OXFORD

Gene expression

A Survey and Systematic Assessment of Computational Methods for Drug Response Prediction

Jinyu Chen and Louxin Zhang*

Department of Mathematics and Computational Biology Programme, National University of Singapore, Singapore 119076

* Corresponding author: Louxin Zhang, Tel: +65-65166579; Email: matzlx@nus.edu.sg

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Drug response prediction arises from both basic and clinical research of personalized therapy, as well as drug discovery for cancer and other diseases. With gene expression profiles and other omics data being available for over 1000 cancer cell lines and tissues, different machine learning approaches have been applied to solve drug response prediction problems. These methods appear in a body of literature and have been evaluated on different datasets with only one or two accuracy metrics. By using nine different metrics, we systematically evaluated 14 representative methods for predicting drug responses in cell lines that have been developed in recent years. This study provides insights and lessons for future research into drug response prediction.

Keywords: Drug response prediction; regression model; low rank matrix factorization based approach; Bayesian inference; benchmarking

Jinyu Chen is currently a postdoctoral research fellow at the Department of Mathematics, National University of Singapore. She completed her PhD at the Academy of Mathematics and Systems Science, Chinese Academy of Sciences, China. Her research interests include integrative analysis of omics data and matrix-factorization methods in bioinformatics.

Louxin Zhang is currently a professor at the Department of Mathematics and the Computational Biology Programme, National University of Singapore. He received his PhD in computer science from the University of Waterloo, Canada. His research interests include bioinformatics in comparative and evolutionary genomics, cancer genomics, cellular protein networks and discrete algorithms.

1 Introduction

Cancer is still an incurable disease. Recent cancer genome studies have suggested that each cancer patient owns a unique profile of gene mutations and that there is no silver bullet for defeating all the cancers. Therefore, how to use the genomic and clinical profiles of cancer patients to design a personalized treatment is vital for effective control of disease progression [1, 2]. This requires accurate prediction of the patients' response to a target-specific drug and/or therapy.

Given the paucity of patients' molecular profiles and their responses to drugs, large-scale drug screening experiments on cultured and genomically profiled cancer cell lines have been used to study drug response [3, 4]. The US National Cancer Institute (NCI) 60 human tumor cell lines screen database provides drug screening data for thousands of drugs in 60 tumor cell lines, for which mRNA, microRNA expression, DNA methylation and mutation profiles are also available [3, 5]. In 2012, another two large-scale pharmacogenomic datasets were made publicly available: the Cancer Cell Line Encyclopedia [6] and the Genomics of Drug Sensitivity in Cancer (GDSC) [7]. The former contains drug sensitivity data from over 11,000 experiments involving 24 drugs and 504 cell lines, and the gene mutational and expression profiles for about 1000 cell lines from 36 tumor sites. The latter includes about 200,000 drug response experiments that tested 266 drugs in over 1000 human cancer cell lines.

With a large amount of drug response data for cell lines being available, various machine learning approaches have been applied to design computational methods for drug response prediction in recent years. These prediction methods, their interpretability, and the importance of the integration of gene expression and other omics profiles for performance improvement are well documented in two survey articles [8, 9]. Additionally, key methods developed before 2014 were evaluated in [10] and [11]. Since 2015, drug response prediction problems have been extensively studied [12–27].

In this study, we aim to systematically assess representative computational methods that have been developed by different research groups since 2015. These methods have often been evaluated by the developers on the datasets that were obtained from the aforementioned three databases through different filtering processes with one or two metrics. Therefore, it remains unclear which are among the best. We evaluated these state-of-the-art methods with common datasets according to nine performance measurements for a comprehensive assessment. On the basis of these analyses, we draw lessons and insights that could hopefully benefit future research into drug response prediction.

2 Brief overview of drug response prediction methods

2.1 Drug response prediction problem

Drug response in a cell line are often measured in terms of either the half maximal inhibitory concentration (IC_{50}) or the area under the concentration response curve (AUC) [28]. In studies into drug response problems, the responses of m drugs in a set of n cell lines are given as an $n \times m$ real matrix, in which the (i, j)-th entry is the IC_{50} or AUC value of the j-th drug in the i-th cell line.

Drug response prediction is a supervised learning problem: given a training drug response dataset $R_{n\times m}$ containing n cell lines and m drugs, a response function $r(x, y)$ is learned from $R_{n\times m}$ that can be used to compute the response of (i) a new cell line x to a known drug y (in the training dataset), (ii) a known cell line x to a new drug y or (iii) a new cell line x to a new drug y. In this study, we assess different methods for the first type of drug response prediction.

2.2 Integration of omics data

The rationale for the predictability of drug responses is that the responses to drugs with similar chemical properties are highly related and cell lines with similar genomic profiles also show similar drug responses. Therefore, prediction methods with good performance often use a few extra omics datasets other than the input drug response matrix. The omics data that are frequently used include:

- Gene expression profiles, which characterize the expression level of a set of genes in a cell line. Currently, both microarray and RNA-Seq gene expression datasets are available in the aforementioned databases.
- Copy number alteration/aberration (CNA) profiles. Copy number alterations are changes in copy number in a tumor cell line.
- Mutational profiles. These refer to a list of mutation types (e.g., base substitutions, short insertions and deletions) and mutation counts in a tumor cell line.
- DNA methylation profiles. DNA methylation is tissue-specific and dynamic. The DNA methylation profiles characterize the patterns of aberrant DNA methylation occurring in a tumor cell line.
- Gene similarity profiles. These comprise curated data characterizing functionally and structurally similarities of the involved genes. Such profiles are often inferred from available genomics and proteomics data.
- Cellular protein networks. These could be protein–protein interaction networks, gene regulatory networks, or signaling pathways.
- Protein expression profiles. These characterize the expression levels of a set of proteins in a cell line. Protein expression levels can be measured on the reverse phase protein lysate microarray platform.
- Drug features. There are two main types of drug features. One is information on the drug's primary target or target pathways, which could be used for drug classification. Another is the molecular structural information. Each drug is characterized by a number of 1-dimensional and 2-dimensional structural properties, such as the number of atoms, the presence of chemical substructures, etc.

By concatenating different omics datasets and encoding non-numeric entries into numbers (if necessary), we can represent multiple omics profiles of n cell lines with a real-valued matrix $C_{n\times s}$, in which each column corresponds to a feature in the omics profiles. Similarly, we can also represent multiple features of m drugs by a real-valued matrix S with m rows and integrate the relationships among s genes such as a gene similarity profile and a cellular protein network in a matrix G. Thus, at a more abstract level, the training datasets for drug response prediction include a cell line–drug response matrix $R_{n\times m}$ and at most three auxiliary profile matrices: $C_{n\times s}$, S and G.

Table 1 summarizes which omics data are used by each of the methods to be evaluated in this work.

2.3 Single-task learning methods

Different regression methods have been developed for drug response prediction. They use the drug response vector $R_{n\times 1}$ of a drug d to learn the linear response function $r_d(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$ with the coefficient vector $\mathbf{w} = (w_1, w_2, \cdots, w_s)^T$, where $\mathbf{x} = (x_1, x_2, \cdots, x_s)^T$ represents the given s-dimensional profile vector of a cell line. For a new cell line with the profile vector $\mathbf{c}$, its predicted response to drug d is $r_d(\mathbf{c})$.

These methods compute the coefficient vector $\mathbf{w}$ in the function $r_d(\mathbf{x})$ by solving the following optimization problem:

$$\min_{\mathbf{w}} \frac{1}{n}\|B_{n\times n}(R_{n\times 1} - C_{n\times s}\mathbf{w})\|_2^2 + \lambda_1\|\mathbf{w}\|_1 + \lambda_2\mathbf{w}^T M\mathbf{w}, \quad (1)$$

where $\|\cdot\|_1$ and $\|\cdot\|_2$ are, respectively, the L_1-and L_2-norm defined for real vectors; M is either the $s \times s$ identity matrix or a $s \times s$ symmetric

Table 1. Summary of key types of auxiliary training data used in the 14 methods in addition to drug response data. GE: Gene expression profile; CNA: Copy number aberration profile; MUT: Mutational profile; PN: Cellular protein network; DM: DNA methylation profile; PE: Protein expression profile; GS: Gene similarity profile; DF: Drug features.

	Omics Data Type							
	GE	CNA	MUT	PN	DM	PE	GS	DF
BMTMKL	√	√	√		√	√	√	
CaDRReS	√							
cwKBMF	√						√	√
DualNets	√							√
ELnet	√	√	√					
GELnet	√			√				
KRL	√							
KRR	√							
MERGE	√	√	√		√			
MVLR	√						√	√
pairwiseMKL	√	√	√		√			√
RF	√							
RWEN	√							
SRMF	√							√

semidefinite matrix derived from an input profile matrix; T denotes the transpose operator for matrices; and $B_{n\times n}$ is the diagonal matrix with n parametric diagonal entries b_j. The free parameters λ_i's and b_j's are set differently in each method.

Elastic net (ELnet)
In this method, $M = \mathbf{I}_{s\times s}$ and $B_{n\times n} = I_{n\times n}$ (i.e., $b_i = 1$) in Eqn. (1) [29]. Thus the third term of the equation simply becomes the L_2-norm of the coefficient vector $\mathbf{w}$. Additionally, ELnet uses a 5- or 10-fold cross-validation procedure to determine the values of λ_1 and λ_2.

Generalized elastic net (GELnet)
In this method, $B_{n\times n} = I_{n\times n}$ in Eqn. (1) [12]. We assessed two versions of this model: **GELnet-P** and **GELnet-NP** in which M is set to the Laplacian matrix and normalized Laplacian matrix of an input network, respectively.

Response-weighted elastic net (RWEN)
These methods use an iterative procedure to set (b_j) and compute the coefficient vector $\mathbf{w}$ in Eqn. (1) [13]. In these methods, the drug response data $R_{n\times 1} = (r_i)_{i=1}^n$ is assumed to follow a normal distribution $N(\mu, \sigma^2)$ with the parameters being estimated as:

$$\mu = \text{median}(R_{n\times 1}),$$

$$\sigma = 1.4826 \times \text{median}_{1\le i\le n}\big(|r_i - \text{median}(R_{n\times 1})|\big).$$

Under this assumption, the drug response value is less than ℓ (resp. larger than γ) with a probability of 0.05, where $\ell = \mu - 1.65 \times \sigma$ and $\gamma = \mu + 1.65 \times \sigma$.

In one RWEN method, we set $(b_j) = (1)$, that is, $b_j^{(0)} = 1$ $(1 \le j \le n)$. Thus in the first step, ELnet is simply used to compute the coefficient vector $\mathbf{w}^{(0)}$. In the t-th iteration step for $t \ge 1$, $b_j^{(t)}$ are set as follows, depending on whether or not $r_j \le \ell$ with the current coefficient vector $\mathbf{w}^{(t-1)}$:

$$b_j^{(t)} = \begin{cases} \exp\big(1 + |r_j - \mathbf{c}_j\mathbf{w}^{(t-1)}|\big), & r_j \le \ell, \\ \delta, & \text{otherwise}, \end{cases} \tag{2}$$

where δ is a small constant close to 0 and $\mathbf{c}_j$ is the j-th row of $C_{n\times s}$. The purpose of setting (b_j) in this way is to find a coefficient vector that gives a good prediction for low drug responses (i.e. cell lines that are sensitive to a drug). The algorithm continues until the difference between the input drug response and the predicted value is small enough. This variant is called the RWEN-left method.

Symmetrically, the method can be modified for predicting high drug responses (i.e. cell lines that are resistant to a drug) by setting $r_j \ge \gamma$ in Eqn. (2), as well as for predicting both head and tail drug responses. These two variants of RWEN are called the RWEN-right and RWEN-both methods, respectively.

Kernel ridge regression (KRR)
This method first maps a cell line vector v to a vector $f(v)$ in a high-dimensional feature space and then computes the coefficient vector $\mathbf{w}$ in the image vector space [20, 30]. More specifically, it solves the following optimization problem:

$$\min_{\mathbf{w}} \frac{1}{n}\|R_{n\times 1} - f(C_{n\times s})\mathbf{w}\|_2^2 + \lambda \mathbf{w}^{\mathrm{T}}\mathbf{w}.$$

This problem can be solved as easily as solving Eqn. (1). The reason is that one only needs to know how to compute the inner product $f(\mathbf{u}) \cdot f(\mathbf{v})$ of image vectors to solve the problem. He *et al.* used the so-called Gaussian kernel function to define the inner product of the vectors $f(\mathbf{u})$ and $f(\mathbf{v})$ as follows [20]:

$$f(\mathbf{u}) \cdot f(\mathbf{v}) = \exp\left(-\frac{\|\mathbf{u} - \mathbf{v}\|_2^2}{\gamma}\right), \tag{3}$$

where γ is a free parameter.

Random forest with gene filtering (RF-g) and with both gene and sample filtering (RF-gs)
Random forest is a very efficient machine learning approach that captures nonlinear relationships between predictors and response variables [31]. It first uses bootstrapping to split the data into several sub-datasets, fits a decision tree to each of them, and finally averages the results. In order to improve the accuracy of the drug response predictions, Riddick *et al.* incorporated gene signature filtering and outlying cell lines removing procedures in the random forest framework [32].

2.4 Multi-task learning methods

The multi-task learning paradigm allows all the responses $r(i, j)$ $(i = 1, 2, \cdots, n; j = 1, 2, \cdots, m)$ to be jointly learned from the training response matrix $R_{n\times m}$ and auxiliary cell line and drug profile matrices. Since some drugs are similar, this approach can achieve high accuracy through providing a statistical boost, particularly when the sample size is small.

The Bayesian multi-task multi-kernel learning method (BMTMKL)
For a matrix X, we use $X[i, *]$ and $X[*, j]$ to denote its i-th row and j-th column, respectively. We can rewrite $R_{n\times m}$ as follows:

$$R_{n\times m} = (R[*, 1], R[*, 2], \cdots, R[*, m]),$$

where $R[*, j]$ is the responses to drug j in the cell lines. Costello *et al.* proposed to decompose $R[*, j]$ into a linear combination of p components $Y_j^{(k)} (k = 1, 2, \cdots, p)$ as follows [11]:

$$R[*, j] \approx \sum_{k=1}^{p} e_k Y_j^{(k)} + \begin{bmatrix} b_j & b_j & \cdots & b_j \end{bmatrix}^T \tag{4}$$

for each $j = 1, 2, \cdots, m$; equivalently:

$$R_{n\times m} \approx \begin{bmatrix} e_1 & e_2 & \cdots & e_p \end{bmatrix} \begin{bmatrix} Y_1^{(1)} & Y_2^{(1)} & \cdots & Y_m^{(1)} \\ Y_1^{(2)} & Y_2^{(2)} & \cdots & Y_m^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ Y_1^{(p)} & Y_2^{(p)} & \cdots & Y_m^{(p)} \end{bmatrix} + \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix}^T \begin{bmatrix} b_1 & b_2 & \cdots & b_m \end{bmatrix} \quad (5)$$

where $[e_1, e_2, \cdots, e_p] \in \mathbb{R}^{1\times p}$ is common to all drugs, and $Y_j^{(i)} \in \mathbb{R}^{n\times 1}$ and b_j only depend on drug j for each pair of i and j.

One may consider $Y_j^{(i)}$ to be the response component of drug j that is inferred from the i-th ($i = 1, 2, \cdots, p$) type of input cell line profiles (such as the DNA methylation profiles or gene expression profiles) for each $j = 1, 2, \cdots, m$. Thus Costello *et al.* also assumed the following connection between $Y_j^{(i)}$ and cell line profiles $K_j^{(i)}$:

$$Y_j^{(i)} = K_j^{(i)} \begin{bmatrix} a_{1j} & a_{2j} & \cdots & a_{nj} \end{bmatrix}^T \quad (6)$$

for each drug j, where $K_j^{(i)}$ is an $n \times n$ similarity matrix that is computed via the Gaussian kernel function in Eqn. (3) or other similarity functions.

Costello *et al.* inferred drug responses by using Bayesian inference. They assumed that each e_i of the parameter vector $e = (e_1, e_2, \cdots, e_p)$ follows a normal distribution $\mathcal{N}(0, \sigma_i^{-1})$ such that σ_i follows a Gaussian distribution:

$$\sigma_i \sim \mathcal{G}(\alpha_\sigma, \beta_\sigma).$$

Similarly, for each drug $j = 1, 2, \cdots, m$, each b_j is sampled from a similar distribution:

$$b_j \sim \mathcal{N}(0, \eta_j^{-1}), \quad \eta_j \sim \mathcal{G}(\alpha_\eta, \beta_\eta).$$

Each component of the vector $(a_{1j}, a_{2j}, \cdots, a_{nj})$ is independently sampled from the following distribution:

$$a_{ij} \sim \mathcal{N}(0, \delta_{ij}^{-1}), \quad \delta_{ij} \sim \mathcal{G}(\alpha_\delta, \beta_\delta).$$

Costello *et al.* further assumed that the response components $Y_j^{(i)}$ ($i = 1, 2, \cdots, p$) are sampled from the following multivariate normal distributions:

$$Y_j^{(i)} \sim \mathcal{N}\left(K_j^{(i)} \begin{bmatrix} a_{1j} & a_{2j} & \cdots & a_{nj} \end{bmatrix}^T, \epsilon_j^{-1}\mathbf{I}_{n\times n}\right),$$
$$\epsilon_j \sim \mathcal{G}(\alpha_\epsilon, \beta_\epsilon)$$

and

$$R[*, j] \sim \mathcal{N}\left(\sum_{k=1}^{p} e_k Y_j^{(k)} + (b_j)_{n\times 1}, \zeta_j^{-1}\mathbf{I}_{n\times n}\right),$$
$$\zeta_j \sim \mathcal{G}(\alpha_\zeta, \beta_\zeta).$$

In summary, this is a very general stochastic model involving the following 10 parameters for prior Gamma distributions:

$$\alpha_\sigma, \beta_\sigma, \alpha_\eta, \beta_\eta, \alpha_\delta, \beta_\delta, \alpha_\epsilon, \beta_\epsilon, \alpha_\zeta, \beta_\zeta. \quad (7)$$

Costello *et al.* applied their method to an instance where the training dataset involved 28 drugs and 35 cell lines and the test data included 18 cell lines. They set the number of components (p) to 22 by considering different combinations of input cell line profiles. It is worth pointing out that they used a form that is more general than Eqn. (6) so that the responses data do not necessarily involve the same set of cell lines for different drugs.

Component-wise kernel Bayesian matrix factorization (cwKBMF)
This was developed on the basis of a matrix factorization that was more general than Eqn. (4)–(6) [18]. Ammad-ud-din et al. took drug profile into account by using a drug similarity matrix.

The CaDRReS method
This method was developed under the assumption that the drug response function can be approximated by the sum of a few low-rank matrices [21] . More specifically, the response matrix $R_{n\times m}$ for m drugs and n cell lines has the following decomposition:

$$R_{n\times m} \approx \mu(R) \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix} + \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \begin{bmatrix} b_1 & b_2 & \cdots & b_m \end{bmatrix} + K_{n\times n} U_{n\times f} W_{f\times m}, \quad (8)$$

where K is a similarity matrix of the cell lines, f is a small positive integer, and $\mu(R)$ is the average value of entries in $R_{n\times m}$.

Let $\bar{R}_{n\times m}((a_i), (b_j), U, W)$ denote the right-hand side of Eqn. (8). Suphavilai *et al.* solved the following minimization problem by using the gradient descent method to predict drug responses [21]:

$$\min_{(a_i),(b_j),U,W} \|R_{n\times m} - \bar{R}_{n\times m}((a_i), (b_j), U, W)\|_F^2,$$

where K is calculated via the Pearson's correlation of gene expression profiles for cell lines and $f = 10$.

Multi-view multi-task linear regression (MVLR)
This approaches the linear regression model via Bayesian inference. For m drugs targeting the same pathways, the linear regression model assumes the following relationship between the drug responses $R_{n\times m}$ and cell line profiles $C_{n\times s}$:

$$R_{n\times m} \approx C_{n\times s} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_s \end{bmatrix} [a_1, a_2, \cdots, a_m]. \quad (9)$$

In [19], the cell line profile was divided into p parts, i.e.,

$$C_{n\times s} = \begin{bmatrix} C_{n\times s_1}^{(1)} & C_{n\times s_2}^{(2)} & \cdots & C_{n\times s_p}^{(p)} \end{bmatrix},$$

where $\sum_{1\le i\le p} s_i = s$. The rationale of this approach is that each $C_{n\times s_i}^{(i)}$ represents a data view, generated for related features; the corresponding w_i's form the regression coefficient vector for each data view; and a_i is a parameter set for each drug i. Note that if $p = 1$, $s_1 = s$, Eqn. (9) is equivalent to the relationship between $R[*, j]$ and $C_{n\times s}$ on which the linear regression model Eqn. (1) is based for each j.

Ammad-ud-din *et al.* inferred the parameters (w_j) $(1 \le j \le s)$ and a_i $(1 \le i \le m)$ by assuming the following distributions [19] :

$$a_i \sim \text{Dir}(\alpha), \quad i = 1, 2. \cdots, m;$$
$$w_k \sim \text{Cauchy}(0, \beta_k\gamma_v), \quad \sum_{i=1}^{v-1} s_i + 1 \le k \le \sum_{i=1}^{v} s_i;$$
$$\beta_k \sim \text{Cauchy}_+(0, \epsilon), \ \gamma_v \sim \text{Dir}(\delta),$$

and

$$R[*,j] \sim \mathcal{N}\left(a_j\left(\sum_{k=1}^{s} c_{ik}w_k\right)_{n\times 1}, \sigma_j \mathbf{I}_{n\times n}\right),$$
$$\sigma_j \sim \text{Cauchy}_+(0,\tau),$$

in which there are four free parameters $(\alpha, \epsilon, \delta, \tau)$ for the Dirichlet prior and Cauchy distribution.

The MERGE method

Lee *et al.* implemented a linear regression method by incorporating multi-omics prior information about genes for all the drugs simultaneously [23]. They used cell line profiles $C_{n\times s} = (c_{ij})$ for inference and driver profiles $D_{s\times k} = (d_{ij})$ for assigning scores for the s genes. Lee *et al.* inferred the regression coefficient matrix $W_{s\times m} = (w_{ij})$ by solving the following optimization problem:

$$\min_{w_{jt},v_\ell} \sum_{j,t=1}^{s,m}\sum_{i=1}^{n}(r_{it} - c_{ij}w_{jt})^2 + \sum_{j,t=1}^{s,m}\left(\lambda - \sum_{\ell=1}^{k} d_{j\ell}v_\ell\right)w_{jt}^2$$
$$- m\sum_{j=1}^{s}\log\left(\lambda - \sum_{\ell=1}^{k} d_{j\ell}v_\ell\right)$$
$$\text{s.t. } \lambda - \sum_{\ell=1}^{k} d_{j\ell}v_\ell > 0, \text{ for } j = 1, 2, \cdots, m.$$

Here, λ is a hyperparameter to be determined on the basis of the data and can be considered as the upper bound of the gene scores $\sum_{\ell=1}^{k} d_{i\ell}v_\ell$. Thus, the second term is used to penalize genes with lower scores.

2.5 Link-prediction-based methods

DualNets

Zhang *et al.* developed this simple but efficient network interpolation method for drug response prediction [24]. They assumed that the response to a known drug d in a new cell line is a weighted combination of the responses of the neighboring cell lines c_i $(i = 1, 2, \cdots, n)$ to drug d as given below:

$$r_{cd} = \sum_{i=1}^{n} w_i r_{c_i d}, \tag{10}$$

where $r_{c_i d}$ is the response level to drug d in cell line c_i and the coefficient w_i is defined as a negative exponential function. Let $s(c, c_i)$ be the gene expression correlation between the two cell lines c and c_i. Then w_i is given as:

$$w_i = \frac{\exp\left(-\frac{(1-s(c,c_i))^2}{2\sigma^2}\right)}{\sum_{j=1}^{n}\exp\left(-\frac{(1-s(c,c_j))^2}{2\sigma^2}\right)}, \tag{11}$$

where σ is the decay rate parameter.

Similarly, the response value of a known cell line to a new drug could be inferred by using drug similarity.

The SRMF method

This method simply assumes that the drug response matrix $R_{n\times m}$ is approximately a product of two low-rank matrices $U_{n\times k}$ (a compression of the cell line profile matrix) and $V_{m\times k}$ (a compression of the drug profile matrix). Wang *et al.* used a cell line similarity matrix S_{cell} and a drug similarity matrix S_{drug} to compute the two low-rank matrices [25] as follows:

$$S_{\text{cell}} \approx U_{n\times k} \times U_{n\times k}^T, \quad S_{\text{drug}} \approx V_{m\times k} \times V_{m\times k}^T.$$

In this way, Wang *et al.* made predictions by solving the following optimization problem:

$$\min_{U,V} \|B \circ (R - UV^T)\|_F^2 + \lambda_1\|S_{\text{cell}} - UU^T\|_F^2$$
$$+ \lambda_2\|S_{\text{drug}} - VV^T\|_F^2 + \lambda_3\left(\|U\|_F^2 + \|V\|_F^2\right) \tag{12}$$

where $B_{n\times m} = (b_{ij})$ is the indicator matrix such that $b_{ij} = 0$ if the drug response value r_{ij} is missing and 1 otherwise, and the operator "$\circ$" is the entry-by-entry matrix product used to mask off the missing entries in R.

2.6 Other methods

Pairwise multiple kernel learning (pairwiseMKL)

Instead of constructing multiple kernels separately for cell lines and drugs, pairwiseMKL computes pairwise kernels by Kronecker multiplying the kernel for each cell line and the kernel for each drug [27].

Kernel rank learning (KRL)

This method focuses on accurately recommending the k most sensitive drugs for each cell line [20]. Therefore it uses discounted cumulative gain, a widely used score for ranking quality, to construct a loss function instead of the square of the loss between observed and predicted values.

3 Assessment of 14 methods

We used the five-fold cross-validation method to evaluate the performance of the following representative methods developed in the past 5 years for the nine metrics in Eqn. (13)–(21):

- CaDRReS,
- cwkBMF,
- DualNets,
- GELnet-P and GELnet-NP),
- KRL,
- KRR,
- MERGE,
- MVLR,
- pairwiseMKL,
- RWEN-left, RWEN-right and RWEN-both,
- SRMF,

together with the two versions of BMTMKL, the best one in the 2014 NCI-DREAM drug sensitivity prediction challenge [11], the two versions of ELnet, RF-g and RF-gs.

All the cell lines were randomly partitioned into five equal parts, four of which were used to train a prediction method and the remaining part was used for test. The predicted responses to each drug in all cell lines were then compared against the observed values by using one of the metrics described below. This test process was repeated 10 times to obtain a robust evaluation.

The assessment results are summarized in Figure 1, Figure S1 and Table S12. Additionally, the one-sided paired Wilcoxon Sign-Rank test (see [33, page 338] for details) was used to assess whether or not a method was significantly better than another under each metric. The heat map of the -$\log_{10}$(p-values) for each test is given in Figure 2 and Figure S2.

The computational time for different methods are provided in Table 2.

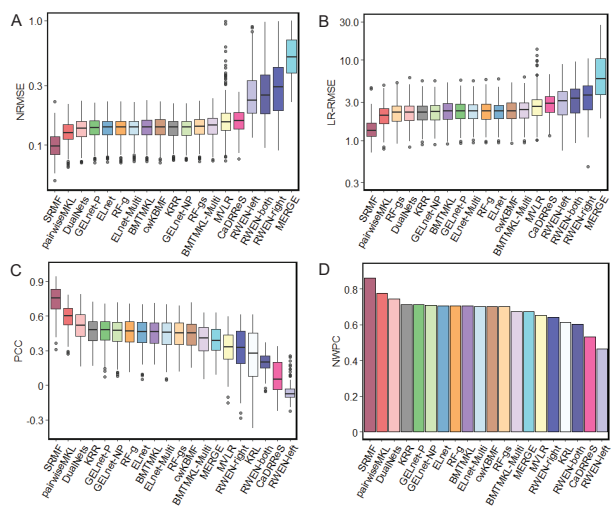

Fig. 1. Prediction performance of 14 methods. Box plots for method performance under the normalized root mean square error (NRMSE) (A), Left-right (LR)-RMSE (B), Pearson correlation coefficient (PCC) (C) and the normalized weighted probabilistic c-index (NWPC) (D) metrics, where the methods are listed from the best (leftmost) to the worst. Each figure shows the average score of 10 repeated cross validation runs across drugs. In (A) and (B), lower values indicate better prediction performance. In (C) and (D), higher values indicate better performance. The kernel rank learning method is not in Panel A and B, as its scores are much larger than those of other methods – its median scores for NRMSE and LR-RMSE are 31 and 309, respectively.

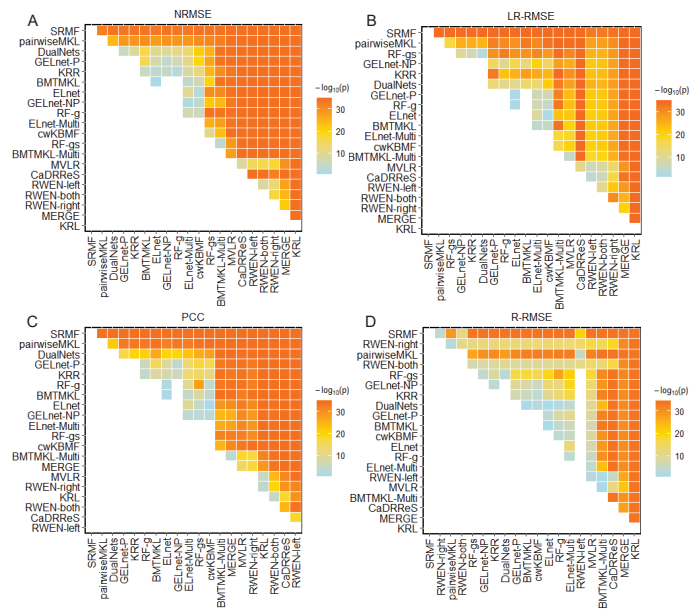

Fig. 2. Prediction performance comparison for each pair of methods. The one-sided paired Wilcoxon Sign-Rank test was conducted based on the data shown in Figure 1 (A–C) and Figure S1, respectively. The value of the entry in the row for Method X and the column for Method Y is -$\log_{10}$(p-value) for the hypothesis that X performs better than Y (i.e., a large value indicates that X significantly outperforms Y.

3.1 Performance of different approaches

The 14 assessed methods can be roughly grouped into three classes: (a) linear regression models (with or without kernel extension), (b) matrix factorization-based Bayesian inference and (c) matrix completion via non-linear optimization methods.

Our results show that the matrix completion approach that is represented by SRMF performed better than the linear regression models, whereas the linear regression models performed better than the different Bayesian inference methods. This has the following implications.

First, the drug response is a complicated non-linear function of the cell line profiles and the drug profiles. When predicting the response to a drug appearing in the training dataset in a new cell line, as one of the best methods, DualNets assumes that the drug response level is linear to the drug profile but non-linear to the cell line profile (Eqn. (10)–(11)), whereas ELnet and GELnet and other regression methods consider the drug response function to be linear to the cell line profiles (Eqn. (1)).

SRMF, another of the best methods, models the drug response matrix as a product of the low-rank proximity of the cell line similarity matrix and the low-rank proximity of the drug similarity matrix, which are both non-linear to the cell line profiles and the drug profiles, respectively.

Second, Bayesian inference is usually powerful for solving machine learning problems. The ineffectiveness of the assessed Bayesian inference methods may have the following causes: the distribution of drug responses may not be as simple as normal distribution; the parameter space in each method may be too large to compute an optimal set of parameter values; the parameter space is so large that the methods may overfit the training dataset.

3.2 Protein–protein interaction networks are informative

GELnet-P and GELnet-NP are the only models that use explicitly cellular protein–protein interaction networks for drug response prediction. They were developed from ELnet through incorporating the (normalized) Laplacian matrix of a protein–protein interaction network into the last regularization term in Eqn. (1). Interestingly, the two methods outperformed ELnet for almost all nine metrics (Figure 1–2, see also Figure S1–S2). This fact may indicate that drug response prediction can benefit from the integration of both protein–protein interaction networks and gene expression profiles with the drug response data.

3.3 Mutation, CNV and methylation profiles of cell lines and drug grouping add little to drug response prediction

We assessed ELnet and BMTMKL by using two sets of auxiliary cell line profiles: (a) only gene expression profile and (b) a combination of gene expression, mutation, CNA and methylation profiles. We did not observe clear improvement in accuracy when multiple cell line profiles were used (Figure 1–2 and Figure S1–S2).

MVLR is the only method that used the information on drug-targeted pathways. It builds a linear regression model for each group of drugs that have the same target pathway. Its predictive performance varies greatly from group to group (Figure 3), where the 23th group that includes three drugs targeting the WNT signaling pathway performs best. It performs well only on about two-third drug groups. Moreover, compared with other methods, MVLR had large variance in performance (Figure S3).

3.4 The performance of a method varies a lot under different measurements

We used nine performance metrics in this study. RMSE and NRMSE measure the performance of a method on all the data points, whereas L-RMSE, R-RMSE and LR-RMSE focus on the performance at the head, tail or both sides of the drug response data distribution. SCC, NDCG and NWPC consider all the data points but measure the rank coherency between the predicted and observed values.

Our study suggests that the performance can be very sensitive to the choice of metrics. One method could significantly outperform another

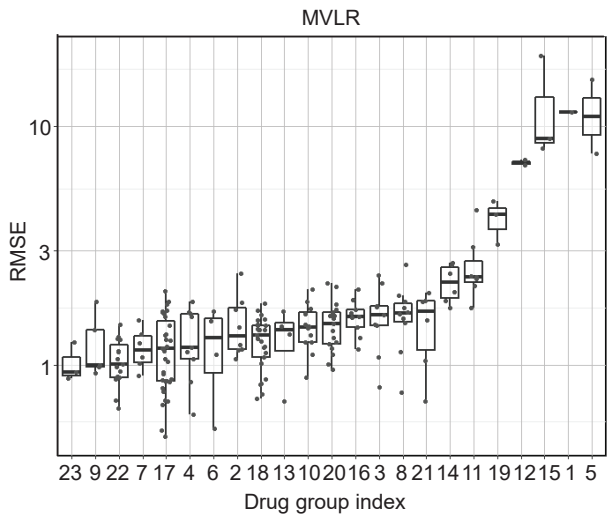

Fig. 3. RMSE distribution of 23 drug groups used in the MVLR method. In MVLR, 23 models were respectively learned for 23 groups of drugs, each of which includes drugs targeting a common pathway. The prediction performances of the last five drug groups are much worse than those of the others.

method for a metric, but perform poorly when comparing other metrics. For instance, cwkBMF significantly outperformed RF-gs, with P-value 10^{-9}, 10^{-10} and 10^{-3} for the RMSE, NRMSE and SCC, respectively; on the other hand, RF-gs significantly outperformed cwkBMF, with P-value 10^{-20}, 10^{-16} and 10^{-28} for the L-RMSE, R-RMSE and LR-RMSE metrics, respectively; there was no significant difference between cwkBMF and RF-gs for the PCC and NDCG metrics.

Table 2. Comparison of computational time for the 14 methods.

Method	Time (min)	Method	Time (min)
Single-task learning		**Multi-task learning**	
KRR	38	CaDRReS	28
ELnet	104	BMTMKL	153
ELnet-Multi	113	BMTMKL-Multi	234
RWEN-right	409	MERGE	260
RWEN-left	609	cwKBMF	387
RWEN-both	1009	MVLR	5839
GELnet-NP	2457		
GELnet-P	6719		
RF-g	6884		
RF-gs	10326		
Link prediction		**Others**	
DualNets	8	pairwiseMKL	239
SRMF	35	KRL	1448

4 Conclusion

We analyzed 14 representative computational methods that have recently been developed for drug response prediction using public cell line–drug response data and nine metrics. The study has provided lessons worth considering ahead of future research.

First, the study reconfirms that the gene expression profiles of cell lines provide crucial information for predicating the responses to a drug of cell lines [10, 11]. How to generate and use the post-treatment and even time point series expression profiles of cell lines is an important issue for drug response predictions in future preclinical and clinical research.

It is commonly accepted that omics data other than gene expression profiles of cell lines often add slight improvements for this purpose [9, 10]. However, the excellent performance of DualNets in our analysis suggests that smart integration of other omics data, particularly drug structure data and protein interaction networks, with gene expression profiles should not be ignored in the study of drug response prediction.

Second, the SRMF method learns drug response functions from the similarity matrices of cell lines and drugs by an iterative approach. This may suggest deep neural network and other AI approaches are likely to be powerful for drug response prediction. Another interesting research question is how to improve the performance of Bayesian inference based methods for drug response prediction.

Third, interpretability and reproducibility are two important factors for drug response models in translation medicine research [11, 34, 35]. For pan-cancer drug response data, different metrics give different pictures of the evaluated methods' performance. In future, computational biologists may focus on cancer-specific drug response data to design robust prediction methods with interpretable features [36].

5 Datasets and methods

5.1 Data

Genomic datasets

We downloaded gene expression, CNA, mutation, and methylation data from the GDSC website (https://www.cancerrxgene.org/downloads, accessed 1 Feb. 2019). The 4 datasets were preprocessed by (i) removing genes with little variation in expression, (ii) extracting both mutations and CNAs for the cancer genes listed in the COSMIC database (https://cancer.sanger.ac.uk/cosmic/download, version 88), and (iii) keeping CpG probes with high variation in methylation levels, respectively. In the end, we obtained:

- A gene expression profile $E_{734\times 8046}$,
- A DNA methylation profile $D_{734\times 8473}$,
- A mutation profile $M_{734\times 636}$, and
- A CNA profile $C_{734\times 694}$

for 734 cancer cell lines (Table S1–S4), of which annotations are provided in Table S5.

Drug response dataset

Drug response data for cancer cell lines were downloaded from the GDSC website, in which the response to a drug in a cell line was measured as the common logarithm of the IC_{50} readout. After removing drugs for which the responses are missing for 80% or more cell lines, we obtained a drug response matrix $R_{734\times 201}$ for 734 tumor cell lines and 201 drugs (Table S6).

Protein–protein interaction network

The 8046 genes in the gene expression profile $E_{734\times 8046}$ encode 7136 proteins. A protein–protein interaction (PPI) network on these proteins (Table S7) was derived by extracting interactions relations between them in the Pathway Commons database (http://www.pathwaycommons.org/, accessed 11 March 2019) [37].

Drug groups and drug similarity matrix

The 201 drugs were divided into 23 groups (Table S8), each containing drugs targeting the same pathway, according to the information available in the GDSC database.

The structural information on the 201 drugs include 1444 1-dimensional, 2-dimensional descriptors and nine types of molecular

fingerprints in PubChem [38] (Table S9). The structural information on the drugs were further processed with the PaDEL-Descriptor [39] to construct a drug similarity network, named $S^{201\times 201}$ (Table S10).

Drug-targeted pathways
We obtained the drug-targeted pathways via the data preprocessing method used in [18]. We downloaded 1329 canonical pathways and 189 oncogenic signature gene sets in C2:CP and C6 collections, respectively, in the MSigDB database (http://software.broadinstitute.org/gsea/msigdb/index.jsp, accessed 1 Feb 2019) [40]. We further extracted only those pathways that contained one or more primary targets of the 201 drugs, leading to 71 drug-targeted pathways. All the genes that were not in any extracted pathways formed a single group named "others". We obtained 72 gene subsets in total (Table S11). Accordingly, a many-to-many relationship among drug groups and gene groups was developed (Table S8).

5.2 Evaluation metrics

In this study, we chose nine widely used metrics to evaluate the 14 methods on the basis of the difference between the observed and predicted responses of all the cell lines to a drug.

Consider a prediction method M. Let $\mathbf{y} = (y_i)_{i=1}^n$ and $\widehat{\mathbf{y}} = (\widehat{y}_i)_{i=1}^n$ denote the vectors of the observed and predicted responses to a drug in n cell lines, respectively.

Root mean square error (RMSE)
The basic RMSE attempts to capture the deviation between the true and predicated responses for M as follows:

$$\mathrm{RMSE}\left(\mathbf{y}, \widehat{\mathbf{y}}\right) = \sqrt{\frac{\sum_i \left(y_i - \widehat{y}_i\right)^2}{n}}. \tag{13}$$

To facilitate a comparison of different drugs or prediction methods with different scales, one may use the normalized RMSE, which is defined as:

$$\mathrm{NRMSE}(\mathbf{y}, \widehat{\mathbf{y}}) = \frac{\mathrm{RMSE}(\mathbf{y}, \widehat{\mathbf{y}})}{\max_i y_i - \min_i y_i}. \tag{14}$$

For a drug, we may be mostly concerned with the sensitive cell lines only or the resistant cell line only. For these purposes, RMSE is further modified as [13]:

$$\text{L-RMSE}(\mathbf{y}, \widehat{\mathbf{y}}) = \sqrt{\frac{\sum_i (y_i - \widehat{y}_i)^2 \cdot \mathbf{1}\left\{y_i \le \ell\right\}}{\sum_i \mathbf{1}\left\{y_i \le \ell\right\}}} \tag{15}$$

$$\text{R-RMSE}(\mathbf{y}, \widehat{\mathbf{y}}) = \sqrt{\frac{\sum_i \left(y_i - \widehat{y}_i\right)^2 \cdot \mathbf{1}\{y_i \ge \gamma\}}{\sum_i \mathbf{1}\{y_i \ge \gamma\}}} \tag{16}$$

$$\text{LR-RMSE}(\mathbf{y}, \widehat{\mathbf{y}}) = \sqrt{\frac{\sum_i \left(y_i - \widehat{y}_i\right)^2 \cdot \mathbf{1}\{(y_i \le \ell)\mathrm{or}(y_i \ge \gamma)\}}{\sum_i \mathbf{1}\{(y_i \le \ell)\mathrm{or}(y_i \ge \gamma)\}}} \tag{17}$$

where $\mathbf{1}\{\}$ denotes the indicator function,

$$\ell = \mathrm{median}(\mathbf{y}) - 1.65 \times \sigma,$$
$$\gamma = \mathrm{median}(\mathbf{y}) + 1.65 \times \sigma,$$
$$\sigma = 1.4826 \times \mathrm{median}_{1\le i\le n}(|y_i - \mathrm{median}(\mathbf{y})|).$$

Note that ℓ and γ are the left and right cutoff points, respectively, and are computed by the lower and upper limit of the 90% confidence interval of the location parameter for Gaussian distributed data with standard deviation σ.

Pearson correlation coefficient (PCC)
Let $\mu(X)$ denote the mean of a random variable. The PCC of $\mathbf{y}$ and $\widehat{\mathbf{y}}$ is defined as:

$$\mathrm{PCC}(\mathbf{y}, \widehat{\mathbf{y}}) = \frac{\sum_i \left(\widehat{y}_i - \mu(\widehat{\mathbf{y}})\right)\left(y_i - \mu(\mathbf{y})\right)}{\sqrt{\sum_i \left(\widehat{y}_i - \mu(\widehat{\mathbf{y}})\right)^2}\sqrt{\sum_i \left(y_i - \mu(\mathbf{y})\right)^2}} \tag{18}$$

Spearman correlation coefficient (SCC)
Assume that $\mathbf{y} = (y_i)_{i=1}^n$ is sorted into $y_{i_1} \ge y_{i_2} \ge \cdots \ge y_{i_n}$ in decreasing order. The rank of the response y_{i_k}, $r(y_{i_k})$, is defined to be k. The integral vector $\mathbf{r(y)} = (r(y_i))_{i=1}^n$ is called the rank vector of $\mathbf{y}$. The rank vector of $\widehat{\mathbf{y}}$ can be defined similarly. The SCC of $\mathbf{y}$ and $\widehat{\mathbf{y}}$ is then defined as:

$$\mathrm{SCC}\big(\mathbf{y}, \widehat{\mathbf{y}}\big) = \mathrm{PCC}\big(\mathbf{r(y)}, \mathbf{r}(\widehat{\mathbf{y}})\big). \tag{19}$$

Normalized discounted cumulative gain (NDCG)
This is a widely used score for ranking quality [21]. The discounted cumulative gain (DCG) for the response prediction made by M is defined as:

$$\mathrm{DCG}\big(\mathbf{s}, \widehat{\mathbf{s}}\big) = \sum_{i=1}^n \frac{2^{s_i} - 1}{\log_2\left(r(\widehat{s}_i) + 1\right)},$$

in which the score vector $\mathbf{s} = (s_i)_{i=1}^n = -\mathbf{y}$ and $\widehat{s}_i = -\widehat{y}_i$. The numerator gives bigger weight to higher sensitivity score whereas the denominator gives preference to higher ranks.

The nomalized DCG is defined as:

$$\mathrm{NDCG}\big(\mathbf{s}, \widehat{\mathbf{s}}\big) = \frac{\mathrm{DCG}(\mathbf{s}, \widehat{\mathbf{s}})}{\mathrm{DCG}(\mathbf{s}, \mathbf{s})} \tag{20}$$

Normalized weighted probabilistic c-index (NWPC)
In the 2014 NCI-DREAM drug sensitivity prediction challenge, NWPC was used to evaluate the performance of the submitted methods. It measures the agreement between a predicted ranked list and the true list, defined as follows:

$$\text{pc-index}\big(\mathbf{s}, \mathbf{r}(\widehat{\mathbf{s}})\big) = \frac{2}{n(n-1)} \sum_{i<j} \mathrm{hp}\big(s_i, s_j, r(\widehat{s}_i), r(\widehat{s}_j), \sigma(\mathbf{s})\big),$$

where $\sigma(X)$ is the standard deviation of a random variable X and the function $\mathrm{hp}\big(s_i, s_j, r(\widehat{s}_i), r(\widehat{s}_j), \sigma(\mathbf{s})\big)$ is defined as

$$\mathrm{hp}\big(s_i, s_j, r(\widehat{s}_i), r(\widehat{s}_j), \sigma(\mathbf{s})\big) = \begin{cases} \frac{1}{2}\Big(1 + \mathrm{erf}\big(\frac{s_i - s_j}{2\sigma(\mathbf{s})}\big)\Big), & r(\widehat{s}_i) < r(\widehat{s}_j), \\ 0.5, & r(\widehat{s}_i) = r(\widehat{s}_j), \\ \frac{1}{2}\Big(1 + \mathrm{erf}\big(\frac{s_j - s_i}{2\sigma(\mathbf{s})}\big)\Big), & r(\widehat{s}_i) > r(\widehat{s}_j), \end{cases}$$

and $\mathrm{erf}(a) = \frac{2}{\sqrt{\pi}} \int_0^a e^{-t^2} dt$.

For a method, the pc-index is calculated separately for each drug d, denoted $\mathrm{pc}_d = \text{pc-index}\Big(\mathbf{s^{(d)}}, \mathbf{r}(\widehat{\mathbf{s}}^{\mathbf{(d)}})\Big)$ where $\mathbf{s^{(d)}}, \widehat{\mathbf{s}}^{\mathbf{(d)}}$ are the true and predicted score vectors for the drug d, respectively. The final score of a method M is calculated as the weighted average of the pc-index scores across all the drugs (i.e. the wpc-index) as follows:

$$\text{wpc-index}(M) = \frac{\sum_d w_d \cdot \mathrm{pc}_d}{\sum_d w_d}$$

The weight w_d is calculated from the empirical null distribution of pc_d. For each drug d, (i) produce a random ranking of n items, denoted $\tilde{\mathbf{r}}^{\mathbf{(d)}}$, and calculate the pc-index$\Big(\mathbf{s^{(d)}}, \tilde{\mathbf{r}}^{\mathbf{(d)}}\Big)$; (ii) repeat this procedure

10,000 times to construct an empirical null distribution with a median of μ_d and a standard deviation of σ_d; (iii) compute the pc-index with the standard ranking $\mathbf{r}(\mathbf{s^{(d)}})$, $\mathrm{pc}_d^* = \text{pc-index}\Big(\mathbf{s^{(d)}}, \mathbf{r}(\mathbf{s^{(d)}})\Big)$; (iv) drug weight $w_d = \frac{\mathrm{pc}_d^* - \mu_d}{\sigma_d}$.

In order to scale the range of the wpc-index into $[0, 1]$, use the transformation:

$$\mathrm{NWPC} = \frac{\mathrm{wpc} - \mathrm{wpc}_{\min}}{\mathrm{wpc}_{\max} - \mathrm{wpc}_{\min}} \tag{21}$$

where $\mathrm{wpc}_{\max}, \mathrm{wpc}_{\min}$ are obtained from the real ranking vector $\mathbf{r}(\mathbf{s^{(d)}})$ and its inverse, respectively, for each drug when computing pc_d.

Unlike the previous eight metrics, the wpc-index measures the overall performance for all the predicted drugs. Therefore, only one value is given for each method.

Lastly, we point out that a smaller value indicates good performance for each of the first five metrics, whereas a value close to 1 suggests good performance for the last four metrics.

Key Points

- The computational methods for drug response prediction can be grouped into three classes: (a) linear regression models (with or without kernel extension), (b) matrix factorization-based Bayesian inference methods and (c) matrix completion via nonlinear optimization methods. The matrix completion approach outperformed the linear regression models, whereas the linear regression models outperformed the different Bayesian inference methods.
- It remains unclear how to improve the performance of Bayesian inference based methods for drug response prediction.
- The SRMF with the best performance learns drug response functions from the similarity matrices of cell lines and drugs by an iterative approach, revealing that the drug response functions of cell lines are non-linear and complicated. This may suggest deep neural network and other AI approaches are likely to be able to achieve high accuracy for drug response prediction.
- The mutation, CNV and methylation profiles of cell lines and drug grouping seem to add little to drug response prediction when they are used with the existing methods. It is worth studying how to use wisely these omics data.
- The performance of a method is sensitive to the choice of measurements. One method could significantly outperform another method for one metric, but performs poorly when comparing other metrics.
- A large set of cell line profiles and drug profiles were curated that can be used for benchmarking for future study into drug response prediction.

Supplementary Data

Supplementary data are available online at https://github.com/Jinyu2019/Supplemental-Data-for-Drug-Response-Prediction.git.

Funding

This work was supported by Singapore National Research Fund [Grant: NRF2016NRF-NSFC001-026].

References

1. Dugger SA, Platt A, Goldstein DB. Drug development in the era of precision medicine. Nat Rev Drug Discov. 2018;17(3):183–196.
2. Roses AD. Pharmacogenetics in drug discovery and development: a translational perspective. Nat Rev Drug Discov. 2008;7(10):807–817.
3. Alley MC, Scudiero DA, Monks A, et al. Feasibility of drug screening with panels of human tumor cell lines using a microculture tetrazolium assay. Cancer Res. 1988;48(3):589–601.
4. Weinstein JN. Drug discovery: cell lines battle cancer. Nature. 2012;483(7391):544–545.
5. Shoemaker RH. The NCI-60 human tumour cell line anticancer drug screen. Nat Rev Cancer. 2006;6(10):813–823.
6. Barretina J, Caponigro G, Stransky N, et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012;483(7391):603–607.
7. Garnett MJ, Edelman EJ, Heidorn SJ, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature. 2012;483(7391):570–575.
8. Azuaje F. Computational models for predicting drug responses in cancer research. Brief Bioinform. 2016;18(5):820–829.
9. Ali M, Aittokallio T. Machine learning and feature selection for drug response prediction in precision oncology applications. Biophys Rev. 2019;11(1):31–39.
10. Jang IS, Neto EC, Guinney J, et al. Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data. Pac Symp Biocomput. 2014;p. 63–74.
11. Costello JC, Heiser LM, Georgii E, et al. A community effort to assess and improve drug sensitivity prediction algorithms. Nat Biotechnol. 2014;32(12):1202–1212.
12. Sokolov A, Carlin DE, Paull EO, et al. Pathway-based genomics prediction using generalized elastic net. PLoS Comput Biol. 2016;12(3):e1004790.
13. Basu A, Mitra R, Liu H, et al. RWEN: response-weighted elastic net for prediction of chemosensitivity of cancer cell lines. Bioinformatics. 2018;34(19):3332–3339.
14. Rahman R, Dhruba SR, Ghosh S, et al. Functional random forest with applications in dose-response predictions. Sci Rep. 2019;9(1):1628.
15. Li Q, Shi R, Liang F. Drug sensitivity prediction with high-dimensional mixture regression. PLoS One. 2019;14(2):e0212108.
16. Su R, Liu X, Xiao G, et al. Meta-GDBP: a high-level stacked regression model to improve anticancer drug response prediction. Brief Bioinform. 2019;bbz022.
17. Yuan H, Paskov I, Paskov H, et al. Multitask learning improves prediction of cancer drug sensitivity. Sci Rep. 2016;6:31619.
18. Ammad-ud-din M, Khan SA, Malani D, et al. Drug response prediction by inferring pathway-response associations with kernelized Bayesian matrix factorization. Bioinformatics. 2016;32(17):455–463.
19. Ammad-ud-din M, Khan SA, Wennerberg K, et al. Systematic identification of feature combinations for predicting drug response with Bayesian multi-view multi-task linear regression. Bioinformatics. 2017;33(14):359–368.
20. He X, Folkman L, Borgwardt K, et al. Kernelized rank learning for personalized drug recommendation. Bioinformatics. 2018;34(16):2808–2816.
21. Supahvilai C, Bertrand D, Nagarajan N. Predicting cancer drug response using a recommender system. Bioinformatics. 2018;34(22):3907–3914.
22. Yang M, Simm J, Lam CC, et al. Linking drug target and pathway activation for effective therapy using multi-task learning. Sci Rep. 2018;8(1):8322.

23. Lee SI, Celik S, Logsdon BA, et al. A machine learning approach to integrate big data for precision medicine in acute myeloid leukemia. Nat Commun. 2018;9(1):42.
24. Zhang N, Wang H, Fang Y, et al. Predicting anticancer drug responses using a dual-layer integrated cell line-drug network model. PLoS Comput Biol. 2015;11(9):e1004498.
25. Wang L, Li X, Zhang L, et al. Improved anticancer drug response prediction in cell lines using matrix factorization with similarity regularization. BMC Cancer. 2017;17(1):513.
26. Chang Y, Park H, Yang HJ, et al. Cancer drug response profile scan (CDRscan): a deep learning model that predicts drug effectiveness from cancer genomic signature. Sci Rep. 2018;8(1):8857.
27. Cichonska A, Pahikkala T, Szedmak S, et al. Learning with multiple pairwise kernels for drug bioactivity prediction. Bioinformatics. 2018;34(13):509–518.
28. Neubig RR, Spedding M, Kenakin T, et al. International union of pharmacology committee on receptor nomenclature and drug classification. XXXVIII. Update on terms and symbols in quantitative pharmacology. Pharmacol Rev. 2003;55(4):597–606.
29. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Statist Soc B. 2005;67(2):301–320.
30. Murphy KP. Machine learning: a probabilistic perspective. MIT Press, Cambridge, MA, USA. 2012;p. 492–493.
31. Breiman L. Random forests. Machine Learning. 2001;45(1):5–32.
32. Riddick G, Song H, Ahn S, et al. Predicting in vitro drug sensitivity using random forests. Bioinformatics. 2010;27(2):220–224.
33. Rosner B. Fundamentals of biostatistics. Boston: Brooks/Cole, Cengage Learning, USA. 2011;.
34. Fey D, Halasz M, Dreidax D, et al. Signaling pathway models as biomarkers: patient-specific simulations of JNK activity predict the survival of neuroblastoma patients. Sci Signal. 2015;8(408):ra130.
35. Zhang F, Wang M, Xi J, et al. A novel heterogeneous network-based method for drug response prediction in cancer cell lines. Sci Rep. 2018;8(1):3355.
36. Fowles JS, Brown KC, Hess AM, et al. Intra-and interspecies gene expression models for predicting drug response in canine osteosarcoma. BMC Bioinformatics. 2016;17(1):93.
37. Cerami EG, Gross BE, Demir E, et al. Pathway Commons, a web resource for biological pathway data. Nucleic Acids Res. 2010;39(Suppl 1):D685–D690.
38. Gindulyte A, Shoemaker BA, Yu B, et al. PubChem 2019 update: improved access to chemical data. Nucleic Acids Res. 2018;47(D1):D1102–D1109.
39. Yap CW. PaDEL-Descriptor: an open source software to calculate molecular descriptors and fingerprints. J Comput Chem. 2011;32(7):1466–1474.
40. Liberzon A, Birger C, Thorvaldsdottir H, et al. The molecular signatures database (MSigDB) hallmark gene set collection. Cell Syst. 2015;1(6):417–425.