Improving Drug Response Prediction Using Dual Similarity Regularization

Ali Reza Ebadi
Islamic Azad University Sanandaj Branch
Ali Soleimani (a.soleimani.uni@iaumalard.ac.ir)
Islamic Azad University of Malard
Abdulbaghi Ghaderzadeh
Islamic Azad University Sanandaj Branch

Research Article

Keywords: Personal Medicine, Matrix Factorization, Anticancer Drug Response Prediction, Therapeutic

Posted Date: June 28th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-50015/v2

Improving drug response prediction using Dual similarity Regularization

ALI REZA EBADI[1], *ALI SOLEIMANI[2], ABDULBAGHI GHADERZADEH[3]

[1]*Department of Computer Engineering, Sanandaj Branch, Islamic Azad University, Sanandaj, Iran*
E-mail: ebadi.phdstudent@iausdj.ac.ir,
**2Department of Computer Engineering, College of Technical and Engineering, Malard Branch, Azad University, Tehran, Iran.*
**E-mail:* a.soleimani.uni@iaumalard.ac.ir
[3]*Department of Computer Engineering, Sanandaj Branch, Islamic Azad University, Sanandaj, Iran*
E-mail: b.ghaderzadeh@iausdj.ac.ir

Abstract
Anti-cancer medicine for a particular patient has been a personal medical goal. Many computational models have been proposed by researchers to predict drug response. But predictive accuracy still remains a challenge. Base on this concept which "Similar cells have similar responses to drugs", we developed the basic method of matrix factorization method by adding fines to similarity. So that the distance of latent factors to two cell lines or (drug) should be inversely related to similarity. This means that two similar drugs or similar cell lines should have a short distance, whereas two similar cell lines or non-similar drugs should have a large gap with their latent factors.
We proposed a Dual similarity-regularized matrix factorization (DSRMF) model, then generated new data for drug similarity from the two-dimensional three-dimensional chemical structure, which were obtained from the CCLE and GDSC databases. In this research, by using the proposed model, and generating new drug similarity data we achieved the average Pearson correlation coefficient (PCC) about 0.96, and average mean square error (RMSE) Root about 0.30, between the observed value and the predicted value for the cell line response to the drug.
Our analysis in this research showed, using heterogeneous data, has better results, and can be obtained with the proposed model, using other panels' cancer cell lines, to calculate similarity between cells. Also, by imposing more restrictions on the similarity between cells, we were able to achieve more accurate prediction for the response of the cell line to the anticancer drug.
Keywords: Personal Medicine, Matrix Factorization, Anticancer Drug Response Prediction, Therapeutic

Background

Personal medicine is a growing as a medical and therapeutic strategy. And it has been able to reached significant achievement as a new solution in the field of treatment of patients. Also, the patient's own genomic and molecular information is used to precisely personalize the patient's response to the drug. As patients respond differently to a different medical treatment, in order to the patient can be treated with the least side effects and the most effective drug treatment, the researchers are developing specific personalized medical solutions to a particular disease. We deal with cancer, a disease that is causing deaths in the world with high complexity in treatment, so researchers are using precision personal medicine to detect cancer. Computational methods for combining genomic profiles and cancer cell lines can be used to create and improve precision and nasal cancer response to anticancer drugs. Researches about the sensitivity of anticancer drugs to cells, are divided into two categories. A number of focusing studies have identified and discovered biomarkers that play an important role in the sensitivity of the drug to the cell. In [1] Using molecular level variations to predict the sensitivity of an anticancer drug to cells. In [2-6] using Elastic net regularization and random forests routines to identify genomic biomarkers, and to predict drug sensitivity to cells Cancer has been used but, in contrast, a large number of researches have been conducted to predict drug susceptibility to gene expression levels. The kernelized Bayesian matrix factorization (KBMF) method was used in [7]. In [8], Weighted graph regularized matrix factorization (WGRMF) algorithm has been proposed to predict the sensitivity of cancer drugs to cell lines. In this way, the likeness between similarity of drugs and the similarity of the cells closest to the neighborhood are alike. In this method the GDSC database was used. In [9], three different deep learning algorithms are compared with random forests and nearest neighbor (knn). And the result shows the combination of RF GE and KNN Residual works better. This method has been suggested to achieve the best Outliers deletion performance, reduce data size, and limited data usage. In [10], Private linear regression model has been proposed to predict the sensitivity of cancer drugs to cell lines. In [11], A Bayesian algorithm that combines kernelized multitasking and dimensionality reduction, called kernelized Bayesian multitasking (KBMT), is used to share a subspace and data in this space. The commonalities between these subsets capture data, are used to learn and improve forecasting performance.[12] have used Multitasking learning on CCLE, CTD4 and NCI60 databases to improve prediction. For their analysis, in [13], a network-based approach called

GloNetDRP was used. In this method, a heterogeneous network between the similarity of drugs like drugs and the similarity of drugs to drugs was used by using a network-based method called GloNetDRP. And the response of responsiveness is used not only on the basis of the neighborhood, but also from the similarity with other drugs in heterogeneous network. In [14], the CCLE and GDSC databases were used. A model based on the similarity of drugs was proposed, so the drug sensitivity profile is given to the new drug if it is structurally similar. In [15], a network based classifier (NBC) method is used to measure the sensitivity of different types of drugs to different types of tumors, as well as a list of apoptotic genes and clinical dose-related predictions were used in this study. The CCLE and GDSC databases were used in [16]. The domain adaptation method called PRECISE were used to collect and predict the shared information that exists between human tumors and preclinical models. In [17], the Support Vector Machine (SVM) method and the recursive feature selection on the CCLE database were combined. Dong divided the cell lines into two sensitive and persistent subsets. Based on the drug response rate, drug responses were used to select features. The SVM model were used in [18]. Consensus p-Median clustering method were used to infer drug response to cell lines on a tumor. In [19], a set of heterogeneous genes that are important for drug response were selected, then Bayesian network model and genomic profiles were used to predict cell line response to drugs. In this study the NCI60 database were used.

Similarity-regularized matrix factorization (SRMF) were used to predict the response to anticancer drugs by cell lines. In this method, the structural similarity of drugs and the level of gene expression of cell lines were used. One of the problems is that the used data were not validated. And all dimensional properties of the two-dimensional chemical structure of the drugs were not used in order to obtain the drug similarity matrix. And also there were an overflow problem and accurate prediction. Working with CCLE database, we used GDSC and NCBI PubChem Repository. In the related work, available drugs were in SDF format. The number of anticancer drugs with a specific chemical structure were less and limited. In this work, the data has been synthesized and the two-dimensional chemical structure has been applied to generate the similarity data between the drugs. Dual similarity-regularized matrix factorization (DSRMF) is also used to improve the prediction accuracy of a computational method. In this study, we used Pearson correlation coefficient (PCC), root mean square error (RMSE) evaluation metrics to evaluate the results. We also deduced the response values of drugs that are missing in the GDSC data. We found that the proposed method has lower RMSE and higher PCC than previous methods.

Methods

In this research, we first used the Genomics of Drug Sensitivity in Cancer project, which its release -5.0 has 790 cell lines and 135 drugs. Then, we used the Genomics of Drug Sensitivity in Cancer project, which its release -5.0 has 790 cell lines and 135 drugs. By initial preprocessing, we found that some drugs were not chemically specific, and their PubChem CIDs and PubMed SDF file were not available in the Cancer Cell Line Encyclopedia (CCLE), so the number of drugs was reduced to 97 for subsequent calculations, and the number of cell lines was reduced to 604 by removing duplicates. In vitro, measuring the response of the drug to an IC50 value indicates that the lower the IC50 value means the greater the sensitivity of the cell line to the drug. The cell lines are identified by genomic features, and the drugs are also encoded by chemical structure in SDF format, and are available in the NCBI PubChem Repository. With using PaDEL software [20], to determine the PubChem Description of the drug fingerprint, we also incorporate two-dimensional and three-determine chemical structures into the computation. And ultimately, using new SDF formulation of drugs, a drug-like matrix was obtained. The Primary Product Response Matrix contains 58588 entries, some of entries, contain missing data. To obtain the similarity matrix of the cell line based on the expression profile of the gene, a Pearson correlation coefficient was obtained between the cell line pairs. And To obtain the similarity matrix between drugs based on the similarity of fingerprints, and the chemical structure between drug pairs, a Jaccard coefficient were obtained.

Problem formulation

In this paper, we use the matrix factorization algorithm to predict the response to the anticancer drug. Similar framework has been adopted to predict drug targets [21]. First, we mapped the drug m and n cells into a shared low dimensionality K latent space in which the properties of each cell line drug are represented by the latent coordinates $U_i$and V_j respectively. We call the drug response watermarks in this space Y. The purpose is to obtain an approximation of the response value of d_j to the cell line c_i, by determining their corresponding latent coordinates. Our objective function here is as follows

$$\min_{u,v} \|W.(Y - UV^T\|_F^2 \qquad (1)$$

Where W is the matrix weigh $W_{ij} = 1$ If Y_{ij} has a certain amount of drug response. Otherwise $W_{ij} = 0$ and $\|.\|_F$ is frobenius norm. And to avoid overflow of U, V training data, L2 (Tikhonov) regularization is added to the latent variables U, V as fines.

$$\min_{u,v} \|W.(Y - UV^T\|_F^2 + \lambda_l(\|U\|_F^2 + \|V\|_F^2) +$$

$$\lambda_d(\|S_d - UU^T\|_F^2) + \lambda_c(\|S_c - VV^T\|_F^2 \qquad (2)$$

According to the work done by [16], Similar drugs and similar cell lines have similar responses to drugs

Lin Wang et al research, helped to enhance the prediction accuracy of drug response by minimizing the objective functions of $\|S_d - UU^T\|_F^2$ and $\|S_C - VV^T\|_F^T$

The major contribution in this research is adding more severe, and more similarity regularization limitation, that the distance of latent factors to the two cell lines or drug should be inversely correlated by analogy. It means that two similar drugs or similar cell lines should have a small distance, whereas two identical cell lines or non-identical drugs should have a large distance with their latent factors. So by using this idea and applying it to the objective function, we were able to improve the correct prediction of drug response compared to previous methods.

We used the following formula to obtain similarity regularization for cell lines:

$$SimReg_c = \frac{1}{8}\sum_{i=1}^{m}\sum_{j=1}^{m}((S_c(i,j) - e^{\gamma\|U_i - U_j\|^2})^2 \qquad (3)$$

And to obtain similarity regularization for drugs:

$\boldsymbol{\gamma}$ is to control the radius of the Gaussian function

$$SimReg_d = \frac{1}{8}\sum_{i=1}^{n}\sum_{j=1}^{n}(S_d(i,j) - e^{\gamma\|V_i - V_j\|^2})^2 \qquad (4)$$

Finally the constraints are added to the main objective function

$$\min_{u,v}\|W.(Y - UV^T\|_F^2 + \lambda_l(\|U\|_F^2 + \|V\|_F^2) + \lambda_d\|S_d - UU^T\|_F^2 + \lambda_c\|S_c - VV^T\|_F^2 + \alpha SimReg_c + \beta SimReg_d \qquad (5)$$

α and β adjust the ratio of similarity regularization term in order on cell lines and drugs. In this research, used SRMF algorithm In[19], were used. To obtain the final drug response, the results of DSRMF were compared with previous works. The results showed improvement on prediction accuracy. In this research we used Python (3.7) 64-bit for our implementation.

Measurements of prediction performance

To measure the efficiency of the method, the Root mean squared error (RMSE) and Pearson correlation coefficient (PCC) are used for each drug in [19]. The RMSE is computed as follow:

$$\text{RMSE}=\sqrt{\frac{\sum_c(R(D,C)-\hat{R}(D,C))^2}{n}} \qquad (6)$$

The value of n refers to the number of cell lines that respond to a particular drug. Also R (D, C) and $\hat{R}$ (D, C) refer to the observed and predicted response values of the cell lines to the drugs.

The hyper parameter settings in the machine learning algorithm used in this paper, which is based on the matrix factorization method, are as follows: First, the matrix data of the response lines of the cell lines to the anticancer drugs were collected from the CCLE and GDSC databases. By dividing the values of the data in the matrix to the maximum absolute value, they are converted to values between range [-1, 1]. Therefore, the regularization parameters λ_c, $\lambda_d$and λ_l were tested in the intervals $[2^{-6}... 2^2]$ respectively. The point is, in [19] paper, which has achieved the best result at λ_d=0 actually, the chemical structure of drugs was ignored. But in our paper, accuracy was higher in DSRMF method at $\lambda_d = 0.0000001$ by applying chemical structure control settings to the loss function. Another difference in this paper is; producing new cell line drug response data, by incorporating two-dimensional and three-dimensional chemical structures in drug similarity and response, to previous drug similarity in previous work. The results of applying the SRMF model to the new production data is shown in this study. The value of k that is the low dimensionality K of the GDSC database, was set as 44, the Iteration training was set as 20. In this research, we compared performance of the SRMF model with the proposed DSRMF model. The average PCC and average RMSE of 100 replicates was used. And the new generated data from the CCLE and GDSC databases, was used in the DSRMF model averaged PCC_S / R (Pearson coefficient correlation). The sensitivity of the cell line to the anticancer drug between the observed and the predicted was approximately 96.0. Comparing with the previous SRMF model in [19], which at the best was 0.71 on its data by λ_d= 0, was about 0.25. For SRMF model on data was about 0.95

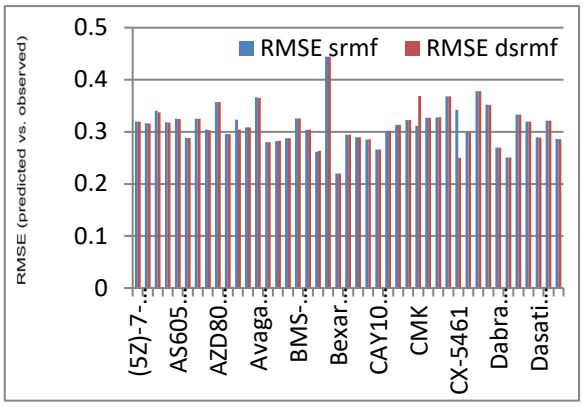

Figure 1. Comparison the mean of PCC_S / R among the samples of drugs between the proposed method (dsrmf) and the previous method

Figure 1. compares the mean PCC_S / R between a samples of drugs, and averaged RMSE_S /R (root mean square error) between observed and predicted cell lines response to drugs in DSRMF model was 0.30, and in SRMF model was about 0.31 for data in this Research. Comparing to the previous work of [16], which had an average RMSE_S / R value of 0.78 on its own data conditions, the proposed method was about 0.47 better than the previous models in RF [22] and KBMF [23]. Also, in Figure 2, we see a compared the proposed method and the previous method for randomly selected anticancer drugs. And in Figure 3, we compared between the previous method and the proposed method by adding a number of algorithm iterations. And Table 1 shows the predictions details and the comparisons of results for different methods.

Table 1. The comparison results of different methods obtained under the10-fold cross validation on GDS dataset

Methods	Drug-averaged PCC_S/R	Drug-averaged RMSE_S/R
DSRMF	96.0	0.31
SRMF	0.71	0.62
KBMF	0.59	0.49
DLN	0.55	0.44
RF	0.50	0.40

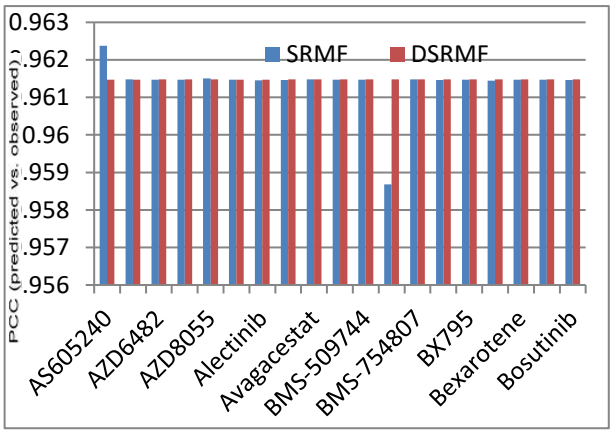

Figure 2. Comparison the average RMSE_S / R among the samples of other drugs between the proposed method (DSRMF) and the previous method.

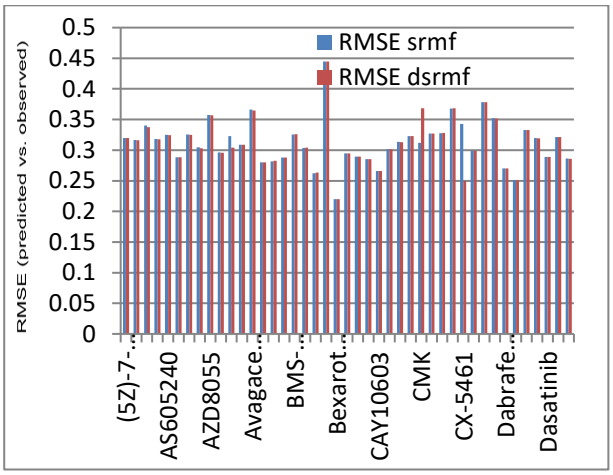

Figure 3. Comparison the average RMSE_S / R between a samples of cell lines drugs between the proposed method (dsrmf) and the previous method.

Figure 4. And Figure 5. The average cell line response to drugs, which is the same IC50 value is evaluated for SRMF and DSRMF (proposed method) and their difference with the observed values. The DSRMF model is more accurate than the SRMF model

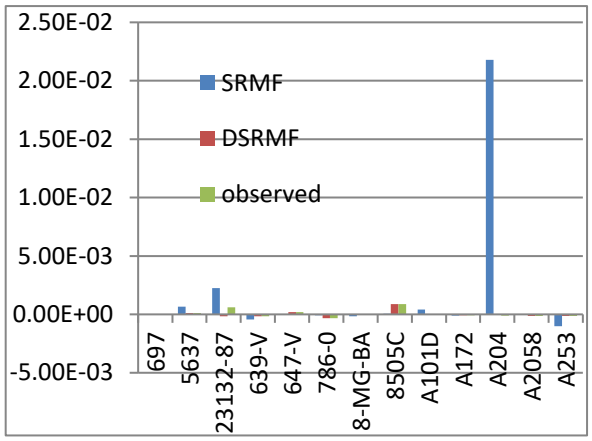

Figure 4. The average cell line response to drugs which is the same IC50 value

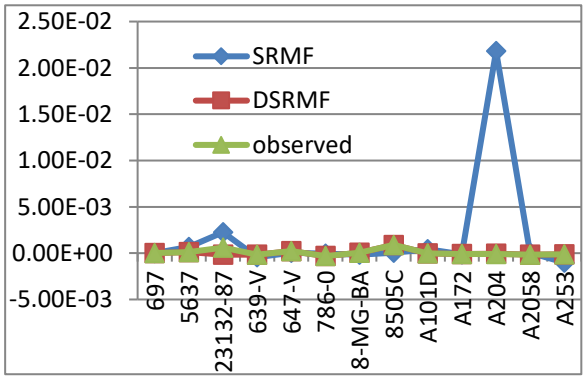

Figure 5. The average cell line response to drugs which is the same IC50 value

Discussion

In this study, we used the CCLE and GDSC databases to compare the similarity between cell lines and the drug webs. To get the resemblance between cells, we used gene expression profile. Whereas, in order to achieve similarity, Cell lines panels could be used for other cancers such as DNA methylation, reverse-phase protein array and, micro RNA expression. And also, by using the genomic properties of cell lines, and other properties such as copy number variation and pathways, and somatic mutation, in the proposed method, DSRMF, could have better results in predicting response to anticancer drugs. It could also be used in other areas of predicting and modeling.

Conclusions

In this study, we used the CCLE and GDSC databases to compare the similarity between cell lines and the drug webs. To get the resemblance between cells, we used gene expression profile. Whereas, in order to achieve similarity, Cell lines panels could be used for other cancers such as DNA methylation, reverse-phase protein array and, micro RNA expression. And also, by using the genomic properties of cell lines, and other properties such as copy number variation and pathways, and somatic mutation, in the proposed method, DSRMF, could have better results in predicting response to anticancer drugs. It could also be used in other areas of predicting and modeling. In this research we developed a Dual similarity-regularized matrix factorization (DSRMF) model to predict response to anticancer drug as measured by IC50 criteria for cell line sensitivity or resistance. We also utilized the CCLE and GDSC databases. And the production of new drug similarity data which incorporated two-dimensional and three-dimensional chemical structures of drugs, and other used properties in previous articles to improve efficiency.

Abbreviation:

Not applicable

Declarations:

- Ethics approval and consent to participate
  This article does not contain any studies with human participants or animals performed by any of the authors.
- Consent to publish
  Agree to be published
- Availability of data and materials
  Used Cell line data in this article was strived from https://portals.broadinstitute.org/ccle
  Used Drug data in this article was strived from https://www.cancerrxgene.org/downloads/anova, and https://pubchem.ncbi.nlm.nih.gov/, also data can be sent emailed as needed.
- Competing interests
  All Authors declares that they have no conflict of interest.
- Funding
  No funding was used.
- Authors' Contributions
  Ali Reza Ebadi is the first author, Ali Soleimani is corresponding author, and Abdulbaghi Ghaderzadeh is the third author.
- Acknowledgements
  -

References

1. P.Geeleher, N.Cox, and R. Stephanie Huang, “Cancer biomarker discovery is improved by accounting for variability in general levels of drug sensitivity in pre-clinical models,” Genome biology, vol.17, 2016, pp. 190-201
2. F.Iorio, Theo A. Knijnenburg,Daniel, J. Vis, G. Bignell et al, “A landscape of pharmacogenomic interactions in cancer,” Cell , vol.166, 2016, pp. 740-754
3. J. Barretina, G. Caponigro, N.Stransky, J.Barretina et al,“The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity,” Nature, vol.13, 2012 , pp. 603–607
4. M.Garnett, E.Edelman, S.Ellis,et al ,“Systematic identification of genomic markers of drug sensitivity in cancer cells,” Nature , vol.483, 2012, pp. 570–575
5. L.C Stetson, T.Pearl, Y.Chen and J.Barnholtz-Sloan,“ Computational identification of multi-omic correlates of anticancer therapeutic response,” BMC Genomics, vol.15, 2014, pp.1-8
6. A.Basu, R.Mitra,H.Liu, S. Schreiber and P. Clemons ,“RWEN: response-weighted elastic net for prediction of chemosensitivity of cancer cell lines,” Bioinformatics,vol. 34, 2018, pp. 3332-3339.
7. A.Muhammad,E.Georgii,M.Gonen,T.Laitinen,O.Kallioniemi,K.Wennerberg,A.Poso,and S.Kaski,“Integrative and personalized QSAR analysis in cancer by kernelized Bayesian matrix factorization,” Journal of chemical information and modeling,vol. 54, 2014, pp. 2347-2359.
8. N.Guan,Y.Zhao,C.Wang,J.QiangLi,X.Chen,and,X.Piao,“Anticancer drug response prediction in cell lines using weighted graph regularized matrix factorization,” Molecular Therapy-Nucleic Acids,vol. 17, 2019, pp. 164-174.
9. K.Matlock,C.DeNiz,R.Rahman,S.Ghosh and R.Pal,“Investigation of model stacking for drug sensitivity prediction,” BMC bioinformatics ,vol.19, 2018, pp.1-8
10. A.Honkela, M.Das, A.Nieminen, O.Dikmen and S.Kask "Efficient differentially private learning improves drug sensitivity prediction." Biology direct, vol.13, 2018, pp.1-12
11. M.Gonen, and A.Margolin,“Drug susceptibility prediction against a panel of drugs using kernelized Bayesian multitask learning ,” Bioinformatics,vol. 30, 2014, pp.556-563
12. H.Yuan, I.Paskov,H.Paskov, A. Gonzále and C. Leslie ,“Multitask learning improves prediction of cancer drug sensitivity,” Scientific reports vol.6, 2016, pp.1-11
13. D. Le and P.van-Huy ,“ Drug Response Prediction by Globally Capturing Drug and Cell Line Information in a Heterogeneous Network,” Journal of molecular biology, vol.430, 2018, pp.2993-3004
14. P.Shivakumar and M.Krauthammer ,“Structural similarity assessment for drug sensitivity prediction in cancer,” In BMC bioinformatics, vol. 10, 2009, pp.1-7
15. S.Kim, V.Sundaresan, L.Zhou, T.Kahveci, “Integrating domain specific knowledge and network analysis to predict drug sensitivity of cancer cell lines,” PloS one,vol. 11, 2016, pp.1-27
16. S.Mourragui, M.Loog, M. van de Wiel, R.Marcel J T, F.Lodewyk,“PRECISE: A domain adaptation approach to transfer predictors of drug response from pre-clinical models to tumors,” Bioinformatics,vol.35, 2019, pp.510-519.

17. Z.Dong , N.Zhang , C.Li, H.Wang, Y.Fang, J.Wang, and X.Zheng ,“ Anticancer drug sensitivity prediction in cell lines from baseline gene expression through recursive feature selection.” BMC cancer, vol.15, 2015, pp. 489-504
18. E. Fersini, E.Messina, and F.Archetti ,“A p-median approach for predicting drug response in tumour cells,” BMC bioinformatics, vol.15, 2014, pp 353-365
19. L.Wang , X. Li, L. Zhang and Q. Gao,“Improved anticancer drug response prediction in cell lines using matrix factorization with similarity regularization,” BMC cancer,vol.17, 2017, pp.513-525
20. Chun Wei Yap,“ PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints,” Journal of computational chemistry,vol.32, 2011, pp.1466-1474.
21. N.Zhang, H.Wang, Y.Fang, J.Wang, X.Zheng, X. Shirley Liu,“Predicting anticancer drug responses using a dual-layer integrated cell line-drug network model ,” PLoS computational biology , vol.11, 2015, pp.1-18
22. I.Cortés.Ciriano,G. Westen, G.Bouvier , M.Nilges, JP. Overington, A.Bender,“Improved large-scale prediction of growth inhibition patternsthe. Using NCI60 cancer cell line panel,” Bioinformatics.vol.32, 2016, pp.85–95
23. M. Ammad-ud-din, E .Georgii, M .Gönen, T. Laitinen, O. Kallioniemi, K. Wennerberg., et al ,“Integrative and personalized QSAR analysis in cancer by kernelized Bayesian matrix factorization." Journal of chemical information and modeling ,vol.54, 2014, pp. 2347-2359

.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementaryFile.docx