CTD2 BeatAML DREAM Challenge: A Multitask Model for Predicting Drug Responses

Ruibo Zhang, Daniel Nolte, Saugato Rahman Dhruba

Abstract

Our model utilizes multitask bagged linear regression with post covariance alignment of the prediction values to predict the area under the curve (AUC) values of individuals inhibitor responses from genomic features.

Introduction

In subchallenge 1, the AUC values of tissue specimens after the application of 122 inhibitors are given for a training set where the goal is to predict the AUC values for the target Leaderboard set using relevant genomic and/or clinical features. In Figure 1, the AUC values are plotted against their ranking within each of the selected inhibitors after subtracting the mean. From Figure 1, we observe significant nonlinearity in the AUC distributions for all the selected inhibitors, specifically in the tails, inferring that the highly responsive and highly resistant specimens exhibit extreme AUC values. This inspired us to use the normalized ranking of the AUCs instead of the actual values to better utilize a linear model, which have been shown to perform well in anticancer drug sensitivity prediction[[1]](#endnote-1).

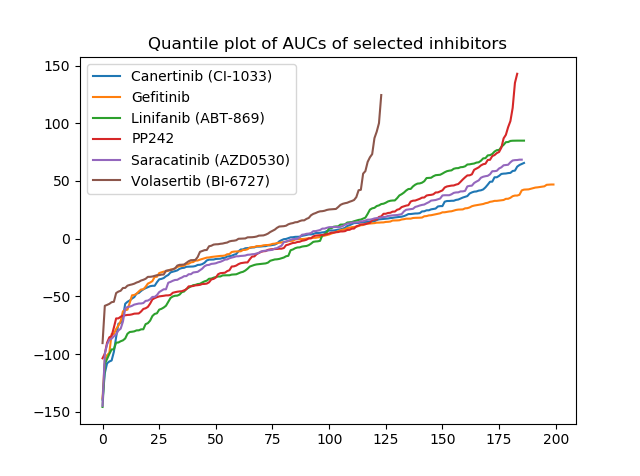


Figure 1. Distribution of the AUC values of 6 selected inhibitors shown in order

Cancer drug design usually follows the logic of pathway control and relationships between chemical groups and biochemical reactions. Therefore, similar inhibitors are assumed to cause similar responses in similar cell specimens. This can be represented by the covariance structures of the AUC values for different inhibitors. The covariance heatmap for the AUC values of 122 inhibitors are shown in Figure 2. Inhibitors are first clustered according to their dissimilarity measures (*i.e.*, 1 – correlation) and shown following the clustering order. As shown in Figure 2, the correlations between inhibitor-pairs show a clear pattern *i.e.*, the clustered inhibitors have high concordance of up to 0.8 while the non-clustered ones have poor concordance demonstrated by values as low as -0.3.

It is, therefore, reasonable to assume that the predicted AUC values for the inhibitors should preserve this covariance structure. To pass the original covariance structure to the predictions from a predictive model, we adopted a covariance alignment method, CORAL, which is well accepted in the cross-domain learning [[2]](#endnote-2). To elaborate, we first remove the covariance structure from our initial model predictions for all inhibitors through “whitening” (*i.e.*, dividing by the square root of the covariance matrix), and then perform “re-coloring” of the whitened prediction values (*i.e.*, multiply by the square root of the covariance matrix of the training set). From another aspect, we are in effect minimizing the higher order prediction residuals (covariance, in the order of 4), besides the mean square error (MSE) in linear regression, which is in the order of 2.

To utilize the multitask intrinsic qualities of our task, we use bagged linear regression as the base model. Bagging compensates the instability of simple linear regression and provides improved prediction accuracy.

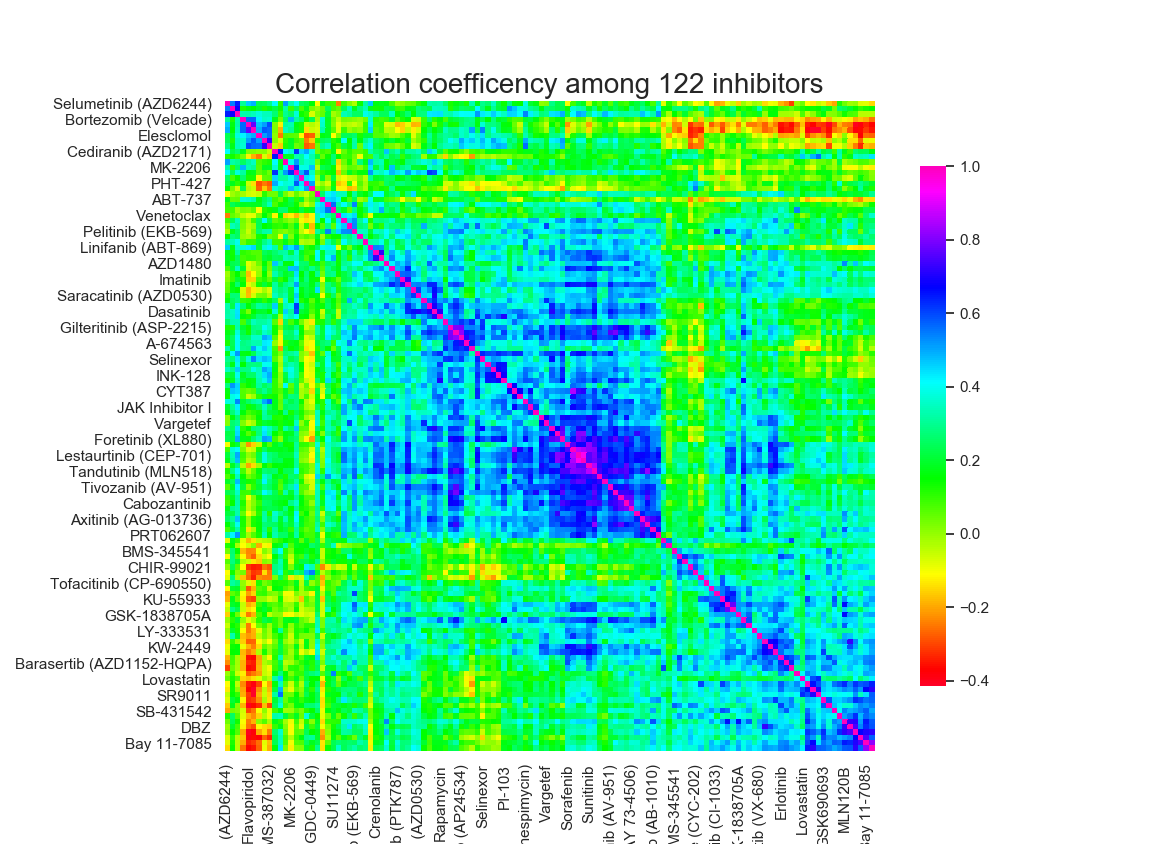


Figure 2. Correlation ecoefficiency among 122 inhibitors

Methods

Data preprocessing

Both training and leaderboard sets are preprocessed in the way described below. In SC1, RNA-seq, DNA-seq, and clinical data (numerical/categorical) are provided as input for modeling. However, we only used RNA-seq gene expression values as predictors. First, the expression values are normalized by dividing the Euclidean norm of each specimen, followed by a feature selection process. Next, the expression values are standardized to achieve zero mean and unity variance for each gene (feature) before modeling. AUC values are provided as the target, however, each inhibitor is not tested on every available specimen, in our case, specimens with more than 30 missing values are dropped out, and the remaining missing values are imputed using KNN imputation with K = 5. Instead of using the actual values directly, we used the normalized AUC rankings of the specimens as the target after subtracting the mean normalized ranking for each inhibitor before model fitting.

Feature selection

In the given RNA-seq data, expression values for 63,677 genes are available which is too high for modeling. First, the dataset is filtered to keep the 30,000 genes with largest correlation with the AUC values. Then, a ridge regression model is fitted and the features with the highest weights are kept. In our case, ~20,000 features are kept and put into the bagged linear regression model for training.

Regression model

A bagged linear regression estimator (M = 5000) is used as our predictive model. To avoid overfitting, we only use a subset of total features (<115) for training each of the 5000 regressors.

Postprocessing

To reduce the covariance discrepancy between training and predicted AUC values, we used the method of covariance alignment (CORAL) described by Sun et al[[3]](#endnote-3). For efficiency, stability, and to avoid matrix singularity, we simply add a regularization term to the covariance matrix. First, predictions are whitened *i.e.*, divided by the square root of its covariance and then re-colored *i.e.*, multiplied by the square root of training data covariance. This can be described as:

where is the regularized prediction covariance with as the regularization parameter, is the regularized training AUC covariance, and is the pristine prediction. For this project, we have used .

Discussion

As shown in Table 1, 9 models are evaluated by 5 fold cross validation using both training and leaderboard data. First, Ridge regression and bagged linear regression on actual AUC values are shown as the baseline. AUC rankings are used instead of AUC values in models #3 and #4. Further, post prediction alignment is shown as model #5 – 8. As a comparison, Bagged LR with Ridge feature selection but not postprocessing is given as #9.

Table 1. Summary of model performances

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Training values | Features kept after 1st feature selection | Features after 2nd feature selection | Post processing | Regression model | Model parameters | Spearman correlation | Pearson correlation |
| #1 | values | 30k | - | - | Ridge |  | 0.322344 | 0.318868 |
| #2 | values | 30k | - | - | Bagged LR | Max feature portion = 0.005 | 0.323710 | 0.312633 |
| #3 | ranks | 30k | - | - | Ridge |  | 0.327609 | 0.314593 |
| #4 | ranks | 30k | - | - | Bagged LR | Max feature portion = 0.005 | 0.328202 | 0.314058 |
| #5 | ranks | 30k | - | CORAL | Ridge |  | 0.336019 | 0.323188 |
| #6 | ranks | 30k | Ridge | CORAL | Ridge |  | 0.330447 | 0.320249 |
| #7 | ranks | 30k | - | CORAL | Bagged LR | Max feature portion = 0.005 | 0.333266 | 0.323085 |
| #8 | ranks | 30k | Ridge | CORAL | Bagged LR | Max feature portion = 0.005 | **0.341710** | **0.326697** |
| #9 | ranks | 30k | Ridge | - | Bagged LR | Max feature portion = 0.005 | 0.333703 | 0.317601 |

References

1. Barretina, Jordi, et al. "The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity." Nature 483.7391 (2012): 603-607. [↑](#endnote-ref-1)
2. Sun B., Saenko K. (2016) Deep CORAL: Correlation Alignment for Deep Domain Adaptation. In: Hua G., Jégou H. (eds) Computer Vision – ECCV 2016 Workshops. ECCV 2016. Lecture Notes in Computer Science, vol 9915. Springer, Cham [↑](#endnote-ref-2)
3. Sun, Baochen, Jiashi Feng, and Kate Saenko. "Return of frustratingly easy domain adaptation." *Thirtieth AAAI Conference on Artificial Intelligence*. 2016. [↑](#endnote-ref-3)