Research Proposal about Drug Synergy Prediction

Ruiqi ZHUANG

Department of Computer Sciences
Hong Kong Baptist University, Hong Kong
February 2023

1 Introduction

Monotherapy is widely practiced as a standard treatment regimen for various diseases. However, the effectiveness of this conventional therapy is restricted by several factors, including the complexity of human diseases (Sun et al., 2020), patient heterogeneity, and drug resistance. One of the most significant limitations of using monotherapy is the development of drug resistance, which can arise due to multiple factors associated with the cancer cell line's system (Torkamannia et al., 2022). As a result, pharmacy researchers use combinatorial drug therapy as an alternative to address these issues. Combinatorial drug therapy is a treatment approach that utilizes more than one medicine or treatment simultaneously to enhance the therapeutic outcomes of individual drugs. Consequently, this approach has been reported to significantly enhance curative efficacy in treating various diseases, particularly cancer (Sun et al., 2020).

Despite the tremendous potential of combinatorial drug therapy, drug combinations that are efficient and safe remain elusive. The search for optimal drug pairs is further complicated by several challenges, including systemic toxicity, and potential side effects. The core of this therapy is the identification of synergy effect, while this procedure can be costly and time-consuming. Therefore scientists have been working for years to tackle this issue. One approach currently being pursued is High Throughput Drug Screening (HTS), a computational method that utilizes large-scale of drug sensitivity data to make predictions on drug response effects.

However, the effectiveness of HTS is dependent on a better understanding of tissue biology and addressing issues of understudied tissue. Ultimately, addressing these issues will foster the development of efficient and effective drug combinations, leading to improved treatment outcomes. To tackle this issue, researchers tried different kinds of ways like transfer learning pseudo data generation and tuning on pre-trained model to eliminate the influence of performance due to the lack of data caused by understudied tissue.

2 Literature review

Traditional techniques for training predictive models to estimate the presence of synergistic effects between two drugs have commonly relied on the premise that the available data is adequate to train models that exhibit high generalizability. For instance, DeepSynergy employs a multi-layer deep neural network that takes the chemical descriptors of two drugs along with genomic data from a cell line as inputs. The embedded information propagates via the neural network, which ultimately yields the synergy score output. MatchMaker presents an alternative approach, utilizing both drug-specific networks and synergy-prediction networks to make predictions. Within the drug-specific network, a single-drug chemical descriptor is concatenated with untreated cells' gene expression data, and these values are propagated through a multiple-layered neural network. This process outputs the drug representation. The synergy-prediction network then takes the output of two drug-specific networks as input and uses drug representation values to propagate the drug representation through a multi-layered neural network, thereby making a prediction for the synergy score.

However, significant issues arise when dealing with understudied tissues where in vivo experiments are difficult to conduct, leading to an insufficiency of data for model training. To overcome the challenge of understudied tissues, researchers have been exploring various methods. For instance, Kim et al., (2021) used transfer learning to solve the data scarcity problem. The model learned the knowledge embedding of drugs, cell lines from SMILES, MACCS fingerprint, FPKM values and disease type, then utilized transfer learning to transfer knowledge from data-rich tissues to understudied tissues, based on the fact that different tissues share certain biological features, with similar features these tissues might have similar drug response. Kim et al., (2021) developed CancerGPT by fine-tuning on generative pretrained transformer (GPT). CancerGPT transformed the original task to a natural language inference task and modified the model of GPT-2 and GPT-3 to generate the answer of synergistic and antagonistic based on the knowledge encoded from tabular input of synergy data containing the information of drug cell line drug sensitivity and synergy score. LiYu (2023) proposed a method namely graph structure learning to predict the drug-targets interaction and drug-drug interaction in understudied tissues, so as to obtain a refined graph with sufficient data for further training. The model, leveraged pre-trained models to extract biological, chemical, and relation information from SMILES amino acid sequence and disease relation data, ultimately generating node embeddings for graphs. The initial graph, representing the relationships among drugs, targets, and diseases, was established using knowledge graph relations. Then utilizing graph structure learning with attention mechanisms, LiYu (2023) was able to obtain the refined graph representation for prediction making. Specifically, attention mechanisms were utilized to calculate the similarity of two drugs or drug-target pairs, determining the existence and type of drug-drug or drug-target edges based on the resulting similarity scores. Additionally, LiYu (2023) employed a self-training strategy to make further improvements.

3 Limitation and improvement

The aforementioned methods take solely the gene expression data before the treatment into account. However, Kong et al., (2022) (this is one paper mentioned in the work of sun kim1) has indicated that the gene expression profile before treatment does not significantly assist in predicting drug response. Furthermore, the experiment conducted by LiYu (2023) did not reveal a strong increase in performance by using pre-trained models. Additionally, the efficacy of solely utilizing graph structure learning has not been specifically examined in any study to date.

4 Method

The proposed framework is comprised of three key inputs, namely drug function embedding, cell line embedding, and disease relation embedding. The drug function embedding pertains to a combination of drug embedding generated by KPGT model and post-treatment gene expression profile generated by NetGP, as proposed by Sum Kim1 (2023). For the cell line embedding, the residual information from the output of EMS-1b model is extracted. EMS-1b is a powerful protein language model that leverages masked amino acid sequence inputs and outputs predictions of missing tokens. To acquire disease information, the study employs RotatE from PrimeKG- a knowledge graph comprising diverse knowledge- to extract disease relations. The proposed framework aims to generate effective predictions related to drug response based on these three inputs.

In this study, we propose a mechanism for drug data augmentation utilizing a Molecular Function Similarity Comparison algorithm and a novel supervised chemical graph mining technique proposed by Sum Kim2 (2023). The graph mining technique is able to generate substructures of drug molecules that have similar functions as the original molecule as candidate drugs. These generated sub-structures are filtered using Molecular Function Similarity Comparison algorithm to retain the remaining drug data for augmentation purposes. Specifically, the algorithm employs a transformer encoder that takes the SMILES as input to encode the embedding of both the original drug and the generated drug, then the similarity between the original drug and generated drug is calculated based on their respective embeddings, finally, a similarity threshold is set for selecting generated drugs.

Self-training strategy represents a straightforward yet highly effective approach. The strategy involves initially training a model with a particular set of labeled data. Subsequently, the model utilizes the trained information to predict labels for a given set of unlabeled data. Finally, the model retrains and refines itself using a combination of the original labeled data and the newly-predicted labeled data generated through the utilization of the trained model. The self-training strategy has demonstrated its efficacy in numerous studies, making it a popular choice for enhancing model performance. Thus, we utilized self-training as an approach to improve the performance of our proposed mechanism for drug data augmentation.

5 Dataset and Evaluation

5.1 Dataset

The dataset of this study integrates the following databases DrugComb database, TTD (Therapeutic Targets Database), NIH-LINC (Library of Integrated Network-Based Cellular Signatures), CCLE (Cancer Cell Line Encyclopedia), COSMIC (Catalogue of Somatic Mutations in Cancer) and CMap LINCS L1000 (Connectivity Map).

DrugComb gathers, analyzes, and disseminates drugs combination screen results on cancer cell lines. This database contains 8397 drugs, 2320 cell lines, 33 tissues, and 739,964 pharmacological combinations. Experiments are conducted on the skin, lungs, ovary, and other organs. TTD contains information on drug targets, such as biological functions, illnesses linked with them, and medications that target them. NIH-LINC is a database of gene expression data from various cellular perturbations, such as pharmacological treatments. Based on the effects on gene expression, it can be used to identify potentially effective drug combinations. CCLE contains genetic and pharmacological information on a considerable number of cancer cell lines. It may be utilized for identifying potential drug targets and screening the effectiveness of drugs. COSMIC is a database of cancer somatic mutations, including those that may affect drug effectiveness. It may be used to identify potential drug targets and to test medications for effectiveness in specific cancer types. CMap LINCS L1000 utilizes the concept of Connectivity Map (CMap) to connect the information of genes, drugs, and diseases as a virtue gene-expression signature. This dataset has around 1.3 million profiles derived from NIH LINCS Consortium.

5.2 Evaluation on drug synergy

In order to evaluate the drug synergy, the following mathematical models could be applied to calculate the synergy value for drug and cell line combinations.

5.2.1 Bliss model

This is an assumption of a stochastic process in which both two drugs act independently. The expectation of drug synergy can be calculated from independent event probabilities (Bliss, 1939).

$$y_{BLISS} = y_1 + y_2 - y_1 * y_2 \tag{1}$$

in which

$$y_{1,2} \in [0, 1] \tag{2}$$

are the effect of a specific drug measured by cell death or cell growth.

5.2.2 Highest Single Agent (HSA)

The expectation of drug synergy is equal to the higher effect of individual drugs (Berenbaum, 1989).

$$y_{HSA} = \max(y_1, y_2) \tag{3}$$

5.2.3 Loewe additivity model

This model takes a different assumption in which the dose-response curves of individual drugs are taken into consideration. The dose-response curves can be described by the 4-parameter log-logistic (4PL) curve (Loewe, 1953).

$$\frac{x_1}{m_1(\frac{y_{LOEWE}-E^1_{min}}{E^1_{max}-y_{LOEWE}})^{\frac{1}{\lambda_1}}} + \frac{x_2}{m_2(\frac{y_{LOEWE}-E^2_{min}}{E^2_{max}-y_{LOEWE}})^{\frac{1}{\lambda_2}}} = 1 \tag{4}$$

in which

$$E_{min}, E_{max} \in [0, 1] \tag{5}$$

are the minimal and maximal synergy effects of the drug, and $m_{1,2}$ are the dose of drugs that cause the midpoint of

$$E_{min} + E_{max} \tag{6}$$

$m_{1,2}$ is also known as EC_{50} or IC_{50}, and

$$\lambda_{1,2}(\lambda > 0) \tag{7}$$

are the shape parameters reflecting the sigmoidicity or slope of dose-response curves. After that, a numerical nonlinear solver may be used to calculate yLOEWE (x1, x2).

5.2.4 Zero Interaction Potency (ZIP)

Given the assumption that the drugs do not enhance each other, this model estimates the expectation of drug synergy of two drugs (Yadav et al., 2015).

$$y_{ZIP} = \frac{(\frac{x_1}{m_1})^{\lambda_1}}{(1+\frac{x_1}{m_1})^{\lambda_1}} + \frac{(\frac{x_2}{m_2})^{\lambda_2}}{(1+\frac{x_2}{m_2})^{\lambda_2}} - \frac{(\frac{x_1}{m_1})^{\lambda_1}}{(1+\frac{x_1}{m_1})^{\lambda_1}} \frac{(\frac{x_2}{m_2})^{\lambda_2}}{(1+\frac{x_2}{m_2})^{\lambda_2}} \tag{8}$$

5.3 Evaluation on drug combination sensitivity

The sensitivity of drug combination is evaluated by the area under the log-scaled dose-response curve (AUC) and combination sensitivity score (CSS).

5.3.1 Area under the log-scaled dose-response curve (AUC)

$$AUC = \int_{c_1}^{c_2} y_{min} + \frac{y_{max} - y_{min}}{1 + 10^{\lambda(\log_{10} IC_{50} - x')}} dx' \tag{9}$$

in which $[c_1, c_2]$ is the concentration of the drug that is tested foreground.

5.3.2 Combination sensitivity score (CSS)

CSS is obtained using relative IC50 values of compounds and the area under the corresponding dose-response curves. To calculate the CSS of a drug pair, one can use a particular concentration for one compound and another using changing concentrations, outputting two CSS values that are then averaged. The dose-response curve for each drug is simulated via a 4-parameter log-logistic curve, so that:

$$u = y_{min} + \frac{y_{max} - y_{min}}{1 + 10^{\lambda(\log_{10} IC_{50} - x')}} \tag{10}$$

in which y_{min}, y_{max} are the minimal inhibition and maximal inhibition respectively, and x' = $\log_{10} x$

6 Summary

This research proposal simply introduced drug synergy and reviewed the related work about drug synergy prediction focusing on the solution for understudied tissue problem. Then we further investigate to discuss the limitation of several works and try to propose a way to make improvements through drug data augmentation. Finally, introduce the dataset and mathematical models to do the evaluation on drug synergy and drug combination sensitivity.

References