Liu *et al. BMC Bioinformatics* (2019) 20:645
https://doi.org/10.1186/s12859-019-3288-1

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Predicting effective drug combinations using gradient tree boosting based on features extracted from drug-protein heterogeneous network

Hui Liu[1†], Wenhao Zhang[1†], Lixia Nie[2], Xiancheng Ding[3], Judong Luo[4*] and Ling Zou[2*]

Abstract

Background: Although targeted drugs have contributed to impressive advances in the treatment of cancer patients, their clinical benefits on tumor therapies are greatly limited due to intrinsic and acquired resistance of cancer cells against such drugs. Drug combinations synergistically interfere with protein networks to inhibit the activity level of carcinogenic genes more effectively, and therefore play an increasingly important role in the treatment of complex disease.

Results: In this paper, we combined the drug similarity network, protein similarity network and known drug-protein associations into a drug-protein heterogenous network. Next, we ran random walk with restart (RWR) on the heterogenous network using the combinatorial drug targets as the initial probability, and obtained the converged probability distribution as the feature vector of each drug combination. Taking these feature vectors as input, we trained a gradient tree boosting (GTB) classifier to predict new drug combinations. We conducted performance evaluation on the widely used drug combination data set derived from the DCDB database. The experimental results show that our method outperforms seven typical classifiers and traditional boosting algorithms.

Conclusions: The heterogeneous network-derived features introduced in our method are more informative and enriching compared to the primary ontology features, which results in better performance. In addition, from the perspective of network pharmacology, our method effectively exploits the topological attributes and interactions of drug targets in the overall biological network, which proves to be a systematic and reliable approach for drug discovery.

Keywords: Drug combination, Random walk, Heterogenous network

Background

Traditional "one drug, one target" treatment can not always lead to desirable therapeutic effect on complex diseases, because biological pathways and networks are often redundant and robust to single point perturbations [1]. Drug combination perturbs the biological network through synergistic and synthetic lethal effects, and inhibits more effectively the activity level of pathogenic genes [2, 3]. Previous studies have shown that combination drugs can effectively inhibit cancer cell growth or promote cancer cell apoptosis, with reduced toxicity and side effects than single target drugs [4]. Even more promising, drug resistance can be decreased or even overcome through combination therapy [5]. Therefore, therapeutic schemes from single- to multi-target drugs play an increasingly important role in the treatment of complex diseases [6].

Despite the increasing successes of combination drugs in inhibiting cancer cell proliferation, most of them are discovered by clinical experience or by occasional chances

*Correspondence: judongluo@163.com; zouling@cczu.edu.cn
†Hui Liu and Wenhao Zhang contributed equally to this work.
[4]Department of Radiation Oncology, the Affiliated Changzhou No.2 People's Hospital of Nanjing Medical University, Changzhou, China
[2]School of Information Science and Engineering, Changzhou University, Jiangsu, China
Full list of author information is available at the end of the article

© The Author(s). 2019 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

[5, 6]. Developing combinations of targeted agents is more difficult than developing a single agent [7], as inhibiting the cross-talks among multiple pathways depends on our insight into the pathway interdependencies underlying the cancer cell proliferation and survival in a specific cancer type [8, 9]. The high-throughput screening (HTS) experiments currently used to evaluate drug combinations are still time- and cost- consuming because they rely heavily on the search for a large number of possible target combinations [10–12]. So, there is an urgent demand for rational and systematically in silico methods to narrow down the candidates for combinatorial drugs for wet-lab experimental validations [13].

Quite a few computational methods have been proposed to predict cancer sensitivity to combinatorial drugs [1, 4, 14–16]. The existing methods can be roughly divided into two categories: system biology-based methods [17] and network-based analysis [15, 18]. System biology-based methods mathematically model the perturbation of drugs using biochemical reactions and kinetic parameters, which are often limited to small scale and well-studied signaling pathway. Network-based methods often exploit genomic, chemical and pharmacological properties to build an overall network composed of the associations among drugs, proteins and pathways, and then adopt scoring rules [19, 20], optimal combination searches [1, 16, 21], machine learning [4, 18] to predict potential drug combinations. As network-based methods integrate various kinds of ontological features and interactions between different subject of interest, some of these methods achieve remarkable performance in predicting drug combinations. For example, Ligeti et al. [20] proposed so-called Target Overlap Score (TOS) prioritization function, which is defined for two drugs as the number of jointly perturbed targets divided by the number of all targets potentially affected by these two drugs, to rank candidate drug combinations. Pang et al. [1] proposed mixed integer linear programming to find balanced target set cover (BTSC) and minimum off target set cover (MOTSC) for combination therapy. Huang et al. [18] propose DrugComboRanker, which first builds a drug functional network based on their genomic profiles, and disease-specific signaling networks based on patients genomic profiles and interactome data, and then prioritize synergistic drug combinations by searching drugs whose targets are enriched in the complementary signaling modules of the disease signaling network. Matlock et al. [21] tried to find drug combinations maximizing sensitivity over tumor cell models while minimizing toxicity over normal cell models, and then proposed a lexicographic search algorithm to find optimal target set. In addition, some methods exploit the concept of synthetic lethality to discover combinatorial drugs [3, 22, 23]. However, most of previous methods are usually limited to the ability to dissect potential molecular mechanisms, or to associate multiple drugs to one disease in huge pharmacological space.

There have been many approaches that integrate multiple heterogeneous networks to infer the associations between biological entities, including lncRNA functions [24–26], lncRNA-disease associations [27], drug-disease associations [28, 29] and gene functions inference [30]. Inspired by heterogeneous network-based inference, we ran random walk with restart on the drug-protein heterogenous network to extract features for drug combinations, and then trained gradient tree boosting classifier using the extracted features to predict new drug combinations. Concretely, we integrated a variety of data sources, including chemical structures of the drugs, protein sequences, and known drug-protein associations, to construct a drug-protein heterogeneous network. The random walk with restart procedure was implemented on the heterogenous network using the combinatorial drug and their targets as the initial probability, respectively. The converged probability distribution was used as feature vector of the drug combination. Based on the probability distribution vectors, we subsequently trained the gradient tree boosting (GTB) classifier, which achieved the AUC of 0.949 by 10-fold cross-validation. We also compared our method to other seven typical classifiers, including kNN, SVM, Logistic regression, Naive Bayes, AdaBoost, Random Forest and LogistBoost. The performance comparison results demonstrate that our proposed model significantly outperformed other traditional methods. From the perspective of network pharmacology, our method effectively make use of the topological attributes and functional interactions of drug targets in the protein-protein network.

Results

Drug combination dataset

The set of effective drug combinations was obtained from DCDB 2.0 [31], a typical drug combination database focused on collecting verified drug combinations to facilitate further exploration, including theoretical modeling and simulation of such beneficial drug combinations. In total, the current version(2.0) of DCDB includes 1363 drug combinations (330 approved and 1033 investigational, including 237 unsuccessful usages), covering 904 individual drugs and 805 targets. We selected those combinations that are approved or under trials in DCDB as positive samples. Note that the number of non-effective drug combinations is actually enormous, much larger than that of effective in real world. Therefore, we generated a number of negative samples of drug combinations by randomly picking up pairwise drugs to balance the positive and negative samples in our benchmark set. The strategy of generation of negative samples has been widely adopted in the prediction of drug-target interactions and drug-disease associations [28, 32]. Importantly, the drug set that we selected pairwise combinations is expanded from the individual drugs in DCDB to their most associated 3 drugs according to STITCH, yielding 3266 drugs in total. Finally, the benchmark drug combination set contains 1359 positive combinations and 1359 negative combinations.

Performance measures

We conducted performance evaluation using 10-fold cross validations. In particular, the training set were randomly divided into ten subsets and each subset had roughly equal size to others. Each subset was in turn used as the test set, and the remaining nine subsets were used as training set. This validation process was repeated ten times and each performance measure was averaged over the ten folds for performance evaluation. A couple of performance measures were used in our experiment, including precision (PRE), recall (REC), F-measure, Matthews correlation coefficient (MCC) and the area under the receiver operating characteristic curve (AUC). They are formally defined as below:

$$Precision = \frac{TP}{(TP + FP)} \tag{1}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{2}$$

$$F - Measure = \frac{(2 * Precision * Recall)}{(Precision + Recall)} \tag{3}$$

$$MCC = \frac{(TP \times TN - FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{4}$$

in which *TP* and *TN* represent the numbers of correctly predicted positive and negative samples, *FP* and *FN* represent the numbers of wrong predicted positive and negative samples, respectively. Additionally, the AUC score is computed by varying the cutoff of the predicted scores from the smallest to the greatest value.

Impact of parameters on performance

To explore the impact of parameter λ, which is the probability of random walker jumping to different type of network, We gradually increased its value from 0.1 to 0.9 at interval of 0.1. The aforementioned metrics obtained by 10-fold cross-validation are shown in Table 1, which demonstrate that λ has a moderate impact on the prediction performance of our proposed method. In terms of AUC, the values fit approximately a parabola which hit the top 0.944 at λ 0.7, which was thus adopted in our subsequent experiments. For other two parameters introduced in random walk with restart, restart probability α and tradeoff η, we conducted similar tuning to determine their optimal values that achieve the best performance. As shown in Additional file 1: Table S1 and S2, the AUC measure reached the highest value when α and tradeoff η were equal to 0.2 and 0.9. According to the results, the restart probability α has a negligible effect on the AUC. Generally, the restart probability is a heuristical parameter without any theoretical guide or justification when selecting [33]. However, the heterogeneous network is established based on drug-drug similarity, protein-protein similarity and known drug-protein associations, resulting in a heterogeneous network with quantitative weighted edges. From this perspective, since the random walk simulates the influence of drugs in protein network, the convergence state will have a bias on higher weighted nodes. Therefore, the restart probability may have a slight effect on the final distribution. As a result, we set the three parameters λ, α and η to 0.7, 0.2 and 0.9 in the following performance comparison experiments.

Table 1 Impact of the parameter λ on the performance of GTB classifer

λ	Precision	Recall	F-Measure	MCC	AUC
0.1	0.861	0.852	0.856	0.715	0.929
0.2	0.865	0.853	0.858	0.720	0.930
0.3	0.870	0.854	0.861	0.726	0.934
0.4	0.878	0.861	0.869	0.738	0.939
0.5	0.883	0.871	0.877	0.755	0.941
0.6	0.885	0.868	0.875	0.755	0.941
0.7	**0.887**	**0.867**	**0.876**	**0.757**	**0.944**
0.8	0.885	0.865	0.875	0.754	0.943
0.9	0.884	0.864	0.874	0.752	0.943

The boldface figures indicate that GTB classifier achieves the best performance at λ equal to 0.7

Performance comparison to typical classifiers

To demonstrate the outstanding performance of our method, we carried out performance evaluation on the benchmark combination set by comparing our method with seven other typical classifiers, including kNN, SVM, Logistic regression, Naive Bayes, Random Forest, Adaboost and LogitBoost. Based on the derived feature distribution vectors, we implemented these competitive classifiers separately using R package [34] so as to conveniently reproduce our work. For Native Bayes, we adopted the R package e1071 [35] and its default setting. Also, logistic regression and SVM are implemented based on the e1071 R package, and logistic regression was run with default settings, while the misclassification penalty coefficient for SVM varied from 10 to 10000 by interval of 500 to achieve best performance. For KNN, R package kknn [36] was used to run the algorithm, in which the parameter k (k=1, 3, 5, 7 and 9) was enumerated to tune its performance. For the distance metric of kNN, we have tried Manhattan distance, Euclidean distance and Chebyshev distance and found that they yield to similar performance, thereby we adopted Chebyshev distance (q=5) in the performance evaluation. The R package randomForest [37] was used to run random forest algorithm and the number of trees varies from 60 to 500 by interval of 20. For boosting methods Adaboost and Logitboost, the R packages Adabag [38] and caTools were used, where the number of training iterations was tuned from 10 to 100 by interval of 5 and 10, respectively. The performance measures of each comparative method, including precision, recall, F1, MCC and AUC, achieved by the fine-tuned parameters, are shown in Table 2. Apparently, our proposed method significantly outperformed other classifiers in terms of almost all performance metrics.

To present clear performance comparison, the ROC curves of GTB and other seven classifiers are also illustrated in Fig 1. It can be demonstrated that GTB classifier greatly outperforms all other competitive methods, which achieves the highest AUC value 0.95, followed by Random forest and Adaboost at 0.86. The performance of Naive Bayes is the worst and gets only 0.508 AUC value.

Performance improvement by heterogenous network-derived features

To validate the effectiveness of the features extracted from drug-protein heterogeneous network, we conducted performance comparison between the primary ontology features and heterogenous network-based features. Due to different number of individual drugs and target proteins involved in drug combinations, we can not directly concatenate the drug fingerprints and protein GO annotations to construct feature vectors that are inconsistent in dimension. Instead, we first unified the chemical fingerprints of individual drugs in a combination, i.e. union of individual fingerprint vectors, as well as the union of GO terms of target proteins of individual drugs. Next, we concatenated the union sets of chemical fingerprints and GO annotations for each pair of drug combinations as the input features of the GTB classifier. The performance measures are shown in Table 3. It can be demonstrated that the performance of GTB classifier with input derived from heterogeneous network-based features is vastly superior to that with primary ontology features. For example, the AUC value increased from 0.528 to 0.949 for GTB classifier. Moreover, we conducted performance comparison for other typical classifiers to validate the advantage of our extracted feature from drug-protein heterogenous network. As shown in Tables 2 and 3, the performance of all these classifiers were greatly boosted by extracting features from the random walk with restart on the heterogenous network.

Table 2 Comparison of GTB with other typical classifiers on heterogenous network-derived features

Method	Precision	Recall	F-Measure	MCC	AUC
GTB	**0.897**	**0.872**	**0.884**	**0.772**	**0.949**
kNN	0.738	0.833	0.783	0.542	0.768
SVM	0.882	0.779	0.840	0.728	0.859
Logistic	0.499	0.527	0.510	0.014	0.520
Naive Bayes	0.504	0.988	0.770	0.086	0.508
Random forest	0.880	0.841	0.862	0.733	0.866
AdaBoost	0.878	0.854	0.863	0.732	0.866
LogitBoost	0.803	0.820	0.811	0.617	0.808

The boldface figures indicate that GTB achieves the best performance compared with other typical classifiers on heterogenous network-derived features

Discussion

In this paper, we proposed a computational method for predicting effective combination drugs based on features derived from drug-protein heterogenous network by random walk with restart. In order to verify our proposed method, we conducted plenty of empirical experiments to compare the performance of our method to other typical classifiers on the benchmark dataset we constructed previously, and the experimental results significantly demonstrated that our method achieves state-of-the-art performance. Note that the input of the GTB classifier is the output of random walk with restart on the heterogeneous network, which is the probability distribution vector only accounting for 6,074 dimensions. Therefore, we believed that the heterogeneous network-derived features are more informative and have been dimension-reduced compared to the high-dimensional primary ontology features that may lead to curse of dimensionality when performing classification. As a result, the performance of GTB and other classifiers are significantly improved. In addition, the majority of current methods to predict drug combinations are limited to their size, in which pairwise drugs are most used. However, our proposed method can expand the size of drug combinations, which appreciably increases the practicality.

It is worth noting that the protein network introduced in the random walk with restart is helpful to dig into the biological mechanism of drug combinations in vivo. In fact, the final probability distribution of certain drug combination derived by random walk with restart strongly suggests the indications of the drug combination to some extent. Taking the pairwise combination Docetaxel and Capecitabine as an example, which has been approved by FDA to treat metastatic breast cancer. We ranked the protein nodes according to the probability distribution, the protein with the highest probability is ENSP00000315644 and the third is ENSP00000252029, which are both encoded by *Tyms* gene. It has been

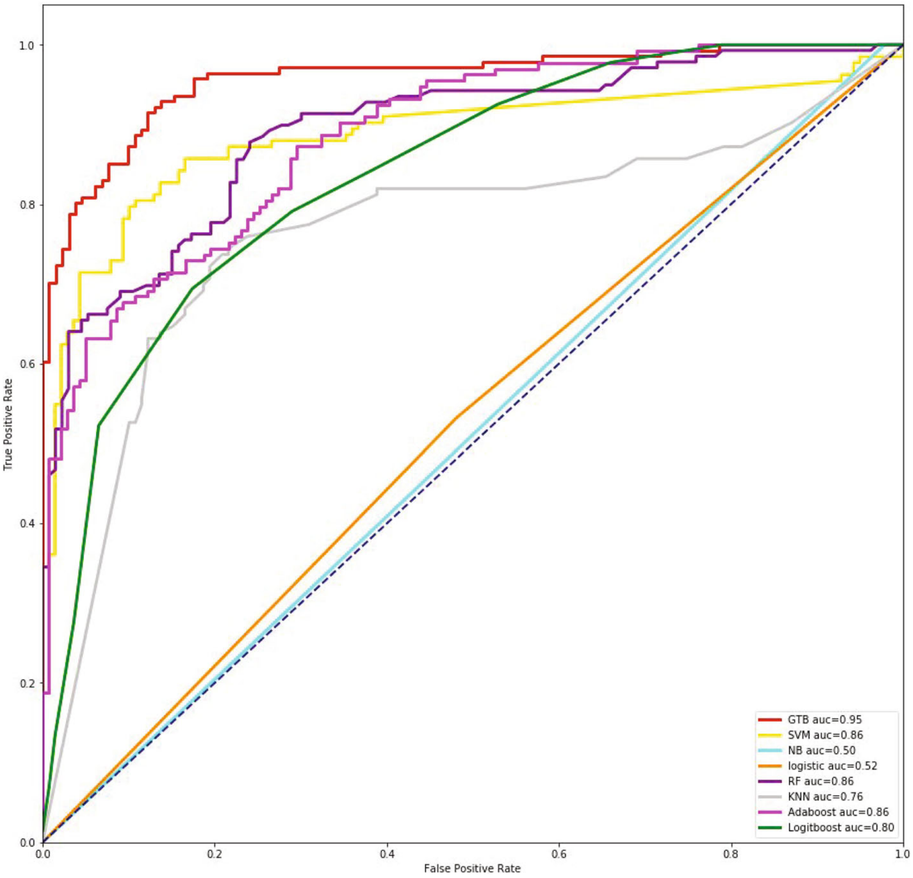


Fig. 1 ROC curves of our method and other typical classifiers on benchmark set

shown that the polymorphisms of *Tyms* gene are associated with etiology of neoplasia, including breast cancer. In addition, the fourth and fifth are ENSP00000269571 and ENSP00000275493, which are encoded by *Erbb2* and *Egfr*, respectively, are all highly linked to breast cancer. To further evaluate the potential of the protein network, we exemplified another pair of drug combination, Atorvastatin and Proguanil, which currently has no official indication. The resulting probability distribution derived by random walk with restart of this drug combination shows that the protein ENSP00000396308 encoded by *Dhfr* has the highest probability value. It has been demonstrated that diseases associated with *Dhfr* include megaloblastic anemia due to dihydrofolate reductase deficiency and megaloblastic anemia. Expectedly, quite a few works have demonstrated the pharmacological effect of Atorvastatin and Proguanil on Anemia. For example, Vahid et al. [39] validated that Atorvastatin can soften human red blood cells, and physical deformation of the red blood cells underlies pathological manifestations of sickle cell anemia and hypercholesterolemia. Another trial demonstrated the effectiveness of Proguanil in treatment of malarial anemia [40]. Therefore, anemia may be a potential indication of the drug combination Atorvastatin and Proguanil. In summary, we draw the conclusion that the probability distribution derived by random walk can effectively reveal the indication of drug combinations.

Table 3 Comparison of GTB with other typical classifiers on primary ontology features

Method	Precision	Recall	F-Measure	MCC	AUC
GTB	**0.526**	**0.53**	**0.523**	**0.052**	**0.528**
kNN	0.514	0.514	0.513	0.028	0.516
SVM	0.509	0.491	0.478	-0.019	0.491
Logistic	0.506	0.506	0.506	0.012	0.504
Naive Bayes	0.479	0.479	0.478	-0.043	0.46
Random forest	0.499	0.499	0.478	-0.002	0.499
AdaBoost	0.501	0.501	0.425	0.002	0.497
LogitBoost	0.499	0.499	0.479	-0.002	0.495

The boldface figures indicate that GTB achieves the best performance compared with other 7 typical classifiers trained on primary ontology features

We further checked the positive samples that are falsely classified, as negative samples are randomly generated. We found that the falsely determined samples by our method have low similarity to other samples. In fact, most existing computational models, which aim at the prediction of drug-target interactions, drug-disease associations, often hold the assumption that similar compounds are likely to interact with similar target proteins and thereby play similar therapeutic efficacy in cellular micro-environment. These computational methods have achieved superior performance, and greatly narrowed down the number of candidate drug targets and reveal new indications of approved drugs. Under this assumption, the prediction accuracy often relies on the close associations of tested samples with known samples that have been validated by wet-lab experiments, such as drug combinations and drug-target interactions. In terms of network medicine, the influence of drug molecule would perturb the cellular network via signal cascade reactions and protein interaction network. Many computational methods have taken into account this consideration, and adopted random walks and diffusion on network to capture the perturbation of the drugs.

However, there are always some samples located far from validated samples in the feature space. For instance, some new drugs have low similarity to other drugs, and some proteins have low similarity to other protein in different protein family. As a result, similarity-based or network diffusion-based computational methods tend to encounter failure in predicting drug combinations or drug-target interactions composed of such drugs or proteins. Fortunately, the emergence of large-scale experimental data derived from high-throughput screening technique can strongly motivate the novelty of methods to predict synergistic drugs or effective drug combinations.

Conclusion

In this paper, we proposed a gradient tree boosting (GTB) classifier based on heterogeneous network-derived features to predict effective drug combinations. The heterogeneous network integrates the drug similarity network, protein similarity network and known drug-protein associations. Next, we ran random walk with restart (RWR) on the heterogenous network using the combinatorial drugs and their associated targets as the initial probability, and obtained the converged probability distribution as the feature vector of each drug combination. The heterogeneous network-derived features introduced in our method are more informative and enriching compared to the primary ontology features. The GTB classifier trained based on the heterogeneous network-derived features outperforms seven typical classifiers and traditional boosting algorithms. Moreover, our case studies show that our method is helpful in revealing the indications of drug combinations. From the perspective of network pharmacology, our method effectively exploits the topological attributes and interactions of drug targets in the overall biological network, which proves to be a systematic and reliable approach for drug discovery.

Methods

Overview of our methodology

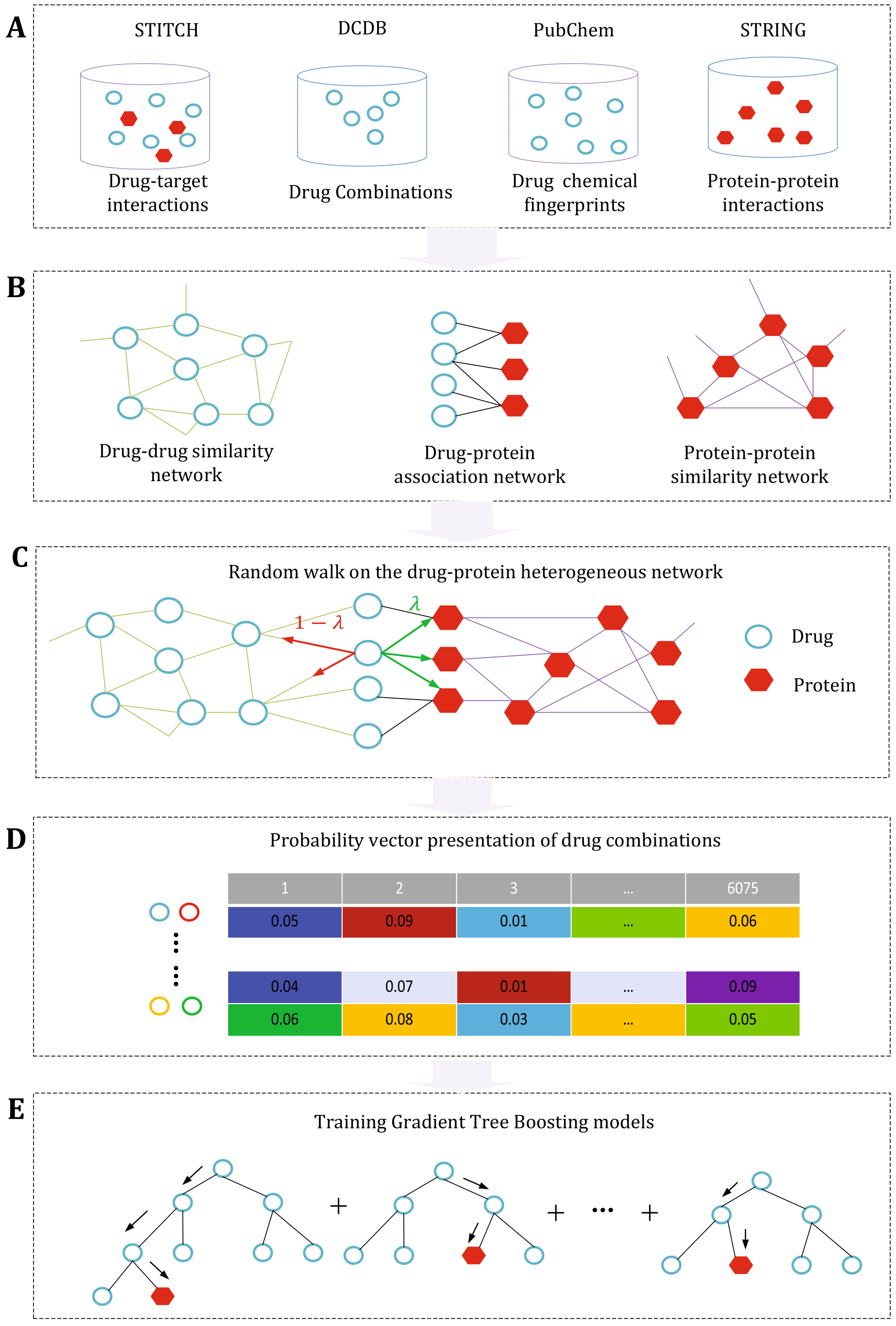
We first constructed the benchmark drug combination set composed of positive samples derived from public databases and negative samples that were randomly generated. For individual drugs included in the benchmark set, we collected a variety of related characteristics, including chemical fingerprints, drug targets and drug-protein associations, as shown in Fig. 2a. These ontology features of drugs and proteins were used to compute the drug-drug similarities and protein-protein similarities. Together with the known drug-protein associations, we constructed the drug-protein heterogeneous network. Next, the random walk with restart on heterogeneous network proposed in our previous work [28] was conducted for each drug combinations as initial state, as illustrated in Fig. 2b-c. The probability distribution when the random walk reaches steady state was used as the feature vector of the drug combination. Based on the feature representation of the drug combinations, the gradient tree boosting (GTB) classifier was trained to predict new effective drug combinations.

Drug-protein associations

We selected drug-protein associations from STITCH database [41], which is a comprehensive database that collected compound-protein interactions from different sources: biochemical experiments, external databases, text mining and computational predictions. STITCH has computed a confidence score for each interaction ranging from 0 to 1,000, which indicates the confidence of the compound-protein interaction supported by four types of evidences. We first used a confidence threshold 0.5 (corresponding to 500 combined score in STITCH) to remove low-confidence target proteins, because we think too low-confidence targets are probable unauthentic ones. Next, we selected top 3 from the rest of target proteins of each drug. If one drug has less than 3 target proteins with confidence score higher than 0.5, we then took only those targets into account. In total, we got 210,235 drug-protein associations regarding to 3,266 unique drugs (drug set are built by selecting top 3 similar drugs, see following subsection for details). Formally, denoted by $D = (d_1, d_2, ..., d_n)$ and $P = (p_1, p_2, ..., p_m)$ the drug and protein node set, and A the adjacent matrix of drug-protein associations with element a_{ij} equal to the confidence score if there is validated interaction between drug i and protein j, and a_{ij}=0 otherwise.

Drug-drug similarity network

We expanded the list of individual drugs by selecting top 10 most similar drugs to each single agent included in DCDB, according to the chemical-chemical combined scores that were derived from STITCH [41]. After removal


A
STITCH
DCDB
PubChem
STRING
Drug-target interactions
Drug Combinations
Drug chemical fingerprints
Protein-protein interactions
B
Drug-drug similarity network
Drug-protein association network
Protein-protein similarity network
C
Random walk on the drug-protein heterogeneous network
1 − λ
λ
Drug
Protein
D
Probability vector presentation of drug combinations
E
Training Gradient Tree Boosting models
+
+ ··· +

1	2	3	...	6075
0.05	0.09	0.01	...	0.06
0.04	0.07	0.01	...	0.09
0.06	0.08	0.03	...	0.05

Fig. 2 Illustrative diagram of the proposed method. **a** Data collection from drug and protein-related databases; **b** Construction of drug-drug similarity network, protein-protein similarity network and drug-protein association network; **c** Random walk with restart on drug-protein heterogenous network; **d** Feature representations of drug combinations via feature extraction process; **e** Training gradient tree boosting classifier

of duplicate drugs, 3266 unique drugs were obtained. Similar compounds are likely to interact with similar target proteins and thereby play similar therapeutic efficacy in cellular micro-environment [42], allowing us to find new drug combinations by introducing similar drugs to known ones. Therefore, we believe that the expanded list of drugs can increase the opportunity for discovery of novel drug combinations.

Next, we generate the chemical fingerprint of the drugs to calculate the similarity measurement of each pair of drugs. Similar to our previous work [28], we applied PaDEL software [43] to compute the chemical fingerprints using the SMILES string of a drug, and obtain an 880-d binary vector for each drug. The element 1 of the binary vector represents that the drug contains the corresponding chemical fingerprint, and 0 otherwise. Subsequently, Jaccard score, a widely used similarity measure, is calculated based on the chemical fingerprints as the chemical similarities for pairwise drugs. The Jaccard score is generally defined as the intersection size divided by the union size of two individual sets, which is shown as follows:

$$S_{ij}^{(d1)} = \frac{|\vec{d}_i \cap \vec{d}_j|}{|\vec{d}_i \cup \vec{d}_j|} \tag{5}$$

Further, the bipartite network projection algorithm, a method inspired by the network-based resource-allocation dynamics [44], was adopted to compute another drug similarity measure based on known drug-protein associations. In the drug-protein bipartite network, each drug node equally allocates the original resource to its associated protein nodes, and successively the assigned resource of each protein node is equally transferred back to its neighborhood drugs. As a result, the proportion of the resource of drug d_i conveyed to drug d_j in such allocation process represents the strength of association between two drugs. Suppose the initial resource of each drug node is one-unit, the second drug similarity measure, denoted by $S_{ij}^{(d2)}$, can be formulated as below:

$$S_{ij}^{(d2)} = \frac{1}{k(d_j)} \sum_{l=1}^{m} \frac{a_{il}a_{jl}}{k(p_l)} \tag{6}$$

in which $k(d_j)$ and $k(p_l)$ are the degrees of drug d_j and protein p_l in the drug-protein association network. Intuitively, more common associated protein nodes the pairwise drugs share, higher similarity the drugs have. Particularly, if the associated proteins of two drugs are not overlapped, i.e. no common associated protein exists, the similarity is denoted by 0.

Finally, these two aforementioned drug-drug similarities were integrated into a comprehensive measurement using the probability disjunction formula as below:

$$S_{ij}^{(d)} = 1 - \left(1 - S_{ij}^{(d1)}\right) * \left(1 - S_{ij}^{(d2)}\right) \tag{7}$$

Protein-protein similarity network

Correspondingly, we constructed the protein-protein similarity network based on two different similarity measures, including protein sequence similarity and GO semantic similarity. By using R package biomaRt (2.40.4) [45, 46], the protein sequences can be readily obtained from Ensembl genome database (2018 updated), which is dedicated to curating gene-related information to encourage genome analysis [47]. The sequence similarity $S_{ij}^{(p1)}$ between protein p_i and protein p_j was computed by using the R package Protr (1.6-2) [48], in which the Smith-Waterman algorithm is applicable.

Similar drugs are supposed to interact with proteins that act in similar biological processes or have similar molecular functions or reside in similar compartments [49]. Therefore, the GO semantic similarity $S_{ij}^{(p2)}$ between protein p_i and protein p_j was calculated using R package GOSemSim (2.10.0) [50]. All three types of ontology features are used in the calculation of semantic similarity.

Likewise, the probability disjunction was used to integrate two aforementioned protein-protein similarities, which is formulated as below:

$$S_{ij}^{(p)} = 1 - \left(1 - S_{ij}^{(p1)}\right) * \left(1 - S_{ij}^{(p2)}\right) \tag{8}$$

where $S_{ij}^{(p)}$ is the comprehensively integrated similarity measurement between protein p_i and protein p_j.

Random walk with restart on heterogenous network

The drug-drug similarity network, protein-protein similarity network and drug-protein association network were combined to construct the drug-protein heterogeneous network $G = (V, E)$. The node set $V = \{D, P\}$, V is the union set of the drug and protein nodes. The edge set $E = \{E_{dd} \cup E_{dp} \cup E_{pd} \cup E_{pp}\}$, where E_{dd}, E_{pp}, E_{dp} and E_{pd} are the drug-drug edge, protein-protein and drug-protein edge collections, respectively.

In order to obtain the feature representations of drug combinations, we extended our previous work in which the random walks with restart on the heterogeneous network was developed for single drug repurposing [28]. More precisely, for a drug combination d_i and drug d_j, we performed random walk with restart on the heterogeneous network in which these two drugs and their known target proteins act as seed nodes, as shown in Fig. 3. Actually, since the initial probability distribution can be easily extended to more drugs and their targets, the number of individual drugs involved in the combination is not limited to 2 in our method. When the random walk process reaches steady state, the probability distribution vector can be regarded as the perturbation on the protein network by the combinatorial drugs. With the drug-protein heterogeneous network, the transition matrix T can be defined as below:

$$T = \begin{bmatrix} T^{(dd)} & T^{(dp)} \\ T^{(pd)} & T^{(pp)} \end{bmatrix} \tag{9}$$

where $T^{(dd)}$ and $T^{(pp)}$ are the probability transition matrix from drug nodes (protein) to drug nodes (protein nodes) during the random walk process; $T^{(dp)}$ denotes the probability transition matrix that drug nodes walk to protein nodes, and $T^{(pd)}$ denotes the probability transition matrix that protein nodes walk to drug nodes.

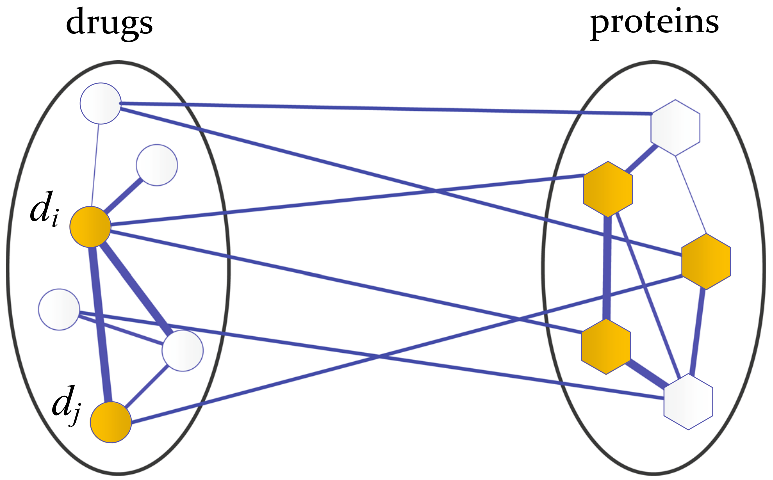


Fig. 3 Illustrative diagram of the random walk with restart on drug-protein heterogenous network, starting from two drugs and their targeted proteins

Suppose that the random walker starts from a drug node, and then visits one of its targeted proteins with probability λ, or visits any other drug nodes with probability (1-λ) in the heterogeneous network. If λ=0, the random walker can only stay within the networks where it starts. According to the drug-drug similarity, the transition probability from drug d_i to drug d_j can be defined as below:

$$T_{ij}^{(dd)} = \begin{cases} S_{ij}^{(d)} / \sum_{k=1}^{n} S_{ik}^{(d)}, & if\ \sum_{l=1}^{m} a_{il} = 0 \\ (1-\lambda) S_{ij}^{(d)} / \sum_{k=1}^{n} S_{ik}^{(d)}, & \text{otherwise.} \end{cases} \tag{10}$$

where S_{ij} is the similarity between ith drug and jth drug, a_{il} is the association confidence score between ith drug and lth protein. The sum of a_{il} equaling to 0 indicates that the drug has no approved or predicted association with any proteins. Similarly, the transition probability from protein p_i to protein p_j can be defined based on the protein-protein similarity as below:

$$T_{ij}^{(pp)} = \begin{cases} S_{ij}^{(p)} / \sum_{k=1}^{m} S_{ik}^{(p)}, & if\ \sum_{l=1}^{n} a_{li} = 0 \\ (1-\lambda) S_{ij}^{(p)} / \sum_{k=1}^{m} S_{ik}^{(p)}, & \text{otherwise.} \end{cases} \tag{11}$$

where S_{ij} is the similarity between ith protein and jth protein, a_{li} is the association score between lth drug and ith protein.

Accordingly, the transition probability from drug d_i to protein p_j is defined as:

$$T_{ij}^{(dp)} = \begin{cases} \lambda a_{ij} / \sum_{l=1}^{m} a_{il}, & if\ \sum_{l=1}^{m} a_{il} \neq 0 \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

The transition probability from protein p_i to drug d_j is defined as:

$$T_{ij}^{(pd)} = \begin{cases} \lambda a_{ji} / \sum_{l=1}^{n} a_{li}, & if\ \sum_{l=1}^{n} a_{li} \neq 0 \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

Provided that $P(t)$ is a $(n+m)$-dimension probability vector at step t, in which $P(t)$[i] represents the probability of the random walker visiting node i(drug or protein), the random walk process can be iteratively calculated as below:

$$P(t+1) = (1-\alpha) T' P(t) + \alpha P_0 \tag{14}$$

where α is the restart probability, and P_0 is the initial probability distribution vector of a set of seed nodes consisting of a combinatorial drugs and their targeted proteins. Take the drug combination d_i and d_j as an example, d_i and d_j are employed as the seed nodes in the drug network and each seed node is given equal probability 1/2. By giving rest drug nodes probability 0, the initial probability matrix with respect to drugs can be constructed. Correspondingly, the protein nodes related to drug d_i and drug d_j are used as seed nodes in protein network and equal probabilities are allocated to these protein nodes so that the sum of the probabilities is 1. As shown in Fig. 3, there are three targeted proteins and thus each protein is given initial probability 1/3. Let $P_0^{(d)}$ and $P_0^{(p)}$ be the initial probability vectors of drugs and proteins separately, the initial probability P_0 for drug-centric random walk can be defined as follows:

$$P_0 = \begin{bmatrix} \eta P_0^{(d)} \\ (1-\eta) P_0^{(p)} \end{bmatrix} \tag{15}$$

where $\eta \in [0, 1]$ is a tradeoff parameter to balance the weight of importance between the drug nodes and protein nodes. In our experiments, η is set to 0.5. If the difference between twice iteration is lower than 1e-10, the random walk is supposed to reach steady state. Once the random walk process converges, the probability distribution is used as the feature vector of the drug combination.

Building gradient tree boosting classifier

Based on the feature vectors produced by the random walk on drug-protein heterogenous network for each pair of drug combination, we built a gradient tree boosting (GTB) classification model, referred to as gradient boosting regression or decision tree (GBRT or GBDT). Gradient tree boosting is an efficacious machine learning method that has achieved desirable performance in both classification and regression problems [51–53]. In fact, Caruana and Niculescu-Mizil have conducted comprehensive performance evaluation on eight different binary classification problems by comparing boosted trees algorithm with other nine typical classifiers, including SVMs, Neural Nets, Logistic regression, Naive Bayes, memory-based learning, Random Forests, Decision Trees, Bagged Trees and Boosted Stumps. Their conclusion showed that boosted tree-based algorithm achieved best performance [54]. Another empirical performance evaluation has also demonstrated that boosted decision trees perform exceptionally well when the dimensionality of the input is not too high [55]. Therefore, we adopted the GTB algorithm to build our classification model.

Formally, the decision function of GTB is initialized as:

$$\theta_0(x) = arg\,min \sum_{i=1}^{N} L(y_i, c) \tag{16}$$

where N is the number of drug combinations contained in the training set. The gradient tree boosting algorithm repeatedly constructs K different classification subtrees $h(x, a_1)$, $h(x, a_2)$,..., $h(x, a_K)$, each of which is separately trained based on a subset of randomly selected samples from the training set, and then iteratively establishes the additive function $\theta_k(x)$:

$$\theta_k(x) = \theta_{k-1}(x) + b_k h(x, a_k) \tag{17}$$

in which b_k and a_k are the weight and parameter vector of the k-th classification subtree $h(x, a_k)$. The loss function $L(y, \theta_k(x))$ is defined as:

$$L(y, \theta(x)) = log(1 + exp(-y\theta(x))) \tag{18}$$

where y is a binary value representing the real class of the combination and $\theta(x)$ is the decision function. In order to minimize the loss function $L(y, \theta_k(x))$, both b_k and a_k are iteratively optimized by applying grid search. In this paper, grid search strategy was adopted to tune the optimal hyperparameters of GTB by 10-fold cross-validation on the constructed drug combination dataset. Finally, the optimal number of trees of the GTB is 300, and the tuned depth of the trees is 13.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12859-019-3288-1.

Additional file 1: Table S1. Impact of the restart probability α on the performance of GTB classifer. **Table S2**. Impact of the tradeoff parameter η on the performance of GTB classifer. **Table S3**. Impact of significant parameters involved in GTB classifier. **Table S4-S8**. Parameter tuning for competitive methods.

Abbreviations

GO: Gene ontology; GTB: Gradient tree boosting; kNN: K-NearestNeighbor; RWR: Random walk with restart; SVM: Support vector machine

Acknowledgements

The authors thank DCDB and STITCH for providing the data.

Authors' contributions

HL, WZ and LN designed the study and conducted experiments. HL and WZ performed statistical analyses. HL and WZ drafted the manuscript. WZ, LN and XD prepared the experimental materials and benchmarks. JL and LZ administrated the project and acquired funding. All authors have read and approved the final manuscript.

Funding

The motivation, design and data analysis and validations of the study was supported by National Natural Science Foundation of China under grants No. 61672113; The analysis and interpretation of data and writing the manuscript is partially supported by the project of Jiangsu Provincial Science and Technology Department (Grant No. BE2018638) and Jiangsu provincial 333 project; The writing of the manuscript is also partially supported by Social Development Projects of Jiangsu Province (BE2018643) and Scientific Program of Changzhou (CE20195048; ZD201919).

Availability of data and materials

The source codes, datasets and additional files used in this work are all available at https://github.com/hliu2016/SynerDrug.

Ethics approval and consent to participate

The drug combination data used to evaluate the GTB model are available at DCDB database which is an announced public data source. The drug-protein associations and drug-drug links are derived from STITCH on reasonable request. The STITCH is supported by European Molecular Biology Laboratory(EMBL), Swiss Institute of Bioinformatics(SIB) and NNF Center for Protein Research(CPR).

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Lab of Information Management, Changzhou University, Jiangsu, China. [2]School of Information Science and Engineering, Changzhou University, Jiangsu, China. [3]Information Center, Changzhou University, Jiangsu 213164, China. [4]Department of Radiation Oncology, the Affiliated Changzhou No.2 People's Hospital of Nanjing Medical University, Changzhou, China.

Received: 1 July 2019 Accepted: 21 November 2019
Published online: 09 December 2019

References

1. Pang K, Wan Y-W, Choi WT, Donehower LA, Sun J, Pant D, Liu Z. Combinatorial therapy discovery using mixed integer linear programming. Bioinformatics. 2014;30(10):1456–63.
2. Lee MJ, Albert SY, Gardino AK, Heijink AM, Sorger PK, MacBeath G, Yaffe MB. Sequential application of anticancer drugs enhances cell death by rewiring apoptotic signaling networks. Cell. 2012;149(4):780–94.
3. Guo J, Liu H, Zheng J. Synlethdb: synthetic lethality database toward discovery of selective and sensitive anticancer drug targets. Nucleic Acids Res. 2015;44(D1):1011–7.
4. Preuer K, Lewis RP, Hochreiter S, Bender A, Bulusu KC, Klambauer G. Deepsynergy: predicting anti-cancer drug synergy with deep learning. Bioinformatics. 2017;34(9):1538–46.
5. Jia J, Zhu F, Ma X, Cao Z, Cao Z, Li Y, Li Y, Chen Y. Mechanisms of drug combinations: interaction and network perspectives. Nat Rev Drug Discov. 2009;8(2):111–28.
6. Al-Lazikani B, Banerji U, Workman P. Combinatorial drug therapy for cancer in the post-genomic era. Nat Biotechnol. 2012;30(7):679.
7. Gioeli D, Wunderlich W, Sebolt-Leopold J, Bekiranov S, Wulfkuhle JD, Petricoin EF, Conaway M, Weber MJ. Compensatory pathways induced by mek inhibition are effective drug targets for combination therapy against castration-resistant prostate cancer. Mol Cancer Ther. 2011;10(9):1581–90.
8. Liu T, Yacoub R, Taliaferro-Smith LD, Sun S-Y, Graham TR, Dolan R, Lobo C, Tighiouart M, Yang L, Adams A, et al. Combinatorial effects of lapatinib and rapamycin in triple-negative breast cancer cells. Mol Cancer Ther. 2011;10(8):1460–9.
9. Smalley KS, Haass NK, Brafford PA, Lioni M, Flaherty KT, Herlyn M. Multiple signaling pathways must be targeted to overcome drug resistance in cell lines derived from melanoma metastases. Mol Cancer Ther. 2006;5(5):1136–44.
10. Nelander S, Wang W, Nilsson B, She Q-B, Pratilas C, Rosen N, Gennemark P, Sander C. Models from experiments: combinatorial drug perturbations of cancer cells. Mol Syst Biol. 2008;4(1):216.
11. Sun X, Vilar S, Tatonetti NP. High-throughput methods for combinatorial drug discovery. Sci Transl Med. 2013;5(205):205–12051.
12. Li P, Huang C, Fu Y, Wang J, Wu Z, Ru J, Zheng C, Guo Z, Chen X, Zhou W, et al. Large-scale exploration and analysis of drug combinations. Bioinformatics. 2015;31(12):2007–16.
13. Costello JC, Heiser LM, Georgii E, Gönen M, Menden MP, Wang NJ, Bansal M, Hintsanen P, Khan SA, Mpindi J-P, et al. A community effort to assess and improve drug sensitivity prediction algorithms. Nat Biotechnol. 2014;32(12):1202.
14. Csermely P, Agoston V, Pongor S. The efficiency of multi-target drugs: the network approach might help drug design. Trends Pharmacol Sci. 2005;26(4):178–82.
15. Iadevaia S, Lu Y, Morales FC, Mills GB, Ram PT. Identification of optimal drug combinations targeting cellular networks: integrating phospho-proteomics and computational network analysis. Cancer Res. 2010;70(17):6704–14.
16. Tang J, Karhinen L, Xu T, Szwajda A, Yadav B, Wennerberg K, Aittokallio T. Target inhibition networks: predicting selective combinations of druggable targets to block cancer survival pathways. PLoS Comput Biol. 2013;9(9):1003226.
17. Ryall KA, Tan AC. Systems biology approaches for advancing the discovery of effective drug combinations. J Cheminform. 2015;7(1):7.
18. Huang L, Li F, Sheng J, Xia X, Ma J, Zhan M, Wong ST. Drugcomboranker: drug combination discovery based on target network analysis. Bioinformatics. 2014;30(12):228–36.
19. Zhao X-M, Iskar M, Zeller G, Kuhn M, Van Noort V, Bork P. Prediction of drug combinations by integrating molecular and pharmacological data. PLoS Comput Biol. 2011;7(12):1002323.
20. Ligeti B, Pénzváltó Z, Vera R, Győrffy B, Pongor S. A network-based target overlap score for characterizing drug combinations: high correlation with cancer clinical trial results. PLoS ONE. 2015;10(6):0129267.
21. Matlock K, Berlow N, Keller C, Pal R. Combination therapy design for maximizing sensitivity and minimizing toxicity. BMC Bioinformatics. 2017;18(4):116.
22. De Raedt T, Walton Z, Yecies JL, Li D, Chen Y, Malone CF, Maertens O, Jeong SM, Bronson RT, Lebleu V, et al. Exploiting cancer cell vulnerabilities to develop a combination therapy for ras-driven tumors. Cancer cell. 2011;20(3):400–13.
23. Roller DG, Axelrod M, Capaldo BJ, Jensen K, Mackey A, Weber MJ, Gioeli D. Synthetic lethal screening with small-molecule inhibitors provides a pathway to rational combination therapies for melanoma. Mol Cancer Ther. 2012;11:2505–15.
24. Zhang Z, Zhang J, Fan C, Tang Y, Deng L. Katzlgo: large-scale prediction of lncrna functions by using the katz measure based on multiple networks. IEEE/ACM Trans Comput Biol Bioinform. 2017. https://doi.org/10.1109/tcbb.2017.2704587.
25. Deng L, Wu H, Liu C, Zhan W, Zhang J. Probing the functions of long non-coding rnas by exploiting the topology of global association and interaction network. Comput Biol Chem. 2018;74:360–7.
26. Deng L, Wang J, Xiao Y, Wang Z, Liu H. Accurate prediction of protein-lncrna interactions by diffusion and hetesim features across heterogeneous network. BMC Bioinformatics. 2018;19(1):370.
27. Zhang J, Zhang Z, Chen Z, Deng L. Integrating multiple heterogeneous networks for novel lncrna-disease association inference. IEEE/ACM Trans Comput Biol Bioinform. 2019;16:396–406.
28. Liu H, Song Y, Guan J, Luo L, Zhuang Z. Inferring new indications for approved drugs via random walk on drug-disease heterogenous networks. BMC Bioinformatics. 2016;17:539.
29. Luo H, Wang J, Li M, Luo J, Peng X, Wu F-X, Pan Y. Drug repositioning based on comprehensive similarity measures and bi-random walk algorithm. Bioinformatics. 2016;32(17):2664–71.
30. Zhang J, Deng L. Integrating multiple interaction networks for gene function inference. Molecules. 2019;24(1):30.
31. Liu Y, Wei Q, Yu G, Gai W, Li Y, Chen X. Dcdb 2.0: a major update of the drug combination database. Database. 2014;2014:. https://doi.org/10.1093/database/bau124.
32. Yamanishi Y, Araki M, Gutteridge A, Honda W, Kanehisa M. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. Bioinformatics. 2008;24(13):232–40.
33. Jin W, Jung J, Kang U. Supervised and extended restart in random walks for ranking and link prediction in networks. PLoS ONE. 2019;14(3):0213857.
34. R Core Team. R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing; 2014. http://www.R-project.org/.
35. Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F, Chih-Chung C, Chih-Chen C. e1071: misc functions of the department of statistics, probability theory group 1.7-3. R package e1071. 2019.
36. Hechenbichler K, Schliep K. Weighted k-Nearest-Neighbor Techniques and Ordinal Classification. Ludwig-Maximilians University Munich; 2004. p. SFB 386. http://nbn-resolving.de/urn/resolver.pl?urn=nbn:de:bvb:19-epub-1769-9.
37. Liaw A, Wiener M. Classification and regression by randomforest. R News. 2002;2(3):18–22.
38. Alfaro E, Gamez M, Garcia N, et al. Adabag: An r package for classification with boosting and bagging. J Stat Softw. 2013;54(2):1–35.
39. Sheikh-Hasani V, Babaei M, Azadbakht A, Pazoki-Toroudi H, Mashaghi A, Moosavi-Movahedi AA, Reihani SNS. Atorvastatin treatment softens human red blood cells: an optical tweezers study. Biomed Opt Expr. 2018;9(3):1256–61.
40. Mulenga M, Malunga F, Bennett S, Thuma PE, Shulman C, Fielding K, Alloueche A, Greenwood BM. A randomised, double-blind, placebo-controlled trial of atovaquone–proguanil vs. sulphadoxine–pyrimethamine in the treatment of malarial anaemia in zambian children. Trop Med Int Health. 2006;11(11):1643–52.
41. Szklarczyk D, Santos A, von Mering C, Jensen LJ, Bork P, Kuhn M. Stitch 5: augmenting protein-chemical interaction networks with tissue and affinity data. Nucleic Acids Res. 2015;44(D1):380–4.
42. Johnson MA, Maggiora GM. Concepts and applications of molecular similarity: Wiley; 1990.
43. Yap CW. Padel-descriptor: An open source software to calculate molecular descriptors and fingerprints. J Comput Chem. 2011;32(7):1466–74.
44. Zhou T, Ren J, Medo M, Zhang Y-C. Bipartite network projection and personal recommendation. Phys Rev E. 2007;76(4):046115.
45. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, Huber W. Biomart and bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics. 2005;21(16):3439–40.

46. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart. Nat Protocol. 2009;4(8):1184.
47. Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Girón CG, et al. Ensembl 2018. Nucleic Acids Res. 2017;46(D1):754–61.
48. Xiao N, Cao D-S, Zhu M-F, Xu Q-S. protr/protrweb: R package and web server for generating various numerical representation schemes of protein sequences. Bioinformatics. 2015;31(11):1857–9.
49. Liu H, Sun J, Guan J, Zheng J, Zhou S. Improving compound–protein interaction prediction by building up highly credible negative samples. Bioinformatics. 2015;31(12):221–9.
50. Yu G, Li F, Qin Y, Bo X, Wu Y, Wang S. Gosemsim: an r package for measuring semantic similarity among go terms and gene products. Bioinformatics. 2010;26(7):976–8.
51. Pan Y, Liu D, Deng L. Accurate prediction of functional effects for variants by combining gradient tree boosting with optimal neighborhood properties. PLoS ONE. 2017;12(6):0179314.
52. Fan C, Liu D, Huang R, Chen Z, Deng L. Predrsa: a gradient boosted regression trees approach for predicting protein solvent accessibility. BMC Bioinformatics. 2016;17:8. BioMed Central.
53. Friedman JH. Stochastic gradient boosting. Comput Stat Data Anal. 2002;38(4):367–78.
54. Caruana R, Niculescu-Mizil A. An empirical comparison of supervised learning algorithms. In: Proceedings of the 23rd International Conference on Machine Learning. ACM; 2006. p. 161–8. https://doi.org/10.1145/1143844.1143865.
55. Caruana R, Karampatziakis N, Yessenalina A. An empirical evaluation of supervised learning in high dimensions. In: Proceedings of the 25th International Conference on Machine Learning. ACM; 2008. p. 96–103. https://doi.org/10.1145/1390156.1390169.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

BMC