January 2016

Data Integration And Targeted Anticancer Drug Synergies Prediction

Xiaoting Gao
Yale University, aprilgao514@gmail.com

DATA INTEGRATION AND TARGETED ANTICANCER DRUG SYNERGIES PREDICTION

by

Xiaoting Gao

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Public Health

Yale University

2016

Approved by _____
Chairperson of Supervisory Committee

_____

_____

_____

Program Authorized
to Offer Degree _____

Date _____

Yale University

Abstract

Data Integration and Targeted Anticancer Drug Synergies Prediction

by Xiaoting Gao

Chairperson of the Supervisory Committee: Professor
Department of Biostatistics

In the past decades, targeted cancer therapies have made considerable achievements in inhibiting cancer progression by modulating specific molecular targets. However, targeted cancer therapies have reached a plateau of efficacy as the primary therapy since tumor cells can achieve adaptability through functional redundancies and activation of compensatory signaling pathways. Therapies using drug combinations have been developed to overcome the bottleneck. Accurate predictions of synergies effect can help prioritize biological experiments to identify effective combination therapies. Data integration can give us a deeper insight into the mechanism of cancer and drug synergies and help to address the challenge in prediction of drug combinations. In this thesis, we illustrate that integrative analysis of multiple types of omics data and pharmacological data can more effectively identify drug synergies, hence improve the prediction accuracy. As part of the AstraZeneca-Sanger Drug Combination Prediction DREAM Challenge, we showed that multiple data integration methods could identify multiple oncogenes and tumor suppressor genes as signature genes. We showed that several models built through data integration outperformed benchmark models without data integration methods.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1. INTRODUCTION

An understanding of the molecular basis of cancer brings the development of targeted anticancer therapies. In the past decades, targeted cancer therapies have made considerable progress in inhibiting cancer progression in some cancer patients by modulating specific molecular targets. However, the development of new drugs has been slowed down in recent years partly because drugs with specific targets often show limited efficacies, poor safety and resistance profiles [1]. One of the reasons of this phenomenon is that most human diseases are results of complex biological processes that are redundant and robust to drug perturbations of a single molecular target. Therefore, in recent years, efforts have been directed to the discovery of compound combinations, as they exhibit several advantages as therapeutics compared to single agent medicines. Drugs in combination may achieve greater effects than the additive therapeutic effect of each drug individually, which is known as synergy [2]. For example, Gefitinib (EGFR tyrosine kinase inhibitor, induces CDK inhibitors p27 and p21, decreases MMP2 and MMP9 enzyme activity [3]), combined with Taxane (disrupts microtubules by binding to β-tubulin, induces tumour suppressor gene p53 and CDK inhibitors p21, downregulates BCL-2, leading to apoptosis [4]) can produce strong synergistic effect in breast cancer MCF7/ADR cells [5]. Studies have also suggested that synergistic drug combinations can achieve therapeutic selectivity by countering biological compensation, allowing reduced dosage of each compound or accessing context-specific multitarget mechanisms [6].

1.1 Targeted anticancer drug synergies and omics data

For single agent therapeutics, predicting the response of a cancer patient to a certain targeted anticancer drug is vital for precision medicine. The efficacy of personalized treatment depends on the ability to identity specific genetic causes for a patient and then use the corresponding targeted therapy. In order to have a

comprehensive understanding of the link between genomic features and treatment effects, large-scale datasets of genomic, proteomic, epigenomic profiling data as well as pharmacological profiling data have been generated from cultured human cell lines, such as the Cancer Cell Line Encyclopedia (CCLE) [7] and Cancer Genome Project (CGP) [8]. In 2012, NCI-DREAM drug sensitivity prediction challenge was held aiming to compare methods for predicting drug sensitivity from multi-omic data in breast cancer cell lines, including copy number variation, gene expression, mutation, DNA methylation and protein abundance [9]. The key to precision medicine will be the ability to "translate large compendia of genomic, epigenomic, and proteomic data into clinically actionable predictions" [9]. Studies from this challenge have illustrated that the incorporation of multiple genomic characterizations could lower the prediction error for single drug sensitivity [9, 43].

Similarly, the accuracy of prediction of drug synergies also largely depends on our comprehension of how different types of omics data can inform us on synergistic effects. Knowledge drawn from multi-omic data can facilitate progress in understanding mechanisms of drug combinations. Omics data have been widely used to investigate the relations among small molecules, genes and diseases and to discover molecules linked to pathological processes. With its ability to reflect biological processes, omics data can also contribute to drug combinations studies. Some research has been done to predict drug synergies or to construct novel molecular networks based on omics data. Chen et al. [28] presented a systematic overview of existing approaches to model drug synergies, including omic-based methods in synergy identification. Sun et al. [29] proposed a model for the prediction of synergistic drug combinations specifically for the treatment of cancer, called Ranking-system of Anti-Cancer Synergy (RACS), which could combine features of targeting networks and transcriptomic profiles. A computational approach, Drug-Induced Genomic Residual Effect (DIGRE) Computational Model, was proposed to predict drug synergies by explicitly modeling drug response curves and gene expression changes after drug treatments [30]. Chen [31] proposed a new method for synergy evaluation by a pathway-pathway interaction network. Vera-

Licona et al. [40] designed a new software, OCSANA (optimal combinations of interventions from network analysis), to identify optimal and minimal combinations of intervention to disrupt the paths between source nodes and target nodes while minimizing the side effects. In Pal & Berlow's article [42], a new approach was presented to predict the sensitivity of a new drug or a drug combination by generating abstract representation of cancer pathways based on known kinases inhibitor targets. These studies have illustrated the possibility of predicting drug synergistic effect from omics data.

1.2 Data integration

Although many methods have been developed for drug combination effect prediction using omics data, most of these methods mainly utilize one type of omics data, usually genomics data or proteomics data. However, with the rise of high-throughput sequencing technologies [10] and large-scale consortia projects, large amounts of heterogeneous datasets have been generated, which make integration of different types of omics data an effective approach to understand the complex interplay of drug combinations against the disease process. Several studies have stated the necessity and benefit of data integration methods in drug related studies [33, 34, 35]. Dorel et al. [39] discussed three integration strategies to predict drug sensitivity and intervention combinations using signaling networks together with high-throughput data: (1) Highthroughput data-based signature retrieval; (2) Inferring intervention points from integrated analysis of interactomes and (3) Interference set finding using topological analysis of networks and mathematical modeling of network rewiring. Zhao et al. [32] presented a computational approach to predict drug combinations. By integrating molecular and pharmacological data, 69% of their top ranked predictions of effective combinations were supported by literature. Azmi et al. [41] showed that combining high-throughput data with network and systems biology-based strategy could facilitate the understanding of the synergy between MI-219 and oxaliplatin at the gene level and aid to identify driver pathways that augment p53 reactivation-mediated events.

In this thesis, data integration refers to the process of selecting important predictors of drug synergies by conducting integrative analysis of genomics, epigenomics, transcriptomics and pharmacological data. To understand drug combination in gene interaction network, there is a need to develop data integration methods because different types of molecules are involved in a biological process and the analysis of one data type may give a partial, and maybe biased, perspective on the biological process. Although microarray technology seems to be the most mature technology from all the omics and many breakthroughs about predicting cancer outcomes from gene expression data have been made, other omics data may contain complementary information not present in gene expression and integrating more than one data source can achieve better prediction of therapy response [38]. It was shown that a set of genes selected by analyzing the correlation of copy number variation and gene expression could discriminate matched adjacent noncancerous samples from gastric cancer samples in an unsupervised two-way hierarchical clustering [11]. Some research found a negative correlation between activities of the enzyme drug L-asparaginase and DNA copy number of genes near asparagine synthetase in the ovarian cancer cells [12].

Assembling all types of omics data into a more comprehensive and complex biological entity can shed light on the biological processes about cancer and targeted drugs that are not fully known to us. In our study, datasets from the AstraZeneca-Sanger Drug Combination Prediction DREAM Challenge, including gene expression, copy number variation, mutation and methylation, were used to illustrate the benefit of applying data integration methods into drug synergy analysis.

CHAPTER 2. RESEARCH DESIGN

2.1 Overview

Our study of the data integration methods in drug synergy study is part of our research for the AstraZeneca-Sanger Drug Combination Prediction DREAM

Challenge, Subchallenge 1. The aim of this Challenge is to "explore – and hopefully reveal - fundamental traits that underlie effective combination treatments and synergistic drug behavior using baseline genomic data, i.e. data collected pretreatment" [13]. It is to "uncover important biomarkers that are predictive of synergistic behavior, yielding a direct path for clinical translation" [13]. Synergy scores of 167 experimentally tested drug combinations over 85 cell lines (primarily colon, lung and breast) were provided. Corresponding monotherapy drug response for each drug and cell line were available as well. Additionally, omics data for 85 cell lines including gene expression, somatic mutation, copy number variation and methylation were offered. Putative drug targets and chemical information about drugs were also known to participants.

In Subchallenge1, participants were asked to predict drug synergy of 167 combinations in the panel of 85 cell lines. The synergy dataset was divided into three sets: a training data set (3/6-50%), a leaderboard set (1/6-16.7%), and a validation set (2/6-33%). In this study, the training dataset and leaderboard dataset were used since validation set has not been released yet. There are two challenges with different restrictions in Subchallenge 1. For Subchallenge 1A (we will use Challenge A in the following text instead), all available data could be used; while for Subchallenge 1B (Challenge B in the following text), only copy number variation, mutation and prior knowledge was allowed [13].

In our study, our main assumption is that combination effects on cell lines with similar molecular features would resemble each other. We implemented our model based on Random Forests, a method widely used and demonstrated to be powerful in various supervised learning problems. Random Forests are a classifier consisting of a collection of tree-structured classifiers [36]. To predict drug synergy scores, Random Forests fitted a combination of regression trees. The predictions from each tree were taken and averaged together to predict new values from this ensemble of trees. We first identified signature genes associated to the cancers and/or influenced by the drugs. Ten sets of signature genes were selected from different

perspectives. To demonstrate the effect of data integration techniques on drug synergy prediction, six of the ten sets of signature genes were selected by integrative analysis while four of them were selected without integrative analysis. Then ten Random Forests models were constructed to automatically choose important molecular features for each combination from the signature genes by including additional covariates such as drug indicators in the regression model. Four models using signature genes not selected by data integration methods were considered benchmark models. The performance of six other models would be compared to the performances of these four models.

2.2 Methods

2.2.1 Data preprocessing

Somatic mutation data were preprocessed using MutSigCV [14], a method to extrapolate the likelihood for each gene of being cancer-associated based on the mutation position, transition status and mutation type. The original chromosome region-based methylation data were converted to gene-based data using R package BioMart [15]. For copy number variation (CNV) data, the maximal copy number information was taken out from GRCh37 CNV file and the non-informative (equal-valued) genes were removed across cell lines, thus CNV data were transformed into a matrix with each entry representing the maximum copy number of a given gene in the corresponding cell line. For single drug sensitivity, the GI50 of a drug on a cell line was calculated using the fitted dose-response curves from the mono-therapy data.

2.2.2 Signature gene selection

The following methods were applied to select signature gene expression/mutation/copy number variation/methylation to be added into the model.

1. DriverNet [16]

   DriverNet is a novel integrated genome/transcriptome analysis approach to identify candidate drivers with aberrant genomic alteration such as mutation or

copy number variation [16]. The assumption of DriverNet is that driver genes with genomic aberration will impact the expression levels of multiple genes rather than a single gene. Genes with more connections to genes with outlying expression are more likely to be driver genes. The associations between mutation/copy number variation and coincident changes in gene expression are analyzed through an influence graph based on prior knowledge about pathways and gene networks obtained from Reactome [17]. Then a greedy algorithm is applied to find the lowest number of genes connected to the most genes with outlying expression.

The expression of a gene in a sample is defined as an outlying expression if it is outside of a predetermined range of gene expression for a given gene across all samples. In this manner, gene expression matrix will be converted to a binary matrix with samples as rows and genes as columns. Mutation matrix and copy number variation matrix are both binary matrices, with cells having a value of 1 if the gene is mutated or has copy number gain/loss in the corresponding sample and 0 otherwise. The associations between gene expression and mutation, or copy number variation are examined by an influence graph, which contains prior knowledge about the protein functional interaction network derived from Wu's study [26].

In our study, DriverNet generated two signature gene lists. One is the result of analyzing mutation and gene expression data while the other is for copy number variation and gene expression.

2. remMap [18]

   RemMap, REgularized Multivariate regression for identifying MAster Predictors, was proposed to fit multivariate response regression models under the high-dimension-low-sample-size setting [18]. The motivation of remMap is to explore the regulatory relationships among different biological molecules from multiple types of high dimensional genomic data, especially the modulation effect of copy

number variation on gene expression [18]. RemMap can build a multivariate linear regression model with an L1 norm penalty to control the overall sparsity of the coefficient matrix and a group lasso penalty [19] to control the total number of predictors entering the model. Consequently, the detection of master predictors can be facilitated.

Peng et al. [18] applied remMap method to gene expression and copy number variation data to investigate the influences of DNA copy number alterations on RNA transcript levels based on 172 breast cancer tumor samples. RemMap can also be utilized to study the relationships between other types of biological molecules [18]. It can be applied to other models as well to select a group of variables in a multiple regression model. In our study, remMap was utilized to select genes whose expression could affect single drug sensitivity.

3. CNAmet [20]

   The assumption of CNAmet is that genes with simultaneous alterations in gene expression, copy number and methylation are likely to be involved in tumor progression. CNAmet can integrate copy number, methylation and gene expression data to detect genes with abnormal expression levels that have concomitant amplification/deletion or hypomethylation/hypermethylation.

   In our study, CNAmet was applied to analyze copy number variation, methylation and gene expression to select signature genes. It was also used to identify corresponding changes in copy number and methylation for genes selected by remMap since remMap was only used to select genes of which expression levels can influence single drug response.

4. COSMIC gene

   COSMIC Cancer Gene Census data [21] contains a catalog of genes for which mutations have been causally implicated in cancer. We selected genes related to the cancer types of the cell lines considered in this challenge, resulting in 81 genes as signature genes. Random forests models with the

expression/mutation/copy number/methylation of COSMIC genes as predictors were benchmark models in our study. Prediction results of other models using signature gene lists generated by data integration methods would be compared to the results of these models.

5. Coefficient of variation

   A total of 300 genes with the highest coefficients of correlation of gene expression levels were selected as signature genes. In our study, a model to predict drug synergies was built with the gene expression, copy number variation and methylation of these genes. This model was another benchmark model to assess the effect of utilizing data integration methods to select signature genes.

2.2.3 Drug synergies prediction

For each drug combination on each cell line, a vector of covariates was constructed. The vector was comprised of the following covariates:

- Drug indicators: which two drugs the combination included;
- Cell line indicators: on which cell line the combination was tested;
- CNV indicators: whether the target genes of the two drugs have copy number variations (e.g. maximal copy number 3) on the cell line;
- Mutation indicators: whether the target genes of the two drugs have mutations on the cell line;
- The CNV, gene expression, and methylation information of the signature genes on the cell line.

Random Forests implemented using R package randomForest [22] was applied to predict drug synergy scores of every drug combination on different cell lines. The response variable, i.e., observed synergy scores, was preprocessed by subtracting the mean of the corresponding combination from them. Then the mean was added

back to the Random Forests predictions. For both Challenge A and Challenge B, benchmark models and models using data integration approaches were built.

2.2.4 Performance evaluation

In our study, the training and leaderboard datasets were combined to train the Random Forests. To assess the performance of each model, 30 rounds of 3-fold cross validations were conducted. In this Dream Challenge, two prediction scoring metrics were designed to compare the prediction accuracy of each model. In our study, we adopted the same metrics to measure each model's prediction precision.

1. Primary metric [27]

   Primary metric is aimed to evaluate the prediction accuracy of each drug combinations. It is a weighted average of Pearson correlation between predicted synergy scores and true synergy scores for each drug combination. Suppose for drug combination i, n_i is the number of cell lines drug combination i was applied to and ρ_i is the Pearson correlation between synergy score prediction and true synergy score for drug combination i, then

   $$Primary\ metric\colon \rho_w = \frac{\sum_{i=1}^{N} \sqrt{n_i - 1} \cdot \rho_i}{\sum_{i=1}^{N} \sqrt{n_i - 1}},$$

   where $N = 167$ is the total number of drug combinations ($\forall\ n_i \geq 2$).

2. Tie metric [27]

   The tie-breaking metric is identical to the primary metric, but only drug combinations with observed synergistic cell lines in the test set are used. For each drug combination, synergistic cell lines are defined as having synergy score greater than 20 in that cell line for a given drug combination.

Since 30 rounds of 3-fold cross validations have been conducted, the mean primary metric and tie metric were used to compare the performances of all models.

CHAPTER 3. RESULTS

3.1 Feature selection

3.1.1 DriverNet

The underlying assumption of the DriverNet method is that aberrant genomic variation will disrupt transcriptional patterns through one or several pathways. The copy number alteration or somatic mutation of one gene may not only lead to the over-representation or under-representation of this gene, but also the expression levels of other genes connected to it through pathways. Similarly, one mutated gene can also affect multiple genes. For our analysis, DriverNet was first applied to mutation and gene expression datasets to select candidate driver mutations. Then it was applied to copy number variation and gene expression datasets again, aiming to detect possible driver copy number variation.

For mutation only, DriverNet identified 326 mutated genes that were associated with abnormal expression levels of other genes. A total of 97 of them were nominated as significant driver candidates with P value below 0.05, including several known oncogenic genes, such as PIK3CA, KRAS, ERBB2, STAG1, as well as tumor suppressor genes TP53 and PTEN. Among 97 significant candidate drivers identified by DriverNet, 21 of them were in the COSMIC cancer gene census (ABL1, AKT1, APC, ATM, BCR, CDH1, EGFR, EP300, ETNK1, FBXW7, JAK1, KRAS, MET, MSH2, PDGFRA, PDGFRB, PIK3CA, PIK3R1, PTEN, SMAD4, TP53). Other candidate genes that were not in COSMIC datasets included FCL1, of which missense mutations, silent mutations and nonsense mutations are often observed in cancers, and RYR3, which may affect the growth, morphology and migration of breast cancer cells [23].

For copy number variation, 96 genes with amplification and 190 genes with deletion were marked as candidate drivers by DriverNet. Among amplified candidate genes, the copy number variation of 15 of them significantly altered the gene expressions of themselves or other genes, while the copy number loss of 32 genes of 190 deleted genes had significant altering effects. Significant candidate drivers ($p<0.05$) with

high-level amplification or homozygous deletion included oncogenes EGFR, PIK3CA, MET and suppressor gene TP53. Oncogenes AKT1, FGFR1 and FGFR2 were also detected but not significant. Most signature genes selected by DriverNet identified were associated with the aberrant gene expression levels of other genes. For example, the amplification of oncogene PIK3CA was found to be associated with the up-regulation of 43 other genes, including EGFR, MAPK9, PRKCA, and EP300. Figure.1 shows the heat map of Spearman correlation between copy number and gene expression of 46 genes selected with p value lower than 0.05.

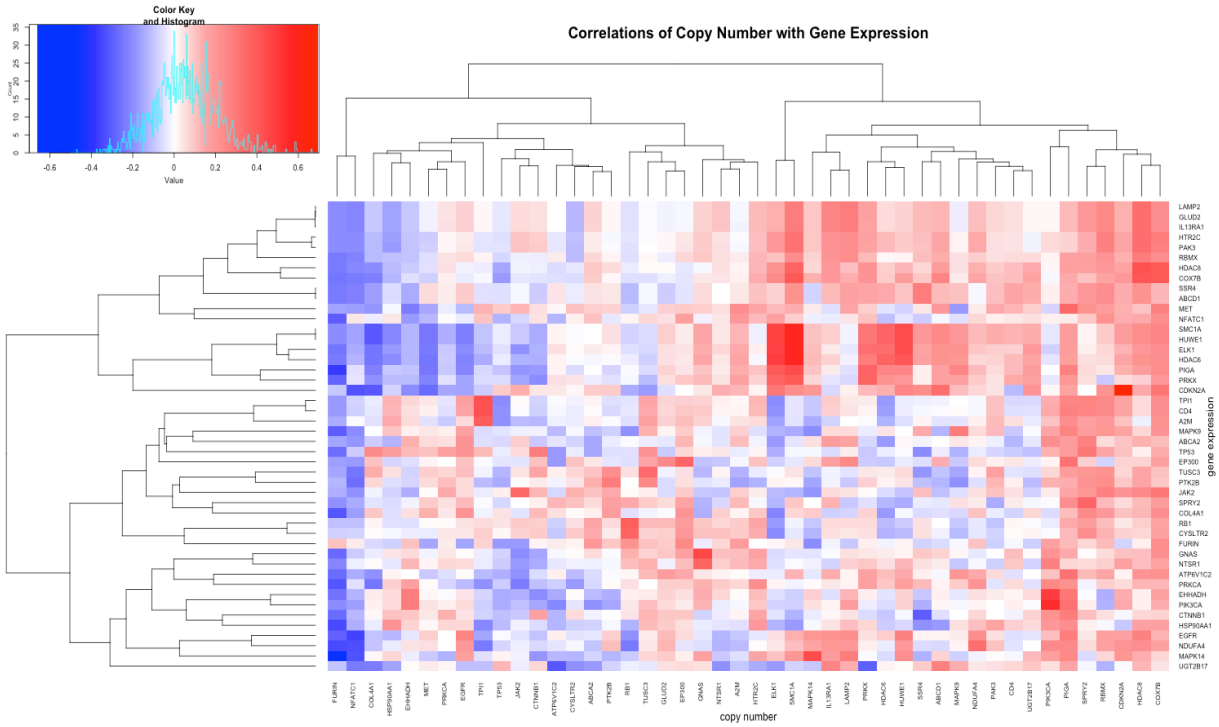
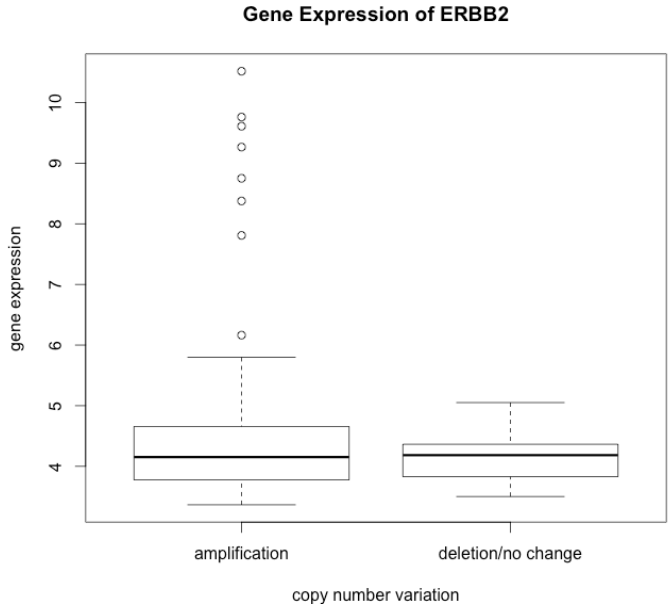


Figure 1 Spearman correlation between CNV and expression: Rows correspond to gene expression and columns correspond to copy number. Red and blue indicate high and low correlations respectively. The plot shows that the copy numbers of selected genes are associated with the expression of other genes beside themselves. The copy numbers of gene FURIN are negatively associated with the expression of almost all the other selected genes. The copy numbers of COX7B are positively associated with the expression of all the selected genes.

Surprisingly, ERBB2, of which the amplification exists in up to 18% of breast cancer patients, was not identified as a candidate driver [24]. However, the average gene expression level of ERBB2 with amplification was significantly higher than that of ERBB2 with deletion or without any copy number variation. Figure 2 is the boxplot displaying the gene expression levels of ERBB2 with and without amplification. The results of t test of equal means of gene expression are presented as well.



Mean (amplification)	**4.87**
Mean (deletion/ no change)	4.19
P value	0.02

Figure 2 Boxplot of the expression of ERBB2: Amplified ERBB2 has relatively higher expression levels than ERBB2 without amplification. The P value of the two-sample t-test for the equal mean is 0.02. Thus, from the results of t test, the expression level varies with the copy number of gene ERBB2.

The candidate driver genes were predictors for both Challenge A and Challenge B. For the models for Challenge B, only the copy number variation and mutation of those genes were added.

3.1.2 CNAmet

CNAmet was applied to select genes with simultaneous expression and copy number (or methylation) alterations. Those selected genes, which may have a key role in tumor progression or drug interaction, were used as predictors for Challenge B.

Our analysis of copy number variation, gene expression and mutation datasets using CNAmet resulted in four gene lists, which corresponded to hypomethylation, hypermethylation, copy number loss, and gain respectively. The copy numbers (methylation) of genes in copy number gain and loss (hypomethylation and hypermethylation) lists were predictors in the model. CNAmet identified 487 genes with both amplification and overexpression, or deletion and underexpression. A total of 369 genes with methylation alterations were detected coexisting with downregulation or upregulation of gene expression. The overlaps between them

were 52 genes, of which both copy numbers and methylations affected their gene expressions. JAK1, whose mutation is often associated with acute lymphoblastic leukemia and one of the drug target genes, was selected as a predictor since its amplification and hypomethylation were linked to the abnormal gene expression levels.

Candidate signature genes nominated by CNAmet included several oncogenes such as FLT3, STAT3, and AKAP9, and 5 drug target genes, JAK1, PIK3CA, RRM1, TYMS, and TOPBP1. Similar to DriverNet, CNAmet did not label ERBB2 as a signature gene. The p value of permutation test for ERBB2 is 0.21.

3.1.3 remMap

Unlike DriverNet and CNAmet, which identify possible driver genes or signature genes without utilizing drug information, remMap allows using single drug responses as the dependent variable to select features that affect drug sensitivity. In theory, drug synergy scores can also be applied as the response variable for variable selection. However, due to the small number of experiments with cell lines for each drug combinations, the limited sample size can lead to low statistical power, hence reduce the chance of detecting signature genes. Therefore, the GI50 value of each drug was used as response variable since there were fewer missing observations for each drug. Figure 3.A and Figure 3.B display two plots, one of which is the heat map of drug synergy scores of all the drug combinations across 85 cell lines and the other is of GI50 values of all drugs across all cell lines.

With GI50 values as response variable, if the predictors in the multivariate linear model were the expression, copy number variation, mutation and methylation of all genes, theoretically remMap could simultaneously select all the important features. Nevertheless, it would take exceptionally long time to run this program. Even using gene expression values along as predictors in the model requires more than five days to run the program. Considering the computational time, only gene expressions with coefficient of correlation greater than 0.2 were added into the model as

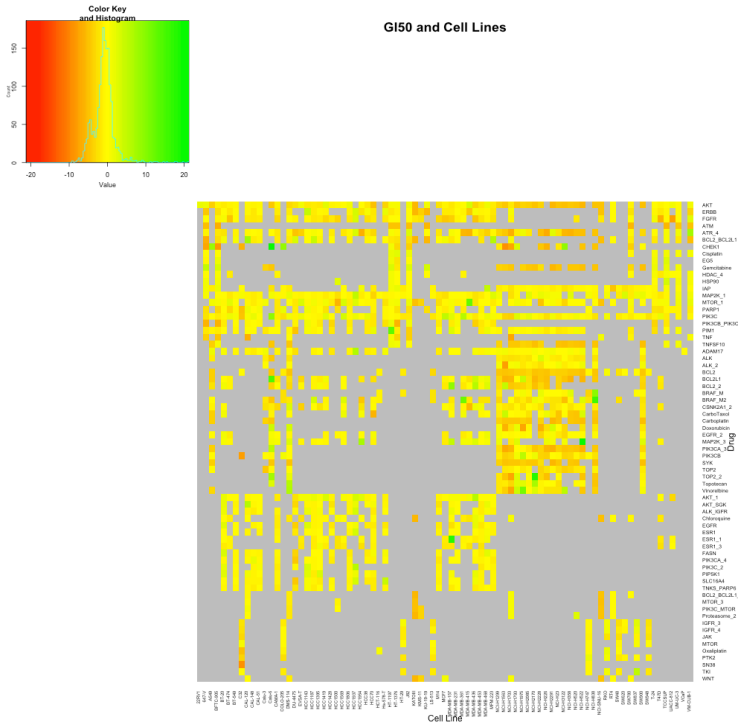


Figure 3.A GI50 and Cell Lines

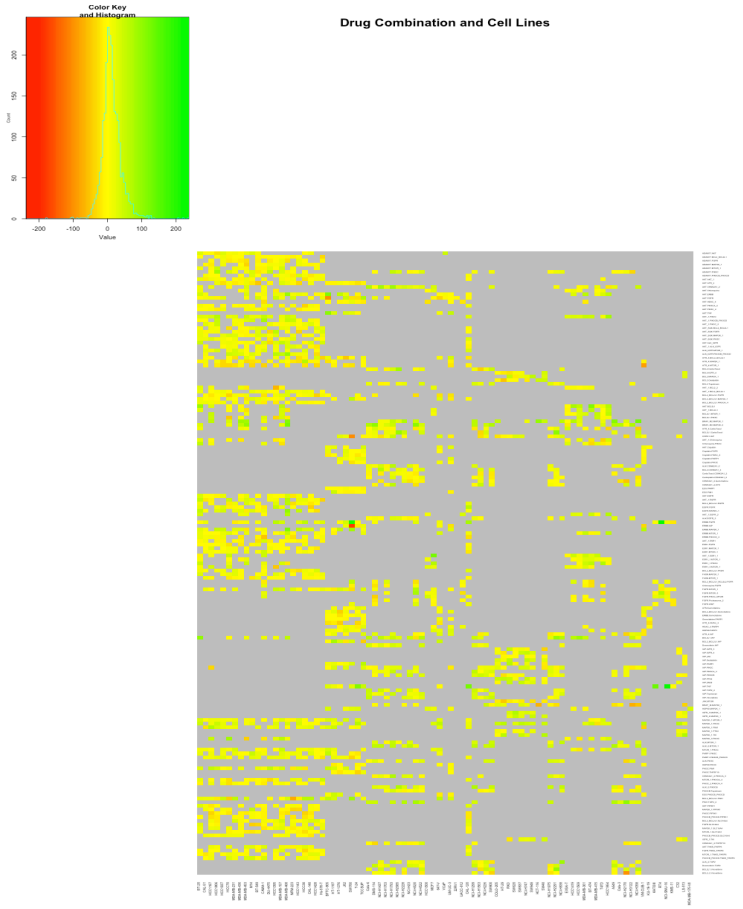


Figure 4.B Drug Synergy Scores and Cell Lines

independent variables. As a result, 2922 predictors were considered in the model. As is shown in Figure 3.A, not all drugs were tested on the same cell lines. Consequently, we divided 69 drugs into five groups that were not mutually exclusive based on the cell lines they tested on and built a model for each group of drugs; hence we had 5 gene lists corresponding to 5 drug lists.

There were 569 genes selected by remMap, including drug target genes PIP5K1B, WNT5B, WNT10A, and others. The results are consistent with existing biological findings on cancer. For example, for drugs targeting MAP2K1 and MAP2K3, this method successfully inferred the association between cell lines' response to those

drugs and the expression of genes FGFR3, MAPK13 and FGF12.

However, as a result of data preprocessing, only a small portion of gene expression data was used, which may result in the failure to capture important information from genes whose expression levels had less variation. Therefore, the gene list selected by remMap was combined with a list of 81 genes from COSMIC data that are related to the cancer types in this challenge and then the gene expression data of this new list served as predictors in the final model. Figure 4 is the heat maps of Pearson correlations between GI50 of five drugs, which were tested on more cell lines than other drugs, and expression levels of genes selected by remMap corresponding to these drugs. Figure 5 shows Pearson correlations between GI50 of the same five drugs and COSMIC gene expressions.

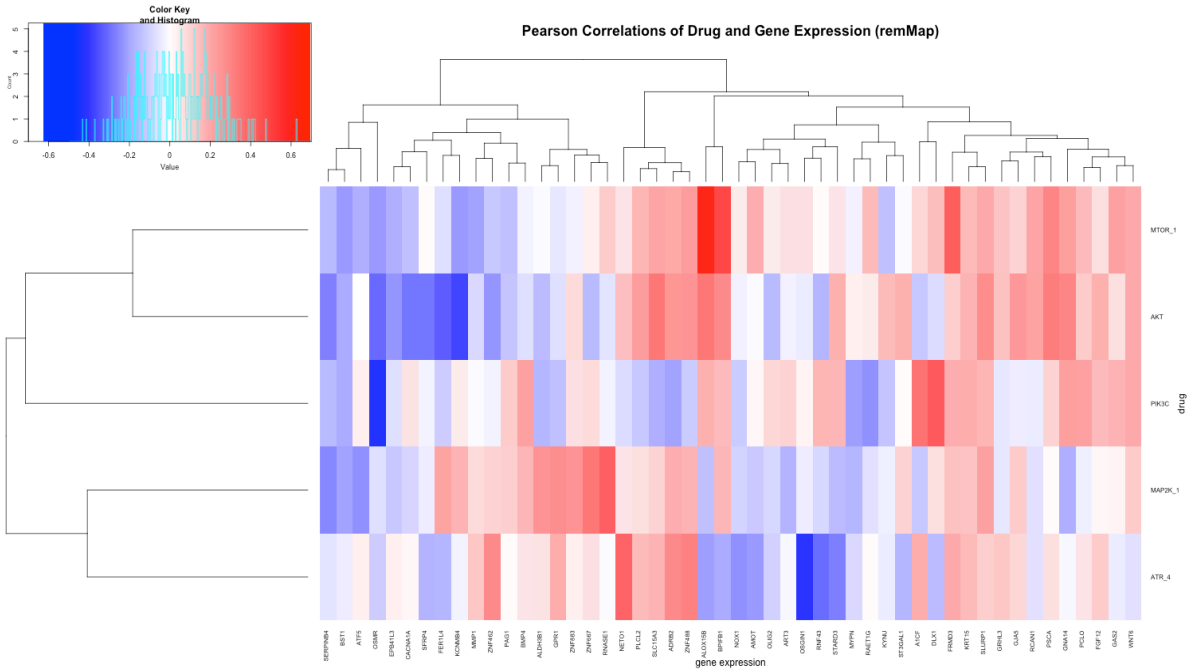


Figure 4 Pearson Correlations between Single Drug Sensitivities and Gene Expressions of 50 genes selected by remMap. Blue and Red indicate low and high correlations. Rows correspond to drugs and columns correspond to gene expression. The plot suggests that the expression levels of selected genes can affect drug responses. The highest positive correlation is detected between the expression of gene ALOX15B and the response of drug "MTOR_1", which is as high as 0.63. The most significant negative correlation is between the expression of gene OSMR and the response of drug "PIK3C", which is -0.43.

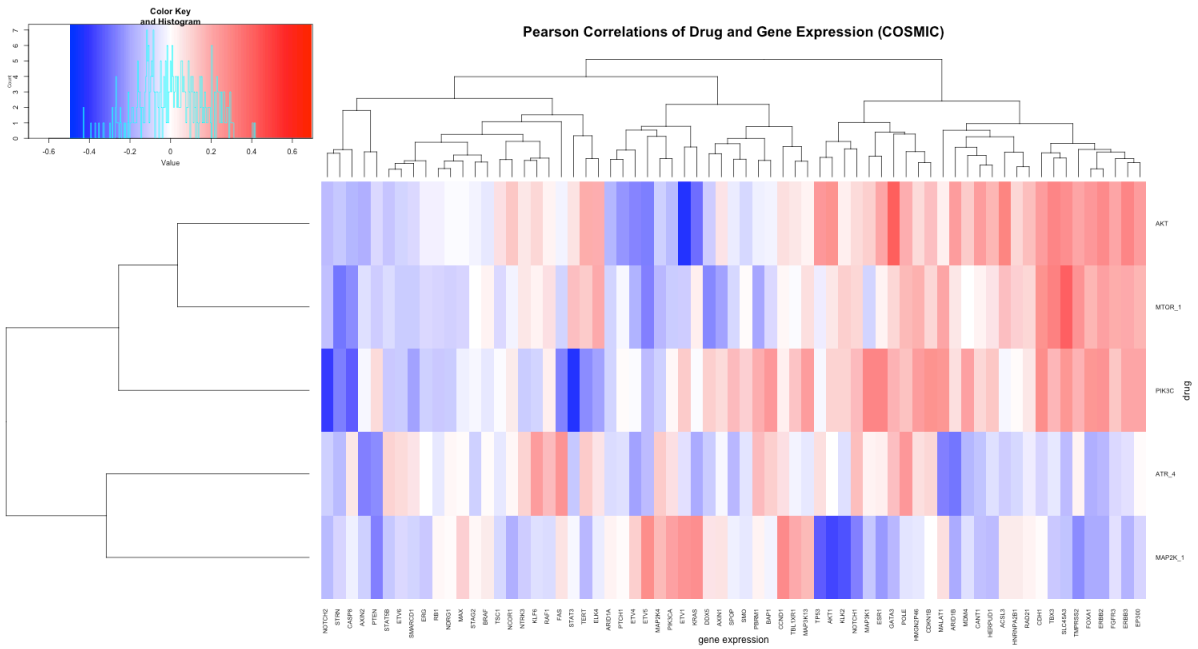


Figure 5 Pearson Correlations between Single Drug Responses and Gene Expression of COSMIC genes. Blue and Red indicate low and high correlations. Rows correspond to drugs and columns correspond to gene expression. Compared to genes selected by remMap, the correlations between drug responses and the expression of COSMIC genes are less strong but strong associations still exist. The expression of ETV1 is negatively correlated with the response of drug "AKT". The Pearson correlation between them is -0.43. The highest positive correlation is 0.41, which is the correlation between the expression of GATA3 and the expression of drug "AKT".

Because only related gene expression features were selected, it is not known whether the copy number variation or methylation of those genes were associated with drug responses. Therefore, CNAmet method was applied again to find those parallel changes. A total of 16 of them were identified as having simultaneous copy number variation, while for methylation, 14 of them were detected with hypermethylation or hypomethylation that can impact drug responses.

3.2 Model Performance

Ten models were built to predict drug synergy scores. Six of them were for Challenge A and the others were for Challenge B, in which only copy number variation and mutation data were allowed to use. For the models of each challenge, the difference of the covariates used lies in the gene expression/copy number variation/methylation of signature genes added into the model. Each model used different signature gene lists generated by multiple methods. Table 1 summarizes which of the CNV, gene expression and methylation information was specifically used in each model.

Table 1.A Features of Signature Genes in Each model for Challenge A

Challenge A	
Model	Covariates
DriverNet1	1. Expression levels of genes selected by DriverNet using mutation dataset and gene expression dataset; 2. Copy numbers of those selected genes; 3. Methylation levels of selected genes.
DriverNet2	1. Expression levels of genes selected by DriverNet using copy number variation dataset and gene expression dataset; 2. Copy numbers of selected genes; 3. Methylation levels of selected genes.
remMap.CNAmet	1. Expression levels of 569 genes selected by remMap; 2. Expression levels of 81 COSMIC genes related to the cancer type in this challenge. (Overlapping genes in two gene lists were removed so that there were no redundant genes). 3. Copy numbers of 16 genes with simultaneous amplification or deletion selected by CNAmet; 4. Methylation levels of 14 genes with parallel hypermethylation or hypomethylation selected by CNAmet.
remMap	1. Expression levels of 569 genes selected by remMap; 2. Expression levels of 81 COSMIC genes related to the cancer type in this challenge. (Overlapping genes in two gene lists are removed so that there will be no redundant variables).
COSMIC (benchmark)	1. Expression levels of 81 genes from COSMIC Cancer Gene Census related to the cancer type in this challenge; 2. Copy numbers of selected genes; 3. Methylation levels of selected genes.

CV300 (benchmark)	1. Expression levels of 300 genes with the highest coefficient of variation of their gene expression levels; 2. Copy numbers of selected genes; 3. Methylation levels of selected genes.

Table 1.B Features of Signature Genes in Each model for Challenge B

Challenge B	
Model	Covariates
DriverNet1 (B)	Copy numbers of genes selected by DriverNet using the mutation dataset and gene expression dataset.
DriverNet2 (B)	Copy numbers of genes selected by DriverNet using the copy number variation dataset and gene expression dataset.
CNAmet	Copy numbers of 487 genes selected by CNAmet with simultaneous amplification or deletion and abnormal gene expression.
COSMIC (B) (benchmark)	Copy numbers of 81 genes from COSMIC Cancer Gene Census related to the cancer types in this challenge.

In order to understand whether the application of data integration would benefit drug synergy prediction, three benchmark models were built without using data integration tools. Two of them, model CV300 and model COSMIC were for Challenge A and model COSMIC (B) was for Challenge B.

The models' performance was evaluated using two metrics designed by challenge organizers, primary metric and tie metric. To compare ten models' prediction accuracy, 30 rounds of 3-fold cross validations were conducted. The mean primary and tie metrics for the entire cross validations were calculated (Figure 6 and Table 2 display the two metrics for each model).

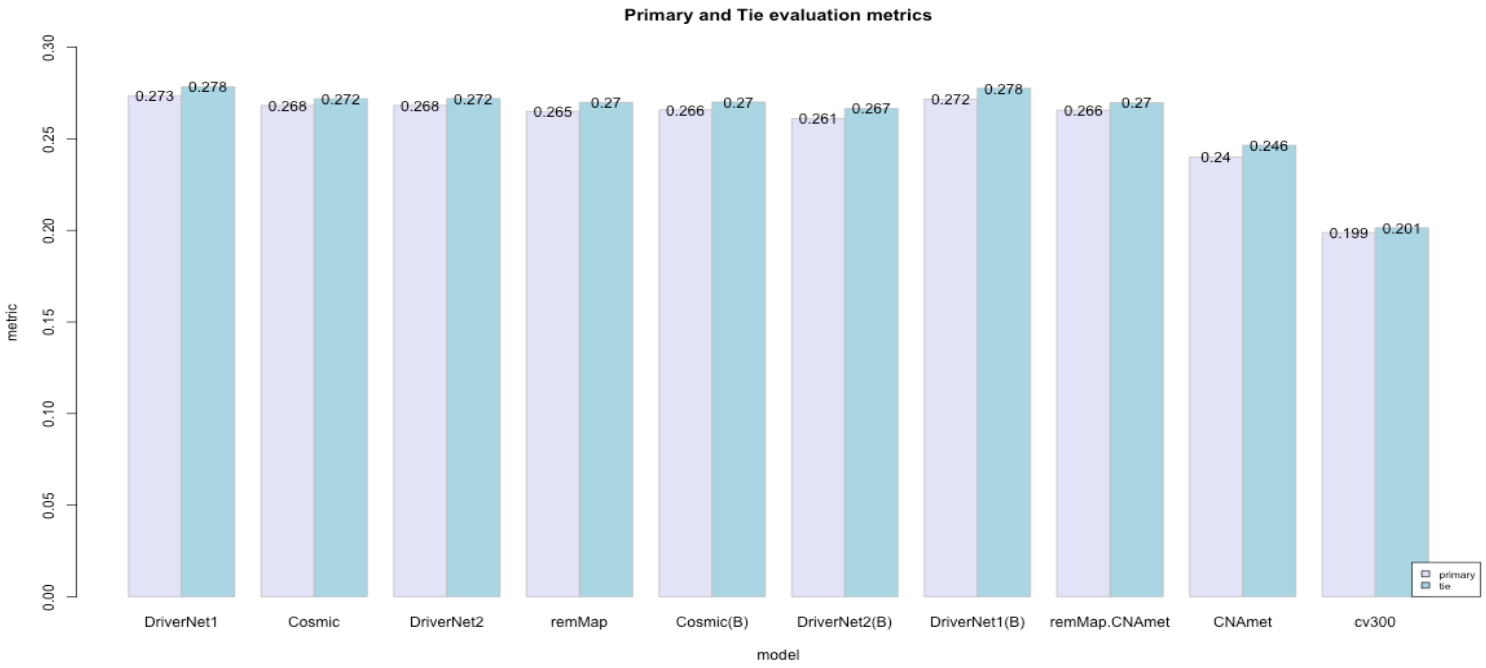


Figure 6 Primary and Tie Metrics for Each Model: Purple bars indicate primary metric and blue bars indicate tie metric. The bars plot shows that model DriverNet1 and model DriverNet1(B) achieve the highest prediction accuracy. Model CV300 has the lowest prediction scores.

Table 2.A Primary and Tie Metrics of Each Model in Challenge A

Model	DriverNet1	COSMIC	DriverNet2	remMap	remMap.CNAmet	CV300
Primary metric	0.273	0.268	0.268	0.265	0.266	0.199
Tie metric	0.278	0.272	0.272	0.270	0.270	0.201

Table 2.B Primary and Tie Metrics of Each Model in Challenge B

Model	COSMIC (B)	DriverNet2 (B)	DriverNet1 (B)	CNAmet
Primary metric	0.266	0.261	0.272	0.240
Tie metric	0.270	0.267	0.278	0.246

Based on the results of cross validations, model DriverNet1 and model DriverNet1 (B) achieved the best performance among models for Challenge A and for Challenge

B respectively. Among six models for Challenge A, model CV300 had the lowest primary and tie metrics. All other models, including models for Challenge B, outperformed model CV300. This indicates that genes in all the other signature gene lists contain more information than the genes whose gene expression have the highest coefficient of variation of gene expression.

For Challenge A, compared to model COSMIC, the only model with better prediction accuracy was model DriverNet1. Model DriverNet2 had the same performances as Model COSMIC. All the other models for Challenge A did not achieve better performance, but the differences in the performances were minimal. Model DriverNet1 and model DriverNet2 both had higher primary metric and tie metric than those of model remMap, even though signature genes used in model remMap were selected with single drug response as dependent variable. Model remMap.CNAmet's performance was slightly better than model remMap. The improvement was minimal and Pearson correlation between the predictions of two models was 0.993. One probable reason is that although CNAmet was employed to pinpoint related changes in copy number or methylation, only 30 genes were detected with concurrent copy number variation or methylation alteration. The effects of extra 30 features can be too limited to bring considerable improvement.

For Challenge B, the prediction of model DriverNet1 (B) was the most accurate. The performances of model COSMIC (B) and model DriverNet2 (B) were very similar, while the primary and tie metrics of model CNAmet were both lower. Therefore, the genes selected by DriverNet using mutation data and gene expression data are most informative for predicting drug synergy scores.

Unexpectedly, the predictions from different models are highly correlated, though the covariates they use have little overlap. For each drug combination, we calculated Pearson correlations between the predictions and true synergy scores. Besides model CV300, model CNAmet, which had relatively lower prediction accuracy, other seven models' prediction were close. Figure 7 shows Pearson correlations of all models besides model CNAmet, model CV300 and model remMap.CNAmet. Since

Pearson correlation between the predictions of model remMap and the predictions of model remMap.CNAmet was 0.993, only model remMap's predictions are used here.

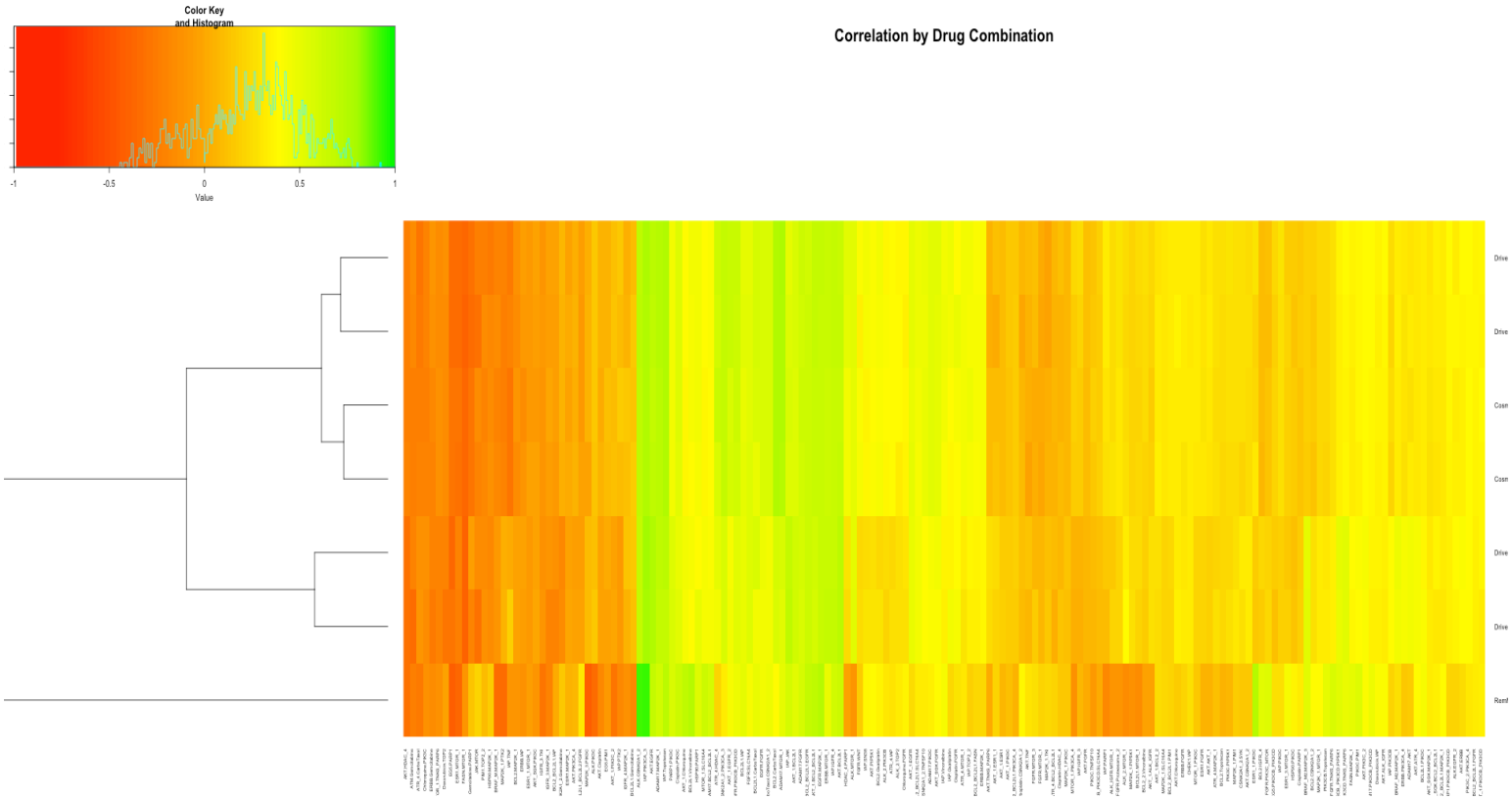


Figure 7 Pearson Correlations between the Predicted Synergy Score and True Score of All the Drug Combinations. As shown in the plot, each model has similar patterns of correlations. The predicted scores of each model are very similar.

CHAPTER 4. CONCLUSIONS

4.1 Summary

In this thesis, we built ten Random Forests models to predict drug synergy scores. Besides three models as benchmarks (model CV300, model COSMIC and model COSMIC (B)), the other seven models utilized the copy number variations/expressions/methylation levels of genes selected by DriverNet, remMap and CNAmet. DriverNet selected 326 mutated genes with altered gene expression using mutation and gene expression data. Another 286 genes with simultaneous copy number variation and gene expression changes were selected by DriverNet after analyzing copy number and gene expression data.

The results of the cross validations suggest that for Random Forests model, signature genes selected by DriverNet using gene expression and mutation, or copy number variation are the most predictive, in that model DriverNet1 and model DriverNet1 (B)'s primary metrics were the highest (0.273 and 0.272). Especially for Challenge B, when gene expression values cannot be used as predictors, DriverNet identifies informative copy number variations successfully.

Although DriverNet cannot integrate drug response or drug synergies information into the variable selection process like remMap, DriverNet1 and DriverNet2 outperformed two models using the gene list generated by remMap (model remMap and model remMap.CNAmet), of which primary metrics were 0.265 and 0.266 respectively. One of the explanations that the genes selected by remMap were not as informative as genes selected by DriverNet might be that due to the limited samples of single drug responses, the power of detecting signature genes was limited. Moreover, unlike DriverNet, which incorporates prior knowledge from existing cancer research through an influence graph, remMap selects signature genes without prior biological knowledges. Thus, when there is much noise in the dataset, uninformative variables may be selected by remMap, hence decreasing model's prediction accuracy.

Compared to DriverNet, CNAmet can conduct integrative analysis of copy number variation, gene expression and methylation. This method is able to distinguish synergistic effect of DNA methylation and copy number variation on gene expression. Model CNAmet's predictions were more accurate than those of Model CV300 with primary metric as 0.240, which reveals that genes selected by CNAmet are more helpful for drug synergies prediction.

For three benchmark models, model CV300, model COSMIC and model COSMIC (B), model CV300 achieved the lowest prediction accuracy with primary metric as 0.199, while the other two models with 81 genes related to the cancer types in this challenge from COSMIC Cancer Gene Census as covariates had the best performances other than model DriverNet1 and model DriverNet1 (B). For Challenge A, model COSMIC had the second highest primary metric, 0.268; while for Challenge B, model COSMIC (B)'s primary metric, 0.266, was the second highest as well. Since all the model's predictions were more precise than that of model CV300, DriverNet, remMap and CNAmet are all effective ways to select signature genes. The fact that model DriverNet1 and model DriverNet1 (B) had the best performance suggests that information from genes selected entirely based on prior knowledge (81 COSMIC genes) can be inadequate to predict drug synergies. Nonetheless, the performance of models utilizing remMap and CNAmet to select features indicates that variable selection methods that are not knowledge-based may also be insufficient.

4.2 Limitations

Although this study has shown that the application of data integration methods can improve model's prediction accuracy, the implementation of data integration methods has some limitations. First of all, each data integration method has its own limitations.

1. DriverNet

The most distinguishing feature of the DriverNet approach is to incorporate prior knowledge about cancer gene networks through the influence graph. Genes outside the influence graph will not be selected, which can reduce the likelihood of including uninformative gene features in the model. Although in this study, drug-gene information was not applied to select signature genes, theoretically it is possible to use such information by designing our own influence graph. However, the use of influence graph can also fail to identify signature genes we have no knowledge of since the graph is inevitably sparse and incomplete.

Another drawback of DriverNet is that it may fail to detect somatic mutations or copy number variations that modulate less extreme but important changes in gene expression, in that a prespecified threshold is used to define outlying gene expression values [16]. In our study, gene expressions outside the two standard deviation values were considered abnormal. This approach may result in the overlook of genes with somatic mutations or copy number variations that modulate important changes in gene expression within the given range [16].

In addition, DriverNet cannot take the directionality of the change in expression into account [16], and it cannot discern the correlation between copy number gain/loss and gene expression is negative or positive.

2. CNAmet

   Compared to DriverNet, which can identify genomic aberrations altering more than one transcriptional network [16], the main limitation of CNAmet is that it can only analyze the one-to-one association between gene expression and methylation (or CNV). It cannot be applied to assess the influence of one gene's CNV or methylation on another gene's expression. For example, in our study, both DriverNet and CNAmet identified PIK3CA as signature gene. Both approaches can recognize the effect of the copy number gain of PIK3CA on its gene expression, but DriverNet can also distinguish the influence of amplified

PIK3CA on gene expression of other genes in the same pathway, such as EGFR, MAPK9, MET, GNAS, CD4, and others.

One of the advantages of CNAmet is that it can conduct integrative analysis of methylation, copy number variation and gene expression. In our study, we only used CNAmet to select genes with important hypermethylation or hypomethylation. (remMap can select alteratinos in methylation with significant effect as well, however, in our study, we only applied remMap to select signature genes based on their expression since adding more variables into the model would cost longer time to run the program). Nevertheless, CNAmet, like DriverNet has the same disadvantage that directionality is not considered.

CNAmet does not employ prior knowledge, which makes it a suitable tool to detect concomitant copy number/methylation and gene expression alteration, but less applicable for signature genes selection. Although there is no clear evidence that knowledge-based integration can outperform data-driven integration, Kim et al.'s paper, in which a graph-based framework for integrating multi-omics data to predict clinical outcomes for cancer patients was proposed, suggested that it was beneficial to incorporate genomic knowledge, such as pathway or GO gene sets, into omics data integration process since it could improve the predictive power and better explain the interplay between different types of data and knowledge [37]. However, CNAmet can serve as a complementary tool for detecting matching alterations in methylation or copy number for signature genes selected by other non-integrating methods. In our study, after selecting signature genes with remMap, CNAmet was applied to search for the parallel changes in methylation or copy number.

3. remMap

   Compared to two other methods, remMap has two main strengths. First, information about drug can be used to select signature genes. Second, remMap treats all sources of genomic information as one coherent dataset rather than

separate ones. Therefore, remMap can carry out one joint analysis by viewing gene expression, methylation, copy number variation, and mutation datasets as one dataset. Although in our study, due to the limited sample size and computational feasibility, only gene expression dataset was used.

Because remMap is time-consuming, proper dimensionality reduction methods should be applied first before using remMap for variable selection. Using original high-dimensional datasets directly can be inefficient. Moreover, it is also hard for remMap to incorporate prior knowledge into the model. Microarray technique suffers from low signal-to-noise ratio, which may cause instability in gene signatures. Utilizing other information may help to reduce the effect of randomly generated differences in expression levels [38]. Without the incorporation of genomic knowledge, remMap may select uninformative genes due to the random noises in microarray data.

Apart from the limitations mentioned above, there are some drawbacks three methods all have.

- Different cancer types and subtypes are not considered. The inability of taking cancer types and subtypes into consideration may lead to the failure of identifying signature genes of certain type or subtype.
- Different omics data have different noise level. For each data integration method, its tolerance to noise in each type of omics dataset is unknown, which may make the results questionable.

4.3 Conclusions

Research in data integration and in drug synergy analysis has mostly remained isolated. Our study suggests that the application of data integration approach may improve our understanding of targeted drug synergies. The emergence of data integration methods will facilitate the process of variable selection for models to predict drug synergies. Although the need for a systematic integrative analysis method has not been fully addressed yet, there are various approaches that can be

implemented in future studies. The major challenge of incorporating data integration analysis into drug synergies study is to combine different types of omics datasets and drug information. As more data are generated across multiple data types, novel integration methodologies of future will further our understanding of important biological processes of gene-gene and drug-gene interactions.

BIBLIOGRAPHY

1. Jia J, Zhu F, Ma X, et al. Mechanisms of drug combinations: interaction and network perspectives[J]. Nature reviews Drug discovery, 2009, 8(2): 111-128.
2. J. Foucquier, M. Guedj. Analysis of drug combinations: current methodological landscape, Pharma Res Per, 3(3), 2015, e00149, doi: 10.1002/prp2.149
3. Lee, E. J., Whang, J. H., Jeon, N. K. & Kim, J. The epidermal growth factor receptor tyrosine kinase inhibitor ZD1839 (Iressa) suppresses proliferation and invasion of human oral squamous carcinoma cells via p53 independent and MMP, uPAR dependent mechanism. Ann. NY Acad. Sci. 1095, 113–128 (2007).
4. Fanucchi, M. & Khuri, F. R. Taxanes in the treatment of non-small cell lung cancer.Treat. Respir. Med. 5, 181–191 (2006).
5. Takabatake, D. et al. Tumor inhibitory effect of gefitinib (ZD1839, Iressa) and taxane combination therapy in EGFR-overexpressing breast cancer cell lines (MCF7/ADR, MDA-MB-231). Int. J. Cancer 120, 181–188 (2007).
6. Lehár J, Krueger A S, Avery W, et al. Synergistic drug combinations tend to improve therapeutically relevant selectivity[J]. Nature biotechnology, 2009, 27(7): 659-666.
7. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, et al. (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 483: 603–607. doi: 10.1038/nature11003. pmid:22460905
8. Garnett MJ, Edelman EJ, Heidorn SJ, Greenman CD, Dastur A, et al. (2012) Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature 483: 570–575. doi: 10.1038/nature11005. pmid:22460902

9. Costello J C, Heiser L M, Georgii E, et al. A community effort to assess and improve drug sensitivity prediction algorithms[J]. Nature biotechnology, 2014, 32(12): 1202-1212.
10. Lappalainen T, Sammeth M, Friedländer M R, et al. Transcriptome and genome sequencing uncovers functional variation in humans[J]. Nature, 2013, 501(7468): 506-511.
11. Cheng L, Wang P, Yang S, Yang Y, Zhang Q, Zhang W, Xiao H, Gao H, Zhang Q (2012) Identification of genes with a correlation between copy number and expression in gastric cancer. BMC Med Genomics 5:14
12. Bussey K J, Chin K, Lababidi S, et al. Integrating data on DNA copy number with gene expression levels and drug sensitivities in the NCI-60 cell line panel[J]. Molecular cancer therapeutics, 2006, 5(4): 853-867.
13. "The AstraZeneca-Sanger Drug Combination Prediction Challenge (Syn4231880)" 2015–2016
14. Lawrence, Michael S, Petar Stojanov, Paz Polak, Gregory V Kryukov, Kristian Cibulskis, Andrey Sivachenko, Scott L Carter, et al. 2013. "Mutational Heterogeneity in Cancer and the Search for New Cancer-Associated Genes." Nature 499 (7457). Nature Publishing Group: 214–18.
15. Kasprzyk, Arek. 2011. "BioMart: Driving a Paradigm Change in Biological Data Management." Database 2011. Oxford University Press: bar049.
16. Bashashati, Ali, Gholamreza Haffari, Jiarui Ding, Gavin Ha, Kenneth Lui, Jamie Rosner, David G Huntsman, Carlos Caldas, Samuel A Aparicio, and Sohrab P Shah. 2012. "DriverNet: Uncovering the Impact of Somatic Driver Mutations on Transcriptional Networks in Cancer." *Genome Biol*13 (12). Springer Science Business Media: R124. doi:10.1186/gb-2012-13-12-r124
17. http://www.reactome.org/
18. Peng J, Zhu J, Bergamaschi A, et al. Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer[J]. The annals of applied statistics, 2010, 4(1): 53.
19. Friedman J, Hastie T, Tibshirani R. A note on the group lasso and a sparse group lasso[J]. arXiv preprint arXiv:1001.0736, 2010.

20. Louhimo R, Hautaniemi S. CNAmet: an R package for integrating copy number, methylation and expression data[J]. Bioinformatics, 2011, 27(6): 887-888.
21. Forbes, Simon A, David Beare, Prasad Gunasekaran, Kenric Leung, Nidhi Bindal, Harry Boutselakis, Minjie Ding, et al. 2015. "COSMIC: Exploring the World's Knowledge of Somatic Mutations in Human Cancer." Nucleic Acids Research 43 (D1). Oxford Univ Press: D805–D811.
22. Liaw A, Wiener M. Classification and regression by randomForest[J]. R news, 2002, 2(3): 18-22.
23. Zhang L, Liu Y, Song F, et al. Functional SNP in the microRNA-367 binding site in the 3′ UTR of the calcium channel ryanodine receptor gene 3 (RYR3) affects breast cancer risk and calcification[J]. Proceedings of the National Academy of Sciences, 2011, 108(33): 13653-13658.
24. Kallioniemi O P, Kallioniemi A, Kurisu W, et al. ERBB2 amplification in breast cancer analyzed by fluorescence in situ hybridization[J]. Proceedings of the National Academy of Sciences, 1992, 89(12): 5321-5325.
25. Shiu KK, Natrajan R, Geyer FC, Ashworth A, Reis-Filho JS. DNA amplifications in breast cancer: genotypic-phenotypic correlations. Future Oncol. 2010;14:967–984. doi: 10.2217/fon.10.56.
26. Wu G, Feng X, Stein L. Research a human functional protein interaction network and its application to cancer data analysis[J]. Genome Biol, 2010, 11: R53.
27. Scoring metrics, https://www.synapse.org/#!Synapse:syn4231880/wiki/235660
28. Chen D, Liu X, Yang Y, et al. Systematic synergy modeling: understanding drug synergy from a systems biology perspective[J]. BMC systems biology, 2015, 9(1): 56.
29. Sun Y, et al. Combining genomic and network characteristics for extended capability in predicting synergistic drugs for cancer. Nat Commun. 2015;6:8481.

30. Yang, J., Tang, H., Li, Y., Zhong, R., Wang, T., Wong, S., Xiao, G. and Xie, Y. (2015), DIGRE: Drug-Induced Genomic Residual Effect Model for Successful Prediction of Multidrug Effects. CPT: Pharmacometrics & Systems Pharmacology, 4: 91–97. doi: 10.1002/psp4.1.
31. Chen, Di and Zhang, Huamin and Lu, Peng and Liu, Xianli and Cao, Hongxin. Synergy evaluation by a pathway-pathway interaction network: a new way to predict drug combination. Molecular BioSystems, 2016, 2: 614—623.
32. Zhao X M, Iskar M, Zeller G, et al. Prediction of drug combinations by integrating molecular and pharmacological data[J]. PLoS Comput Biol, 2011, 7(12): e1002323.
33. Louie B, Mork P, Martin-Sanchez F, et al. Data integration and genomic medicine[J]. Journal of biomedical informatics, 2007, 40(1): 5-16.
34. Ritchie M D, Holzinger E R, Li R, et al. Methods of integrating data to uncover genotype-phenotype interactions[J]. Nature Reviews Genetics, 2015, 16(2): 85-97.
35. Kristensen V N, Lingjærde O C, Russnes H G, et al. Principles and methods of integrative genomic analyses in cancer[J]. Nature Reviews Cancer, 2014, 14(5): 299-313.
36. Breiman L. Random forests[J]. Machine learning, 2001, 45(1): 5-32.
37. Kim D, Joung J G, Sohn K A, et al. Knowledge boosting: a graph-based integration approach with multi-omics data and genomic knowledge for cancer clinical outcome prediction[J]. Journal of the American Medical Informatics Association, 2014: amiajnl-2013-002481.
38. Gevaert O, De Moor B. Prediction of cancer outcome using DNA microarray technology: past, present and future[J]. Expert opinion on medical diagnostics, 2009, 3(2): 157-165.
39. Dorel M, Barillot E, Zinovyev A, et al. Network-based approaches for drug response prediction and targeted therapy development in cancer[J]. Biochemical and biophysical research communications, 2015, 464(2): 386-391.

40. Vera-Licona P, Bonnet E, Barillot E, et al. OCSANA: optimal combinations of interventions from network analysis[J]. Bioinformatics, 2013, 29(12): 1571-1573.
41. Azmi A S, Wang Z, Philip P A, et al. Proof of concept: network and systems biology approaches aid in the discovery of potent anticancer drug combinations[J]. Molecular cancer therapeutics, 2010, 9(12): 3137-3144.
42. Pal R, Berlow N. A kinase inhibition map approach for tumor sensitivity prediction and combination therapy design for targeted drugs[C]//Pac Symp Biocomput. 2012, 351: 62.
43. Wan Q, Pal R. An ensemble based top performing approach for NCI-DREAM drug sensitivity prediction challenge[J]. PloS one, 2014, 9(6): e101183.