Using Copy Number, Gene Expression, Methylation, and Tissue Type to Predict Efficacy of Chemotherapies for Diverse Primary Tumors

SCHOOL OF PUBLIC HEALTH
BIG DATA SUMMER INSTITUTE
UNIVERSITY OF MICHIGAN

Christopher Archer, Ritoban Kundu, Corriene Sept, Jacob Shkrob

Cornell University, Indian Statistical Institute, Case Western Reserve University, McGill University

Introduction

- Cancer is the abnormal growth of cells capable of spreading to other parts of the body. Cancer is the 2nd leading cause of death in the United States[1]. In addition, it is estimated that 38.4% of Americans will be diagnosed with cancer at some point in their lives[2].
- Sensitivity to chemotherapy is patient-specific and hypothesized to be heavily influenced by the genetic makeup of the tumor.
- Chemotherapies are incredibly toxic to healthy tissue, as well as cancer cells. Minimizing damage to healthy tissue and harmful side effects by choosing the best chemotherapy first is essential to improving patient outcomes.
- Drug effectiveness can be confounded by the tissue location of the tumor

Objective: To classify chemotherapies as effective or ineffective for a particular cancer cell line given the site of a patient's tumor and the methylation, copy number, and gene expression of cancer cells in the primary tumor. In addition, increase the interpretability of the classifier using estimated coefficients from ensemble models.

Table 1: Sample sizes of genetic data by drug after data cleaning and imputation.

		Number of Covariates		
Drug	Number of Patients / Drug	Copy Number	Expression	Methylation
1001	152	45179	17,737	475,915
1006	269	43410		
1007	246	42861		
1008	364	43801		
1011	328	45159		
1014	246	42551		
1015	166	44152		
1016	233	44143		
1026	306	41550		
1037	115	44479		
1053	170	42635		
1054	243	43608		
1058	191	42490		
1060	251	42128		
1066	179	42561		

Materials

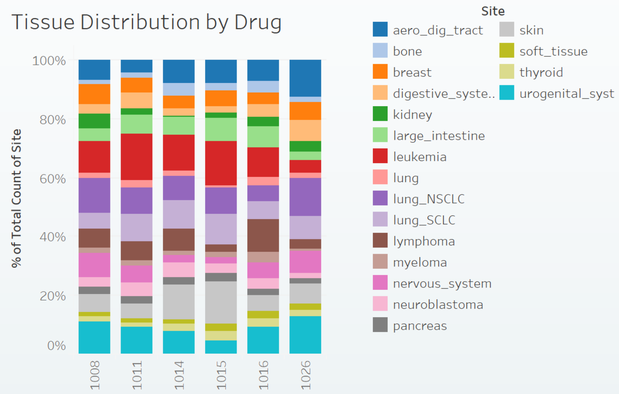
Study Population: Between 115 and 364 established cancer cell lines (number depends on the chemotherapy) from the Genomics of Drug Sensitivity in Cancer (GDSC) database.

Variables of Interest:

Outcomes: We defined our outcomes as binary: effective or ineffective for a particular chemotherapy applied to a specific cancer cell line. Effective drugs had a relative IC50 less than 0.2 and ineffective drugs had a relative IC50 greater than 8.0.

Predictors: Copy number, gene expression, and methylation of the cancer cell lines tested with the given treatment. In addition, categorical data describing site of the primary tumor.

Figure 1: Distribution of Tumor Site by Drug for Drugs 1008, 1011, 1014, 1015, 1016, 1026.



Methods

Statistical Approach:

- *Missing Data*
  - Methylation: Columns with missing data in the methylation dataset (<5% of columns) were removed
  - Copy Number: 21.2% of genes contained missing values. Genes missing more than 5% of observations, followed by insignificant genes with missing values, were removed. Imputation was performed for remaining genes using missForest[3] (at maximum, 10% of genes) using the assumption of missing at random
- *Dimensionality reduction*
  - Differential Expression: Selected genes with lowest p-values from two-sample t-test
  - Principal Component Analysis: Used singular value decomposition (SVD) and 5-fold cross-validation to obtain the best q principal components and model parameters
- *Combination of data sets*
  - To eliminate collinearity between datasets, we used orthogonal projections of the first *m* principal components from each dataset
  - Note: For this dataset we did not use missForest for imputation on Copy Number but instead performed SVD on CC' to obtain pc-scores

Classification while accounting for tissue type:

- Different drugs have different proportions of cancer tissues they screen against (Figure 1), so we wanted to know if **utilizing the cancer tissue type in our classification would yield significant improvements**
- We re-centered our genomic data by subtracting the mean value for each tissue type and used **one-hot encoding** on our tissue type data to create an indicator matrix showing the tissue type of each cell line
- This procedure in effect **regresses out** the effect of tissue type
- Fit classification models (LDA, Random Forest, SVM, Lasso, Ridge, Naïve Bayes, Logistic Regression) on both tissue-centered data and indicator matrix, then used the probability vectors from these methods on the testing set using two procedures:
  - **Logistic Regression:** Using the genomic data and indicator matrix as covariates, we fit a logistic regression model
  - **Weighted Average:** Combined information from genetic data (coefficient: **α**) and from tissue type (coefficient: **1 - α**) to predict efficacy of treatments. Alpha was obtained through fitting different values and choosing the alpha which minimized error

Results

Using the weighted average method for centered gene expression generally yields similar or slightly improved error rates, compared to running the best classifier on the uncentered combined data and the centered combined data. In addition, the weighted average method increases interpretability and enables us to separate out the influence of tissue type and genetics on classifying efficacy of chemotherapies. Furthermore, the median alpha coefficients vary from drug to drug, suggesting different **levels of gene-drug interaction** (high α corresponds to higher weighting of genetics data compared to tissue type with regard to classification.)

Table 2: Median α coefficients and average error rates for centered combined, uncentered combined, and centered expression models. Alpha corresponds to the weight of the genetics variables using the weighted average method. Red error rates indicate the best error rate found for each drug using each of the three methods tested.

Drug	Median Alpha, Centered Combined	Average Error (Combined, Centered)	Average Error (Combined, Uncentered)	Median Alpha, Centered Expression	Average Error (Expression, Centered)
1001	0.64	0.2074	0.1267	0.35	0.1407
1006	0.00	0.2113	0.1296	0.56	0.1774
1007	0.36	0.1551	0.1400	0.1	0.1429
1008	0.00	0.0493	0.0548	0.62	0.04
1011	0.42	0.1242	0.0758	0.23	0.1061
1014	0.68	0.1106	0.0800	0.36	0.0596
1015	0.47	0.1857	0.1176	0.31	0.1143
1016	0.00	0.1489	0.0936	0.05	0.1277
1026	0.01	0.1475	0.0645	0.26	0.1016
1037	0.00	0.1300	0.1304	0.14	0.06
1053	0.00	0.2563	0.2059	0.58	0.1875
1054	0.00	0.2298	0.1837	0.24	0.1745
1058	0.54	0.2944	0.2308	0.49	0.2444
1060	0.44	0.225	0.0784	0.41	0.1083
1066	0.00	0.2229	0.1111	0.17	0.1400

Results Continued

The following averaged ROC curves show the results for drug 1008. **Figure 2** shows the ROC curves using the best models found on the uncentered data sets. **Figure 3** shows 2 different ROC curves for each data set: the *dotted* ROC curve shows the results of running the best classifiers on centered genomic data, while the *solid* ROC curve shows results from weighted average method discussed in the methodology.

Figure 2: ROC Curve for Non-Centered Data, Drug 1008.

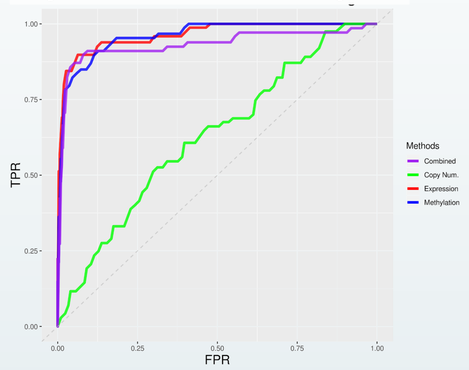
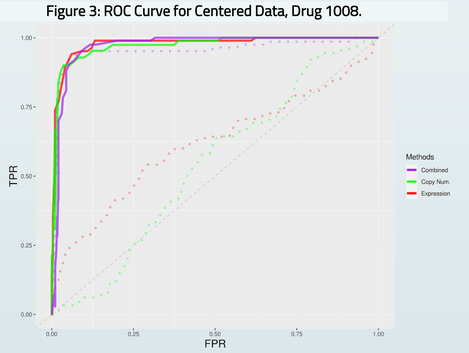


Figure 3: ROC Curve for Centered Data, Drug 1008.



Conclusion

- We determined a more interpretable model using tissue-centered and weighted average methods, with comparable errors to the uncentered combined method.
- This tissue-centered model enables us to determine whether our classification algorithm is operating solely off tissue type or a combination of tissue type and genetic data.
- This information will better inform future research into drug-gene interactions and predicting efficacy of cancer treatments.
- Future directions would include incorporating more robust outcomes (obtained through a method such as bimodal mixture model), improving accuracy by using five-fold cross validation to select the best α in our centered combined model, and using a more complex weighted average of the probabilities from the top 5 classifiers (SVM, Lasso, LDA, etc)

References & Acknowledgements

1. Centers for Disease Control and Prevention. Leading Causes of Death (2017).
2. National Cancer Institute. Cancer Statistics (2018).
3. https://cran.r-project.org/web/packages/missForest/index.html