we first performed the hyperparameter tuning for an elastic net regression. We set the tolerance in elastic net regression to 0.01 because the default tolerance for Sklearn elastic net did not converge.

Again, we performed a 10-fold cross-validation on the training set to determine the optimal α and λ using the same 17 values of α and 10 values of λ (see section 3.2). We performed regression by regressing one label at a time. We obtained the average MSE of the validation set and the average R^2 value of the cross-validation training set to determine the optimal λ and α for each label. We manually selected the λ and α for a model that limited the overfitting, but still fit the data well.

Next, we performed bootstrapping on our training data. We generated 200 re-sampled datasets by sampling from the training data with replacement. Each re-sampled dataset consists of 250 examples, about (1-1/e) or 63% of the training data size. We then used elastic net on each bootstrapping dataset with the optimal λ and α found previously to create a model. The regression coefficients for each model were used to create a matrix where each row represents the solution for one bootstrapped dataset and each column represents the weight of a feature. With this matrix, we calculated the percentage of times each feature had a nonzero weight across all the bootstrapped datasets. The top ten features with the highest bootstrapped frequency were selected as significant. Finally, we ran each of the 200 boot-

strap models on the test data and computed the MSE between the true label and the average output vector of the 200 models applied to the test data.

4 Results

4.1 GCP and drug sensitivity modeling

For each of the 3 labels, we looked at the combinations of λ and α that yielded the ten lowest MSEs and the top ten R^2 values. In figure 2, the top 5 lowest MSEs for each label is shown. After running 10-fold cross-validation with all possible combinations of alpha and lambda outlined in our methods section, we have discovered that all our R^2 values are $\lessapprox 0.1$. The highest R^2 was obtained by using activity area (ActArea) as our label, namely $R^2 = 0.112$. This was the R^2 value associated with mean $MSE = 2.312$, $\alpha = 0.20$ and $\lambda = 0.001$. This low λ indicates that the model prefers no regularization as it was most likely underfitting and the low α indicates that our elastic net model prefers very little L1 regularization. In other words, our model does not want to make weights go to zero. When compared to a linear regression model without regularization, we obtained a mean $R^2 = 0.14473$ and mean MSE of 2.347.

Figure 2: Hyperparameters obtained with GCP data

IC 50

	Lambda	Alpha	MSE Test	MSE Train	R2
101	0.3	1.00	12.993113	12.524075	0.021803
84	0.1	1.00	13.019230	11.893680	0.071047
83	0.1	0.95	13.025692	11.879694	0.072139
100	0.3	0.95	13.032337	12.536065	0.020867
82	0.1	0.90	13.036028	11.866107	0.073199

A_max

	Lambda	Alpha	MSE Test	MSE Train	R2
118	1.0	1.00	841.694422	792.672591	0.041268
135	3.0	1.00	842.483756	827.473250	-0.000799
117	1.0	0.95	851.585493	800.317666	0.032023
116	1.0	0.90	868.419184	814.514103	0.014848
100	0.3	0.95	880.875913	761.221342	0.079336

Activity Area

	Lambda	Alpha	MSE Test	MSE Train	R2
84	0.1	1.00	2.086050	2.020410	0.016703
83	0.1	0.95	2.087535	2.018359	0.017698
82	0.1	0.90	2.090324	2.015804	0.018938
81	0.1	0.85	2.093176	2.013152	0.020227
80	0.1	0.80	2.096937	2.010303	0.021611

We also performed linear regression on each histone modification independently with respect to each of our 3 labels and found no features with significant p-values after Benjamini Hochberg multiple hypothesis correction. Furthermore, plotting the top 2 principal components did not show any distinction between the features and labels; nor did the Spearman correlation heatmap show any strong monotonic correlation between the features (fig. 3).

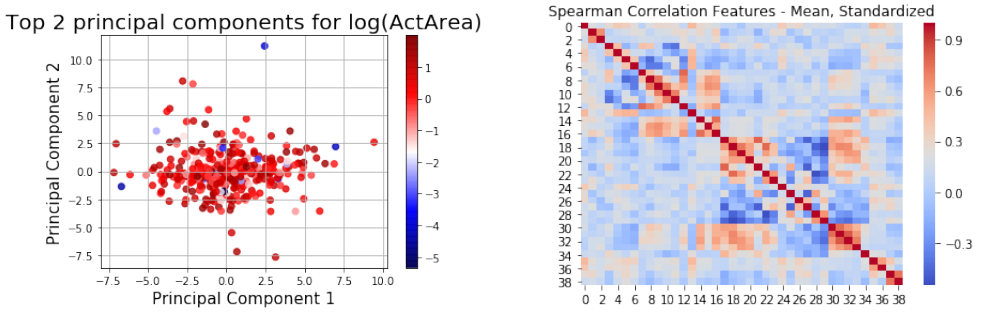


Figure 3: Left: PCA. Right: Spearman correlation heatmap

Overall, our results indicate that our features (histone modifications) are not linearly correlated with our drug sensitivity.

4.2 Gene expression and drug sensitivity modeling

Similar to GCP, we looked at the combinations of λ and α that yield the 10 lowest testing MSEs for all labels. The top 10 combinations all had a R^2 greater than 0.5 and we manually selected the hyperparameter combination that had a Train MSE that did not overfit with high R^2 to show that there was a good correlation. The final hyperparameters were $\alpha = 0.8$, $\lambda = 0.03$ for activity area, $\alpha = 1.0$, $\lambda = 0.3$ for IC50, and $\alpha = 1.0$, $\lambda = 1.0$ for A_{max}. Our top 5 combinations with the smallest MSEs are shown in figure 4.

Figure 4: Hyperparameters obtained with gene expression data

	Lambda	Alpha	MSE Val	MSE Train	R2
135	3.0	1.00	535.081636	317.188125	0.630173
118	1.0	1.00	550.770765	68.105051	0.920636
117	1.0	0.95	610.349074	69.802693	0.918654
101	0.3	1.00	635.723392	7.717824	0.991002
134	3.0	0.95	645.080313	358.647555	0.581793

A_max

	Lambda	Alpha	MSE Val	MSE Train	R2
101	0.3	1.00	8.542118	4.143462	0.675630
100	0.3	0.95	8.604842	3.955840	0.690318
99	0.3	0.90	8.668106	3.755339	0.706022
98	0.3	0.85	8.760968	3.548933	0.722189
97	0.3	0.80	8.895009	3.360213	0.736963
84	0.1	1.00	9.030159	0.789759	0.938199

IC 50

	Lambda	Alpha	MSE Val	MSE Train	R2
67	0.03	1.00	0.158408	0.047683	0.813496
66	0.03	0.95	0.159082	0.044421	0.826257
65	0.03	0.90	0.160146	0.041204	0.838838
64	0.03	0.85	0.161186	0.037957	0.851542
63	0.03	0.80	0.162848	0.034817	0.863821

Activity Area

The top 10 genes that show up during bootstrapping for each label are shown in figure 5. For future analysis, we will only be looking at the model

predicting activity area as it gave us the lowest MSE and highest R^2 values out of the three labels. The gene name, bootstrapped frequency, and average weight across 200 bootstraps is reported in table 1.

Our final Train MSE and Test MSE are 0.039 and 1.63 respectively.

Figure 5: Top 10 genes from bootstrapping for each label

	Feature	Pct Sigf
3837	153478_at	0.835
7924	29978_at	0.745
16646	84102_at	0.535
2261	114885_at	0.520
9033	388272_at	0.485
10794	51619_at	0.445
15343	7837_at	0.375
16306	81928_at	0.375
8637	3569_at	0.365
9232	3995_at	0.360

IC50

	Feature	Pct Sigf
15505	79192_at	0.715
7924	29978_at	0.680
3837	153478_at	0.650
836	100507224_at	0.645
14195	6506_at	0.625
8055	3096_at	0.615
4361	1748_at	0.600
1957	11147_at	0.595
2261	114885_at	0.575
10794	51619_at	0.575

A_{max}

	Feature	Pct Sigf
1218	10253_at	1.000
2061	112616_at	0.995
6772	27006_at	0.815
4862	2118_at	0.800
2216	114757_at	0.775
2488	1192_at	0.775
3254	140825_at	0.730
5003	2207_at	0.705
2129	11322_at	0.695
7908	29952_at	0.645

Act Area

Table 1: Top 10 genes from bootstrapping for activity area

Gene ID	**Gene Name**	**Frequency**	**Ave Weight**
10253_at	SPRY2	1	0.655
112616_at	CMTM7	0.995	0.728
27006_at	FGF22	0.815	0.464
2118_at	ETV4	0.8	0.723
114757_at	CYGB	0.775	0.697
1192_at	CLIC1	0.775	0.445
140825_at	NEURL2	0.73	0.482
2207_at	FCER1G	0.705	0.649
11322_at	TMC6	0.695	0.492
29952_at	DPP7	0.645	0.450

5 Discussion

5.1 GCP and drug sensitivity modeling

We were not able to find any correlation between our GCP data with drug sensitivity to the MEK inhibitor PD-0325901. There are many reasons why we think such results were obtained. The first is that we do not have enough features for patterns or relations to be captured. In addition, in this specific case, these global abundances of histone modifications may not be representative of the level of effects of a signalling pathway inhibitor. In other words, globally measured post-translational histone modifications may be such a downstream effect that patterns may be muffled by confounding factors. In this perspective, histone modifications may still correlate with drug sensitivity to the MEK inhibitor PD-0325901, but we just do not have the data to show it. Another reason for our results is that there could be no correlation that exists between histone data and drug sensitivity.

5.2 GCP and drug sensitivity modeling

We were able to find a correlation between our gene expression data with drug sensitivity to the MEK inhibitor PD-0325901 when choosing the hyperparameters. The paper did not state their chosen hyperparameters, so we have no basis for comparison, but we do know that they chose the hyperparameters that resulted in the smallest Test MSE values. In relation to the results we obtained, this would mean that their alphas would be equal to 1. What that would mean in terms of an elastic net model is that only the effects of the L1 regularization term will be considered.

When looking at which genes came up frequently during the bootstrap for predicting activity area, we noticed that the top 3 genes are related to cancer progression[7]:

- SPRY2 is a regulator of MAPK output.
- CMTM7 is a tumor suppressor gene.
- FGF22 is involved in mitogenic and cell survival activities.
- The other genes listed in table 1 are not all necessarily related with cancer progression; furthermore, their bootstrapped frequency is less than 80%, which was the cutoff in Barretina et al.'s study.

Looking at Barretina et al.'s results for validation, we found that 2 genes overlapped with their 11 genes that had a bootstrapping frequency greater than 80%[2]. These genes were SPRY2 and NEURL2 (which regulates myofibril organization). Interestingly, however, we found that their reported weights were the opposite sign of our average weight. When computing the Pearson correlation between standardized SPRY2 expression and activity area, however, our Pearson correlation coeffecient was 0.015 with a p-value of 0.741. Furthermore, to see if the α parameter may be affecting the weight, we ran the bootstrap with $\alpha = 0$ but still got a positive average weight for SPRY2. Thus, there was a positive correlation from the three methods we used; but since the Pearson correlation was relatively low and the p-value so high, we cannot draw strong conclusions.

Overall, we have shown that there is promise for gene expression profiles to help predict drug sensitivity in cancer cell lines. Further work would more extensively explore this relationship across multiple drugs with various effects.

6 Contribution by each team member

- Elysia Chou led the team with her expertise on the biological side of the project and contributed to the data processing, PCA analysis, and the writing of the problem, data, and discussion sections of the reports as well as editing and compiling all the final deliverables.

- Keefer Chern used his background in both Computer Science and BME to contribute to data processing, regression and correlation analysis (using QQ plots, heatmaps, etc.), writing and editing all the reports, and debugging.

- Alexander Chang led the technical side of the project with his background in machine learning. He contributed to the model selection, hyperparameter tuning, regression analysis, bootstrapping, and the writing of the methods, results, and discussion sections of the report as well as editing and compiling all the final code to be submitted.

7 Reference

1. Gillet, J. P., Varma, S., & Gottesman, M. M. (2013). The clinical relevance of cancer cell lines. Journal of the National Cancer Institute, 105(7), 452-458.

2. Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., ... & Reddy, A. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature, 483(7391), 603.

3. Minshull, T. C., & Dickman, M. J. (2014). Mass spectrometry analysis of histone post translational modifications. Drug Discovery Today: Disease Models, 12, 41-48.

4. Litichevskiy, L., Peckner, R., Abelin, J. G., Asiedu, J. K., Creech, A. L., Davis, J. F., ... & Jaffe, J. D. (2018). A library of phosphoproteomic and chromatin signatures for characterizing cellular responses to drug perturbations. Cell systems, 6(4), 424-443.

5. Jaffe, J. D., Wang, Y., Chan, H. M., Zhang, J., Huether, R., Kryukov, G. V., & Stegmeier, F. (2013). Global chromatin profiling reveals NSD2 mutations in pediatric acute lymphoblastic leukemia. Nature genetics, 45(11), 13861391. doi:10.1038/ng.2777

6. Rahman, R., & Pal, R. (2016, February). Analyzing drug sensitivity prediction based on dose response curve characteristics. In 2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI) (pp. 140-143). IEEE.

7. https://www.ncbi.nlm.nih.gov/gene/

8. Data retrieved from: https://portals.broadinstitute.org/ccle