An Integrative Analysis of Anti-Cancer Drug Response

Bo Wang **Yuchi Liu**

Yusi Chen

Abstract

In this project, we perform an integrative analysis of 24 kinds of anti-cancer drug response on around 400 cancer patients. First, we use Lasso to build a prediction system for each individual drug response. Then a graph Lasso is employed to utilize the correlation between multiple tasks in order to build a more robust prediction system. Finally, we further perform supervised and unsupervised categorization of patients into multiple cancer types. Our analysis shows gene expression is most discriminative and useful in cancer categorization and drug response prediction.

1 Introduction

Human cancer cell lines contain valuable information with respect to tumour biology and drug discovery. Annotated with both genetic and pharmacological data, cell-line panels are widely used to address the problem of a tumour lineage or across multiple cancer types. However, these studies are limited in their depth of genetic characterization and pharmacological interrogation. Recently, a large-scale genomic dataset has been generated to facilitate the research of finding key genetic variations that have the highest correlation or casualty with anti-cancer drug response to multiple complex cancers. The goal of our project is to develop a novel machine learning algorithm that is capable of detecting the relevant genetic variations that can well predict the drug response to multiple cancer types.

We formulate this problem as a multi-view multi-task sparse regression problems. Given multiple data types, and multiple outputs, we want to apply a sparse version of regression model that can automatically detect the important factors in the data. Different from traditional regression model that assume equal correlation between multiple outputs and inputs, we need to address the structural correlation between multiple genetic data and various drug reponses. In this project, we propose a novel framework that can make use of the structural informations with the constraints of sparsity. The contributions are at least two-fold: 1) We formulate the drug response regression problems in a multi-view multi-task sparse regression model. This model can well characterize the correlation between genetic data and anti-cancer drug response. 2) We consider the structures information in the sparse coefficient spaces and therefore can well capture the high-order biological interaction between genotypes and phenotypes.

2 Regression prediction of drug efficacy

2.1 Method

What we are considering is a multi-task regression problem, previous methods such as l_1-regularized multi-task regression assume that all of the output variables are equally related to the inputs, but in

the drug efficacy analysis problem, the drugs are likely to affect similar groups of genes. It is reasonable to believe that closely related outputs (pharmarcological characterization of the drugs) tend to share a common set of relevant inputs (genomic characterization of each genes). So instead of using normal lasso method, we use graph-guided lasso which introduce a fusion penalty in order to identify shared relevant covariates for related output variables.

Given sample of N instances, each represented as a J-dimensional input vector and K-dimensional output vector. Let $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N) \in \mathbb{R}^{N \times J}$. Let $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_K\} \in \mathbb{R}^{N \times K}$ represent the output matrix. For each of the K output variables, the usual linear model is as follows:

$$\mathbf{y}_k = \mathbf{X}\beta_k + \epsilon_k$$

where $\beta_k = (\beta_{1k}, \beta_{2k}, \ldots, \beta_{Jk}) \in \mathbb{R}^J$ is the vector of regression coefficients for the k-th output variable and ϵ_k is a vector of N independent zero-mean Gaussian noise.

Let $\mathbf{B} = (\beta_1, \beta_2, \ldots, \beta_K)$ denote the $J \times K$ matrix of regression coefficients, then a traditional Lasso obtains $\widehat{\mathbf{B}}_{lasso}$ by solving the following optimization problems:

$$\widehat{\mathbf{B}}_{lasso} = \arg\min_{\mathbf{B}} \frac{1}{2}\|Y - \mathbf{X}\mathbf{B}\|_F^2 + \lambda\|\mathbf{B}\|_1$$

where $\|\cdot\|_F$ means the matrix Frobenius norm, $\|\cdot\|_1$ denotes the entry-wise matrix l_1 norm, and λ is a regularization parameters that controls the sparsity level.

One problem with Lasso is the implicit assumption that, each task is independent of each other. This assumption is not true in our data because most of anti-cancer drug can be highly related because they target a subset of same cancer-related genes. We show the correlations between drug responses in Fig(1(b)). High correlations are observed between the 24 drug response. To better utilize these correlations, we employs a novel multi-task prediction method, GFlasso [2].

GFlasso takes into account the complex dependency structure in the output variables represented as a graph when estimating the regression coefficients. The graph G has a set of nodes $\mathbf{V} = \{1, \ldots, K\}$ and edges E. The graph is constructed by computing pairwise Pearson correlation based on $\mathbf{y}_k$'s and two nodes $\mathbf{y}_m$ and $\mathbf{y}_l$ are connected with an edge if their Pearson correlation r_{ml} is above a given threshold ρ.

Given the graph G, GFlasso introduses an additional constraint over the standard lasso by fusing the β_{jm} and β_{jl} if $(m, l) \in E$ as follows:

$$\widehat{\mathbf{B}}_{GF} = \arg\min_{\mathbf{B}} \frac{1}{2}\|Y - \mathbf{X}\mathbf{B}\|_F^2 + \lambda\|\mathbf{B}\|_1 + \gamma \sum_{(m,l)\in E} |r_{ml}| \sum_{j=1}^{J} |\beta_{jm} - sign(r_{ml})\beta_{jl}|$$

where γ is a regularization parameters that denotes fusion penalty magnitude. Outputs that have a stronger correlation would receive higher fusion effect because they have larger $\|r_{ml}\|$. The overall effect is that each subset of output variables with a stronger correlation tend to have common relevant covariates.

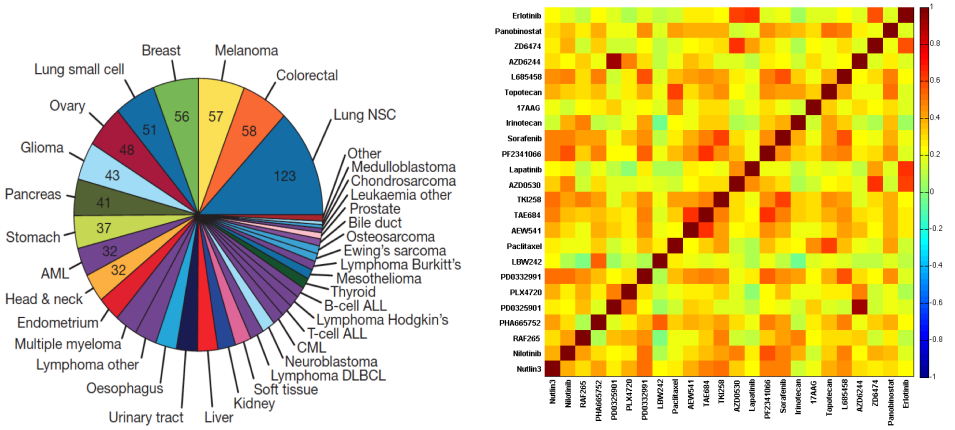
For optimization of the GFlasso, we use the proximal-gradient method [2]

2.2 Data

We utilize the gene expression, gene copy number, and gene mutation value data taken from 432 patients with different types of cancer. And get dose response data for the 24 types of drugs. The distribution of different cancer is shown in Fig(1(a)):

2.2.1 Genomic characterization

Gene Expression mRNA expression data was obtained using Affymetrix Human Genome U133 Plus 2.0 arrays. Genecentric expression values were objtained using updated Affymetrix probe set definition files from Brainarray[3] and background correction was accomplished using RMA[4] and quantile normalization.



(a) Cancer Distribution.

(b) Correlation map for the 24 drugs.

Figure 1: Data Overview

Gene Copy Number The raw CEL files were normalized to copy number estimates using a GenePattern pipeline[5] and hg18 Affymetrix probe annotations. Normalized copy number estimates were segmented using the Circular Binary Segmentation algorithm, followed by median centering of the segment values to a value of zero in each sample.

Gene mutation values Mutation data was generated for specific cancer gene loci using the mass spectrometric genotyping based OncoMap platform and data analysis was performed as previously described[6].

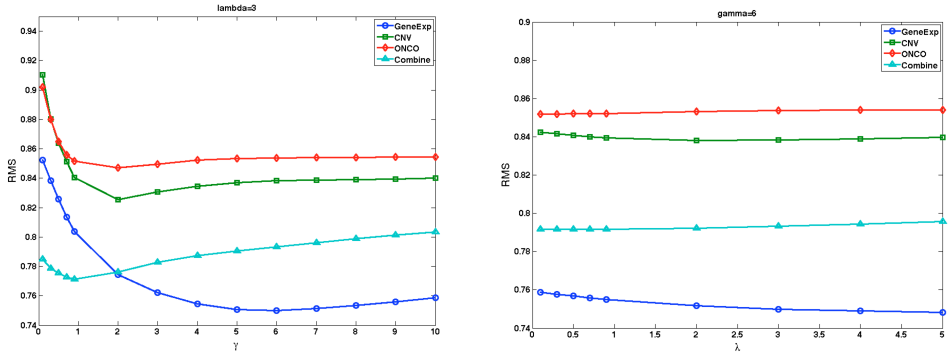
2.2.2 Pharmacological characterization

All dose-response data was fitted to one of three models depending on the statistical quality of the fits measured using a Chi-squared test. One approach was the 4 parameter sigmoid model:
$y = A_{inf} + \frac{A_0 - Ainf}{1+\left(\frac{x}{EC_{50}}\right)^{Hill}}$
Alternatively, a constant model $y = A_{inf}$ was employed;or a non-parametric spline interpolation of the data points was performed. In these models, A_0 and A_{inf} are the top and bottom asymptotes of the response; EC_{50} is the inflection point of the curve; and $Hill$ is the Hill slope. Key parameters derived from the models include IC_{50} and *Activity area*. IC_{50} is the concentration where the fitted curve crosses -50%.The *Activity area* is calculated as the sum of differences between the measured A_i at concentration i and the reference level.

2.3 Prediction of Drug Response

We have $N = 432$ cell line samples, represented as a J=dimensional input vector and K-dimensional output vector, where J equals to 15701, 23124 and 1667 respectively for gene expression, copy number variation and gene mutation value, and $K = 24$ for 24 types of drugs. The correlation map is generated for the 24 drugs. We can see that some of the drugs are strongly correlated.

We use 2/3 of the cell line sample for training and the rest 1/3 for testing. Using the aforementioned GFlasso method on the three features gene expression, copy number variation and gene mutation value separately and collectively, we generate regression coefficients with different choices of values for the regularization parameter λ, γ and then compared the root mean squared error of prediction on the testing sample. We find that (1) gene expression would have a better estimation than the other features, making it most useful for cancer categoriztion. (2)With a higher value of γ, gene expression prediction would have a smaller RMS error, which implies that the by considering the correlation between different drug outputs, we would have a better estimation of drug response. With $\lambda = 3$ and $\gamma = 6$, we would achieve the best estimation of drug response.



(a) RMS curve with different choice of gamma (b) RMS curve with different choice of lambda

Figure 2: RMS error on test sample

3 Cancer Categorization

We want to find the most discriminative type of genomic data. As a result, we implement classification for cancer types using GE, CNV and GM, respectively. The most discriminative type of features should give us the highest accuracy. 3 methods were implemented: 3-fold supervised SVM using LIBLINEAR[7], unsupervised k-means clustering[8] and spectral clustering[9]. Considering the high skewness of the cancer types (some cancers have only one patient, while other cancers have more than 80 patients), we only choose the cancers with more than 20 patients. The results can be seen from Fig(3).

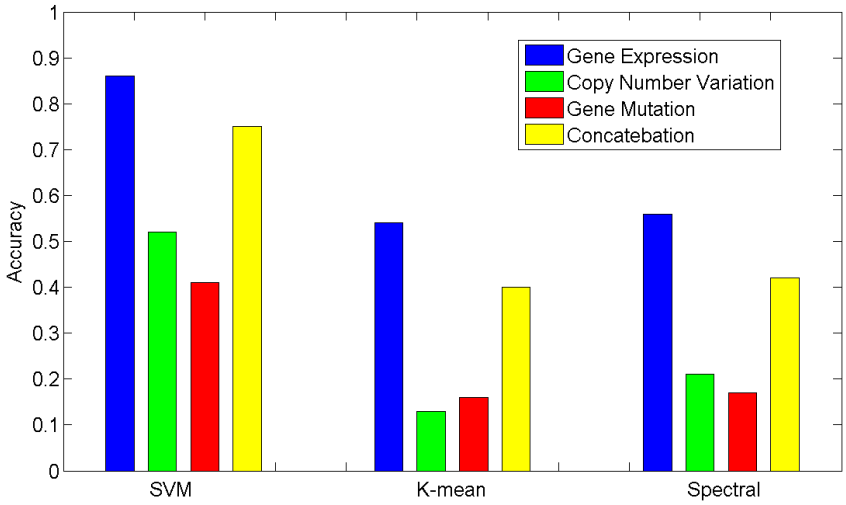


Figure 3: Cancer Categorization

From our results, we can see that for all 3 methods, GE gives us the highest accuracy, which means it is the most discriminative feature. This can be explained as below. Typically, a certain type of cancer is only related to a small number of gene mutations. For CNV and GM, only the few features relate to the few genes that mutate will be different for different types of cancers, while the rest of thousands of features are identical. Therefore, to use GE and CNV, we need to pre-select the effective genes out. However, this is still an open question. Therefore, using CNV and GM as features may not be effective. In contrast, although only a few genes mutate for a certain kind of cancer, however, all of the Gene Expression will be changed correspondingly. Therefore, GE is highly discriminative.

What's more, we test the effect of concatenation of all 3 types of features (GE, CNV and GM). From Fig(3), we can see that concatenation is worse the GE alone. That is because concatenation will diminish some correlation information, which is very important for our case. Also, concatenation will augment the noise. As a result, it is better for us to use Gene Expression alone.

4 Conclusion

In this project, we performed an integrative analysis on a large cancer data set with multiple anti-cancer drug response. We use graph-based multi-task regression method to build a robust drug response prediction system. It is observed that task correlation can improve the prediction accuracy than Lasso which assume independence between tasks. Also, we achieved the finding that gene expression is mostly discriminative in the task of cancer categorization. Last, a better integration method of multiple genomic data is needed since data correlation can be diminished and noise can be augmented by simple concatenation.

5 Acknowledgements

We sincerely thank David Knowles for help throughout the project.

References

[1] Jordi Barretina, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam A Margolin, Sungjoon Kim, Christopher J Wilson, Joseph Lehár, Gregory V Kryukov, Dmitriy Sonkin, et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603–607, 2012.

[2] Xi Chen, Seyoung Kim, Qihang Lin, Jaime G Carbonell, and Eric P Xing. Graph-structured multi-task regression and an efficient optimization method for general fused lasso. *arXiv preprint arXiv:1005.3579*, 2010.

[3] Manhong Dai, Pinglang Wang, Andrew D Boyd, Georgi Kostov, Brian Athey, Edward G Jones, William E Bunney, Richard M Myers, Terry P Speed, Huda Akil, et al. Evolving gene/transcript definitions significantly alter the interpretation of genechip data. *Nucleic acids research*, 33(20):e175–e175, 2005.

[4] Rafael A Irizarry, Bridget Hobbs, Francois Collin, Yasmin D Beazer-Barclay, Kristen J Antonellis, Uwe Scherf, and Terence P Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, 2003.

[5] Roger McLendon, Allan Friedman, Darrell Bigner, Erwin G Van Meir, Daniel J Brat, Gena M Mastrogianakis, Jeffrey J Olson, Tom Mikkelsen, Norman Lehman, Ken Aldape, et al. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061–1068, 2008.

[6] Julia Brettschneider, François Collin, Benjamin M Bolstad, and Terence P Speed. Quality assessment for short oligonucleotide microarray data. *Technometrics*, 50(3), 2008.

[7] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

[8] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, page 14. California, USA, 1967.

[9] Andrew Y Ng, Michael I Jordan, Yair Weiss, et al. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 2:849–856, 2002.