Original Paper

Clinical Drug Response Prediction by Using a Lq Penalized Network-Constrained Logistic Regression Method

Hai-Hui Huang^{a,b} Jing-Guo Dai^a Yong-Liang^b

^aSchool of Information Science and Engineering & Provincial Demonstration Software Institute, Shaoguan University, Shaoguan, ^bFaculty of Information Technology & State Key Laboratory of Quality Research in Chinese Medicines, Macau University of Science and Technology, Macau, China

Key Words

Personalized medicine • Drug response prediction • Regularization • Variable selection

Abstract

Background/Aims: One of the most important impacts of personalized medicine is the connection between patients' genotypes and their drug responses. Despite a series of studies exploring this relationship, the predictive ability of such analyses still needs to be strengthened. **Methods:** Here we present the Lq penalized network-constrained logistic regression (Lq-NLR) method to meet this need, in which the predictors are integrated into the gene expression data and biological network knowledge and are combined with a more aggressive penalty function. Response prediction models for two cancer targeting drugs (erlotinib and sorafenib) were developed from gene expression data and IC50 values from a large panel of cancer cell lines by utilizing the proposed approach. Then the drug responders were tested with the baseline tumor gene expression data, yielding an *in vivo* drug sensitivity prediction. **Results:** These results demonstrated the high effectiveness of this approach. One of the best results achieved by our method was a correlation of 0.841 between the cell line *in vitro* drug response and patient's *in vivo* drug response. We then applied these two drug prediction models to develop a personalized medicine approach in which the subsequent treatment depends on each patient's gene-expression profile. **Conclusion:** The proposed method is much better than the existing approach and can capture a more accurate reflection of the relationship between genotypes and phenotypes.

© 2018 The Author(s)
Published by S. Karger AG, Basel

Hai-Hui Huang School of Information Science and Engineering & Provincial Demonstration Software Institute, Shaoguan University, Shaoguan; Faculty of Information Technology & State Key Laboratory of Quality Research in Chinese Medicines, Macau University of Science and Technology, Macau (China); E-Mail tomyhwang@163.com

Introduction

Personalized medicine (PM), the use of DNA- or protein-based diagnosis and subsequent treatment decisions based on an individual's molecular profile, will influence the way medicine is practiced. PM promises us individually tailored medicine and health care, and this development is actively gaining momentum.

With advances in high-throughput genetic profiling techniques, high-throughput sequencing has become cheaper and more available. There are a growing number of different popular molecular resources, for example, GEO [1], Cancer Genome Project (CGP) [2], BioGrid [3], and KEGG [4]. Each of these are individually providing insights into how cells work from a system-level perspective with a high level of molecular detail [5]. Integrating all this valuable knowledge together in an appropriate way could provide useful prior information for detecting noise and removing confounding factors from biological data [6, 7]. Moreover, this can lead to the discovery of a stable link between genes, proteins, drugs, and disease. Among these relationships, one is particularly important to PM (Fig. 1). That is the link between each patient's genotype and drug response. However, it is hard to extract this relationship from humans directly for practical or ethical reasons. Some interesting research has been suggested to address this problem. For example, the CGP project used baseline gene expression microarray data and pharmacological profiles of 138 anticancer drugs in a panel of over 700 human cancer cell lines to identify the predictive biomarkers of drug responses by applying the elastic net (or $L_1 + L_2$ sparse method) model. Menden et al. constructed models that combined genomic features of cell lines with chemical properties of drugs and used a neural network to predict *in vitro* drug sensitivity [8]. Geeleher et al. used the Ridge regression (or L_2 sparse method) model to capture the relationship between each cell line's gene expression variation and drug response by analysing the same CGP dataset [9]. Zhang et al. recommend a multi-layer combined cell line-drug network method, which uses both a cell line similarity network and a drug similarity network to predict the drug response of a given cell line using a weighted model [10]. Chen and Zhang employed a L_1 network-regularized partial least squares method to capture the nonlinear relationships between drug responses and cell line genomic features [11]. Even though the above methods were quite interesting, gaps remain in the utility of these tools in predicting drug responses.

In this study, we proposed the L_q ($0 < q < 1$) penalized network-constrained strategy, where predictors are based on gene expression data with biological network knowledge. The L_q ($0 < q < 1$) penalty is often achieved with more sparsity and enjoys better accuracy

Fig. 1. Genotype–drug relationships. The figure shows the relationship between each patient's baseline tumor gene expression and effect of various drugs. This network suggests appropriate drug selection for each patient. Edges between patients and medicines are colored based on higher (red) to lower (green) drug sensitivity.

Fig. 2. Overviews of predicting *in vivo* drug response by using our method. The cell lines and clinical microarray data are (1a) combined and homogenized using ComBat. (1b) Then, the gene expression data of cell lines is integrated with a biological interaction network. (2) Subsequently, we use the Lq-NLR regression method to fit the integrated data against the *in vitro* drug IC50 estimate to build the Lq-NLR drug response predictor. (3) This model is next applied to the microarray data from the clinical trial to yield drug sensitivity estimates. Lq-NLR is the Lq penalized network-constrained logistic regression.

than the L_1 , L_2 or $L_1 + L_2$ methods [12–15]. The network-constrained strategy integrates biological prior interaction knowledge, like gene or protein regulatory networks with gene expression data. Such a strategy can identify biomarkers that only have weak effects but often play a significant role in diverse biological processes [6, 7, 16, 17]. We then applied the proposed method to a logistic regression model to build two drug response predictors, erlotinib and sorafenib for non-small cell lung cancer (NSCLC). The drug response predictors were developed from CGP data, which contain cell line genome data and recorded responses to 138 drugs across nearly 700 cancer cell lines (Fig. 2). To test our approach, we identified clinical trial datasets that had assessed tumor gene expression before drug treatment (using expression microarrays) and had subsequently measured a clear drug-response phenotype. Using these data, we can test whether our models derived from cell lines capture a significant proportion of the variability in drug responses among patients.

Materials and Methods

Data Preparation

The cell line gene expression data can be downloaded from ArrayExpress (accession number E-MTAB-783). The IC50 values for the drugs are given in the Cancer Genome Project (CGP). The erlotinib and sorafenib data are available from GEO series accession number GSE33072 [18].

We downloaded the biological interaction (gene, protein, or cancer-specific pathways) network from the BioGrid (<https://thebiogrid.org>) and KEGG (<http://www.genome.jp/kegg>). The network is comprised of 14,355 genes or proteins and 507,757 interactions.

Data Process

Following Geleher et al. [9], we processed raw CGP gene expression microarray data with background correction, quantile normalization, and median polish summarization. CGP cell lines (treated with erlotinib or sorafenib) were divided into two groups that were the most responsive (labeled 1) and most resistant (labeled 0). Then, training samples for erlotinib included sensitive and resistant individuals ($n = 15$ and $n = 55$, respectively). Training samples for sorafenib also consisted of sensitive and resistant individuals ($n = 15$ and $n = 56$, respectively). We use every cancer type cell line that was treated with erlotinib or sorafenib instead of using only lung cancer type cell lines because of the small sample size for the latter.

Cell line data and clinical trial data (GSE33072) were mapped to a unique official gene symbol. We summarized multiple probes sets that mapped to the same gene by their mean expression value. We used the subset of common genes represented among them. Next, these datasets were homogenized with ComBat (a function in *sva*, R), and among both erlotinib and sorafenib data, 11, 884 genes were remaining.

Lq penalized network-constrained logistic regression method

Gene selection is a major problem in genomic research. The regularization method is one of the popular methods for gene selection [19]. In general, a common framework of the regularization in machine learning has the for

$$\beta^* = \arg\min_{\beta} \{R(\beta) + \lambda P(\beta)\} \quad (1),$$

in which $R(\beta)$ is the loss function and $P(\beta)$ is the regularization term, and λ is a tuning parameter. A common regularization term is the L_1 (Lasso) method [20], which has a penalty function $P(\beta) = \sum |\beta|$. Lasso and its variants (or the Lasso type algorithms), such as Elastic net [21], SCAD [22], MCP [23], adaptive Lasso [24], and stage-wise Lasso [25] have been extensively studied and applied in recent years in the fields of statistics and machine learning owing to their convex optimization problems that can be easily solved. However, the L_1 type penalty may yield insufficient sparsity for biomarker selection. Furthermore, the penalty may not consistently select the correct model and is not necessarily asymptotically normal [26, 27]. It is well known that L_0 regularization where $P(\beta) = \sum |\beta|^0$ is ideally sparsest for variable selection. Unfortunately, this is an NP-hard problem. One of the alternatives for gaining a better result is the L_q ($0 < q < 1$) penalty, where $P(\beta) = \sum |\beta|^q$. Smaller q values result in sparser and better solutions found through L_q regularization [27].

To this point, we observed dense molecular interplay records of disease-associated biological processes and accumulating them through databases focused on many aspects of biological structures. For example, BioGRID investigated over 44, 000 publications to collect diverse biological interactions. These relationships are typically represented as a network. Integrating such networks extracted from biological processes with an analysis of the gene expression data could offer beneficial combinational signals to increase biomarker identification accuracy for several classification and regression models. For instance, Li and Li developed a network-constrained regularization procedure to enable variable selection and regression analysis for genomic data [28]. Zhang et al. proposed graph Laplacian regularized logistic regression to integrate biological networks into disease classification and pathway association problems [29]. Chen and Zhang incorporated a molecular network into the partial least squares model to identify common modular patterns using large-scale pairwise gene-expression and drug-response data [11]. In these studies, the network-based penalty is defined in the same way as a quadratic form of the Laplacian matrix associated with the interaction network.

Inspired by the ideas above, we introduce a L_q penalized network-constrained logistic regression (L_q -NLR) method, in which the predictors are integrated into the gene expression data and biological network knowledge. The proposed method is aimed at identifying some more accurate biomarkers of diseases or drug responses.

Suppose the predictor $X_i = (x_{i1}, x_{i2}, \dots, x_{ip})$, in which the i^{th} sample with p genes and dependent variable y_i that takes a value of 0 or 1, then the L_q -NLR method can be formulated as follows:

$$L(\lambda_1, \lambda_2, \beta) = -\sum_i \{y_i \log [f(X_i' \beta)] + (1 - y_i) \log [1 - f(X_i' \beta)]\} + \lambda_1 |\beta|_q + \lambda_2 \beta' L \beta. \quad (2)$$

Equation (2) is composed of three key terms. The first one is the logistic regression function, where the function is a classical method for classification. The second one is the L_q penalty, that is $|\beta|_q = \sum |\beta_j|^q$, where the sparsity of the variable β is enforced such that the results will have a better biological interpretation. The last term captures the key prior knowledge, which ensures the connected genes in the network are likely to be placed in the same module or result. The L is the normalized Laplacian matrix defined as

$$L = D^{(-1/2)}(D - A)D^{(-1/2)} = I - D^{(-1/2)}AD^{(-1/2)},$$

where $A = (a_{ij})_{p \times p}$ is the binary (or weight) adjacency matrix of the gene-gene interaction network G , a_{ij} equals 1 or a value ranging from 0 to 1 if genes i and j are linked in the network and $a_{ij} = 0$ otherwise; $D = (d_{ii})_{p \times p}$ is the degree matrix of graph G , where $d_{ii} = a_{i1} + a_{i2} + \dots + a_{ip}$, and $d_{ij} = 0$, for $i \neq j$. The tuning parameters λ_1 and λ_2 control the amount of regularization for sparsity and smoothness, respectively.

Algorithm for the Lq penalized network-constrained logistic regression method

We adopted the generalized soft-thresholding (GST) function algorithm [30] combined with the coordinate descent algorithm (CDA) [31] to solve the Lq-NLR model. The GST is the global solution projection of Lq which is theoretically more robust and more efficient to implement, and it can converge to a more accurate solution compared with the state-of-the-art methods, like the iteratively reweighted least squares method [32] and look-up table method [33]. The idea of the coordinate descent algorithm is to find local optima of a multivariate optimization problem by solving a sequence of univariate optimization problems.

To beginning with the solution, we first need to transform the Lq-NLR model into a linear model. Following the Li and Li [28] approach, Equation (2) can be expressed as

$$L(\lambda_1, \lambda_2, \beta) = L(\gamma, \beta) = -\sum_i \{y_i^* \log [f(X_i^* \beta)] + (1 - y_i^*) \log [1 - f(X_i^* \beta)]\} + \gamma |\beta|^q \quad (3)$$

where $X_{(n+p) \times p}^* = (1 + \lambda_2)^{(1/2)} [X; \lambda_2^{(1/2)} K']$, which because the Laplacian matrix L is non-negative definite, can be expressed as $L = KK'$ by Cholesky decomposition, $Y_{(n+p)}^* = [Y; 0]$, $\beta^* = (1 + \lambda_2)^{(1/2)} \beta$ and γ is the regularization parameter, which can be expressed as $\gamma = \lambda_1 (1 + \lambda_2)^{(1/(2q) - 1/2)}$.

By one-term Taylor series expansion, Equation (3) can be written as:

$$L(\gamma, \beta) \approx (1/2n) \sum_i (Q_i - X_i^* \beta)' W_i (Q_i - X_i^* \beta) + \gamma |\beta|^q, \quad (4)$$

where $Q_i = X_i^* \beta^* + (y_i^* - f(X_i^* \beta^*)) / f(X_i^* \beta^*) (1 - f(X_i^* \beta^*))$ is the estimated response and $W_i = f(X_i^* \beta^*) (1 - f(X_i^* \beta^*))$ is the weight for the estimated response. Additionally, $f(X_i^* \beta^*) = \exp(X_i^* \beta^*) / (1 + \exp(X_i^* \beta^*))$ is the evaluated value under the current parameters. Thus, we can redefine the partial residual for fitting the current β^* as $Q_i^{-j} = \sum_{k \neq j} x_{ik} \beta_k^*$ and $\omega_j = \sum_i W_i x_{ij}^* (Q_i - Q_i^{-j})$.

Then, the algorithm for the Lq-NLR model is:

GST function algorithm: $\beta_j = GST(\omega_j, \lambda, q, E)$

if $|\omega_j| \leq (2\lambda(1-q))^{1/(2-q)} + \lambda q (2\lambda(1-q))^{(q-1)/(2-q)}$

$\beta_j = 0$

else

$i = 0, x(i) = |\omega_j|$

Iterate on $i = 0, 1, \dots, E$

$x(i+1) = |\omega_j| - \lambda q (x(i))^{q-1}$

$i = i + 1$

$\beta_j = \text{sgn}(\omega_j) x(i)$

CD algorithm: $\beta_j = CD(X^*, y^*, \omega_j, \gamma, q, E)$

Step 1: Initialize all $\beta_j(m) = 0$ ($j = 1, 2, \dots, p$), y^* and X^* ,

set $m = 0$, γ and q are chosen by cross-validation, E ;

Step 2: Calculate $Q(m)$ and $W(m)$ and approximate the loss function (4) based on the current $\beta(m)$;

Step 3: Update each $\beta_j(m)$, and cycle over $j = 1, \dots, p$

Step 3.1: Compute:

$Q_i^{-j}(m) = \sum_{k \neq j} x_{ik} \beta_k(m)$ and $\omega_j(m) = \sum_i [W_i x_{ij}^* (Q_i(m) - Q_i^{-j}(m))]$;

Step 3.2: Update $\beta_j(m+1) = GST(\omega_j(m), \gamma, q, E)$;

Step 4: Repeat Steps 2 and 3 until $|\beta(m+1)| - \beta(m)| < 10^{-5}$.

The development of Lq-NLR-drug sensitivity models

Overview of the model development is presented in Fig. 2. Cell line data (treated with erlotinib) was combined and homogenized with clinical trial data (GSE33072, treated erlotinib) as described in "Data Process". By integrating the gene expression data of cell lines with the prepared network, the final network L includes 8693 genes and 360,964 edges. We then use the Lq-NLR model to fit the integrated data against the *in vitro* drug response estimates (IC50), to build the Lq-NLR-erlotinib responder. For Lq-NLR-sorafenib model development, the procedure is the same setting as that for Lq-NLR-erlotinib except we used the data with the sorafenib treatment. We choose the model with the lowest 10-CV error.

Lasso, Elastic Net, Ridge, and InterNet

Lasso, Elastic Net, and Ridge were fitted using the MATLAB (2014a version) function *lassoglm*. We used ten-fold CV to tune the Lasso penalty parameter. Elastic Net and Ridge penalty parameters (alpha

and lambda) were selected using cross-validation on two-dimensional surfaces [21, 34]. InterNet was implemented using the R package *regnet* [35–38].

Results

In this study, we aim to predict clinical drug responses according to baseline tumor gene expression. Toward this end, two biomarkers-based drug response models (for erlotinib and sorafenib respectively) were developed from recently released CGP, BioGrid, and KEGG data by utilizing the L_q -NLR method. We then applied the models to a subset of clinical trial data, the Biomarker-Integrated Approaches of Targeted Therapy for Lung Cancer Elimination, for validation [18, 39]. These data are based on patients treated with erlotinib ($n = 25$), a growth factor receptor (EGFR) inhibitor, and sorafenib ($n = 37$), a vascular endothelial growth factor (VEGFR) inhibitor. This clinical trial gene expression and drug sensitivity data were obtained from GEO (GEO: GSE33072).

To develop the model, we followed the main steps outlined by Gleeher et al. [9] to deal with the cell line data and clinical trial data, such as combining, homogenizing, and dividing cell line data into sensitive or resistant groups to reduce noise in the cell line data (see Materials and methods). As a result, for erlotinib, the training data ($N = 70$) included individuals that were sensitive and resistant to erlotinib ($n = 15$ and $n = 55$, respectively). For sorafenib, the training samples ($N = 71$) included individuals who were sensitive and resistant to sorafenib ($n = 15$ and $n = 56$, respectively).

We compared our method with four state-of-the-art algorithms, including logistic regression with Lasso, Ridge, Elastic net, and interconnection network constrained (InterNet) methods [35]. Among them, the proposed method consistently performed the best, given its prevailing generalization ability, which is crucial for clinical application.

We also performed a convergence test for each method, and the convergence times amongst the network methods (L_q -NLR and InterNet) were similar. For example, for erlotinib data, it took MATLAB or R 153, 155, 2, 93, and 2 s to converge for the L_q -NLR, InterNet, Lasso, Ridge, and Elastic net methods, respectively.

Erlotinib and sorafenib response in non-small cell lung cancer

We used the L_q -NLR method to build two drug response predictors (see Tables S1 & S2 for details - For all supplemental material see www.karger.com/10.1159/000495826), L_q -NLR-erlotinib (L_q -NLR-E) contains 22 biomarkers, and L_q -NLR-sorafenib (L_q -NLR-S)

Fig. 3. Prediction of drug responses in NSCLC patients. (A) Months-to-progression plotted against the predicted index of erlotinib sensitivity. KRAS mutations were colored (red), and a linear regression line is shown. (B) The predicted index of sensitivity to sorafenib plotted against months-to-progression for patients treated with sorafenib. NSCLC, non-small cell lung cancer.

Table 1. Relationships between erlotinib response index and months-to-progression. This table describes the probability of months-to-progression treatment with erlotinib from historical clinical data (patients treated with erlotinib) combined with the results predicted by the Lq-NLR-E model

Class	Drug Response Index	n = Months			Total
		0-2	2-4	4-6	
1	≤450	100% (10/10)	- (0/10)	- (0/10)	10
2	450-500	36.4% (4/11)	54.5% (6/11)	9.1% (1/11)	11
3	>500	- (0/4)	25% (1/4)	75% (3/4)	4

Table 2. Relationships between sorafenib response index and months-to-progression. This table describes the probability of months-to-progression under treatment with sorafenib from historical clinical data (patients treated with sorafenib) integrated with the outcomes tested by the Lq-NLR-S model

Class	Drug Response Index	n = Months					Total
		0-2	2-4	4-6	6-8	8-12	
1	≤25	81.2% (13/16)	18.8% (3/16)	- (0/16)	- (0/16)	- (0/16)	16
2	25-40	25% (3/12)	50% (6/12)	16.7% (2/12)	8.3% (1/12)	- (0/12)	12
3	≥40	- (0/9)	33.3% (3/9)	11.1% (1/9)	11.1% (1/9)	44.5% (4/9)	9

includes 26 biomarkers. In the model training period, our Lq-NLR-E model achieved 98.6% classification accuracy and an AUC of 0.996 with 3.7% 3-fold CV classification error. The Lq-NLR-S model gained 97.18% accuracy and an AUC of 0.985 AUC with 4.2% 3-fold CV error.

Geeleher et al. [9] observed high variation in the *in vivo* erlotinib response (Spearman's rank correlation coefficient 0.64 with $p = 5.3e-04$). To compare our method with this result, we applied the Lq-NLR-E model to the same clinical trial data (patients treated with erlotinib, whose validation data had been combined and homogenized with cell line gene expression data), yield an *in vivo* erlotinib response against individuals' months-to-progression (Fig. 3A; $\rho = 0.841$ with $p = 1.44e-07$ from a Spearman's correlation test). This outcome was much better than the results achieved by Geeleher et al. [9]. Our Lq-NLR-E model predicted one particular patient would be the most sensitive individual in the cohort, and in fact, this patient was the most sensitive individual. A further interesting observation is that KRAS mutations are known as EGFR inhibitor resistance biomarkers. However, our result shows patients with a KRAS mutation did not strongly coincide with the conclusion. This result is consistent with those reported previously [9]. One reasonable interpretation of this phenomenon is a different type of KRAS mutation may be predicted for different results [40].

We then used the Lq-NLR-S model to fit the clinical trial data (patients treated with sorafenib and associated test data combined and homogenized with cell line training data) to predict individuals' *in vivo* sorafenib responses against the patients' months-to-progression (Fig. 3B; Spearman's correlation test, $\rho = 0.788$, $p = 2.20e-9$). It is notable that the patient with the longest months-to-progression treated with sorafenib was the one with the highest sensitivity to sorafenib as predicted by Lq-NLR-S. These results revealed the high effectiveness of the proposed method.

Table 3. Personalized medicine examples. We applied the erlotinib and sorafenib response prediction model to patients' baseline gene expression data respectively, to yield an *in vivo* drug response, coupled with the information in Tables 1 and 2, to yield the probability of months-to-progression

Patients	Erlotinib			Sorafenib				
	0-2	2-4	4-6	n = Months		4-6	6-8	8-2
GSM677324	36.4%	54.5%	9.1%	0-2	2-4	4-6	6-8	8-2
GSM677325	36.4%	54.5%	9.1%	81.3%	18.8%	-	-	-
GSM677328	36.4%	54.5%	9.1%	25.0%	50.0%	16.7%	8.3%	-
GSM677329	100.0%	-	-	81.3%	18.8%	-	-	-
GSM677330	36.4%	54.5%	9.1%	25.0%	50.0%	16.7%	8.3%	-
GSM677331	100.0%	-	-	25.0%	50.0%	16.7%	8.3%	-
GSM677332	100.0%	-	-	25.0%	50.0%	16.7%	8.3%	-
GSM677334	100.0%	-	-	33.3%	11.1%	11.1%	44.4%	-
GSM789978	100.0%	-	-	81.3%	18.8%	-	-	-
GSM789979	-	25.0%	75.0%	81.3%	18.8%	-	-	-
GSM789981	-	25.0%	75.0%	-	33.3%	11.1%	11.1%	44.4%
GSM789983	36.4%	54.5%	9.1%	25.0%	50.0%	16.7%	8.3%	-
GSM789986	100.0%	-	-	25.0%	50.0%	16.7%	8.3%	-
GSM789987	100.0%	-	-	81.3%	18.8%	-	-	-
GSM790019	-	25.0%	75.0%	81.3%	18.8%	-	-	-

A personalized medicine application

The information visualized in Fig. 3 is also shown in Tables 1 and 2. Table 1 is the summary of Fig. 3A (erlotinib), while Table 2 is the outline of Fig. 3B (sorafenib). These two tables describe the probability of months-to-progression treated with a specific drug from historical clinical data, combined with the outcomes predicted by our drug response prediction models. For example, in Table 1, we divided the months-to-progression into three categories (i.e., 0–2, 2–4, 4–6), and the predicted erlotinib sensitivity index is divided into three classes (i.e., 1, 2, and 3). The last column indicates the number of people within a certain sensitivity index class. For example, in class 3 (>500), there are four people with erlotinib sensitivity indexes that are over 500 as predicted by the Lq-NLR-E model. In the real clinical setting, among these four patients, one had recurrence at 2–4 months and the other three had recurrence at 4–6 months (see Fig. 3A). Therefore, the corresponding results are 25% (1/4) and 75% (3/4) respectively. These results imply that if an NSCLC patient is in erlotinib sensitivity index class 3, as predicted by the model, the patient is predicted to have a 100% chance of progression occurring within 2 months.

We randomly selected 15 patients who received neither treatment with erlotinib nor sorafenib from GSE33072. We then applied our two drug response prediction models to the patient's gene expression data, respectively, to yield an *in vivo* drug response to erlotinib and sorafenib (details in Supplementary Tables S4 and S5). By integrating all the known patient information (i.e., the predicted value of *in vivo* drug responses and the historical data in Tables 1 and 2), predictions can be summarized clearly (Table 3) and in a way that may aid physicians in making appropriate patient care decisions. For example, patient GSM790019 may benefit more from erlotinib rather than sorafenib treatment. GSM677334 perhaps preferred using sorafenib rather than erlotinib to extend the time of cancer recurrence. GSM789987 may resist both erlotinib and sorafenib.

Fig. 4. Gene network of genes selected by the Lq-NLR-E model from the *in vitro* erlotinib data. An edge between two nodes indicates that the two genes belong to the same pathway. Each color indicates a different pathway.

Fig. 5. Gene network of genes selected by Lq-NLR-E model from the *in vitro* sorafenib data. An edge between two nodes means that the two genes belong to the same pathway. Different colors are used for different pathways.

Brief biological analyses of the selected genes

We mapped our selected genes to the Reactome (<https://reactome.org>). For erlotinib, as shown in Fig. 4, the genes selected by the *Lq*-NLR-E model were involved in many vital pathways, such as Negative regulation of the PI3K/AKT network and Cargo recognition for clathrin-mediated endocytosis. Moreover, the genes selected by our method might be beneficial to investigators whose goal it is to look for the possible biomarkers for tumor sensitivity to erlotinib. For example, the biomarkers selected by our method include EGFR (epidermal growth factor receptor), and erlotinib itself acts on the EGFR. Therefore, the selected genes that are connected with EGFR, such as TRIB3, MET, and COPS7B may play a critical role in erlotinib-sensitive prediction. Notably, Ayllón et al. [41] determined TRIB3 and MET are correlated with erlotinib activity. SerpinB5 (serpin peptidase inhibitor 5) is a novel tumor suppressor biomarker that inhibits the growth, invasion, and metastasis of tumor cells, induces apoptosis and impedes neovascularization [42]. Thus, it is implied that SerpinB5 is a promising responder to erlotinib.

For sorafenib, as shown in Fig. 5, the genes selected by the *Lq*-NLR-S model were involved in many important pathways, such as the RAF/MAP kinase cascade and Interleukin-4 and Interleukin-13 signaling. Moreover, the biomarkers selected by our method include VEGFA (vascular endothelial growth factor A), which plays a key role in angiogenesis. Some research suggests hepatocellular carcinoma (HCC) patients with elevated levels of VEGFA are distinctly sensitive to sorafenib [43], and VEGFA can therefore be used as a biomarker for personalized treatment of HCC with sorafenib [44]. Our result implies that VEGFA may also play a critical role in NSCLC patients with sorafenib. FGF3 (fibroblast growth factor 3), is involved in many biological processes, such as cell growth, tumor growth, and invasion. Additionally, FGF3 is an effective responder to sorafenib [45].

Discussion

Discovery of responsive drug biomarkers from pre-clinical data is a critical step in drug development and personalized medicine. To identify some more accurate biomarkers regarding diseases or drug responses, in this study, we proposed the *Lq*-NLR method, where the predictors are integrated into the gene-expression data and biological network knowledge, and embed a more aggressive penalty function. We used our method to build two cancer-targeted drug response prediction models. The drug responders built using gene expression data and IC50 values from a large panel of cancer cell lines that were treated with a specific drug. These results demonstrated the high effectiveness of the approach.

We showed strong relationships between drug responses, gene expression changes in cell lines, and patient baselines. The results of our method are quite promising (Fig. 3, erlotinib $\rho = 0.841$ with $p = 1.44\text{e-}07$; sorafenib, $\rho = 0.788$ with $p = 2.20\text{e-}9$ from a Spearman's correlation test). In the similar study conducted by Gleeleher et al. [9], the authors utilized the ridge logistic regression model to build the drug responder. The performance of our method is 24.5% (88.5% versus 64%) higher than the authors' approach to assessing erlotinib. It may seem surprising that the established approach has such strong performance in drug response prediction. This effectiveness may stem from the synergistic action of the *Lq* penalty function and the network-constrained function, which combine to capture a more accurate and specific assessment of genome variation.

In personalized medicine, the choice of therapy depends on each patient's gene expression profile instead of primarily relying on the cancer subtypes identified. This idea holds promise as it enables the screening of any chemotherapeutic compound against a panel of cell lines, coupled with methods that adjust the intrinsic gap in gene expression between cell lines and tumor biopsies, thus simplifying treatment decisions after disease diagnosis [46].

Conclusion

In this study, we proposed a novel biomarker selection method. We constructed two biomarker-based models by our method to aid in a personal treatment selection (i.e., erlotinib versus sorafenib). The proposed method outperforms several existing approaches (see Table S3). These findings advance the clinical application of personalized medicine. Prospective work will focus on blood-based biomarker research for early lung cancer detection and further verification in future clinical trials.

Abbreviations

CGP (Cancer Genome Project); NSCLC (non-small cell lung cancer); *Lq*-NLR (*Lq* penalized network-constrained logistic regression); *Lq*-NLR-E (*Lq*-NLR-erlotinib); *Lq*-NLR-S (*Lq*-NLR-sorafenib).

Acknowledgements

This research was supported by National Natural Science Foundation of Guangdong (2018A030307033), MOE (Ministry of Education in China) Project of Humanities and Social Sciences (18YJCZH054), “Computer Science and Technology” Key Breeding Discipline of Guangdong Province 2016, Research Fund of Shaoguan University (SY2017KJ05), High-level College’s Talent Project of Guangdong (2013-178), and Macau Science and Technology Development Funds (003/2016/AFJ) of Macau SAR of China.

Hai-Hui Huang developed the method, performed the analysis and drafted the article. Yong-Liang and Jing-Guo Dai initiated, and supervised the project. All authors have read the final version of the article and have approved submission of the manuscript to “Cellular Physiology and Biochemistry”.

Disclosure Statement

The authors declare to have no competing interests.

References

- 1 Edgar R, Domrachev M, Lash AE: Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* 2002;30:207–210.
- 2 Garnett MJ, Edelman EJ, Heidorn SJ, Greenman CD, Dastur A, Lau KW, Greninger P, Thompson I R, Luo X, Soares J, Liu Q, Iorio F, Surdez D, Chen L, Milano RJ, Bignell GR, Tam A T, Davies H, Stevenson JA, Barthorpe S et al.: Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* 2012;483:570–575.
- 3 Chatr-Aryamontri A, Oughtred R, Boucher L, Rust J, Chang C, Kolas NK, O'Donnell L, Oster S, Theesfeld C, Sellam A, Stark C, Breitkreutz B-J, Dolinski K, Tyers M: The BioGRID interaction database: 2017 update. *Nucleic Acids Res* 2017;45:D369–D379.
- 4 Kanehisa M, Goto S: KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 2000;28:27–30.
- 5 Ma'ayan A, Rouillard AD, Clark NR, Wang Z, Duan Q, Kou Y: Lean Big Data integration in systems biology and systems pharmacology. *Trends Pharmacol Sci* 2014;35:450–460.
- 6 Tian Z, Hwang T, Kuang R: A hypergraph-based learning algorithm for classifying gene expression and arrayCGH data with prior knowledge. *Bioinformatics* 2009;25:2831–2838.

- 7 Huang H-H, Liang Y, Liu X-Y: Network-based logistic classification with an enhanced $l_{1/2}$ solver reveals biomarker and subnetwork signatures for diagnosing lung cancer. *Biomed Res Int* 2015;2015:713953.
- 8 Menden M P, Iorio F, Garnett M, McDermott U, Benes CH, Ballester PJ, Saez-Rodriguez J: Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. *PLoS One* 2013;8:e61318.
- 9 Geeleher P, Cox NJ, Huang RS: Clinical drug response can be predicted using baseline gene expression levels and *in vitro* drug sensitivity in cell lines. *Genome Biol* 2014;15:R47.
- 10 Zhang N, Wang H, Fang Y, Wang J, Zheng X, Liu XS: Predicting anticancer drug responses using a dual-layer integrated cell line-drug network model. *PLOS Comput Biol* 2015;11:e1004498.
- 11 Chen J, Zhang S, C. S: Integrative analysis for identifying joint modular patterns of gene-expression and drug-response data. *Bioinformatics* 2016;32:1724–1732.
- 12 Liu Y, Helen Zhang H, Park C, Ahn J: Support vector machines with adaptive penalty. *Comput Stat Data Anal* 2007;51:6380–6394.
- 13 Liang Y, Liu C, Luan XZ, Leung KS, Chan TM, Xu ZB, Zhang H: Sparse logistic regression with a $L_{1/2}$ penalty for gene selection in cancer classification. *BMC Bioinformatics* 2013;14:198.
- 14 Huang H-H, Liu X-Y, Liang Y, Chai H, Xia L-Y: Identification of 13 blood-based gene expression signatures to accurately distinguish tuberculosis from other pulmonary diseases and healthy controls. *Biomed Mater Eng* 2015;26 Suppl 1:S1837-1843.
- 15 Huang H-H, Liu X-Y, Liang Y: Feature Selection and Cancer Classification via Sparse Logistic Regression with the Hybrid $L_{1/2} + 2$ Regularization. *PLoS One* 2016;11:e0149675.
- 16 Avey S, Mohanty S, Wilson J, Zapata H, Joshi SR, Siconolfi B, Tsang S, Shaw AC, Kleinstein SH: Multiple network-constrained regressions expand insights into influenza vaccination responses. *Bioinformatics* 2017;33:i208–i216.
- 17 Huang HH, Liu XY, Li HM, Liang Y: Molecular pathway identification using a new $L_{1/2}$ solver and biological network-constrained mode. *Int J Data Min Bioinform* 2017;17:189.
- 18 Byers LA, Diao L, Wang J, Saintigny P, Girard L, Peyton M, Shen L, Fan Y, Giri U, Tumula PK, Nilsson MB, Gudikote J, Tran H, Cardnell RJG, Bearss DJ, Warner SL, Foulks JM, Kanner SB, Gandhi V, Krett N et al.: An epithelial–mesenchymal transition gene signature predicts resistance to egfr and PI3K inhibitors and identifies axl as a therapeutic target for overcoming EGFR inhibitor resistance. *Clin Cancer Res* 2013;19:279–290.
- 19 Huang H-H, Liang Y: Hybrid $L_{1/2} + 2$ method for gene selection in the Cox proportional hazards model. *Comput Methods Programs Biomed* 2018;164:65–73.
- 20 Tibshirani R: Regression shrinkage and selection via the lasso. *J R Stat Soc B* 1996;267–288.
- 21 Zou H, Hastie T: Regularization and variable selection via the elastic net. *J R Stat Soc Ser B (Statistical Methodol)* 2005;67:301–320.
- 22 Fan J, Li R: Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *J Am Stat Assoc* 2001;96:1348–1360.
- 23 Zhang C-H: Nearly unbiased variable selection under minimax concave penalty. *Ann Stat* 2010;38:894–942.
- 24 Zou H: The adaptive lasso and its oracle properties. *J Am Stat Assoc* 2006;101:1418–1429.
- 25 Zhao P, Yu B: Stagewise lasso. *J Mach Learn Res* 2007;8:2701–2726.
- 26 Knight K, Fu W: Asymptotics for lasso-type estimators. *Ann Stat* 2000;1356–1378.
- 27 Malioutov D, Cetin M, Willsky AS: A sparse signal reconstruction perspective for source localization with sensor arrays. *Signal Process IEEE Trans* 2005;53:3010–3022.
- 28 Li C, Li H: Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics* 2008;24:1175–1182.
- 29 Zhang W, Wan YW, Allen GI, Pang K, Anderson ML, Liu Z: Molecular pathway identification using biological network-regularized logistic models. *BMC Genomics* 2013;14 Suppl 8:S7–2164–14–S8–S7. Epub 2013 Dec 9.
- 30 Zuo W, Meng D, Zhang L, Feng X, Zhang D: A Generalized Iterated Shrinkage Algorithm for Non-convex Sparse Coding 2013;217–224.
- 31 Friedman J, Hastie T, Tibshirani R: Regularization paths for generalized linear models via coordinate descent. *J Stat Softw* 2010;33:1–22.
- 32 Chartrand R, Wotao Yin: Iteratively reweighted algorithms for compressive sensing; in : 2008 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2008, pp 3869–3872.

- 33 Krishnan D, Fergus R: Fast Image Deconvolution Using hyper-Laplacian Priors; in : Proceedings of the 22Nd International Conference on Neural Information Processing Systems. USA, Curran Associates Inc., 2009, pp 1033–1041.
- 34 Peng X, Yang Y: Algorithms for interval-valued fuzzy soft sets in stochastic multi-criteria decision making based on regret theory and prospect theory with combined weight. *Appl Soft Comput* 2017;54:415–430.
- 35 Ren J, He T, Li Y, Liu S, Du Y, Jiang Y, Wu C: Network-based regularization for high dimensional SNP data in the case-control study of Type 2 diabetes. *BMC Genet* 2017;18:44.
- 36 Wu C, Ma S: A selective review of robust variable selection with applications in bioinformatics. *Brief Bioinform* 2015;16:873–883.
- 37 Wu C, Shi X, Cui Y, Ma S: A penalized robust semiparametric approach for gene-environment interactions. *Stat Med* 2015;34:4016–4030.
- 38 Wu C, Jiang Y, Ren J, Cui Y, Ma S: Dissecting gene-environment interactions: A penalized robust approach accounting for hierarchical structures. *Stat Med* 2018;37:437–456.
- 39 Kim ES, Herbst RS, Wistuba II, Lee JJ, Blumenschein GR, Tsao A, Stewart DJ, Hicks ME, Erasmus J, Gupta S, Alden CM, Liu S, Tang X, Khuri FR, Tran HT, Johnson BE, Heymach JV, Mao L, Fossella F, Kies MS, Papadimitrakopoulou V, Davis SE, Lippman SM, Hong WK: The battle trial: personalizing therapy for lung cancer. *Cancer Discov* 2011;1:44–53.
- 40 Blumenschein GR, Saintigny P, Liu S, Kim ES, Tsao AS, Herbst RS, Alden C, Lee JJ, Tang X, Stewart D J, Kies MS, Fossella FV, Tran HT, Mao L, Hicks ME, Erasmus J, Gupta S, Girard L, Peyton M, Diao L et al.: Comprehensive Biomarker Analysis and Final Efficacy Results of Sorafenib in the BATTLE Trial. *Clin Cancer Res* 2013;19:6967–6975.
- 41 López-Ayllón BD, De Castro-Carpeño J, Rodríguez C, Pernía O, Ibañez De Cáceres I, Belda-Iniesta C, Perona R, Sastre L: Biomarkers of erlotinib response in non-small cell lung cancer tumors that do not harbor the more common epidermal growth factor receptor mutations. *Int J Clin Exp Pathol* 2015;8:2888–2898.
- 42 Mccarroll JA, Gan PP, Erlich RB, Liu M, Dwarte T, Sagnella SS, Akerfeldt MC, Yang L, Parker AL, Chang MH, Shum MS, Byrne FL, Kavallaris M: TUBB3/ β III-tubulin acts through the PTEN/AKT signaling axis to promote tumorigenesis and anoikis resistance in non-small cell lung cancer. *Cancer Res* 2015;75:415–425.
- 43 Llovet JM: Focal gains of VEGFA: candidate predictors of sorafenib response in hepatocellular carcinoma. *Cancer Cell* 2014;25:560–562.
- 44 Horwitz E, Stein I, Andreozzi M, Nemeth J, Shoham A, Pappo O, Schweitzer N, Tornillo L, Kanarek N, Quagliata L, Zreik F, Porat RM, Finkelstein R, Reuter H, Koschny R, Ganten T, Mogler C, Shibolet O, Hess J, Breuhahn K et al.: Human and mouse VEGFA-amplified hepatocellular carcinomas are highly sensitive to sorafenib treatment. *Cancer Discov* 2014;4:730–7fig43.
- 45 Arai T, Ueshima K, Matsumoto K, Nagai T, Kimura H, Hagiwara S, Sakurai T, Haji S, Kanazawa A, Hidaka H, Iso Y, Kubota K, Shimada M, Utsunomiya T, Hirooka M, Hiasa Y, Toyoki Y, Hakamada K, Yasui K, Kumada T et al.: FGF3/FGF4 amplification and multiple lung metastases in responders to sorafenib in hepatocellular carcinoma. *Hepatology* 2013;57:1407–1415.
- 46 Zhang CH, Zhang YP: Maximizing the commercial value of personalized therapeutics and companion diagnostics. *Nat Biotechnol* 2013;31:803–805.