Artificial Intelligence in Medicine xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

journal homepage: www.elsevier.com/locate/aiim

Prediction of synergistic anti-cancer drug combinations based on drug target network and drug induced gene expression profiles

Xiangyi Li [a,b,1], Yingjie Xu [c,1], Hui Cui [b,e], Tao Huang [d], Disong Wang [b], Baofeng Lian [b], Wei Li [b], Guangrong Qin [b,*], Lanming Chen [a,*], Lu Xie [b,*]

[a] Key Laboratory of Quality and Safety Risk Assessment for Aquatic Products on Storage and Preservation (Shanghai), China Ministry of Agriculture, College of Food Science and Technology, Shanghai Ocean University, 999 Hu Cheng Huan Road, Shanghai 201306, PR China
[b] Shanghai Center for Bioinformation Technology, Shanghai Academy of Science and Technology, 1278 Keyuan Road, Shanghai 201203, PR China
[c] Tongren Hospital Shanghai Jiao Tong University School of Medicine, 1111 Xianxia Road, Shanghai 200336, PR China
[d] Institute of Health Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 320 Yueyang Road, Shanghai 200031, PR China
[e] School of Life Science and Technology, ShanghaiTech University, Shanghai 201210, PR China

ARTICLE INFO

Article history:
Received 31 December 2016
Received in revised form 18 April 2017
Accepted 11 May 2017

Keywords:
Synergistic drug combinations
Cancer
Network features
Gene expression profiles
Random forest

ABSTRACT

Objective: Synergistic drug combinations are promising therapies for cancer treatment. However, effective prediction of synergistic drug combinations is quite challenging as mechanisms of drug synergism are still unclear. Various features such as drug response, and target networks may contribute to prediction of synergistic drug combinations. In this study, we aimed to construct a computational model to predict synergistic drug combinations.
Methods: We designed drug physicochemical features and network features, including drug chemical structure similarity, target distance in protein–protein network and targeted pathway similarity. At the same time, we designed fifteen pharmacogenomics features using drug treated gene expression profiles based on the background of cancer-related biology network. Based on these eighteen features, we built a prediction model for Synergistic Drug combination using Random forest algorithm (SyDRa).
Results: Our model achieved a quite good performance with AUC value of 0.89 and Out-of-bag estimate error rate of 0.15 in training dataset. Using the random anti-cancer drug combinations which have transcriptional profile data in the Connectivity Map dataset as the testing dataset, we identified 28 potentially synergistic drug combinations, three out of which had been reported to be effective drug combinations by literatures.
Conclusions: We studied eighteen features for drug combinations and built a computational model using random forest algorithm. The model was evaluated using an independent test dataset. Our model provides an efficient strategy to identify potentially synergistic drug combinations for cancer and may help reduce the search space for high-throughput synergistic drug combinations screening.

1. Introduction

Cancer has been a global health threat and one of the leading causes of death worldwide. Resection, chemotherapy, and radiotherapy are the most common therapies in cancer. Also, target drugs and immune therapy have been promising therapies. Dozens of chemotherapeutic agents and targeted anti-cancer drugs have been approved. Although chemotherapy has significantly enhanced cancer treatment, the death rate and recurrence rate are still high [1]. Drug resistance and side effect have been considered as the main reasons for the failure of cancer chemotherapy [2,3]. Synergistic drug combination therapy could be an ideal solution [4]. Synergistic drug combinations can significantly decrease the drug dosage but enhance or maintain the same efficacy to avoid toxicity and minimize or slow down the development of dug resistance. However, available synergistic anti-cancer drug combinations are limited and experimental screening is expensive, thus, the computational prediction of synergistic anti-cancer drugs could be crucial and urgent.

In recent decades, with rapid accumulation of drug-target interactions, protein–protein networks, drug physicochemical property information, tumor omics data and pharmacogenomics informa-

* Corresponding authors.
E-mail addresses: grqin@scbit.org (G. Qin), lmchen@shou.edu.cn (L. Chen), xielu@scbit.org (L. Xie).
[1] These authors contributed equally.

http://dx.doi.org/10.1016/j.artmed.2017.05.008
0933-3657/

tion, the research on the relationship among drugs, target genes and cancers have made great progress. Many synergistic drug combination prediction models based on above mentioned information have been proposed [5,6]. The NLLSS platform integrated known synergistic drug combinations, drug target interactions, and drug chemical structure to predict potentially antifungal synergistic drug combination [7]. According to drug similarity scores, a classifier based on framework of Laplacian Regularized Least Square was built to distinguish synergistic drug combinations ($drug_{sy}$) from non-synergistic drug combinations ($drug_{non\text{-}sy}$). Similarly, Zhao et al. collected information such as drug target proteins, the drug Anatomical Therapeutic Chemical (ATC) Classification System code, drug side effect and drug indication from approved drug combinations. Based on enrichment analysis of these information collected from individual drug of approved drug combinations, they computed an enrichment score for each query drug combination. Drug combinations with enrichment score above a certain threshold would be more likely to be new drug combinations [8]. These models mainly depend on comparison of similarities between each query drug combination and known drug combinations, however, these features could not illustrate the mechanism of drug–drug interaction. Network-based computational biology is a booming and promising strategy in the area of biomedicine research [9–11]. Predicting synergistic drug combinations based on biology molecular network may be more complex but can benefit the research of mechanism of drug–drug interaction [12]. PEA model not only included drug properties such as drug chemical structure, ATC code, side effect, but also studied drug target protein sequence, connection of drug targets in PPI network and similarity of drug targets based on Gene Ontology for synergistic drug combinations [13]. Then drug similarity score of the above six features and the likelihood ratio (LR) of query drug pairs to the approved drug combinations were calculated by a Bayesian network approach. Di Chen et al. built a model in the pathway–pathway interaction perspective [14]. They studied the characteristics of pathways perturbed by synergistic drug combinations and deduced that synergistic drug combinations might be more likely to affect the same pathway through different targets or members of synergistic drug combinations might tend to regulate a few highly-associated pathways.

However, features mentioned above can only reflect the similarity between drugs to an extent and cannot directly reflect the drug effect on cells. As a drug with different dose may have different efficacy, the efficacies of drug combinations with different doses may also differ. Gene expression profiles following drug treatment with different doses can reflect the biological response to drug treatment, which may further provide clues for recognizing the mechanism of drug action and predicting synergistic drug combinations [15,16]. Zikai Wu et al. built a prediction model to identify effective cocktail drugs based on drug induced gene expression profiles. They assumed that a sub-network or pathway would be affected in the network cellular system after administrating a drug and proposed a method to predict the gene expression profiles of drug combinations based on the gene expression profiles of individual drug. Their model could output a score which could reflect the overall effect of the individual drug or drug combinations [15]. Yin-Ying Wang et al. constructed a genetic interaction network and studied the relationship between drug combinations and their target proteins. They found that effective drug combinations might tend to have a smaller effect radius (i.e. average shortest distance of drug targets) in the genetic interaction network and effective drug combinations might be more likely to modulate functionally related pathways [17]. Similarly, in the synergistic drug prediction model RACS [18], gene expression profiles following drug treatment in different dosage were also considered. By involving synergistic drug combinations and random drug combinations, the authors applied a semi-supervised learning algorithm and achieved an excellent performance after applying RACS to test dataset. This model was built based on seven features which were related to properties of drug target network. Adding two features which were used to describe the correlation between differential expression genes (DEGs) following drug treatment can significantly increase the prediction sensitivity.

In this manuscript, we built a simple prediction model of synergistic anti-cancer drug combinations based on drug target network features and pharmacogenomics features using random forest algorithm. As is shown in Fig. 1, the gene expression profiles of drug perturbation from DREAM Challenge 7 sub-challenge 2 (hereinafter abbreviated as DREAM) was used as the training dataset [19], while the gene expression profiles of anti-cancer drug perturbation from Connectivity Map [20] (CMap) dataset was used as an independent test dataset. All the pharmacogenomics features are newly designed. After performing feature selection, we chose the model with best performance on training dataset to predict potentially synergistic anti-cancer drug combinations which consisted of drugs in CMap dataset.

2. Materials and methods

2.1. Collection of drug targets

In total, seventeen approved anti-cancer drugs which are also included in CMap database were collected (supplementary file 1). The target information for drugs in DREAM was collected from the work of Goswami et al. [21], which can be found in supplementary file 2, and the target information of the approved anti-cancer drugs in CMap was collected from DrugBank [22], as listed in supplementary file 3.

2.2. Identification of significantly differential expression genes after drug perturbation

The DREAM provided two types of data: the gene expression profiles which were generated from the perturbation of 14 drugs at two concentration (IC_{20} at 24 h and 48 h) for three time points (6 h, 12 h and 24 h) with DMSO as control on the diffuse large B-cell lymphoma (DLBCL) cell line, and the gold standard data generated from 91 ($C_{14}^2 = 91$) drug combinations treated DLBCL cell line at their respective IC_{20} at 60 h for 60 h [19]. IC_{20} value represents the concentration of each drug to suppress 20% viability of the cell line. The gene expression profile data can be downloaded from Gene Expression Omnibus (GEO) and its identifier ID is GSE51068. The gold standard data of this project can be downloaded from http://www.nature.com/nbt/journal/v32/n12/full/nbt.3052.html. To get the differential expression genes (DEGs), we compared the gene expression profiles of drug treated samples (IC20 at 48 h, 24 h) with control (DMSO) samples, The training dataset was firstly performed with quality control by a R package called arrayQualityMetrics [23]. Then the training dataset was normalized with RMA normalization method using R package affy [24]. Finally the DEGs induced by each drug were screened by comparing the gene expression changes between the drug treatment group and respective control group through *t*-test. The threshold for DEGs screen was set as P-value < 0.05, $|\log_2 \text{FoldChange}| > 1$. A total of 170 approved anti-cancer drugs were collected from U.S. Food and Drug Administration (FDA, https://www.fda.gov/). By mapping these drugs to CMap dataset, we found 17 out of 170 were used to treat MCF7 cell line. We collected twenty instances generated from treatment of the above mentioned 17 anti-cancer drugs. Here one instance denotes one experiment that one drug treats cell line. The treatment of same drug with different concentration denotes different instances. Then significant DEGs were screened depending on the

Please cite this article in press as: Li X, et al. Prediction of synergistic anti-cancer drug combinations based on drug target network and drug induced gene expression profiles. Artif Intell Med (2017), http://dx.doi.org/10.1016/j.artmed.2017.05.008