BMC Bioinformatics

METHODOLOGY

Open Access

Drug response prediction using graph representation learning and Laplacian feature selection

Minzhu Xie[1,2*], Xiaowen Lei[1], Jianchen Zhong[1], Jianxing Ouyang[1] and Guijing Li[1]

From 16th International Symposium on Bioinformatics Research and Applications
Virtual. 1-4 December 2020. https://isbra.confreg.org/

*Correspondence:
xieminzhu@hunnu.edu.cn

[1] College of Information Science and Engineering, Hunan Normal University, Changsha, China
[2] Key Laboratory of Computing and Stochastic Mathematics (LCSM) (Ministry of Education), School of Mathematics and Statistics, Hunan Normal University, Changsha, China

Abstract

Background: Knowing the responses of a patient to drugs is essential to make personalized medicine practical. Since the current clinical drug response experiments are time-consuming and expensive, utilizing human genomic information and drug molecular characteristics to predict drug responses is of urgent importance. Although a variety of computational drug response prediction methods have been proposed, their effectiveness is still not satisfying.

Results: In this study, we propose a method called LGRDRP (Learning Graph Representation for Drug Response Prediction) to predict cell line-drug responses. At first, LGRDRP constructs a heterogeneous network integrating multiple kinds of information: cell line miRNA expression profiles, drug chemical structure similarity, gene-gene interaction, cell line-gene interaction and known cell line-drug responses. Then, for each cell line, learning graph representation and Laplacian feature selection are combined to obtain network topology features related to the cell line. The learning graph representation method learns network topology structure features, and the Laplacian feature selection method further selects out some most important ones from them. Finally, LGRDRP trains an SVM model to predict drug responses based on the selected features of the known cell line-drug responses. Our five-fold cross-validation results show that LGRDRP is significantly superior to the art-of-the-state methods in the measures of the average area under the receiver operating characteristics curve, the average area under the precision-recall curve and the recall rate of top-*k* predicted sensitive cell lines.

Conclusions: Our results demonstrated that the usage of multiple types of information about cell lines and drugs, the learning graph representation method, and the Laplacian feature selection is useful to the improvement of performance in predicting drug responses. We believe that such an approach would be easily extended to similar problems such as miRNA-disease relationship inference.

Keywords: Drug response, Learning graph representation, Laplacian feature selection, Network topology feature

Background

Personalized medicine focuses on finding appropriate drugs for individual patients. Since the same drugs have different effects on different patients, knowing the responses to drugs for each individual is a prerequisite of personalized medicine [1]. Since clinical drug response experiments are time-consuming and expensive, computational drug response prediction methods based on the related information of drugs and cell lines are of urgent practical importance and have attracted many researchers [2]. A variety of drug response prediction methods have been proposed, and they are mainly based on existing biological databases [3], of which Genomics of Drug Sensitivity in Cancer (GDSC) [4] and Cancer Cell Line Encyclopedia (CCLE) [5] are the two most famous. GDSC contains known cancer cell-drug responses and the corresponding cell lines' profiles [4]. CCLE provides public access to the gene expression, gene methylation and mutation data of over 1100 cell lines [5]. These databases provide researchers with benchmark data to test drug response prediction methods.

Based on cell line gene expression data, Torkamani et al. [6] used PCA to extract gene expression features, and constructed a linear regression model to predict drug responses. Gupta et al. [7] proposed a prediction model based on genomic characteristics such as copy number variations of cancer cell lines. Based on the CCLE dataset, Fang et al. [8] used a quantile regression forest method to predict drug response. Based on a support vector machine and a recursive feature selection tool, Dong et al. [9] used the gene expression and drug sensitivity data in CCLE to build a drug response predictor. Using the same data set, Geeleher et al. [10] proposed a ridge regression prediction model. Liu et al. [11] proposed an ensemble learning method that integrated a low-rank matrix completion model and a ridge regression model to predict drug responses. By integrating the pathways of drug targets and the related gene sets, Ammad et al. [12] proposed a kernelized Bayesian matrix factorization with component-wise multiple kernel learning to predict drug responses.

Based on the gene expression features of cell lines and the chemical features of drugs, Li et al. [13] developed a deep learning architecture to learn a prediction model. Yan et al. [14] proposed an interpretable model to predict drug responses, which integrated drug features, cell line features and drug responses using triple matrix factorization, and Guvencpaltun et al. [15] proposed a framework of Bayesian importance-weighted tri-matrix and two-matrix factorization to predict drug responses. These methods mainly considered the basic information of cell lines and drugs, and obtained good prediction performance for some certain drugs. However, they neglected other useful information such as the relationship between different cell-lines and the relationship between different drugs [16].

Based on the assumption that similar cell-lines tend to respond similarly to similar drugs, a lot of network-based drug response prediction methods have been proposed recently. For example, after integrating different kinds of information such as

gene mutation, DNA copy number and mRNA expression data of cell lines and compound molecular properties, ATC-codes and side-effects of drugs, Wang et al. [17] built similarity networks for cell lines and drugs, and proposed an SVM classifier model. Stanfield et al. [18] integrated cell line gene mutations, known cell line-drug responses and protein-protein interactions (PPIs), and built a heterogeneous network consisting of the

genes, cell lines and drugs. They utilized a random walk with restart (RWR) in the network to predict drug responses. Similarly, Zhang et al. [19] used cell line gene expression data to build a cell line similarity network, used drug chemical structures to build a drug similarity network, and used PPIs to build a gene-gene interaction network. They combined the networks with known cell line-drug associations and drug-target (gene) interactions into a heterozygous network and proposed a prediction model. Based on a cell line similarity network and a drug similarity network, Liu et al. [20] adopted a neighbor-based collaborative filtering with global effect removal method, Zhang et al. [21] adopted a hybrid interpolation weighted collaborative filtering method, and Guan et al. [22] utilized weighted graph regularized matrix factorization to predict drug responses.

In most network based drug response prediction methods, heterogeneous data are integrated using a weighted graph (i.e., a network), how to capture useful topological information of a graph is important for the efficiency of drug response prediction. Recently many learning graph representation methods have been introduced, and GraRep [23] is a state-of-the-art one that could learn a global representation of a graph, which contains the topological information of the graph and is convenient to use as input features of machine learning methods.

In the paper, we formulate the drug response prediction problem as a classification task as most of the existing methods: for each drug, classify cell lines into two groups: sensitive and resistant according to the cell lines' features. To improve drug response prediction performance, we integrate several available related data such as known cell line-drug associations, miRNA expression profiles of cell lines, chemical structures of drugs, PPIs, cell line gene sequence variations and hypermethylation informtion into a heterogeneous network. Then GraRep [23] and Laplacian Feature Selection are used to learn the cell lines' features in the network and features reduction, respectively. Finally, an SVM model [24] for the classification task is trained on the network. Our method is called LGRDRP (a Learning Graph Representation method for Drug Response Prediction) and is illustrated in Fig. 1.

Results and discussion

We conducted a series of 5-fold cross-validation experiments to test the performance of LGRDRP and some other state-of-the-art drug response prediction methods. The cross-validations were done for each drug. When a query drug was selected, all the cell lines with known responses (sensitive or resistant) to the drug were randomly divided into 5 groups. We randomly select one group as the test data and the other four as the training data. The heterogeneous network of the train data was obtained by removing the edges between the query drug vertex and the test cell line vertices.

For each cell line, a drug response prediction method calculated a score, and the test cell lines were sorted according to their scores. With a fixed threshold, if the score of one cell line is below the threshold, it is labeled as negative (resistant), and if it is known sensitive to the query drug, it is a false negative; if it is known resistant to the drug, it is a true negative. When the prediction score of a cell line is equal to or above the threshold, it is viewed as positive, and if it is known sensitive to the query drug, it is a true positive; if it is known resistant to the drug, it is a false positive. The true positive rate (TPR), the false positive rate (FPR), the precision ratio (Prec) and the recall ratio (Rec)

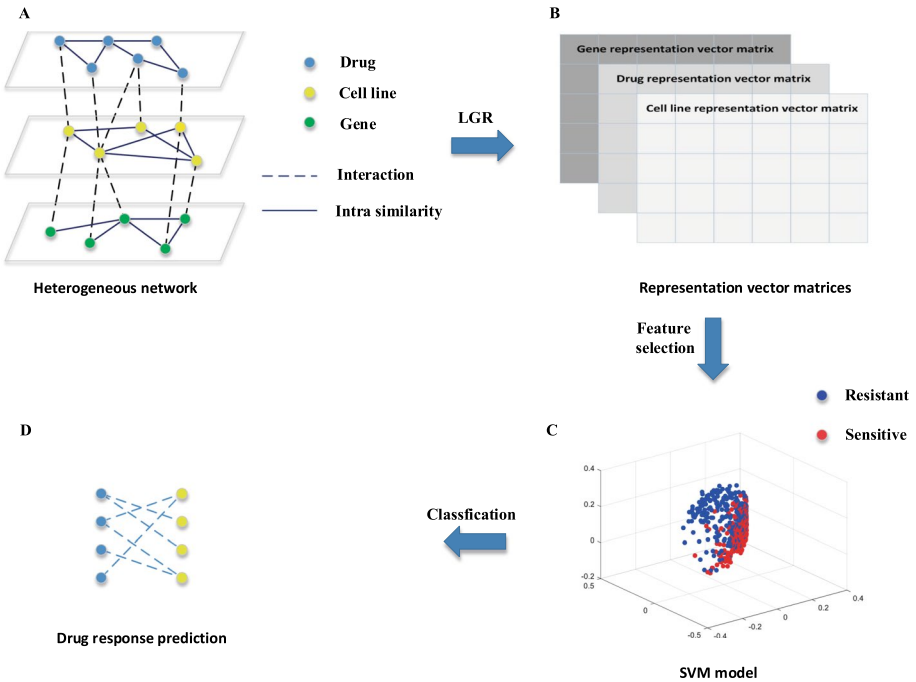

Fig. 1 The flowchart of LGRDRP. LGRDRP consists of four steps: **A** Construction of a heterogeneous network. **B** Learning representation vectors using a learning graph representation (LGR) method. **C** Feature selection and SVM model training. **D** Drug response prediction

can be computed as follows: $TPR = TP/(TP + FN)$, $FPR = FP/(FP + TN)$, $Prec = TP/(TP + FP)$, $Rec = TP/(TP + FN)$, where TP, FN, FP and TN are the numbers of cell lines that are true positive, false negative, false positive and true negative, respectively.

With the threshold increases from the smallest score to the highest score, a receiver operating characteristic (ROC) curve is drawn according to the varying TPRs and FPRs as X-axis values and Y-axis values respectively.

The area under the ROC curve (AUC) is calculated to evaluate the prediction performance. Since in our data set, the number of resistant responses is much larger than the sensitive responses. To better measure the prediction performance, we also used another metric: the area under the precision-recall (PR) curve (AUPR). A PR curve is a trajectory of the performance at a plane with the precision ratio as Y-axis value and recall ratio as X-axis value when the threshold changes.

Furthermore we may be more interested in the cell lines at the top of the sorted list. Therefore, the percentages of the true sensitive cell lines in the top 10, 20, 50, 100 of the sorted cell lines according to their response scores to the query drug were also used to evaluate the prediction performance.

Parameters selection

There are three parameters K, d and t need to set for LGRDRP, where K denotes the maximal transit steps, d determines the dimension of representation vectors and t is the number of features remained after the feature selection procedure. We tested the performance of LGRDRP with different K, d and t. Figure 2 shows the average AUC over different combinations of K, d and t. In the left panel, t was set 64. Please note, when the

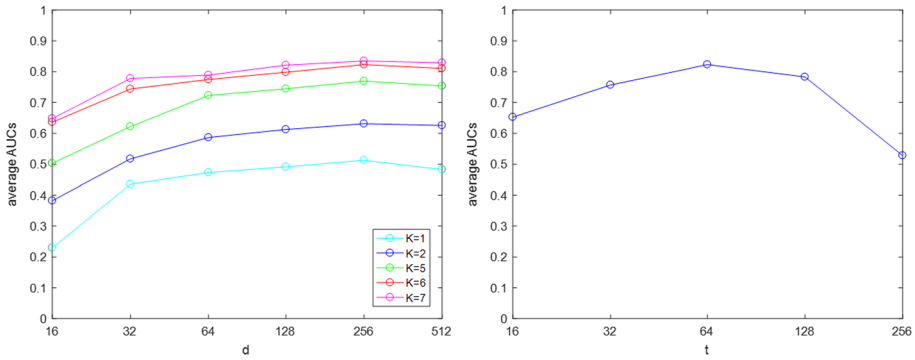

Fig. 2 Performances of LGRDRP with different values of parameters *K*, *d* and *t*. The value of *t* is set 64 in the left panel, and $K = 6$ and $d = 256$ in the right panel

length $K \times d$ of the final representation vector is less than 64, all features are kept. It can be observed when K increased, the average AUC of LGRDRP increased accordingly. However, when K increased from 6 to 7, the performance of LGRDRP didn't improve too much, but the computing time increased significantly. When d increased from 16 to 256, AUC increased accordingly. However, when d is set 512, AUC begins to decrease. The reason may be that the overfitting model results in poor generalization.

With $d = 256$ and $K = 6$, experiments were also conducted with varying t, and the results are shown in the right panel of Fig. 2. It indicates that the average AUC reached the best when $t = 64$. When t is too small, the left features can't capture enough structure information, but when t is too large, too many features may include some unimportant information which may disturb the prediction ability [25].

In the following tests, $K = 6, d = 256$ and $t = 64$ without specific description.

Performance evaluation

We compared the prediction performance of LGRDRP with three other art-of-the-state methods: HNMDRP [19], SVMDRP [17] and Stanfield's method [18]. The parameters of HNMDRP, SVMDRP and Stanfield's method were set as recommended by the corresponding literature. We compared their performances over 226 drugs of the GDSC dataset via the same five-fold cross-validation experiments, and the results are shown in Fig. 3. Figure 3A–C displays their ROC curves on three drugs: VX-680, Erlotinib and Nilotinib. It can be observed that for each case the ROC curve of LGRDRP is clearly above those of the others, which implies the prediction performance of LGRDRP is the best. Figure 3D illustrates the AUCs over all drugs with the comparison results reported as boxplots, which shows LGRDRP is generally more accurate than other methods. HNMDRP performs slightly better than SVMDRP, and SVMDRP better than Stanfield's method, which indicates that protein information is not sufficient to reveal the cell line-drug association and integration of multi biological information could improve the predictive power. The average AUC of LGRDRP over all drugs is 0.8131, and achieves the highest value 0.9422 as regarding to drug SNX-2112, an oral anti-tumor drug as a Hsp90 inhibitor. For 50% drugs, the AUCs of LGRDRP are larger than 0.8229, and for 25% drugs, the AUCs of LGRDRP are larger than 0.8734.

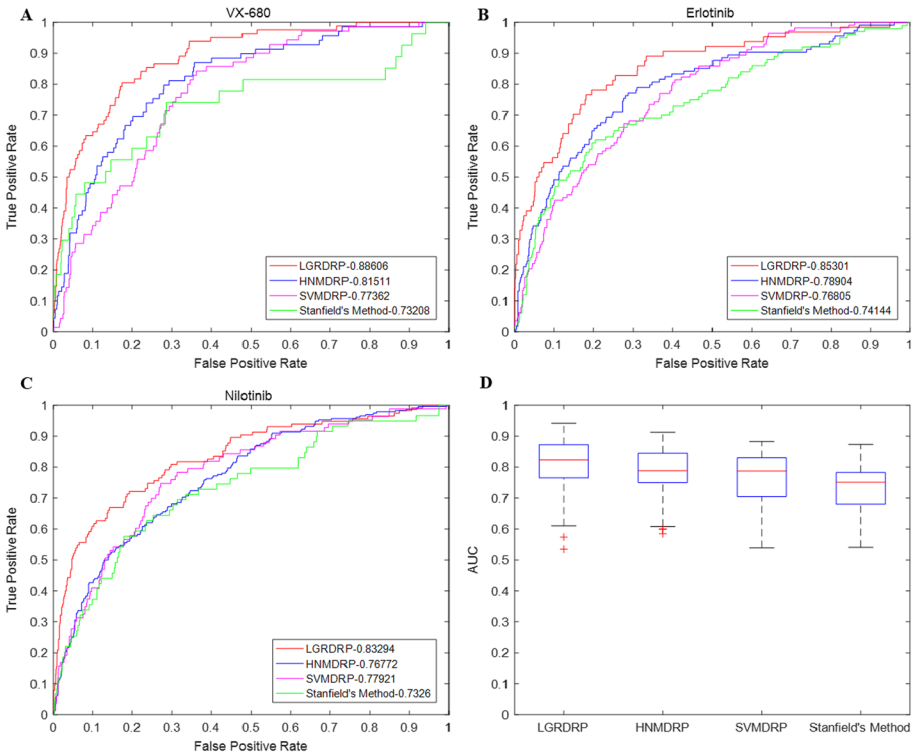

Fig. 3 Prediction performances of LGRDRP and the other three methods (HNMDRP, SVMDRP and Stanfield's method) based on their ROC curves and AUCs. **A**–**C** The ROC curves of four methods on drugs: VX-680, Erlotinib and Nilotinib, respectively. **D** The AUCs over all drugs

Figure 4A–C illustrates the PR curves of LGRDRP, HNMDRP, SVMDRP and Stanfield's method on three drugs FMK, AP-24534 and BMS-345541. The PR curves of LGRDRP also lie above those of the other methods. The average AUPR of LGRDRP over all drugs is 8.52%, 11.63%, 16.79% higher than those of HNMDRP, SVMDRP and Stanfield's method respectively. The AUPRs of the methods over all drugs are shown in Fig. 4D. The experiment results indicate that LGRDRP is successful to accomplish the prediction task even with the greatly unbalanced data set, demonstrating its reliability and prediction capability.

Figure 5 shows the retrieved number of real sensitive cell lines in the predicted top 10, 20, 50, 100 sensitive cell lines for drugs CAY10603 and NVP-BHG712, and LGRDRPP shows a significant advantage over the other methods again.

Conclusion

In this paper, we propose a drug responses prediction method called LGRDRP. It first uses the cell line miRNA expression profile to build a cell line similarity network, drug chemical structures to build drug similarity network, cell line gene variations and methylation data to build cell line-gene interaction network. By integrating the known cell line-drug responses into the above networks, LGRDRP constructs a heterogeneous network. Then LGRDRP uses a learning graph representation method GraRep to obtain the representation vectors as the topology structure features of vetices in the network. To avoid overfitting causing by using too many features, a Laplacian score method is

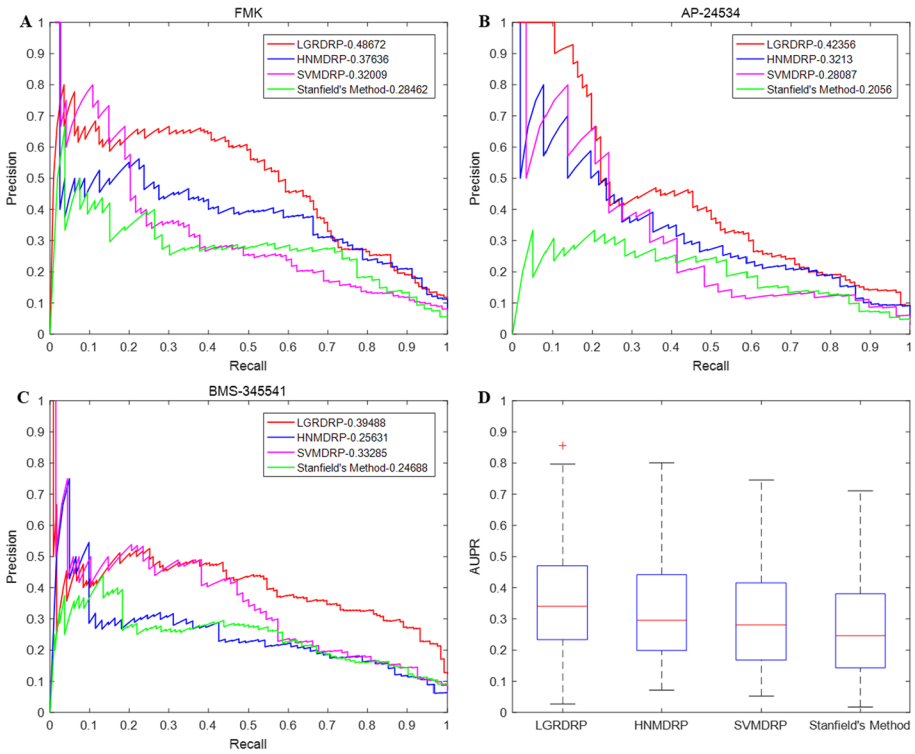

Fig. 4 Prediction performances of LGRDRP, HNMDRP, SVMDRP and Stanfield's method based on their PR curves and AUPRs. **A**–**C** The PR curves on drugs: FMK, AP-24534 and BMS-345541. **D** The AUPRs over all drugs

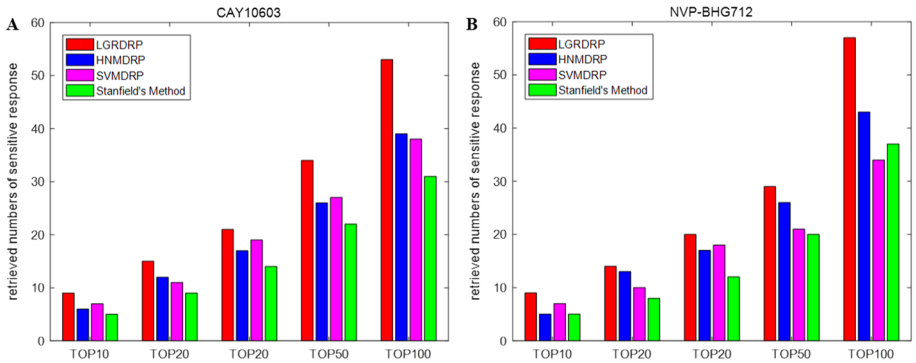

Fig. 5 The retrieved number of sensitive response cell lines in the TOP10, TOP20, TOP30, TOP50, TOP100 predicted sensitive cell lines

adopted to pick out some important features. Finally, LGRDRP learns an SVM model which is used to predict drug responses. Extensive 5-fold cross-validation experiments showed that LGRDRP was generally superior to three art-of-the-state methods HNMDRP, SVMDRP and Stanfield's method. The success of our method is based on the effective integration of diverse biological information, the good graph representation of the topology structure of the network, and the effective feature selection. After minor modifications or simple extensions, LGRDRP can also be employed in other biological predictions such as gene-disease [26], drug-target [27] and microRNA-disease [28], and the

prediction performance can be further improved by admitting other appropriate biological information, such as gene function annotations and drug semantic annotations. In clinical practice, some combinations of multiple drugs can increase treatment efficacy, and the response prediction of a cell line response to a drug combination is an important extension of the single drug response prediction [29]. In the future, we are going to improve LGRDPR so that it could deal with the drug combination response prediction.

Methods

Construction of the heterogeneous network

The heterogeneous network consists of a drug similarity network, a cell line similarity network, a gene similarity network, a cell line-drug interaction network, and a cell line-gene interaction network, as shown in Fig. 1A.

The drug similarity network is based on the chemical structure data of 226 frequently used drugs, which consists of a 3D structure similarity matrix of the drugs and was downloaded from PubChem (http://pubchem.ncbi.nlm.nih.gov/). To avoid disturbing from noises and make sure the network has a clear biological meaning, the elements smaller than 0.2 in the similarity matrix were set as 0. The drug similarity network consists of 226 vertices and 24456 edges, where each vertex denotes a drug and each edge has a similarity score weight (≥ 0.2).

miRNA expression information of cell lines could be used to classify cancer cell lines into subtypes [30], and our cell line similarity network was built on the miRNA expression data of 968 cancer cell lines, which was from CCLE (http://www.broadinstitute.org/ccle). The Pearson correlation coefficient of the miRNA expressions of two cell lines is regarded as the similarity between them, and is used as the weight of the corresponding edge in the network. Protein-protein interactions (PPIs) have been extensively studied, and we used the interactions between the proteins to represent the interactions between the genes coding the proteins. The gene-gene similarity network is based on the PPI data from iRefIndex [31], which contains 2981 genes and 53409 gene-gene interactions, with each gene possessing at least 5 interactions.

The cell line-drug interaction network is based on the drug response data of the 968 cell lines and the 226 drugs from GDSC (http://www.cancerrxgene.org/). The responses have been divided into two types sensitive and resistant according to the log-normalized IC50 threshold, and there are 20346 sensitive responses and 155277 resistant responses. Accordingly, there are 20346 interaction edges with weight 1 in the cell line-drug interaction network.

The copy number variation, somatic mutation and hypermethylation are called Cancer Functional Events (CFEs). The CFE data of the cell lines have been downloaded from GDSC and were used to build the cell line-gene interaction network. Similar to previous literature [32], we classified the cell line-gene relationship into associated and unassociated according to whether the coverage percentage of CFEs in the gene of the cell line is higher than 5%. Finally, we obtained a cell line-gene interaction network of 14330 associations between the 968 cell lines and the 2981 genes. The network is a bipartite graph consisting of gene vertices and cell line vertices and edges with weight 1 indicating corresponding cell line-gene associations.

Finally, we constructed a heterogeneous network including 226 drugs vertices in the drug similarity network, 2981 gene vertices in the gene similarity network, and 968 cell line vertices in the cell line similarity network. The network is represented as a weighted graph $G = (V, E)$. The vertex set of G is $V = \{v_1, v_2, \ldots, v_n\}$ whose element denotes a drug, a cell line or a gene. The edge set of G is $E = \{e_{i,j}\}$ whose element denotes the relationship between vertex v_i and vertex v_j, and the weight of an edge $e_{i,j}$ is set as described above.

Learning graph representation

As most similar works, we assume that similar cell lines tend to have similar responses to the same drug, and predict the response of a query cell line to a certain drug by utilizing the similarity between the query cell line and other cell lines with known responses to the drug. Since the heterogeneous network has integrated multiple types of information related with cell-lines and drugs, the neighbourhood of cell-lines in the network could be used to measure the similarity between cell-lines. Based on the idea, we use the learning graph representation method GraRep [23] to obtain representation vectors as the topology structure features of the vertices in the network.

Given a graph G, Learning Graph Representation (LGR) aims to learn a feature vector $F_i \in R^d$ for vertex v_i such that the global topology structure information (i.e. the neighbourhood) of the vertex is captured in the vector. In our method, the global topology structure information is represented by the distinct connections in different transitional steps between vertices, which is calculated by the process of LGR and is described as follows.

We use an $n \times n$ adjacent matrix M to represent the heterogeneous network G, and element M_{ij} is the weight of the edge between the vertices v_i and v_j. Based on M, a weighted degree matrix W is calculated according to Eq. (1).

$$W_{ij} = \begin{cases} \sum_{p=1,..,n} M_{ip}, & \text{if } i = j \\ 0, & \text{if } i \neq j \end{cases} \tag{1}$$

An edge of G implies an association relation. Considering transition between vertices, larger M_{ij} means larger transition probability from v_i to v_j. Hence the 1-step probability transition matrix T is calculated according Eq. (2).

$$T = W^{-1}M, \tag{2}$$

where the element T_{ij} is the transition probability from vertex v_i to v_j with exact one step. A k-step probability transition matrix T^k is calculated according Eq. (3).

$$T^k = \underbrace{T \cdots T}_{k}, \tag{3}$$

where T_{ij}^k is the transiting probability from vertex v_i to v_j with exact k steps.

For a drug d and a cell line c, let the representation vectors of d and c are $\vec{d}$ and $\vec{c}$ respectively, and we model the possibility $P(E = 1 \mid d, c)$ that c is sensitive to d (*i.e.* there is an edge between the vertices d and c in G) as follows:

$$P(E=1 \mid d,c) = \sigma(\vec{d}\cdot\vec{c}) = \frac{1}{1+\mathrm{e}^{-\vec{d}\,\vec{c}}}, \tag{4}$$

where $\sigma(.)$ is the sigmoid function. Accordingly, $P(E=0 \mid d,c)$ denotes the possibility that c is resistant to d (*i.e.* there is no edge between the vertices d and c in G) and $P(E=0 \mid d,c) = \sigma(-\vec{d}\cdot\vec{c})$.

Our objective is to maximize $P(E=1 \mid d,c)$ for the observed edge (d, c) in G while maximizing $P(E=0 \mid d,c)$ for the resistant response that there is no edge between the vertices d and c. Therefore, the following log-likelihood ℓ of G is our global objective function.

$$\ell = \sum_{d\in V_D}\sum_{c\in V_C} E(d,c)\left(\log\sigma(\vec{d}\cdot\vec{c}) + N\mathbb{E}_{c_N\sim P_E}\left[\log\sigma\left(-\vec{d}\cdot\vec{c}_N\right)\right]\right), \tag{5}$$

where V_D and V_C are the drug vertex set and the cell line vertex set respectively, and $E(d, c)$ indicates whether there is an edge between the drug vertex d and the cell line vertex c: $E(d,c)=1$ if there is an edge, otherwise, it is 0. N is the number of resistant responses, and $\mathbb{E}$ is the expectation value of the log-likelihood of resistant responses and is defined as Eq. (6).

$$\begin{aligned}\mathbb{E}_{c_N\sim P_E}\left[\log\sigma\left(-\vec{d}\cdot\vec{c}_N\right)\right] &= \textstyle\sum_{c_N\in V_C} p_E(c_N)\log\sigma\left(-\vec{d}\cdot\vec{c}_N\right)\\ &= p_E(c)\log\sigma(-\vec{d}\cdot\vec{c}) + \textstyle\sum_{c_N\in V_C\setminus\{c\}} p_E(c_N)\log\sigma\left(-\vec{d}\cdot\vec{c}_N\right),\end{aligned} \tag{6}$$

where $p_E(c)$ is the transiting probability from the drug vertex d to the cell vertex c.

Let $x = \vec{d}\cdot\vec{c}$, and maximizing ℓ requires the derivative of ℓ with respect to x be 0, therefore Eq. (7) follows.

$$x = \vec{d}\cdot\vec{c} = \log\left(\frac{p_E(c \mid d)|E|}{p_E(d)p_E(c)}\right) - \log N \tag{7}$$

Let D_i denotes the representation vector of ith drug vertex, C_j denotes the representation vector of jth cell line vertex, and $Y_{ij} = D_i \cdot C_j$. According to Eq. (7), we have:

$$Y_{ij} = D_i\cdot C_j = \log\left(\frac{T_{ij}}{\sum_{p=1}^{n} T_{ij}}\right) - \log N, \tag{8}$$

where T is the probability transition matrix of graph G.

Considering k-step random walks and based on Eqs. (3) and (8), we obtain Eq. (9).

$$Y_{ij}^k = \log\left(\frac{T_{ij}^k}{\sum_{p=1}^{n} T_{ij}^k}\right) - \log N \tag{9}$$

In order to obtain the representation vectors of the drug vertices and the cell line vertices, we apply a popular singular value decomposition (SVD) method to factorize Y.

$$\left[U^k, \Sigma^k, \left(V^k\right)^T\right] = \mathrm{SVD}\left(Y^k\right) \tag{10}$$

The representation vector matrix of k-step random walks F^k is calculated as follows.

$$F^k = U_d^k \left(\Sigma_d^k\right)^{1/2} \tag{11}$$

In Eqs. (10) and (11), Σ_d^k is the matrix composed by the top d singular values and U_d^k is the first d columns of U^k, which are the first d eigenvector of $Y^k(Y^k)^T$. For $k = 1$ to K, we calculate K representation vector matrices, and concatenating them obtains an $n \times Kd$ matrix F, whose rows are the representation vectors of the vertices in G. F will be used as the input features in the following classification.

Classification via support vector machine

As we have assumed before, cell lines with similar topology structures in the network tend to have similar responses to the same drug. Since cell lines with similar topology structures are more similar with respect to the representation vectors, we construct a binary classification model using the representation vectors of the cell lines as the input features and output their responses (sensitive or resistant) to a query drug. To integrate comprehensive similarity information between cell lines, we consider all k-step representations with $k = 1, 2, \ldots, K$. Using a larger K could capture more distant similarity information but also introduce more noise.

In case that drug vertices and genes vertices have few edges with cell line vertices in the heterogeneous network, there would be a large number of poor features for some cell lines. When K and d are large numbers, the number of features (i.e., the length $K \times d$ of the representation vector) will be too large, and the overfitting problem will occur. To deal with the problem, we use a Laplacian score method [24] to select out most valuable features from the $K \times d$ features. For each feature, the Laplacian score method assesses the ability to represent the graph structure and calculate a corresponding score. We only use the features with top t Laplacian scores for the classification. The values of K, d, and t are determined by 5-fold cross-validation experiments as described earlier in the subsection Parameters Selection.

In the heterogeneous network, the number of sensitive drug responses is 20346, while the number of resistant drug responses is 155277, which is a great imbalance of positive and negative samples. Since the support vector machine (SVM) method could effectively deal with the imbalance problem by assigning different weights to positive and negative samples, we chose SVM to conduct the classification task. In the following experiments, we employed LIBSVM [24] to do the classification. LIBSVM is an integrated software package including diverse SVM models, which can be chosen by setting some options. We set the options of LIBSVM as follows: the SVM_type was set as the default C-SVC, the kernel type was set as polynomial

and the order of the polynomial kernel took the default value 3, the weights for the positive class and the negative class were set as the number of negative samples, and positive samples, respectively, and the other options were left default. For each drug, LIBSVM can learn an SVC model on the training data, and for each cell line, the model outputs a decision score. If the decision score is larger than 0, the cell line is predicted sensitive to the drug, and a larger score indicates that the prediction is more convinced.

Abbreviations

AUC	Area under receiver operating characteristic curve
AUPRC	Area under precision-recall curve
SVM	Support vector machine

Acknowledgements
We would like to thank the anonymous reviewers for their helpful and constructive comments.

About this supplement
This article has been published as part of BMC Bioinformatics Volume 23 Supplement 8, 2022: Selected articles from the 16th International Symposium on Bioinformatics Research and Applications (ISBRA-20): bioinformatics. The full contents of the supplement are available online at https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-23-supplement-8.

Author contributions
M.X. and J.O. designed the model, algorithm and experiments. J.O. implemented the algorithm and X.L. conducted the experiments. M.X., G.L. and J.Z. wrote the paper. All authors read and approved the final manuscript.

Funding
This work was supported by the National Natural Science Foundation of China (Nos. 62172028, 61772197).

Availability of data and materials
The datasets used and analysed during the current study available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
Not Applicable.

Consent for publication
Not Applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 6 November 2022 Accepted: 22 November 2022
Published online: 09 December 2022

References

1. ...Costello JC, Heiser LM, Georgii E, Gonen M, Menden MP, Wang NJ, Bansal M, Ammad-ud-din M, Hintsanen P, Khan SA, Mpindi JP, Kallioniemi O, Honkela A, Aittokallio T, Wennerberg K, Community ND, Collins JJ, Gallahan D, Singer D, Saez-Rodriguez J, Kaski S, Gray JW, Stolovitzky G. A community effort to assess and improve drug sensitivity prediction algorithms. Nat Biotechnol. 2014;32(12):1202–12. https://doi.org/10.1038/nbt.2877.
2. Eisenstein M. Personalized medicine: special treatment. Nature. 2014;513(7517):8–9. https://doi.org/10.1038/513S8a.
3. Mirnezami R, Nicholson J, Darzi A. Preparing for precision medicine. N Engl J Med. 2012;366(6):489–91. https://doi.org/10.1056/NEJMp1114866.
4. Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, Bindal N, Beare D, Smith JA, Thompson IR, Ramaswamy S, Futreal PA, Haber DA, Stratton MR, Benes C, McDermott U, Garnett MJ. Genomics of drug sensitivity in cancer (gdsc): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Res. 2013;41(Database issue):955–61. https://doi.org/10.1093/nar/gks1111.
5. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehar J, Kryukov GV, Sonkin D, Reddy A, Liu M, Murray L, Berger MF, Monahan JE, Morais P, Meltzer J, Korejwa A, Jane-Valbuena J, Mapa FA, Thibault J, Bric-Furlong E, Raman P, Shipway A, Engels IH, Cheng J, Yu GK, Yu J, Aspesi JP, de Silva M, Jagtap K, Jones MD, Wang L, Hatton C, Palescandolo E, Gupta S, Mahan S, Sougnez C, Onofrio RC, Liefeld T, MacConaill L, Winckler W, Reich M, Li N, Mesirov JP, Gabriel SB, Getz G, Ardlie K, Chan V, Myer VE, Weber BL, Porter J, Warmuth M, Finan P, Harris JL, Meyerson M, Golub TR, Morrissey MP, Sellers WR, Schlegel R, Garraway LA. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012;483(7391):603–7. https://doi.org/10.1038/nature11003.
6. Torkamani A, Schork NJ. Background gene expression networks significantly enhance drug response prediction by transcriptional profiling. Pharmacogenomics J. 2012;12(5):446–52. https://doi.org/10.1038/tpj.2011.35.
7. Gupta S, Chaudhary K, Kumar R, Gautam A, Nanda JS, Dhanda SK, Brahmachari SK, Raghava GP. Prioritization of anticancer drugs against a cancer using genomic features of cancer cells: a step towards personalized medicine. Sci Rep. 2016;6:23857. https://doi.org/10.1038/srep23857.
8. Fang Y, Xu P, Yang J, Qin Y. A quantile regression forest based method to predict drug response and assess prediction reliability. PLoS One. 2018;13(10):0205155. https://doi.org/10.1371/journal.pone.0205155.
9. Dong ZL, Zhang NQ, Li C, Wang HY, Fang Y, Wang J, Zheng XQ. Anticancer drug sensitivity prediction in cell lines from baseline gene expression through recursive feature selection. BMC Cancer. 2015;15:489. https://doi.org/10.1186/s12885-015-1492-6.

10. Geeleher P, Cox NJ, Huang RS. Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines. Genome Biol. 2014;15(3):47. https://doi.org/10.1186/gb-2014-15-3-r47.
11. Liu CY, Wei D, Xiang J, Ren FQ, Huang L, Lang JD, Tian G, Li YS, Yang JL. An improved anticancer drug-response prediction based on an ensemble method integrating matrix completion and ridge regression. Mol Ther Nucleic Acids. 2020;21:676–86. https://doi.org/10.1016/j.omtn.2020.07.003.
12. Ammad-ud-din M, Georgii E, Gonen M, Laitinen T, Kallioniemi O, Wennerberg K, Poso A, Kaski S. Integrative and personalized QSAR analysis in cancer by kernelized Bayesian matrix factorization. J Chem Inf Model. 2014;54(8):2347–59. https://doi.org/10.1021/ci500152b.
13. Li M, Wang Y, Zheng R, Shi X, Li Y, Wu FX, Wang J. Deepdsc: a deep learning method to predict drug sensitivity of cancer cell lines. IEEE/ACM Trans Comput Biol Bioinform. 2021;18(2):575–82. https://doi.org/10.1109/TCBB.2019.2919581.
14. Yan XY, Zhang SW, Yiu SM, Shi JY. Interpretable prediction of drug-cell line response by triple matrix factorization. Quantit Biol. 2021;9(4):426–39. https://doi.org/10.15302/j-qb-021-0259.
15. Guvenc Paltun B, Kaski S, Mamitsuka H. Diverse: Bayesian data integrative learning for precise drug response prediction. IEEE/ACM Trans Comput Biol Bioinform. 2021. https://doi.org/10.1109/tcbb.2021.3065535.
16. Creighton CJ. Molecular classification and drug response prediction in cancer. Curr Drug Targets. 2012;13(12):1488–94. https://doi.org/10.2174/138945012803530143.
17. Wang Y, Fang J, Chen S. Inferences of drug responses in cancer cells from cancer genomic features and compound chemical and therapeutic properties. Sci Rep. 2016;6:32679. https://doi.org/10.1038/srep32679.
18. Stanfield Z, Coskun M, Koyuturk M. Drug response prediction as a link prediction problem. Sci Rep. 2017;7:40321. https://doi.org/10.1038/srep40321.
19. Zhang F, Wang M, Xi J, Yang J, Li A. A novel heterogeneous network-based method for drug response prediction in cancer cell lines. Sci Rep. 2018;8(1):3355. https://doi.org/10.1038/s41598-018-21622-4.
20. Liu H, Zhao Y, Zhang L, Chen X. Anti-cancer drug response prediction using neighbor-based collaborative filtering with global effect removal. Mol Ther Nucleic Acids. 2018;13:303–11. https://doi.org/10.1016/j.omtn.2018.09.011.
21. Zhang L, Chen X, Guan NN, Liu H, Li JQ. A hybrid interpolation weighted collaborative filtering method for anticancer drug response prediction. Front Pharmacol. 2018;9:1017. https://doi.org/10.3389/fphar.2018.01017.
22. Guan NN, Zhao Y, Wang CC, Li JQ, Chen X, Piao X. Anticancer drug response prediction in cell lines using weighted graph regularized matrix factorization. Mol Ther Nucleic Acids. 2019;17:164–74. https://doi.org/10.1016/j.omtn.2019.05.017.
23. Cao S, Lu W, Xu Q. Grarep: Learning graph representations with global structural information. In: Proceedings of the 24th ACM international on conference on information and knowledge management. New York, NY, United States: Association for Computing Machinery; 2015. pp. 891–900. https://doi.org/10.1145/2806416.2806512.
24. Chang CC, Lin CJ. Libsvm: a library for support vector machines. Acm Trans Intell Syst Technol. 2011. https://doi.org/10.1145/1961189.1961199.
25. Awan SE, Bennamoun M, Sohel F, Sanfilippo FM, Chow BJ, Dwivedi G. Feature selection and transformation by machine learning reduce variable numbers and improve prediction for heart failure readmission or death. PLoS One. 2019;14(6):0218760. https://doi.org/10.1371/journal.pone.0218760.
26. Le DH. Machine learning-based approaches for disease gene prediction. Brief Funct Genomics. 2020;19(5–6):350–63. https://doi.org/10.1093/bfgp/elaa013.
27. Chen X, Yan CC, Zhang XT, Zhang X, Dai F, Yin J, Zhang YD. Drug-target interaction prediction: databases, web servers and computational models. Brief Bioinform. 2016;17(4):696–712. https://doi.org/10.1093/bib/bbv066.
28. Chen X, Xie D, Zhao Q, You ZH. Micrornas and complex diseases: from experimental results to computational models. Brief Bioinform. 2019;20(2):515–39. https://doi.org/10.1093/bib/bbx130.
29. Chen X, Ren B, Chen M, Wang Q, Zhang L, Yan G. NIlss: predicting synergistic drug combinations based on semi-supervised learning. Plos Comput Biol. 2016;12(7):1004975. https://doi.org/10.1371/journal.pcbi.1004975.
30. Yu CW, Dai DJ, Xie J. Molecular subtype classification of papillary renal cell cancer using mirna expression. Onco-targets Therapy. 2019;12:2311–22. https://doi.org/10.2147/Ott.S193808.
31. Razick S, Magklaras G, Donaldson IM. irefindex: a consolidated protein interaction database with provenance. BMC Bioinformatics. 2008;9:405. https://doi.org/10.1186/1471-2105-9-405.
32. Iorio F, Knijnenburg TA, Vis DJ, Bignell GR, Menden MP, Schubert M, Aben N, Goncalves E, Barthorpe S, Lightfoot H, Cokelaer T, Greninger P, van Dyk E, Chang H, de Silva H, Heyn H, Deng XM, Egan RK, Liu QS, Mironenko T, Mitropoulos X, Richardson L, Wang JH, Zhang TH, Moran S, Sayols S, Soleimani M, Tamborero D, Lopez-Bigas N, Ross-Macdonald P, Esteller M, Gray NS, Haber DA, Stratton MR, Benes CH, Wessels LFA, Saez-Rodriguez J, McDermott U, Garnett MJ. A landscape of pharmacogenomic interactions in cancer. Cell. 2016;166(3):740–54. https://doi.org/10.1016/j.cell.2016.06.017.

Publisher's Note