Chapter 7

Dissecting the Genome for Drug Response Prediction

Gerardo Pepe, Chiara Carrino, Luca Parca, and Manuela Helmer-Citterich

Abstract

The prediction of the cancer cell lines sensitivity to a specific treatment is one of the current challenges in precision medicine. With omics and pharmacogenomics data being available for over 1000 cancer cell lines, several machine learning and deep learning algorithms have been proposed for drug sensitivity prediction. However, deciding which omics data to use and which computational methods can efficiently incorporate data from different sources is the challenge which several research groups are working on. In this review, we summarize recent advances in the representative computational methods that have been developed in the last 2 years on three public datasets: COSMIC, CCLE, NCI-60. These methods aim to improve the prediction of the cancer cell lines sensitivity to a given treatment by incorporating drug's chemical information in the input or using a priori feature selection. Finally, we discuss the latest published method which aims to improve the prediction of clinical drug response of real patients starting from cancer cell line molecular profiles.

Key words Drug sensitivity prediction, Feature selection, Cancer cell lines, Machine learning, Deep learning, Drug screening, Precision medicine

1 Introduction

Precision medicine is a relatively young and growing field and represents an ongoing challenge in recent years. One of the main purposes of precision medicine is to move from "one-size-fits-all" to a personalized drug administration based on the needs of an individual patient. To achieve this goal, the characterization of the patients' genomic alterations plays a key role. The main challenges in this field are represented by sensitivity response prediction, drug repositioning, and precision oncology.

Recent advances in high-throughput sequencing technologies [1] and more accurate machine learning (ML) approaches allow us to identify treatments based on the molecular profile of patients' tumors [2–6]. Considering the lack of the molecular profiles and responses to drugs of cancer patients, cell lines large-scale drug

Oliviero Carugo and Frank Eisenhaber (eds.), *Data Mining Techniques for the Life Sciences*,
Methods in Molecular Biology, vol. 2449, https://doi.org/10.1007/978-1-0716-2095-3_7,

screening experiments which capture both molecular features of cancer and differences in therapeutic responses have become widely used [7–9].

To date three large-scale genomic projects on cancer cell lines are available, Cancer Cell Line Encyclopedia (CCLE) [10] and the Sanger's Catalogue of Somatic Mutations in Cancer (COSMIC) [11] both contain genomic and expression data for about 1000 cell lines. Moreover, CCLE contains drug sensitivity data for 24 drugs on 504 cell lines, while about 266 drugs were tested on nearly 1000 cell lines and drug response data are publicly available in the COSMIC database. NCI-60 is the third resource which characterizes 60 cancer cell lines. It is designed to screen up to 3000 small molecules per year for potential anticancer activity making the results available to the scientific community [12, 13].

The accumulated data such as expression, copy number alteration, single nucleotide mutation, and methylation status allow researchers to link the molecular features with the sensitivity/resistance to a given drug. Considering the nature of these datasets where the number of cell lines is much smaller than the number of genes in the -omics profiles of cell lines, machine learning methods often face the "small n, large p" problem [14]. This tends to limit the performance of the traditional machine learning algorithms, especially deep learning-based methods that require more observations to train the model. To overcome this problem, several approaches are used for the feature prioritization, thus discarding features that are not very useful for learning and that can only increase the level of noise. Some algorithms identify the set of informative features using a correlation-based method [15, 16], while some, on the other hand, are based on the sample variance [16, 17]. Contrary to supervised methods there are other deep learning-based unsupervised methods that allow to automatically prioritize genomic features of cell lines to improve anticancer drug responsiveness prediction [18–20].

Therefore, methods that use a proper feature selection before training the model outperform the methods without an a priori feature selection.

With this review, we aim to describe and systematically assess the representative computational methods that have been developed in the last 2 years (Table 1) on three public datasets: COSMIC, CCLE, and NCI-60. These methods aim to improve the drug response prediction performance and can be divided into two groups: (i) methods which include drugs' chemical information in the input and (ii) methods which use a priori informative feature selection (Fig. 1).

Although these methods are able to successfully predict drug response of preclinical samples, they have had limited success in predicting the clinical drug response of real patients [21, 22], with some exceptions [23, 24]. So, in the last part of this review we

Table 1
Published studies that have used machine learning or deep learning for drug sensitivity prediction in cancer cell lines or real patients

Study	Model	Input data	Prediction task
GraphDRP [14]	RDKit (drugs chemical information) + 1D CNN (genomic features) + Graph Convolutional Network	Drug's structure, gene expression	Cell line drug sensitivity
ADRML [26]	Similarity matrices + Manifold learning	Drug's structure, gene expression, mutation, copy number variation	Cell line drug sensitivity
Auto-HMM-LMF [37]	Autoencoder networks (gene expression and copy number variation) + hidden Markov model (single-nucleotide mutation) + logistic matrix factorization	Gene expression, copy number variation, single nucleotide mutation	Cell line drug sensitivity
Ahmed et al. [39]	Network-based feature selection + two graph-based neural network models	Gene expression	Cell line drug sensitivity
ISIRS [40]	Feature selection through iterative sure independent ranking and screening	Copy number alteration, gene expression and mutation status	Cell line drug sensitivity
TG-LASSO [46]	Linear and nonlinear regression model	Gene expression, tissue of origin	Clinical drug response

discuss the latest published method which aims to improve the prediction of clinical drug response of real patients starting from cancer cell line molecular profiles, developing an accurate computational "preclinical-to-clinical" model (Fig. 1).

2 Bioinformatic Approaches to Improve the Drug Sensitivity Prediction in Cancer Cell Lines and Real Patients

2.1 Considering the Chemical Properties of Drugs Improves Drug Response Prediction

Nguyen and collaborators propose a new model to predict drug response (GraphDRP); this model takes as input both chemical information of drugs and genomic features of cell lines (mutations, copy number alterations, genomic alterations) [14]. The authors, in this work, used the RDKit (Open-source cheminformatics; http://www.rdkit.org) to build a molecular graph reflecting interactions between the atoms inside the drug, unlike the method published a few years earlier by Liu and collaborators (tCNN), where drug molecules were represented as SMILES strings [25]. Several graph convolutional network models were used to learn the features of the drug (i.e., Graph Convolutional Networks

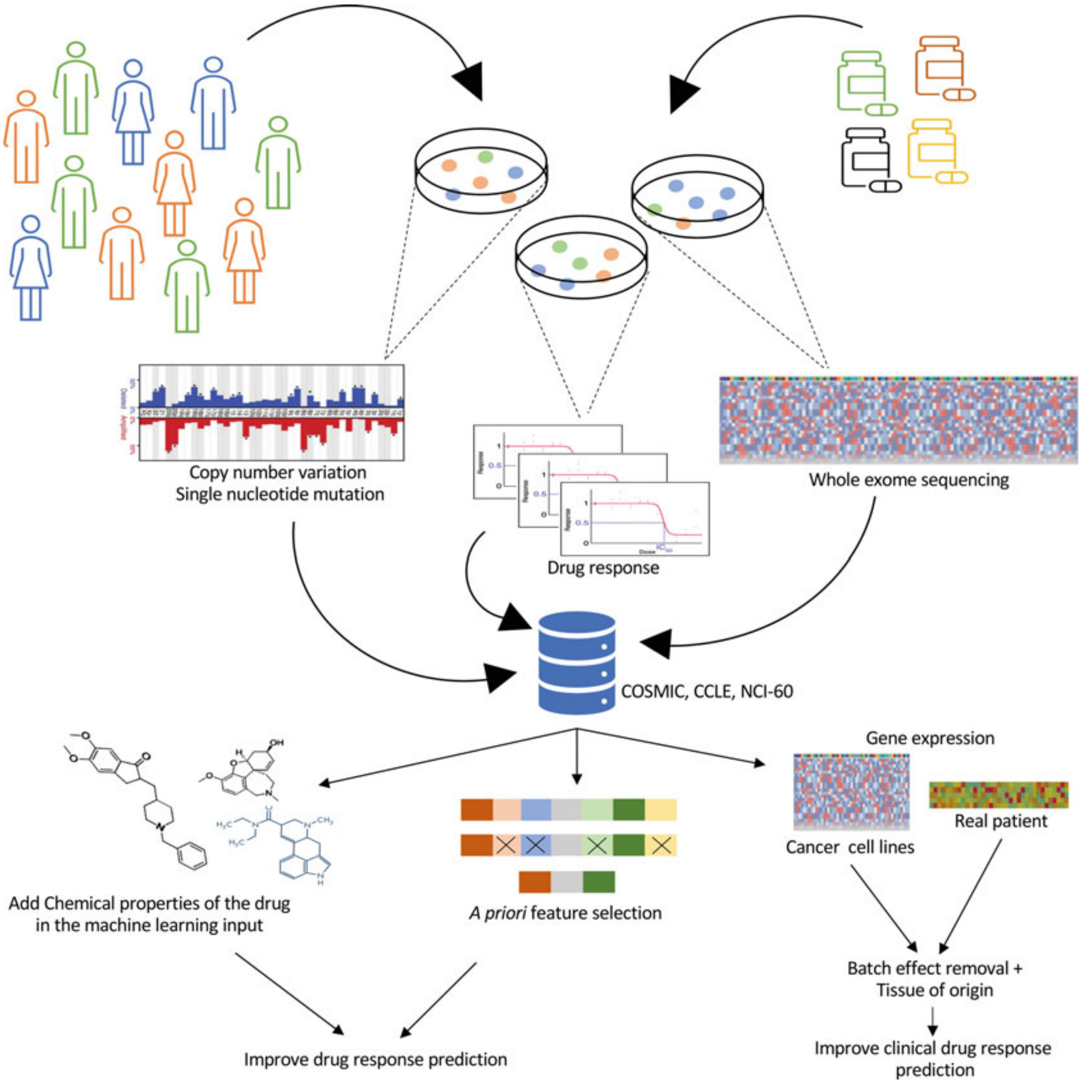
Copy number variation
Single nucleotide mutation
Drug response
Whole exome sequencing
COSMIC, CCLE, NCI-60
Add Chemical properties of the drug in the machine learning input
A priori feature selection
Gene expression
Cancer cell lines
Real patient
Batch effect removal + Tissue of origin
Improve drug response prediction
Improve clinical drug response prediction

Fig. 1 Characterization of the molecular profiles of human-derived cancer cell lines and massive drug screening provide a large amount of genomic and pharmacogenomic data that are collected and publicly accessible in the COSMIC, CCLE, and NCI-60 databases. These data are used by machine learning or deep learning methods in order to predict the cancer cell lines drug sensitivity. Two new approaches allow an improvement of the predictors' performance, respectively: (i) introducing the information about the structure of the drug in the model input and (ii) using an a priori selection of the features. On the other hand, a significant improvement in the quality of the clinical drug prediction is achieved using batch effect removal and tissue of origin information

(GCN), Graph Attention Networks (GAT), Graph Isomorphism Network (GIN), and combined GAT-GCN) and a fully connected layer was used to convert the result to 128 dimensions.

The genomic features of cell lines, represented in one-hot encoding, were used as input of a 1D convolutional neural network (CNN) layer to learn latent features, then the output was flattened to a 128-dimension vector of cell line representation. The combination of both the drug's feature and the cell line's feature was used to get a 256-dimension vector used later to predict the drug response. The authors adopted root mean square error (RMSE) and Pearson correlation coefficient (CC_p) to measure the performance of the model and compared their results with the results obtained using the tCNN model [25]. The experimental results indicate that the GraphDRP method achieves better performance in terms of both the root mean square error and the Pearson correlation coefficient, compared with the method that used only the SMILE string to represent the drug's properties. These results suggest that representing drugs in graphs is more suitable than in strings format since it considers the nature of chemical structures of the drugs.

Similar to the work described above, Moughari and Eslahchi propose ADRML, a model for Anticancer Drug Response Prediction based on Manifold Learning [26], which takes into account the drugs' characteristics as well as the cancer cell line molecular features as input for the predictive model. First step of this method is the construction of similarity matrices between cell lines (or drugs). Similarity matrices were computed for gene expression, copy number variation, mutations, and drugs, respectively. Then, a bipartite graph with two parts is used (drugs and cell lines) and later the manifold learning was used to factorize the drug response matrix in two latent matrices with lower rank. Manifold learning is useful to reduce the space dimensionality and some studies highlight how this method can conserve the topological structure of data [27, 28]. The authors achieved a better performance than other already existing methods that use both gene expression and drug chemical information as prediction model input (CDCN [29], SRMF [30], CaDRReS [31], KNN [32]). The predicted drug response values revealed high correlation with observed drug responses and several evidence in the literature supports the predictions of ADRML about novel cell line-drug pairs.

2.2 A Priori feature Selection to Improve Drug Sensitivity Prediction Performances

The identification of an optimal subset of features from a large number of candidate features is a crucial point for predicting drug response. In fact, it has been shown that a proper selection of the input features results associated with an improvement in the drug response prediction [33]. Thus, to improve the selection of informative features, many algorithms have been proposed by different research groups [16, 34–36]. One of the last published methods, in this field, highlights how using two different strategies could select proper features that significantly improved the drug response prediction. The authors propose a two step method named Auto-HMM-LMF [37]. In the first step, they apply a feature selection based on autoencoder networks to build two different similarity matrices using gene expression and copy number variation data, respectively. In the next step, they build the single-nucleotide mutation similarity matrix using the hidden Markov model and multinomial mixture model. Moreover, two similarity matrices are built using IC50 values and tissue type data, respectively.

Finally, the logistic matrix factorization method was applied for constructing the latent vectors for each cell line and drug and predicting the cell line's sensitivity or resistance to a given drug.

Auto-HMM-LMF shows better overall prediction power than the state-of-the-art prediction algorithms. Moreover, this innovative feature selection method returns better results in terms of drug response prediction when compared with the Ensemble Feature Selection, published by Neumann and collaborators [38].

In a more recent work, also describing network-based feature selection models [39], the authors showed an improvement of the drug response prediction accuracy using a priori feature selection. They introduce a network-based feature selection method and two graph-based neural network models. These methods analyze the modular co-expression structures along with gene discriminative power across lung cancer cell lines, in order to provide more reliable representative features for prediction performances. First, they compare the prediction power of the genes identified by the network-based feature selection model and the genes identified by graph-based deep neural network models, then they apply four canonical prediction methods (i.e., Elastic net, Partial least squares regression, Random forest, Support vector regression) and Deep Neural Network (DNN) to the previously selected features to evaluate drug sensitivity prediction performances. As a first step they introduce a network-based learning model that is based on the idea that the relations between the genes are more robust and stable for low sample size genomic datasets, with respect to the correlation between each individual gene and the drug response. They also introduce two graph-based models for drug sensitivity prediction. The former graph-based model proposes a network-based embedding method based on local neighborhood structure information to learn the gene expression level of the target gene. The latter graph-based model is a multi-layer graphical neural network model (GNN) that considers the global structure of the network to learn representative features. What emerges is that the network-based method provides better consistency in genomic feature identification and it is able to extract useful genomic information necessary to ensure the construction of efficient predictive models for drug sensitivity prediction. On the contrary, graph-based models show to be affected by small sample size and their performance is better than the one shown by DNN, but worse with respect to the other canonical methods. Next, they compare the different canonical prediction methods and DNN, used as a comparison, to evaluate drug sensitivity performance of the models previously analyzed. They report that Random Forest has the best overall performance and performs better than DNN due to the limited number of cell lines. The authors suggest a larger sample size to further increase the prediction accuracy and they highlight the possibility of multi-omics data integration in predictive algorithms to provide more accurate molecular signatures for drug response prediction.

Integration of multi-omics data is necessary in order to understand the molecular basis of a patient's disease. This type of data include genetic mutations, gene expression, and protein concentration and they can be efficiently integrated by each other and translated into predictive models for assessing patient specific therapy. At this scope An et al. propose a new method to select drug response-associated features called Iterative Sure Independent Ranking and Screening (ISIRS) [40]. This new method takes into account given genomic features and measures the conditional distribution of drug response, using an iterative procedure that overcomes the marginal utility measure drawbacks of missing marginally insignificant response features that are closely related with the response. As a first step the authors estimate the marginal correlation between all genomic features and drug responses by ranking those features; this step takes into account all candidate features and drug response values and produces a ranked list of candidate predictors, from which the top set of features are selected. Subsequently they perform the lasso regression based on a linear model for variable selection obtaining shrinkage estimates. Then they fit the drug response with these estimates of the features, by using a linear regression model and obtain the residuals. In order to consider important features with weak marginal correlation, but also to be sensitive to outliers and considering asymmetric distribution for most drug sensitivity data, they subsequently apply sure independent ranking and screening, known as SIRS [41], following a linearity assumption in modeling drug response and by using residual of response to do the iterations. Afterward they also cross-validate this method with other canonical methods like Iterative Sure Independent Screening (ISIS) [42], Simple Top Features (STF), and Sure Independent Ranking and Screening (SIRS) [41] for prediction accuracy. In evaluating the predicting performance they follow the Pearson correlation coefficient criterion and also evaluate mean squared errors (MSE) of the averaged predicted values. What emerges is that ISRS is robust to outliers and it is able to detect some new drug response related genomic features, marginally weak but biologically important, that have strong combination effects on drug response. Furthermore, ISIRS showed much higher correlations between predicted and true drug sensitivities than other canonical methods.

2.3 Prediction of Clinical Drug Response of Cancer Patients Using In Vitro Experiments on Preclinical Cancer Cell Lines

The final purpose of the drug response prediction is to be able to discriminate sensitive patients from the resistant ones based on their molecular features. Unfortunately, nowadays there are few data regarding patients' clinical drug responses and therefore they cannot be used to efficiently train machine learning models. Taking this into account, several drug predictor methods use cancer cell lines data to train the model and then try to predict the sensitivity/resistance of each cancer patient, obtaining poor results [23, 43–45]. In this context, Huang and collaborators developed a new method named Tissue-Guided LASSO (TG-LASSO) [46], which is trained using cancer cell line molecular features and is used to predict the patients' drug responses outperforming the already existing methods. In this study, the authors compare a variety of linear and nonlinear single-task and multi-task machine learning methods proving that the clinical drug response of many drugs can be predicted using regularized linear models trained on cancer cell lines. The proposed method differs from the other already published methods essentially in two aspects: (i) batch-effect removal in the input data and (ii) inclusion of the tissue origin in the input training data. The authors used ComBat for batch-effect removal [47] in order to homogenize the gene expression between cancer cell lines (microarray) and patients (RNA-seq). Using auxiliary information such as the tissue origin the authors achieve better results than all other methods that try to predict the patients' drug sensitivity training the models with the cancer cell lines features. However, a big step remains before bringing these models into medical practice. Recent advances in developing human-derived xenografts [48, 49] and 3D human organoids [50, 51] may enable developing a more accurate predictive model of clinical drug response in cancer.

3 Conclusion

In this article, we reviewed some of the latest studies that have employed machine learning and deep learning algorithms to predict the effects of a single drug on cancer cell lines. The results of these studies are encouraging, demonstrating that a proper feature selection or adding auxiliary information such as drugs' chemical details allow to outperform the existing machine learning- or deep learning-based methods. All the described methods are trained and predict the response to drug treatment on cancer cell lines which are publicly available in COSMIC, CCLE, and NCI-60 resources. On the other hand, predicting the drug response of real patients still remains a complicated goal, whereas pharmacogenomics data for real patients are currently limited. Using preclinical cell line data to train models does not always yield accurate drug response predictions for real patients. This is due to the fact that machine learning algorithms assume that the training and the test samples come from the same distributions. Homogenizing data and removing batch-effects could help alleviate this problem and, moreover, advances in more realistic preclinical models of cancer can be very useful in improving drug response predictions for real patients.

References

1. Reuter JA, Spacek DV, Snyder MP (2015) High-throughput sequencing technologies. Mol Cell 58:586–597
2. Iorio F, Knijnenburg TA, Vis DJ et al (2016) A landscape of pharmacogenomic interactions in cancer. Cell 166:740–754
3. Garnett MJ, Edelman EJ, Heidorn SJ et al (2012) Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature 483:570–575
4. Azuaje F (2017) Computational models for predicting drug responses in cancer research. Brief Bioinform 18:820–829
5. Menden MP, AstraZeneca-Sanger Drug Combination DREAM Consortium, Wang D et al (2019) Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen. Nat Commun 10
6. Huang C, Mezencev R, McDonald JF, Vannberg F (2017) Open source machine-learning algorithms for the prediction of optimal cancer drug therapies. PLoS One 12:e0186906
7. Weinstein JN (2012) Drug discovery: cell lines battle cancer. Nature 483:544–545
8. Wilding JL, Bodmer WF (2014) Cancer cell lines for drug discovery and development. Cancer Res 74:2377–2384
9. Yamori T (2003) Panel of human cancer cell lines provides valuable database for drug discovery and bioinformatics. Cancer Chemother Pharmacol 52:74–79
10. Barretina J, Caponigro G, Stransky N et al (2012) The cancer cell line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 483:603–607
11. Forbes SA, Beare D, Boutselakis H et al (2017) COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res 45:D777–D783
12. Alley MC, Scudiero DA, Monks A et al (1988) Feasibility of drug screening with panels of human tumor cell lines using a microculture tetrazolium assay. Cancer Res 48:589–601
13. Shoemaker RH (2006) The NCI60 human tumour cell line anticancer drug screen. Nat Rev Cancer 6:813–823
14. Nguyen T, Nguyen GTT, Nguyen T, Le D-H (2021) Graph convolutional networks for drug response prediction. bioRxiv. 2020.04.07.030908
15. Blessie EC, Chandra Blessie E, Karthikeyan E (2012) Sigmis: a feature selection algorithm using correlation based method. J Algorithms Computl Technol 6:385–394
16. Parca L, Pepe G, Pietrosanto M et al (2019) Modeling cancer drug response through drug-specific informative genes. Sci Rep 9:15222
17. Sánchez-Maroño N, Caamaño-Fernández M, Castillo E, Alonso-Betanzos A (2006) Functional networks and analysis of variance for feature selection. Intell Data Eng Autom Learn 2006:1031–1038
18. Chang Y, Park H, Yang H-J et al (2018) Cancer drug response profile scan (CDRscan): a deep learning model that predicts drug effectiveness from cancer genomic signature. Sci Rep 8:8857
19. Chiu Y-C, Chen H-IH, Zhang T et al (2019) Predicting drug response of tumors from integrated genomic profiles by deep neural networks. BMC Med Genet 12:18
20. Li M, Wang Y, Zheng R et al (2021) DeepDSC: a deep learning method to predict drug sensitivity of cancer cell lines. IEEE/ACM Trans Comput Biol Bioinform 18:575–582
21. Ali M, Aittokallio T (2019) Machine learning and feature selection for drug response prediction in precision oncology applications. Biophys Rev 11:31–39
22. Gillet J-P, Varma S, Gottesman MM (2013) The clinical relevance of cancer cell lines. J Natl Cancer Inst 105:452–458
23. Geeleher P, Cox NJ, Huang R (2014) Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines. Genome Biol 15:R47
24. Huang H-H, Dai J-G, Liang Y (2018) clinical drug response prediction by using a Lq penalized network-constrained logistic regression method. Cell Physiol Biochem 51:2073–2084

25. Liu P, Li H, Li S, Leung K-S (2019) Improving prediction of phenotypic drug response on cancer cell lines using deep convolutional network. BMC Bioinformatics 20:408
26. Moughari FA, Eslahchi C (2020) Author correction: ADRML: anticancer drug response prediction using manifold learning. Sci Rep 10:22360
27. Ma Y, Fu Y (2011) Manifold learning theory and applications. CRC Press
28. Wang JJ-Y, Huang JZ, Sun Y, Gao X (2015) Feature selection and multi-kernel learning for adaptive graph regularized nonnegative matrix factorization. Expert Syst Appl 42:1278–1286
29. Wei D, Liu C, Zheng X, Li Y (2019) Comprehensive anticancer drug response prediction based on a simple cell line-drug complex network model. BMC Bioinformatics 20:44
30. Wang L, Li X, Zhang L, Gao Q (2017) Improved anticancer drug response prediction in cell lines using matrix factorization with similarity regularization. BMC Cancer 17:513
31. Suphavilai C, Bertrand D, Nagarajan N (2018) Predicting cancer drug response using a recommender system. Bioinformatics 34:3907–3914
32. Garreta R, Moncecchi G (2013) Learning scikit-learn: machine Learning in Python. Packt Publishing Ltd.
33. Koras K, Juraeva D, Kreis J et al (2020) Feature selection strategies for drug sensitivity prediction. Sci Rep 10:9377
34. Kursa MB, Jankowski A, Rudnicki WR (2010) Boruta – a system for feature selection. Fundamenta Inform 101:271–285
35. Xu X, Gu H, Wang Y et al (2019) Autoencoder based feature selection method for classification of anticancer drug response. Front Genet 10:233
36. Dong Z, Zhang N, Li C et al (2015) Anticancer drug sensitivity prediction in cell lines from baseline gene expression through recursive feature selection. BMC Cancer 15:489
37. Emdadi A, Eslahchi C (2021) Auto-HMM-LMF: feature selection based method for prediction of drug response via autoencoder and hidden Markov model. BMC Bioinformatics 22:33
38. Neumann U, Genze N, Heider D (2017) EFS: an ensemble feature selection tool implemented as R-package and web-application. BioData Min 10:21
39. Ahmed KT, Park S, Jiang Q et al (2020) Network-based drug sensitivity prediction. BMC Med Genet 13:193
40. An B, Zhang Q, Fang Y et al (2020) Iterative sure independent ranking and screening for drug response prediction. BMC Med Inform Decis Mak 20:224
41. Zhu L, Li L, Li R, Zhu L (2011) Model-free feature screening for ultrahigh dimensional data. J Am Stat Assoc 106:1464–1475
42. Fang Y, Qin Y, Zhang N et al (2015) DISIS: prediction of drug response through an iterative sure independence screening. PLoS One 10:e0120408
43. Majumder B, Baraneedharan U, Thiyagarajan S et al (2015) Predicting clinical response to anticancer drugs using an ex vivo platform that captures tumour heterogeneity. Nat Commun 6(1):1–14
44. Ding Z, Zu S, Gu J (2016) Evaluating the molecule-based prediction of clinical drug responses in cancer. Bioinformatics 32:2891–2895
45. Turki T, Wei Z, Wang JTL (2018) A transfer learning approach via procrustes analysis and mean shift for cancer drug sensitivity prediction. J Bioinforma Comput Biol 16:1840014
46. Huang EW, Bhope A, Lim J et al (2020) Tissue-guided LASSO for prediction of clinical drug response using preclinical samples. PLoS Comput Biol 16:e1007607
47. Johnson WE, Evan Johnson W, Li C Adjusting batch effects in microarray experiments with small sample size using empirical Bayes methods. Batch Effects Noise Microarray Exp:113–129
48. Marangoni E, Poupon M-F (2014) Patient-derived tumour xenografts as models for breast cancer drug development. Curr Opin Oncol 26:556–561
49. Tentler JJ, Tan AC, Weekes CD et al (2012) Patient-derived tumour xenografts as models for oncology drug development. Nat Rev Clin Oncol 9:338–350
50. Weeber F, Ooft SN, Dijkstra KK, Voest EE (2017) Tumor organoids as a pre-clinical cancer model for drug discovery. Cell Chem Biol 24:1092–1100
51. Rae C, Amato F, Braconi C (2021) Patient-derived organoids as a model for cancer drug discovery. Int J Mol Sci 22:3483