Drug Combination Prediction Challenge

Chieh Lin
Machine Learning Department
Carnegie Mellon University
chiehl1@cs.cmu.edu

Po-Wei Wang
Machine Learning Department
Carnegie Mellon University
poweiw@cs.cmu.edu

Yan Xia (Not in the class)
Machine Learning Department
Carnegie Mellon University
yanxia18@cmu.edu

Ziv Bar-Joseph (Supervisor)
Machine Learning Department
Carnegie Mellon University
zivbj@cs.cmu.edu

Abstract

The multi-drug combination effects is always an important issue because it can help develop new therapies for diseases and prevent deadly drug-combination side-effects. However, it takes substantial efforts and resources for $in\ vivo$ methods to verify the effects since the number of combinations is too large. Therefore, the $in\ vitro$ method of machine learning for predicting the multi-drug combination effect is necessary. The AstraZeneca-Sanger Drug Combination Prediction Challenge [1] provides datasets for modelling two-drug synergy effect. With the datasets, the goal of this project is to find the best machine learning strategy to solve the two-drug synergy prediction problem, hoping to lay the foundation of future multi-drug combination effect prediction.

1 Introduction

To excel in this challenge, the information from drugs, proteins, and genes are indispensable. First of all, almost all of the biological functions are performed by proteins in most organisms. Based on the importance of protein functions, drugs are usually developed to regulate the functions of specific proteins in order to cure diseases. Therefore, proteins and drugs information are crucial to this problem. However, the challenge does not provide any protein data. Instead, the gene expression data is provided and the protein information may be indirectly given from gene expression data. Proteins are produced by translation from RNAs, and RNAs are produced by transcription from DNA, in which genes locates. Also, genes usually work in groups to produce proteins. Hence it is also important to find a good way to group genes together. Note that the groups can be overlap since the same gene may work in different groups at the same time. While the challenge does not provide any direct protein data, it is encouraged to use external datasets as ”prior knowledge” to help boost the models. Based on the above facts, models are designed to incorporate features from the data of genes, proteins, and drug information. External datasets are also used if needed.

2 Problem Definition

Given single drug trails of all the drugs and some combinatorial drug tails, the goal of this challenge is to predict whether the combinatorial effect exceeds the standard model on given drug pairs. Here we give more details of this problem.

The effect of drugs is measured by the cell living rate E. For each single trail, different dose v of drug A is applied to the cell line i, and the resulting cell living rate $E_A(v)$ is given. Specifically,

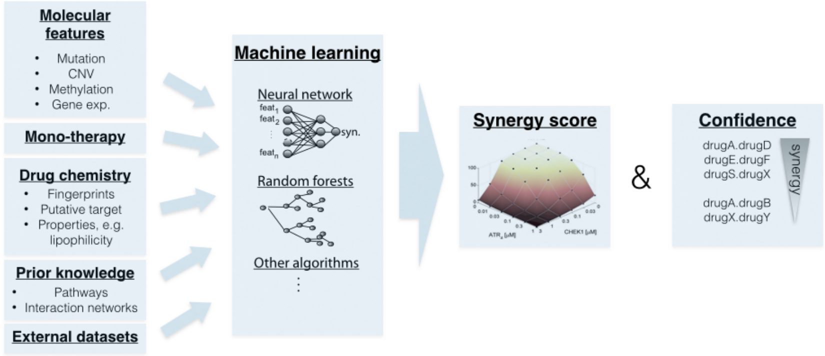

Figure 1: Problem Definition from Contest Website [1]

$E_{i,A}(v)$ could be fitted to the following standard model.

$$\hat{E}_{i,A}(v) = 100 + \frac{E^{\infty}_{i,A} - 100}{1 + (\frac{IC^{50}_{i,A}}{v})^{H_{i,A}}}, \tag{1}$$

where $IC^{50}_{i,A}$ is the half-living rate, $E^{\infty}_{i,A}$ is the maximum kill (effect on infinite dose), and H^A is a slope parameter. For combinatorial trails, the experimental effect $E_{i,A,B}(v_A, v_B)$ is given w.r.t. the combination of dose v_A of drug A and dose v_B of drug B on cell line i. Let's temporarily omit i for simplicity. Consider the doses X_A and X_B that cause equivalent effect on single trails w.r.t drug A and B. That is,

$$E_{A,B}(v_A, v_B) = E_A(X_A) = E_B(X_B). \tag{2}$$

The Loewe additivity [2] assumes that, when the drugs are similar,

$$1 = \frac{v_A}{X_A} + \frac{v_B}{X_B}. \tag{3}$$

Substituting E_A, E_B with the model $\hat{E}_A$ into (2), $\hat{E}_B$, we have the following equation

$$\frac{E^{\infty}_A - 100}{E^{\infty}_B - 100} = \frac{1 + (\frac{IC^{50}_A}{X_A})^{H_A}}{1 + (\frac{IC^{50}_B}{X_B})^{H_B}} \tag{4}$$

Combining equation (3) and (4), we can predict the combinatorial effect as

$$\hat{E}_{i,A,B}(v_A, v_B) \equiv \hat{E}_A(X_A), \text{where} X_A \text{is the solution of (4) and (3).} \tag{5}$$

Note that X_A can be obtained by numeric method given the parameters of $\hat{E}_A$ and $\hat{E}_B$.

Sometimes drugs have interactions that will boost the effect, which we called synergy. Define the synergy to be

$$S_{A,B}(v_A, v_B) \equiv E_{A,B}(v_A, v_B) - \hat{E}_{A,B}(v_A, v_B). \tag{6}$$

which means the difference between the experiment result and standard model. The goal of this competition is to predict synergy. However, for each prediction, we only need to submit the total synergy,

$$TS_{A,B} = \text{Volume under} S_{A,B} \text{under log axis of} v_A, v_B. \tag{7}$$

The score of each submission is given by the Pearson correlation between the real and predicted total synergy.

3 Dataset

The AstraZeneca-Sanger Drug Combination Prediction Challenge provides many data sets. In this project, the provided dataset of cell, drug, mono-therapy, combined-therapy, and gene expression will be used. The cell data include 85 different cell lines with basic information such as what kind of disease it is related to. The Drug data contains 119 drugs with 81 of them have chemical structures information. The mono-therapy data have 11759 cell experiment data showing the cell live rates under different drug concentrations. The combined-therapy (synergy) data is similar to mono-therapy data, but only have 2199 combinations of two-drug experiment data. Gene expression data is given as a matrix with size 17419 by 83 (gene by cell), showing how much is the extend of genes that are expressed in each cell line. This challenge also highly encourages participants to use external datasets to improve the predicting power of models. In this project, Gene Ontology database (GO)[3] contains the information of each genes and its corresponding biological functions and their relationship. Therefore, GO will be used to group genes according to their biological functions. Other external datasets may be added into this project in the future.

4 Methods

Using Loewe addivitity standard model to estimate the effect of drug combination
By the dose-response curve (1), Loewe additivity (3) and (4), the combined cell live rate $E^{loewe}_{A,B}(v_A, v_B)$ of Loewe model and the equivalent drug concentrations X_A, X_B can be determined. We call it the "Loewe additivity standard model". Further, we showed that the combined equation has unique solution and can be solved exactly by line-searched Newton method.

Dimension reduction by gene clustering

The contest provides high-dimensional gene expression features. The dimension of genes, 17419, is relatively large compared to the number of samples. Thus, instead of using these gene expression values directly we first sought to learn a low dimension representation of them. The idea takes advantage of the fact that gene expressions are highly correlated, i.e., genes with similar cellular functions most likely also have similar expression patterns. The main idea of our dimension reduction methodology, therefore, is to group genes with similar cellular functions together and compute the means of gene expression values for these groups. As a result, we reduce the dimension from the number of genes to the number of gene groups.

Given that drugs are designed to regulate protein (target) functions, we choose to use the target genes as the initial seeds for gene groups. There are about 100 unique target genes and for each of them we identify its similar genes, and thus we obtain about 100 groups. Subsequently, we compute the mean of expression values for genes in a group and use this single value for that group. To qualitatively measure the functional similarity between two genes, we used the Fastsemsim tool [4], which is a python library that evaluates the semantic similarity measure using gene annotations from Gene Ontology Consortium [3].

In other words, to generate one gene group we compute the pairwise similarity scores between one target gene and every other gene in the pool of 17419 genes from the microarray experiments. We then build the group by including the target gene and 20 other genes with the highest similarity scores.

Additive Damage Model
Jones, Leslie Braziel, et al. proposed the additive damage model [5] which can fitted into experiment data better than any existing mathematical synergy models. The reason to choose this model is because this mathematical model can be directly applied to our dataset using only the mono-therapy data. The concept of this model is to assume that the cell live rate curve is a sigmoidal function of damage and they define a set of damage parameters of drugs to fit the experiment data. This model serves as a baseline model for our work.

Kernelized logistic regression with lasso feature selection for drug coefficients
For a given (drug, cell) pair, we assume

$$W_d \cdot S_c = \sum_g W_{dg} * S_{cg} = \text{maximum dead rate of cell c by drug d} \tag{8}$$

where d is the drug, c is the cell, and g is the gene group, S_{cg} is the top 30% mean gene expression value of gene group g in cell c, W_{dg} means the importance of drug d to the gene group g.

```
Data: $cell \in \mathbb{R}^{m\times n}$
Result: $drug \in \mathbb{R}^{m\times l}$
while not converge do
    Generate a random permutation $perm[]$ for all $m \times l$ coordinate in $drug$;
    for coordinate $a, k$ in permutation $perm[]$ do
        Solve the one-variable sub-problem of (12) for variable $drug[a][k]$ exactly.;
        Update all linear or bi-linear terms in (12) involving $drug[a][k]$;
    end
end
```

Algorithm 1: Coordinate descent method for L1 regualrized locality model

Given that each drug only has effects on a limited number of proteins, it is assumed that the vector of drug coefficients W_{dg} on gene groups is sparse. Therefore, lasso is used to identify the sparse important drug coefficients as a feature selection step for our synergy model. For the target of the lasso regression problem, the maximum cell dead rate is chosen as the target because we believed that it characterizes the max efficacy of a drug on a cell. The drug coefficients and the gene expression values are combined as a feature vector. Then, libsvm is applied to fit synergy scores with RBF kernel and polynomial kernel (degree 2).

Locality model

The concept of locality model is demonstrated in Figure 2. It can be observed that each entry of the cell live rate matrix is highly dependent to the mono therapy of the drugs. The idea of locality model is that the cell live rate can be predicted by using one of the cell live rate of the mono therapy data as a reference starting point and adjust the value based on the curve of the other mono therapy data. Here, we choose the minimum of the two mono therapy data as the reference starting live rate and modify the live rate by a scaling factor times the damage of the other mono therapy data. To be more specific, we predict the cell live rate $\hat{E}_{\alpha,\beta}$ by

$$\hat{E}_{\alpha,\beta} := \mathrm{E}_{\alpha} - (\sum_k a_k b_k c_k + \boldsymbol{b}\cdot\boldsymbol{c})\cdot(E_\beta - E_{\beta 0}) \tag{9}$$

where α, β are the dosage of drug a and b, k is the index of gene groups, and $E_\alpha \leq E_\beta$, $(\sum_k a_k b_k c_k + \boldsymbol{b}\cdot\boldsymbol{c})$ is the scaling factor, $(E_\beta - E_{\beta 0})$ is the damage and we want to learn the drug coefficients a_k, b_k. Also, we want the drug vector to be sparse, and want to fit the total synergy. Therefore, we derived the following optimization problem to solve:

$$\min_{drug} \|drug\|_1 + \lambda^{mo} \sum_{i\in\text{mono}} \sum_{\alpha} (\hat{E}^i_\alpha - E^i_\alpha)^2 + \lambda^{co} \sum_{i\in\text{combined}} (\sum_{\alpha,\beta} w_{\alpha,\beta}(\hat{E}^i_{\alpha,\beta} - E^i_{\alpha,\beta}))^2, \tag{10}$$

$$\text{where} \quad \hat{E}^i_{\alpha,\beta} := \mathrm{E}^i_\alpha - (\sum_k a^i_k b^i_k c^i_k + \boldsymbol{b}^i\cdot\boldsymbol{c}^i)\cdot(E^i_\beta - E^i_{\beta 0}), \tag{11}$$

$$\hat{E}^i_\alpha := E^i_{\alpha 0} - (\boldsymbol{a}^i\cdot\boldsymbol{c}^i)(E^i_\alpha - E_{\alpha 0}). \tag{12}$$

$w_{\alpha,\beta}$ is the numeric integral constant for calculating total synergy score, $\lambda^{mo}, \lambda^{co}$ is the optimization parameter we need to specify, denoting the importance of mono therapy and combined therapy). This problem is not convex, but is convex quadratic in every single coordinate. A proximal coordinate descent algorithm is developed to solve this problem. $\hat{E}_{\alpha,\beta}$ is maintained during each iteration to save computation cost. The time complexity for the algorithm is $O(\#data\#feature)$ per iteration. The original gene grouping and the poly2 expansion of the gene groups are used as the cell features for this model. See Algorithm 1 for details.

5 Results

It should be noticed that redoing the cell live rate experiment with exact the same drug combination will have a Pearson correlation around 0.5. This information is from the contest holder and we regard this value as the maximum performance of any model dealing with this problem. For the models that use combined therapy data, five fold cross validation is performed to measure the testing performance. The mean Pearson correlation results of each combination of drug pairs are in table 1.

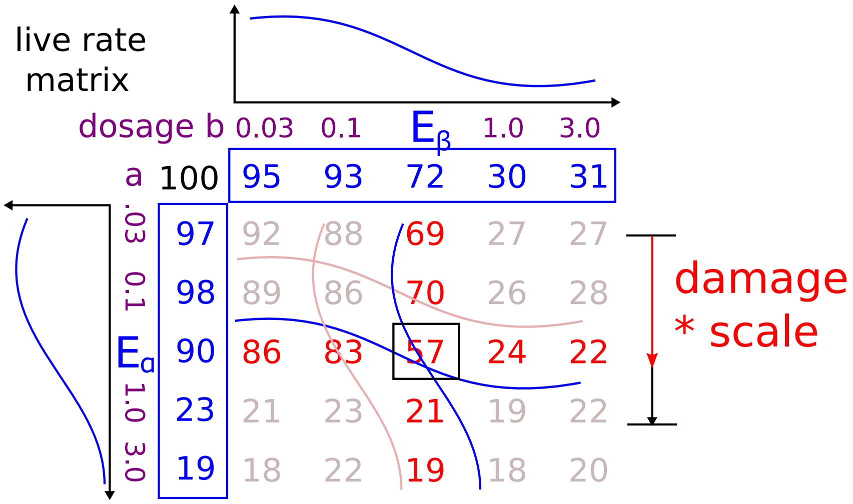

Figure 2: The concept of locality model

Table 1: The Pearson correlation of synergy models

Model/Performance	Mean Correlation	Average Non-Zero Terms
Additive Damage Model (mono-only)	0.16	X
Naive Locality Model (mono-only)	0.21	X
Logistic Regression: Poly2 Kernel (5-CV)	0.22	X
Logistic Regression: RBF Kernel (5-CV)	0.23	X
Locality: original grouping (5-CV)	0.28	8.32
Locality: poly2 grouping (5-CV)	0.28	1.03

It can be observed that the models using only mono therapy data have less performance, which is reasonable because they didn't use any combined therapy data. The naive locality model, which only use the minimum mono therapy data and ignore the scalar-adjustment term, have significantly better performance than additive damage model. This shows that the assumption of naive locality model is more suitable for this dataset. For naive kernelized logistic regression, the performance improves but not much. For the locality model, the performance increases significantly without using too many non-zero terms in drug features. It is surprising that the poly2 expansion of the gene groups can reach the same level of performance with only 1.03 non-zero terms in drug features.

6 Conclusion

For now, we have tried three strategies to deal with drug synergy problem: curve fitting, naive regression, and our newly developed locality model. We found that assuming the dataset is a sigmoidal curve and doing curve fitting results in the lowest performance. This tells us that although the sigmoidal curve may best characterize the behavior of dosage-response curve theoretically, the noises accumulated in each steps of experiment might still have a great impact to the performance. Logistic regression gives better results, but still no better than locality model. The locality model is the new way to predict live rate matrix from mono therapy without fitting any parametric curve. We developed new algorithm of proximal coordinate descent to solve the optimization problem of locality model with complexity linear to the number of features. Also, the L1 regulation term in regression plus poly2 expansion leads to good result with controllable non-zero terms, which shows that this technique can help keep the model from overfiting with stable performance in cross validation. This study is a systematic way to model drug synergy effect without using too much domain knowledge.

7 Future Work

Different gene grouping strategies (rewrite)
For now we only choose the gene groups that are built from the given drug target information and GO gene ontology database. However, the resulting 100 groups may not be enough to cover all the targets because drugs can have undiscovered targets. We will add other groupings strategies to cover as much possible target as we can and at the same time not increase the dimension of gene groups too much.

Adding non-linear term in optimization problem
We assume that, in each rows or columns of the live rate matrix, the differences corresponding to an anchoring point are similar to the difference in a mono-therapy. However, it is not true when one drug is dominating the other. In such scenario, the differences of the non-dominating drug are smaller than the differences of the dominating drug because of saturation. To address this issue, we need to add an non-linear saturation terms in the scaling factor that's related to the dosages of both drug.

Adding more domain knowledge
Now we use L1 regularization and consider only the sparsity of the drug parameters. If we wish to model the similarity between drugs or to use the relationship between proteins, we can use a different regularization term. Example includes using group-Lasso regularization, graph-Laplacian normalization [6], and graph-induced norms.

Discover biological factors to predict synergy
We found that locality model can predict drug synergy by using a very little number of features. Our next goal is to find biological factors related to the features and verify it to see if these factors actually have important impact to synergy effect in real world. We hope that our model can not only have good predicting power but also help to discover new important biological factors.

8 References

[1] Contest Website https://www.synapse.org/#!Synapse:syn4231880/wiki/235645

[2] Fonnum, Frode, and Espen Mariussen. "Mechanisms involved in the neurotoxic effects of environmental toxicants such as polychlorinated biphenyls and brominated flame retardants." Journal of neurochemistry 111.6 (2009): 1327-1347.

[3] Gene Ontology Database Website http://geneontology.org/page/go-database

[4] Fastsemsim tool http://pythonhosted.org/fastsemsim/

[5] Jones, Leslie Braziel, et al. "The additive damage model: A mathematical model for cellular responses to drug combinations." Journal of theoretical biology 357 (2014): 10-20.

[6] N. Rao, H. Yu, P. Ravikumar, I. Dhillon. "Collaborative Filtering with Graph Information: Consistency and Scalable Methods" Neural Information Processing Systems (2015).

Group Formation

This project is supervised by Prof. Ziv Bar-Joseph. One of the group member, Yan Xia, is not in the class. He is a MS student in Machine Learning Department and has a solid background in Computational Biology. Yan Xia and Chieh Lin are both in Prof. Ziv Bar-Joseph's research group. Because the actual number of people working on this project is three, we didn't recruit the third person from class as a group member.