Computer Methods and Programs in Biomedicine 178 (2019) 105–112

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Computational models for predicting anticancer drug efficacy: A multi linear regression analysis based on molecular, cellular and clinical data of oral squamous cell carcinoma cohort

Beaulah Mary Robert[a], G.R. Brindha[b,**], B. Santhi[b], G. Kanimozhi[c], Nagarajan Rajendra Prasad[a,*]

[a] Department of Biochemistry and Biotechnology, Annamalai University, Annamalainagar 608 002, Tamilnadu, India
[b] School of Computing, SASTRA Deemed to be University, Tirumalaisamudram, Thanjavur 613401, Tamilnadu, India
[c] Department of Biochemistry, Dharmapuramn Gnanambigai Government Arts and Science College for Women, Mayiladuthurai, Tamilnadu, India

ARTICLE INFO

Article history:
Received 25 February 2019
Revised 15 April 2019
Accepted 11 June 2019

Keywords:
Oral squamous cell carcinoma
Multi linear regression
Precision medicine
Computational models
Drug efficacy prediction

ABSTRACT

Background and objectives: The computational prediction of drug responses based on the analysis of multiple clinical features of the tumor will be a novel strategy for accomplishing the long-term goal of precision medicine in oncology. The cancer patients will be benefitted if we computationally account all the tumor characteristics (data) for the selection of most effective and precise therapeutic drug. In this study, we developed and validated few computational models to predict anticancer drug efficacy based on molecular, cellular and clinical features of 31 oral squamous cell carcinoma (OSCC) cohort using computational methods.

Methods: We developed drug efficacy prediction models using multiple tumor features by employing the statistical methods like multi linear regression (MLR), modified MLR-weighted least square (MLR-WLS) and enhanced MLR-WLS. All the three developed drug efficacy prediction models were then validated using the data of actual OSCC samples (train-test ratio 31: 31) and actual Vs hypothetical samples (train-test ratio 31: 30). The selected best statistical model i.e. enhanced MLR-WLS has then been cross-validated (CV) using 341 theoretical tumor data. Finally, the performances of the models were assessed by the level of learning confidence, significance, accuracy and error terms.

Results: The train-test process for the real tumor samples of MLR-WLS method revealed the drug efficacy prediction enhancement and we observed that there was very less priming difference between actual and predicted. Furthermore, we found there was a less difference between actual apoptotic priming and predicted apoptotic priming for the tumors 6, 8, 21 and 30 whereas, for the remaining tumors there were no differences between predicted and actual priming data. The error terms (Actual Vs Predicted) also revealed the reliability of enhanced MLR-WLS model for drug efficacy prediction.

Conclusion: We developed effective computational prediction models using MLR analysis for anticancer drug efficacy which will be useful in the field of precision medicine to choose the choice of drug in a personalized manner. We observed that the enhanced MLR-WLS model was the best fit to predict anticancer drug efficacy which may have translational applications.

1. Introduction

Oral cancer is defined as a malignant tumor of the oral cavity, and is the sixth most common cancer worldwide, with an annual incident of 400,000 new cases accounting for 4% of cancers in men and 2% of cancers in women [1,34]. Oral cancer is a serious global health problem, with estimated more than 300,000 deaths annually [2,13]. It is a major health problem particularly in developing countries and in the Indian subcontinent. It is a foremost leading cancer sub-type among men (16.1%) than women (10.4%) of all cancers [6,39]. In India, 20 per 100,000 populations are affected by oral cancer which accounts for about 30% of all subtypes of cancer [18]. Chemotherapy remain the only option for advanced oral cancer whenever salvage surgery or re-irradiation is not feasible. However, the oral cancer is associated with high

* Corresponding author at: Department of Biochemistry and Biotechnology, Annamalai University, Annamalainagar 608 002, Tamilnadu, India.
** Corresponding author at: School of Computing, SASTRA Deemed to be University, Tirumalaisamudram, Thanjavur 613401, Tamilnadu, India.
E-mail addresses: brindha.gr@ict.sastra.edu (G.R. Brindha), drprasadnr@gmail.com (N.R. Prasad).

https://doi.org/10.1016/j.cmpb.2019.06.011
0169-2607/

incidence of loco-regional recurrences, which account for the treatment failures [37]. Relapsed oral cancer represents a major clinical challenge in part due to their aggressive and invasive behaviors. More importantly, the efficacy of chemotherapeutic agents is limited due to the development of multidrug resistance[1] (MDR). Furthermore, oral cancer patients respond differently to each anticancer drug owing to disease diversity, genetic factors and environmental causes [36].

A significant challenge in oral cancer medicine is to assign a drug that will be precise and specific to the patients [25]. In oncology, the therapeutic decision making was assisted by anatomic location of the tumor, histology, cytogenetics and the presence of surface antigens [15]. However, the forecasters like histological grade and lymph node status often failed to predict the drug response due to the heterogenic nature of tumors [14]. The identification of a patient's drug response pattern provides the opportunity to guide the selection of rational therapeutics in an attempt to improve the therapeutic outcome of patients with advanced cancers. Precision medicine entails the design of therapies that are matched for each individual patient. Large-scale clinical trials were conducted to evaluate the efficacy of chemotherapeutic drugs [23]. Considering the cost, adverse effects, time consumption and the number of experiments of clinical trials there is an urgent need for a computation based models to effectively predict the efficacy of chemotherapeutic drugs even before therapy.

There were numerous computational algorithms proposed to advance the prediction of drug sensitivity in cancer patients [3,10,47]. Computational biologists have recently carried out large-scale data analyses that measure the sensitivity of molecularly well characterized cancer cells to different chemotherapeutic agents. Resources like the NCI-60 drug sensitivity database [38], the Cancer Cell Line Encyclopedia [16], the Cancer Target Discovery and Development small molecule screening data set [5], and the Genomics of Drug Sensitivity in Cancer [44] were the few scientific ventures which predict drug sensitivity based on cancer cells molecular profiling. Interestingly, many of these studies employ machine learning methods based on pharmacological data sets that can predict drug response from the genomic and transcriptomic features of cancer cells [7,46]. The machine learning-based predictive statistical methods of drug responses might enable the choice of personalized therapy. In general, the accuracy and interpretability of drug prediction models are limited. Therefore, multitask learning approaches have been proposed for the improved and accurate drug efficacy prediction [30]. Multiple tumor features like molecular, cellular and various tumor specific clinical data can be used for machine learning process and thereby an effective predictive algorithm could be developed. The drug response-linked gene expression data (molecular) has widely been employed for various statistical analyses in order to effectively predict anticancer drug efficacy [22] (Bernard et al., 2017). Recently, the dynamic BH3 profiling data (cellular), an analytical method for predicting apoptotic sensitivity, has also been successfully employed to measure early changes in net pro-apoptotic signaling induced by chemotherapeutic agents in cancer cells [15]. For example, the carboplatin efficacy in 16 ovarian adenocarcinoma patients has been predicted based on BH3 profiling by Kaplan–Meier plot and Mantel-Cox statistical analysis [15].

The computational algorithms available for drug efficacy prediction have been diverse [12]. Costello et al. [10] developed 44 drug sensitivity prediction algorithms based on breast cancer profiling datasets. Geeleher et al. [27] developed a ridge regression model based on baseline gene expression levels and the drug IC_{50} estimates to understand drug sensitivity prediction. Logistic regression effectively predicted cytarabine treatment on the patients' overall survival based on BH3 profiling and other clinical features like age and cytogenetics status [42]. The multilinear regression[2] (MLR) models can effectively be used to estimate the drug efficacy due to its straightforward approach and goodness of fit [3,4].

Recently, we reported the drug of choice for the effective treatment of 31 oral squamous tumors (OSCC) samples based on MDR-linked gene expression pattern and % Δ apoptotic priming [28]. We found that the confidence level of the drug of choice prediction was only 33% (low *R*-Square value, 0.3331) due to insufficient tumor samples. In order to improve the prediction accuracy, herein, we applied three MLR-based computational models by accounting other valuable predictive supporting clinical data such as age, sex, tumor-stage, tumor-grade, clinical-staging along with drug-response gene expression pattern and% apoptotic priming. We employed MLR, modified MLR-weighted least square[3] (WLS) and enhanced MLR-WLS for the prediction of anticancer drug efficacy. To prove the robustness of the developed models, we used the data of actual 31 OSCC samples, 30 hypothetical test samples (fictional samples whose feature values were nearer to the real samples) and 341 theoretical samples (created sample set in which each sample reflects the real samples 11 times). Finally, the developed computational models were validated by train-test method (31 Vs 31 samples and 341 Vs 341 samples) and train-test hypothetical samples (31 Vs 30). The developed best statistical model i.e. enhanced MLR-WLS was then cross-validated[4] (CV) using 341 theoretical tumor data.

2. Materials and methods

2.1. Tumor data

We collected 31 oral squamous cell carcinoma[5] (OSCC) cohort samples prior to radiotherapy and chemotherapy. The collected tumor samples were classified on the basis of sex, age, anatomical location of the tumor, and histological grade and tumor stages before the computational analysis. The quick plot depicted the categorical features of all tumor samples (Supplementary Fig. 1). The features such as drug response gene expression pattern (molecular), apoptotic sensitivity (cellular), age, sex, tumor-grade, tumor-stage and clinical-stages (clinical) were used as the data set for developing drug efficacy prediction models (Table 1). The categorical clinical features like clinical_stage, tumor_stage and tumor_grade were converted as numeric values, i.e. 1, 2, 3. We used the values of Best-Drug-Efficacy[6] (BDE) data of actual 31 OSCC, based on our recent publication [28], as the dependent variable to develop the computational models in order to computationally predict the% apoptotic priming of unknown test samples (Supplementary Table 1).

To predict the drug response pattern, the molecular, cellular and clinical data of 31 OSCC samples were accounted for machine learning; a training process (Supplementary Fig. 2 - Step-1). The same 31 OSCC samples were given to the testing process (prediction) as well (Supplementary Fig. 2 -Step-2a) after excluding the actual% apoptotic sensitivity (priming) to the anticancer drugs (cellular BH3 profiling data).

The predicted values provided by MLR were compared with the actual drug response values (cellular apoptotic priming data) to get the performance measures of the model (Supplementary Fig. 2 - Step-2b). In continuation with the training-test process, 30 new

[1] MDR: multidrug resistance.

[2] MLR: multi linear regression.

[3] WLS: weighted least square.

[4] CV: cross validated.

[5] OSCC: oral squamous cell carcinoma.

[6] BDE: best drug efficacy.

Table 1
Drug efficacy prediction (% apoptotic priming) based on tumor data of real 31 OSCC samples. The MLR, modified MLR-WLS, and enhanced MLR-WLS were developed using Eqs. (1)–(3), respectively. The features used in Fit-1, Fit-2, and Fit-3 and their predictive performances were compared. The developed Fit-3 showed enhanced performance for drug efficacy prediction.

Method → (Testing known trained samples)	MLR (Fit-1)	Modified MLR-WLS (Fit-2)	Enhanced MLR–WLS (Fit- 3)	Criterion
Features used	Molecular + Clinical	Molecular + Clinical + polynomial degree 2 (Non-factor features)	Molecular + Clinical + Square (Less significant drug response-linked gene expression)	Continuous
Dependent variable to be predicted	BDE-Response (Best_Priming_Drug)			Continuous
Tumor Count	31	31	31	More the better
Residual standard error	13.92	6.16	5.36	Lower the better
Multiple-*R*-squared	0.77	0.99	0.99	Higher the better (>0.70)
Adjusted-*R*-squared	0.51	0.90	0.93	Higher the better
F-statistic	2.96 on 21	11.48 on 27	15.78 on 26	Higher the better
p-value	0.0239	0.03351	0.007854	To be less than 0.05
Root means square error	9.3	1.92	1.92	Lower the better
Mean square error	87.56	3.7	3.7	Lower the better
Min–max-accuracy	0.175	0.223	0.229	Higher the better
AIC	262.6	186.33	184.54	Lower the better
BIC	288.43	227.9153	224.7	Lower the better

hypothetical (Supplementary Table 2) tumor data were given (Supplementary Fig. 2 - Step-3a) to the model in order to understand the model efficiency (Supplementary Fig. 2 - Step-3b). Then, we compared the apoptotic priming values of real data with the hypothetical data of 341 samples for validating the prediction efficacy for unknown test samples (Supplementary Fig. 2 Step-1, Step-2a and Step-2b). The drug efficacy prediction depends on the tumor sample size and the clinical features of tumor samples. Since, the features of the tumor samples (count) used in this study was only 16 (molecular, cellular and clinical), we theoretically increased the sample size to 341 by assuming each tumor features were present in 10 more theoretical tumors (totally 11 tumor data). Therefore, the actual 31 tumor data were theoretically considered as 341 tumor data (11×31 samples). Further, we applied the 341 sample data for 10 fold Cross Validation (CV) process wherein 9 folds were used to train the model and one fold was used as the test sample to assess the efficiency and accuracy of the developed model (Supplementary Fig. 3).

2.2. Statistical methods and computational models

All the statistical analysis and computations were performed using R version-3.3.2 [29]. We developed MLR model by using BDE data of OSCC tumor samples; then developed modified MLR-WLS model by quadratic terms of molecular features and finally developed enhanced MLR-WLS which excluded squaring of significant (near or higher) features. Then, we developed a model for individual drug efficacy prediction using enhanced MLR-WLS model (Supplementary Fig. 4).

The MLR analysis was employed for drug response prediction by accounting the 11 drugs response-linked gene expression pattern as dependent variables (X). The MLR analysis for drug response prediction model was calculated as follows:

$$Y = a + b.X + e \tag{1}$$

where Y was the variable to be predicted (% apoptotic priming), a was the intercept, b was the slope of the line, e represented the error term and X was the drug expression pattern of 31 tumor samples (Supplementary Data 1).

Combining biologically meaningful features would increase additional structure in the drug response prediction [3]. In order to attain linear relationship between drug response-linked gene expression and% apoptotic priming, we also accounted other clinical features like age, sex, tumor_grade, tumor_stage, and clinical_stage in place of X (Eq. (1)) and drug efficacy was predicted [29,40].

Quadratic term of features in MLR had been considered as an optimal method to attain the best fit of the samples [40]. We applied squaring process to continuous predictor variables (drug response-linked gene expression; $j = 1–11$) since, squaring categorical variables (age, sex, tumor_grade, tumor_stage, clinical_stage; $i = 1–5$) was not reasonable (Supplementary Data 1). The fact behind the best fit was its learning the regression line that had the shortest average distance to all data samples.

The modified MLR-WLS was calculated as follow (Eq. (2)):

$$Y = a + b.X_{i+j} + c.X_j^2 + e \tag{2}$$

Here, c was the regression coefficient.

To improve the drug efficacy prediction, we employed enhanced MLR-WLS method by the p-value of each independent variable for statistical significance. The variables (k = any drug response-linked gene expression from 1 to 11) that had significance or near significance (0–0.1) were excluded ($j - k$) from squaring process (Eq. (3)) to get the enhanced MLR-WLS (squaring process applied only to the less significant features).

$$Y = a + b.X_{i+j} + c.X_{j-k}^2 + e \tag{3}$$

When the one particular independent variable so strongly depended on other variables, it could not contribute to prediction [35]. In the enhanced MLR-WLS model, the higher gene expression pattern (x_j) and its quadratic term (x_j^2) was found to be mutually depended (Supplementary Data 1). Hence, we removed the quadratic feature from the independent variable set in order to increase the drug efficacy prediction.

2.2.1. Stability assessment of the developed models

The performance of the models was compared and analyzed based on the fitting parameters and error terms. To assess the stability of the developed models, the following performance measures were used: adjusted-*R*-square, *F*-statistics, mean square error[7] (MSE), root mean square error[8] (RMSE) and min-max-accuracy. Multiple-*R*-square measured the percentage of the total uncertainty in the samples that was explained by the regression line (Supplementary Data 2).

[7] MSE: mean square error.
[8] RMSE: root mean square error.

Multiple-*R*-square was calculated using sum square error[9] (SSE) and total sum squares[10] (SST). The *R*-squared value would be between 0 and 1; closer to 1, the greater proportion of variance was judged by the model [35]. The adjusted-*R*-square value was less than or equal to 1; when it was closer to 1, the inference was considered as a 'good fit'. If the value of MSE was closer to 0 then the fit was more helpful for prediction.

The standard error of the regression was measured by RMSE. The Akkaike's information criterion[11] (AIC) and Bayesian information criterion[12] (BIC) depended on the maximized value of the likely hood function for the comparison of models. Hence, when AIC value was lower, the model would closer to the actual values. When the BIC value was lower, the performance of a fit will be higher [33].

Cook's distance was used to identify samples that showed the disproportionate impact on the determination of the model parameters [11]. We analyzed Cook's distance to find data samples with large residuals (outliers) and to ascertain the accuracy of regression. We also measured the min-max accuracy from the correlation between the actual observations and predicted values.

2.2.2. Feature validation

Backward selection started with the inclusion of all relevant features for model building and then one after other features were removed and the performance measures were verified. During model creation, the quadratic term of each drug-response linked gene expression feature was excluded one by one to get the significant feature set.

3. Results and discussion

Drug sensitivity prediction is an integral part of personalized medicine that refers to therapy tailored to an individual patient, rather than a one-size-fits-all approach. A prime prospect in precision cancer medicine is to identify tumor features that are predictive of drug treatment responses in cancer cells. Although there are several computational models available for accurate drug response prediction based on tumor data, these algorithms often lack the ability to infer precise drug efficacy prediction when we employ fewer input features and the sample size is minimum [12,28]. As increasing amounts of diverse genome-wide data and cellular level treatment responses are becoming available, there is a need to build new powerful computational models that can effectively combine these data sources and identify maximal predictive efficacy. Linear regression model for cellular drug response prediction has successfully been employed based on multiple input data sources (molecular features and clinical features) and the criterion variable (cellular drug responses). The motivation of the present study was to employ multitask machine learning in order to precisely predict the drug response. We applied multitask learning strategy for drug response prediction and feature interaction using R Script. The algorithm tries to learn multiple tasks (molecular, cellular and clinical data) simultaneously and reveal the common feature i.e. apoptotic sensitivity that can benefit each individual learning task. Similarly, Yang et al. (2018) applied the Genomics data of Drug Sensitivity in Cancer2 (GDSC) cell line panel with drug response (IC50) of 265 drugs on 990 cell lines along with protein targets and signalling pathways' activities to support a personalized treatment strategy in order to achieve the goal to maximize the drug response [44].

We developed an MLR model for drug efficacy prediction using the data of 11 drug response gene expression pattern and% apoptotic sensitivity ("priming") of five anticancer drugs in OSCC tumor samples. The variance of samples learned by the developed model (*R*-squared value) was found to be less (0.68) and the adjusted-*R*-square value (0.498) conveys that the contribution of features was not significant (Supplementary Table 3). The error rates (RMSE 11 and MSE121.89), AIC (262.87) and BIC (281.5) were also found to be high. The *F*-statistic (3.71) and min-max accuracy (0.14) were not to be sufficient. Further, we observed that there was a very weak linear relationship between the drug-response linked gene expression and% apoptotic priming which suggests that drug response is not easy to interpret using the available 31 tumor samples and 11 gene features (Supplementary Fig. 5). This outcome conveyed that the molecular information alone was not sufficient to predict drug response pattern. Safikhani et al. (2017) [32] found that drug response prediction was inconsistent when they employ gene expression data alone; it suggested that using new additional data allow the sensitivity of drug efficacy prediction [32].

3.1. MLR model based on drug response-linked gene expression and apoptotic priming data

In order to further substantiate the predictive accuracy of the developed computational model, we accounted actual cellular apoptotic priming data (as the dependent variable) and other clinically relevant categorical features apart from drug response gene expression data for MLR, modified MLR-WLS and enhanced MLR-WLS analysis. Gender affects the efficacy of chemotherapy and, therefore, we accounted sex as a clinical categorical feature to predict drug efficacy. Xiaoming et al. (2016) [43] showed that drug response gene expression pattern of certain gender-specific cancers such as breast cancer and ovarian cancer [43]. Similarly, age is an another main factor for drug response prediction and many studies have shown that people in different age groups vary significantly in drug response and treatment outcome [20]. Moreover, clinical categorical features like tumor size, histological subtype, tumor grade, symptomatic presentation, and pathological stages were also been used to predict treatment outcome in a cohort of patients with renal cell carcinoma [41]. Therefore, we accounted all possible tumor features such as molecular (drug response gene expression pattern), clinical categorical features (age, sex, tumor grade, clinical stage, and tumor stage) and the cellular (apoptotic sensitivity) data for training the computational model. The relationship of independent and dependent variables has initially learned by the model using the data of 31 actual tumor samples. Then, the same 31 samples data were given as test data and the derived predicted data were compared with the actual apoptotic priming efficacy data of five anticancer drugs in order to verify the performance of the model. Further, the suitability of the developed model was verified by molecular and clinical data of the additional 30 hypothetical test samples.

The outcome of MLR (Fit-1) using the clinical features combined with molecular data (Eq. (1)) was found to be suitable for drug efficacy prediction. Homoscedasticity (the variance of the predictor did not differ with the levels of target variables) was observed through the random band around the line (Supplementary Fig. 6a). Though the *p*-value (0.02) was lesser than 0.05 (significant), the other measures like multiple-*R*-squared (0.77), adjusted-*R*-squared (0.51), *F*-statistic (2.96), RMSE (9.3) and MSE (87.56) represented the necessity for improvement in the MLR-based drug efficacy prediction model (Table 1; Supplementary Fig. 6a). The RMSE (9.27), MSE (85.97) and min-max accuracy (0.175) showed that there was a further need for the improvement in drug efficacy prediction model (Table 2). [31] clearly illustrated the aware of

[9] SSE: sum square error.
[10] SST: total sum square.
[11] AIC: Akkaike's information criterion.
[12] BIC: Bayesian information criterion.

Table 2
Drug response prediction by error comparison using hypothetical test-samples. The drug efficacy was predicted by providing a set of test-samples to the learned models. The enhanced MLR-WLS (Fit-3) showed decreased RMSE and MSE and increased min-max accuracy.

Method → (Testing unknown Samples)	MLR (Fit-1)	MLR-WLS (Fit-2)	Enhanced MLR–WLS (Fit-3)	Criterion
Features	Molecular + Clinical	Molecular + Clinical	Molecular + Clinical + Square (Less significant drug response-linked gene expression, except ABCC3)	Continuous
Dependent variable to be predicted	BDE- Response (Best_Priming_Drug)			Continuous
Tumor count	30	30	30	Better when less than the train –samples
Root means square error	9.27	3.065	2.96	Lower the better
Mean square error	85.97	9.39	8.81	Lower the better
Min-max-accuracy	0.175	0.186	0.19	Higher the better

error measures for the validation of predictive models for chemometrics and intelligent laboratory systems.

We observed that Fit-2 (Table 1) based on modified MLR-WLS showed improved drug efficacy prediction (Eq. (2)). The multiple-*R*-squared (0.99) has been increased by 22% when compared with Fit-1 which confirmed that the addition of features by squaring process of genomic data enhances the drug efficacy prediction. José et al. [17] showed that the integrated Brier score (IBS; a measure of the model's accuracy) increases by squaring the differences between the patient primary outcome at a particular point in time. Squaring of features in MLR had been considered as an optimal method to attain the best fit of the samples [37]. We observed increased *p*-value (0.03) when compared to Fit-1. Error measurements like AIC (186.33) and BIC (227.91) were found to be reduced drastically. Homoscedasticity has also arrived and all the downgraded points like *F*-statistics (11.48), min–max accuracy (0.223), RMSE (1.92) and MSE (3.7) were improved significantly in Fit-3 (Supplementary Fig. 6c) when compared to Fit-2 (Supplementary Fig. 6b). These train results were also supported by test results of 30 theoretical samples (Table 2). The RMSE (3.065) and MSE (9.39) were found to be decreased drastically (by the difference of 6.205 and 76.58 points, respectively) and the min-max-accuracy showed the negligible difference (0.011 points) between Fit-1 and Fit-2.

The individual *p*-value of all the molecular and clinical features was verified for its significance using Eq. (1). The enhanced MLR-WLS method was introduced by the second-order polynomial process based on the *p*-value. The features which had less or no significance were considered for second order polynomial. One of the features, ABCC3 had significance (Supplementary Table 4), so it was excluded from the squaring process to get the results of enhanced MLR-WLS. We noticed that Fit-3 showed 99% of total variation (*R*-square value) with the less residual error (5.36). Therefore, the Fit-1 and Fit-2 models for the prediction of apoptotic priming by anticancer drugs could be rejected and the Fit-3 was selected for drug efficacy prediction due to its significant *p*-value (0.007). The increase in the value of adjusted-*R*-square (0.93) and the value of *F*-statistic (15.78) support the robustness of the Fit-3. Further, the AIC (184.54) and BIC (224.7) has also been found to be decreased in Fit-3 (Table 1). Therefore, the best model for drug efficacy prediction based on the values of AIC and BIC were in the order of Fit-1 > Fit-2 > Fit-3. The strength of a prediction model could be validated by its residuals which had the normal distribution with mean (zero) and constant standard deviation [19,22,24]. The decreased value of polynomial adjusted *R*-square value indicates enhanced MLR_WLS is the best method for drug response prediction when compared to other two methods such as MLR and MLR-WLS. Similarly, Li et al. (2016) compared the robustness of MLR methods and showed that diagonally weighted least squares was less biased and more accurate than MLR in estimating the factor loading across nearly every condition [23].

We observed that the variation did not alter with increasing drug response gene expression and hence there was no violation of homoscedasticity (Constant variance assumption was met) in the plot (Supplementary Fig. 6c). The testing of hypothetical samples also conveyed the robustness of the developed Fit-3 with the decrease in error terms (RMSE-2.96, MSE-8.81) and increase in min-max accuracy (0.19) (Table 2). MLR-WLS were generally less biased than those obtained by mean-adjusted ML, irrespective of the number of categories [9,45].

The proposed model has been validated by unknown samples. We observed that the values of actual priming (fictional 30 tumor samples) compared with the predicted values by Fit-3 (Fig. 1; Supplementary Table 5). Though the priming performance of Fit-2 (brown) and Fit-3 (orange) were more or less same, the priming was better in Fit-3. We noticed that the predicted values of Fit-3 (orange) were found to be nearer to the actual (blue) values for 9 tumors (tumors 1, 4, 5, 6, 9, 10, 18, 19 and 25). The comparison of efficacy values predicted by Fit-3 and actual conveyed that predicted values were equal with actual values for the tumors 1, 4, 5, 9 and 10; slightly higher efficacy were predicted for the tumors 3, 6, 14, 15, 16, 19, 21, 23, 25, 26, 28, and 30; whereas, slightly lower than the actual drug efficacy was predicted for the tumors 2, 7, 8, 11, 12, 13, 17, 18, 20, 22, 24, 27 and 29.

The difference value of actual and predicted while testing the same trained samples is known as train error and testing with unknown sample is test error [19]. The train-test process for the real tumor samples of Fit-3 revealed the drug efficacy prediction enhancement and there was very less priming difference between actual and predicted (Supplementary Fig. 7a). We found there was a less difference between actual apoptotic priming and predicted apoptotic priming (tumors 6, 8, 21 and 30) whereas, for the remaining tumors there were no differences between predicted and actual priming data. The error terms (Actual Vs Predicted) also revealed the reliability of enhanced MLR-WLS model (Fig. 2) for drug efficacy prediction. The actual priming values and predicted values of new hypothetical test samples (Supplementary Fig. 7b) and their error terms (Fig. 2) conveyed that Fit-3 performed well for the new unknown clinical tumor samples [30].

Studentized residual illustrate the extreme outlier using overall and leave-one-out evaluation of the error variance [8]. We observed extreme studentized residual tumor samples in all the Fit models (Supplementary Table 6). Extreme outliers (far from the middle of the data distribution) and outliers (upper side of the data distribution) were tumor data whose deviation of predicted value was greater from actual value and they increased the prediction error. For Fit-3, the tumor sample 6 was found to

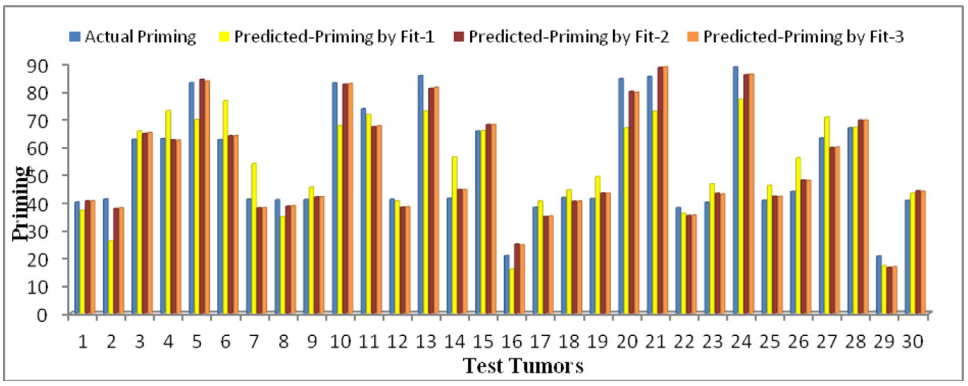

Fig. 1. Actual priming Vs predicted priming by the developed computational models (Fit-1, Fit-2, and Fit-3) for unknown hypothetical test samples. The blue bar represents the actual drug efficacy value by BH3 profiling of the know samples; the Yellow bar represents the predicted drug efficacy value by Fit-1, the brown bar represents the predicted drug efficacy by Fit-2 and orange bar represents the drug efficacy values predicted by the Fit-3. The enhanced MLR-WLS (Fit-3) predicted the priming values which were closest to the actual values.

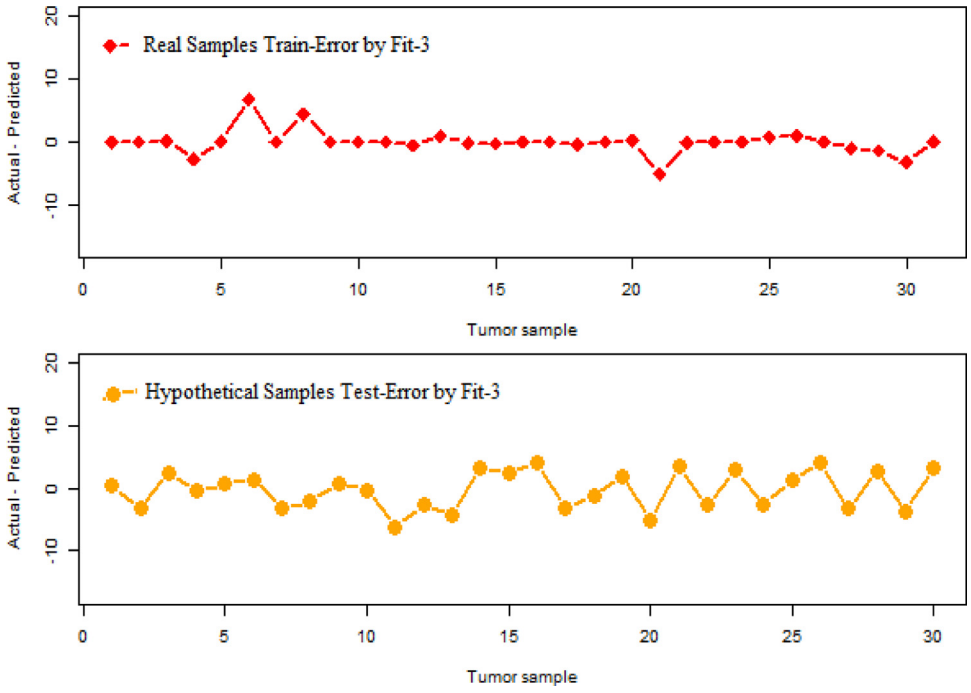

Fig. 2. The error difference between actual priming values and predicted values by the developed Fit-3 for the real samples and hypothetical samples. The red and orange lines showed the deviation in drug efficacy prediction.

be an extreme outlier; tumor samples 21 and 30 were the other outliers. Tumor sample 19 was the influential sample, which greatly affected the slope of the regression curve (Supplementary Table 6).

The quantile-quantile (Q-Q) plot is a graphical technique for determining if two data sets come from populations with a common distribution [21]. Quantiles were the points in samples below which a certain amount of data fall. The Q-Q probability plot assessed the normality of assumption based on studentized residual and showed the extreme points (Supplementary Fig. 8a–c). The linear pattern confirmed that the estimates by all the models were valid. All the points fall approximately along the red/blue dotted reference line to confirm normality. Compared with Fit-1, the Fit-2 and Fit-3 had all points along the reference line (Supplementary Fig. 8a–c).

Based on the results of a train-test process the Fit-3 was found to the best model. However, it should be proved using the backward selection that the features included in the model were relevant to learn the tumor characteristics. The error terms, *R*-square and adjusted-*R*-square values were obtained to confirm the inclusion of features. The removal of quadratic terms of drug response linked-genes ABCC1, ABCC3 and ABCG2 resulted in good learning confidence of the model (higher *R*-squared values) with lower error rates when compared to other features removal (Supplementary Fig. 9a and b). Among these three, ABCC3 had best adjusted-*R*-squared values. This outcome was a supportive proof for the enhanced MLR-WLS (squaring term of less significant features).

A large sample size is required for constructing a robust drug efficacy model [35]. Hence, 341 samples were used to train the model (Eq. (3)) and the same has been used to test the model (Table 3). The proposed model works better for 341 samples than 31 samples in the train-test process of Fit-3 model which was revealed in the drastic reduction of residual error (difference of 3.335

Table 3
Performance measures comparison of Fit-3 train-test for 31 samples and 341 samples. The enhanced MLR-WLS provided the acceptable outcome for the train-test process of 31 samples and also satisfactory outcome for 341 samples. Further, cross-validation results showed the stability and robustness of the developed enhanced MLR-WLS.

Method→	Enhanced MLR–WLS (Fit-3) (Testing the 31 trained Samples)	Enhanced MLR–WLS(Fit-3) (Testing the 341 trained Samples)	Enhanced MLR–WLS (Fit-3) (Cross Validation)
Features	Molecular + Clinical + Square (Less significant drug response-linked gene expression, except ABCC3)		
Dependent variable to be predicted	BDE-Response(Best_Priming_Drug)		
Tumor count	31	341	341
Residual standard error	5.36	2.005	0.078
Multiple *R*-squared	0.99	0.99	0.99
Adjusted-*R*-square	0.93	0.9895	0.989
F-statistic	15.78 on 26	1239	1126
p-value	0.007854	0.00000	0.00000
Root means square error	1.92	1.92	1.64
Mean square error	3.7	3.7	3.7
Min–max-accuracy	0.229	0.229	0.454

points) and an increase in *F*-statistic. Other measures remain stable without any degradation in the model.

Cross-validation is an inevitable proof for the prediction model [26]. When the samples were increased, the model got enough description of features of tumors and able to reduce the residual error and also a RMSE error. Though *F*-statistic was reduced, still it was in the higher level. To our surprise, min-max-accuracy was doubled with significance. This reveals that the observation capability of the model was improved because of the increased samples. The other error terms, multiple *R*-square value, and min-max accuracy were stable (Table 3).

4. Conclusion

Summarily, we developed computational models using clinical, molecular features to predict anticancer drug efficacy. The enhanced MLR-WLS model was found to be the most effective and accurate method for selection of anticancer drugs based on tumor features of the patients in a personalized manner. Increasing more features like the mode of action of the drugs, the chemical structure of anticancer agents as additional features will improve the robustness of the developed model and might have translational application. Further, the enhanced MLR_WLS model can also be used for the efficacy prediction of individual drugs. The individual drug prediction can be done effectively using methods such as bagging- boosting and random forest. This study has not concerned the intertumor and intratumour heterogeneity in pathogenesis and subsequent correlation with the drug response outcome. The availability of molecular profiling technologies such as next-generation sequencing coupled with advances in bioinformatics might reveal molecular pathogenesis and this could be used as an another important feature to predict accurate treatment response and prognosis. Without a full understanding of the spectrum of a patient's mutations, we may risk expending large resources on the development of fundamentally flawed approaches to biomarker-directed therapeutics.

Declaration of Competing Interest

No conflict of Interest.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.cmpb.2019.06.011.

References

[1] A. Jemal, et al., Cancer statistics, 2009, CA Cancer J. Clin. 59 (2009) 225–249.
[2] A.A. Hussein, M.N. Helder, Jan G. de Visscher, C.R. Leemans, B.J. Braakhuis, H.C.W. de Vet, T. Forouzanfar, Global incidence of oral and oropharynx cancer in patients younger than 45 years versus older patients: a systematic review, Eur. J. Cancer 82 (2017) 115e127.
[3] M. Ammad-ud-din, et al., Systematic identification of feature combinations for predicting drug response with Bayesian multi-view multi-task linear regression, Bioinformatics 33 (14) (2017) i359–i368.
[4] B. Western, Concepts and suggestions for robust regression analysis, Am. J. Pol. Sci. 39 (3) (1995) 786–817.
[5] A.A. Bülent, et al., CTD2 Dashboard: a searchable web interface to connect validated results from the Cancer Target Discovery and Development Network, Database 2017 (2017) bax054.
[6] Cancer Statistics: http://cancerindia.org.in/cancer-statistics/.
[7] C. Huang, et al., Machine learning predicts individual cancer patient responses to therapeutic drugs with high accuracy, Sci. Rep. 8 (1) (2018) 16444.
[8] J.M. Chambers, Linear models, in: J.M. Chambers, T.J. Hastie (Eds.), Chapter 4 of Statistical Models in S, Wadsworth & Brooks/Cole, 1992.
[9] L. Cheng-Hsien, Confirmatory factor analysis with ordinal data: comparing robust maximum likelihood and diagonally weighted least squares, Behav. Res. Methods 48 (3) (2016) 936–949.
[10] J.C. Costello, et al., A community effort to assess and improve drug sensitivity prediction algorithms, Nat. Biotechnol. 32 (12) (2014) 1202–1212.
[11] J. Fox, S. Weisberg, An R Companion to Applied Regression, second ed., Sage Publications, 2011.
[12] A. Francisco, Computational models for predicting drug responses in cancer research, Brief. Bioinformat. 5 (1) (2017) 820–829.
[13] Globocan: http://globocan.iarc.fr/Pages/summary_table_pop_sel.aspx
[14] S. Goodison, et al., Derivation of cancer diagnostic and prognostic signatures from gene expression data, Bioanalysis 2 (5) (2010) 855–862.
[15] M. Joan, et al., Drug-induced death signaling strategy rapidly predicts cancer response to chemotherapy, Cell 160 (5) (2015) 977–989.
[16] J. Barretina, et al., The Cancer Cell Line Encyclopedia enables predictive modeling of anticancer drug sensitivity, Nature 483 (7391) (2012) 603–607.
[17] J.M. Lezcano-Valverde, et al., Development and validation of a multivariate predictive model for rheumatoid arthritis mortality using a machine learning approach, Sci Rep. 7 (1) (2017) 10189.
[18] R.C. Ken, Challenges of the oral cancer burden in India, J. Cancer Epidemiol. 2012 (2012) 701932.
[19] J. Keum, H. Nam, Self-blm: prediction of drug-target interactions via self-training svm, PLoS One 12 (2) (2017) e0171839.
[20] J. Koch-Weser, et al., Drug disposition in old age, N. Engl. J. Med. 306 (18) (1982) 1081–1088.
[21] M. Kozak, H.P. Piepho, What's normal anyway? Residual plots are more telling than significance tests when checking ANOVA assumptions, J. Agro. Crop Sci. 204 (2018) 86–98.
[22] D.H. Le, D. Nguyen-Ngoc, Multi-task regression learning for prediction of response against a panel of anti-cancer drugs in personalized medicine, in: Multimedia Analysis and Pattern Recognition (MAPR), 2018 1st International Conference on IEEE, 2018, pp. 1–5.
[23] H. Li, et al., A novel multi-target regression framework for time-series prediction of drug efficacy, Sci. Rep. 7 (1) (2017) 40652.
[24] H. Lu, et al., A hybrid feature selection algorithm for gene expression data classification, Neurocomputing 256 (2017) 56–62.
[25] B. Majumder, et al., Predicting clinical response to anticancer drugs using an ex vivo platform that captures tumor heterogeneity, Nat. Commun. 6 (1) (2015) 6169.

[26] M.A. Little, et al., Using and understanding cross-validation strategies. Perspectives on Saeb et al., Gigascience 6 (5) (2017) 1–6.
[27] P. Geeleher, et al., Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines, Genome Biol. 15 (3) (2014) R47 R47-R47.
[28] B.M. Robert, et al., Predicting tumor sensitivity to chemotherapeutic drugs in oral squamous cell carcinoma patients, Sci. Rep. 8 (1) (2018) 15545.
[29] I.K. Robert, R in Action, Data Analysis and Graphics With R, Manning Publications Co, 2011 ISBN: 9781935182399<.
[30] Rodolfo, et al., Transfer and multi-task learning in QSAR modeling: advances and challenges, Front. Pharmacol. 9 (2018) 74.
[31] K. Roy, et al., Be aware of error measures. Further studies on validation of predictive QSAR models, Chemom. Intell. Lab. Syst. 152 (15) (2016) 18–33.
[32] Z. Safikhani, et al., Revisiting inconsistency in large pharmacogenomic studies, F1000*Research* 5 (2017) 2333.
[33] Y. Sakamoto, et al., Akaike Information Criterion Statistics, D. Reidel Publishing Company, 1986.
[34] Cancer R. Sankaranarayanan, K. Ramadas, H. Amarasinghe, S. Subramanian, N. Johnson, Oral cancer: prevention, early detection, and treatment, in: H. Gelband, P. Jha, R. Sankaranarayanan, S. Horton (Eds.), Disease Control Priorities, 3 World Bank, Washington, DC, 2015. Cancer.
[35] A. Schneider, et al., Linear regression analysis: part 14 of a series on evaluation of scientific publications, Deutsches Ärzteblatt Int. 107 (44) (2010) 776–782.
[36] S. Ahmed, et al., Pharmacogenomics of drug metabolizing enzymes and Transporters: relevance to precision medicine, Genom. Proteom. Bioinformat. 14 (5) (2016) 298–313.
[37] S. Da, et al., Recurrent oral cancer: current and emerging therapeutic approaches, Front. Pharmacol. 3 (2012) 149.
[38] S. Wang, et al., CellMiner Companion: an interactive web application to explore CellMiner NCI-60 data, Bioinformatics 32 (15) (2016) 2399–2401.
[39] S. Sharma, et al., Oral cancer statistics in India on the basis of first report of 29 population-based cancer registries, J. Oral Maxillofac. Pathol. 22 (1) (2018) 18–26.
[40] S. Weisberg, Computing Primer for Applied Linear Regression, Using R, 2005.
[41] P.P. William, J.C. Cheville, I. Frank, H.B. Zaid, C.M. Lohse, S.A. Boorjian, B.C. Leibovich, R. Houston Thompson, Application of the Stage, Size, Grade, and Necrosis (SSIGN) score for clear cell renal cell carcinoma in contemporary patients, Eur. Urol. 71 (4) (2017) 665–673.
[42] E. William, et al., BH3 profiling discriminates response to cytarabine-based treatment of acute myelogenous leukemia, Mol. Cancer Ther. 12 (12) (2013).
[43] Liu Xiaoming, et al., A systematic study on drug-response associated genes using baseline gene expressions of the Cancer Cell Line Encyclopedia, Sci. Rep. 6 (1) (2016) 22811.
[44] W. Yang, et al., Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells, Nucleic Acids Res. 41 (Database issue) (2013) D955–D961.
[45] F. Yang-Wallentin, K.G. Jöreskog, H. Luo, Confirmatory factor analysis of ordinal variables with misspecified models, Struct. Equ. Model. 17 (3) (2010) 392–423.
[46] Y. Chang, et al., Cancer Drug Response Profile scan (CDRscan): a deep learning model that predicts drug effectiveness from cancer genomic signature, Sci Rep. 8 (1) (2018) 8857.
[47] K. Zhang, et al., Sparse multitask regression for identifying the common mechanism of response to therapeutic targets, Bioinformatics 26 (12) (2010) i97–i105.