Life Science Informatics Publications

Research Journal of Life Sciences, Bioinformatics, Pharmaceutical and Chemical Sciences

Journal Home page http://www.rjlbpcs.com/

RJLBPCS

ISSN 2454-6348

Original Research Article **DOI:10.26479/2019.0501.52**

A QUANTITATIVE PREDICTIVE MODELLING FOR INHIBITION OF TRANSGLUTAMINASE-2 USING PENALIZED REGRESSION APPROACH

Prachi P. Parvatikar*, Shivkumar B. Madagi

Department of Bioinformatics, Akkamahadevi Women's University

(Formerly; Karnataka State Women's University), Vijayapura, Karnataka, India.

ABSTRACT: Screening potential protein targets and designing novel inhibitors is a major hurdle in drug development pipeline consuming enormous time and cost. Transglutaminase 2 (TG2) is a vital target expressed in malignant cells, especially during lung cancer. The biochemical pathway of TG2 is crucial for apoptosis in humans. In this view of its impact, identification and prediction of novel drug candidates against TG2 may provide valuable insights for combating tumor growth. This study focuses on screening novel drug molecules engrossed in three classes of TG2 inhibitors. Molecular descriptors are generated for each inhibitor based on their physicochemical characteristics. After feature reduction, seven molecular descriptors are selected for designing Quantitative Structure Activity Relationship(QSAR) models. Four different regression models are developed for the dataset, with elastic net regularization model yielding better performance by demonstrating Mean Square Error value(MSE) of 0.36656, followed by ridge model with MSE of 2.4014. The performance of QSAR models is further evaluated by cross-validation and statistical parameters like RSS and error analysis. The results reflect that regression models can predict the relationship between drug descriptors and their activity for TG2 ligands. This model so developed can be implemented into a data driven system for identifying novel anti-TG2 molecules.

KEYWORDS: TG2, molecular descriptors, QSAR, regression, elastic net.

Corresponding Author: Mrs. Prachi P. Parvatikar*

Department of Bioinformatics, Akkamahadevi Women's University

(Formerly; Karnataka State Women's University), Vijayapura, Karnataka, India.

Email Address:prachisandeepk@gmail.com

1.INTRODUCTION

Lung cancer has become one of the prominent threats to global healthcare, with a death rate of one out of four deaths in both men and women. According to American Cancer Society, lung cancer contributes around 14% of new cancer prototypes, with the incidence of 222,500 new cases and 155, 870 deaths in the United States [1]. Consequently, in India, the incidence of lung cancer has increased by around 11.3% of new cancer cases prototypes and by contributing to 13.7% of total deaths caused by cancer [2]. Depending on tumor size and stages of lung cancer several therapeutic regimens are available to treat lung cancer. Yet conventional treatments like chemotherapy and radiation therapy are associated with several adverse effects on the human system. It thereby becomes necessary to understand the pathology behind the diseases states for providing better insights to improve diagnosis. In this direction, identification of novel therapeutic targets expressed in lung cancer turns out to be an effective approach for designing better drug molecule [3]. Transglutaminase 2 (TG2) is the most widely expressed gene among the enzymes of Transglutaminase family. It is distributed among major cell types and tissues and plays a significant role in several biological processes. TG2 mediates post-translational modification of proteins by cross-linkage forming covalent bonds between lysine and glutamine groups in the Ca^{2+}dependent mechanism [4]. Activation of Ca^{2+} is regulated by GTP binding via signaling pathways. Activated Ca^{2+}, in turn, activates TG2 in intracellular environment. Up-regulation of TG2 levels within the cell can lead to malfunctioning resulting in Alzheimer's, Parkinson's disease, multiple sclerosis, celiac sprue along with tumor development. Overexpression of TG2is associated with malignant effects leading to tumor genesis, invasion, cell differentiation, and apoptosis [5]. As a consequence, TG2 is known to have a significant impact on lung carcinoma, especially towards non-small cell lung carcinoma [6]. Targeting TG2 protein may be valuable in identifying potential inhibitors against lung cancer.Several molecules are available in the literature which has potent inhibiting activities against TG2 isoforms. Some of the prominent ones are derivatives of Bromodihydroisoxazole (DHI), peptidomimetic derivatives along with natural and synthetic derivatives [7]. Consequently, new molecules are screened for their pharmacological properties to design novel drugs against TG2. Such molecules are being tested continuously under physiological conditions in pre-clinical and clinical settings. In this scenario, it is encouraging to design advanced strategies to screen molecules against TG2.Anti-cancer drugs against TG2 can be screened using experimental techniques like mass spectroscopy and chromatographic approaches. These techniques require manual intervention and are also cost-effective. Subsequently, it is also observed that there are limited TG2 inhibitor complexes derived from crystallographic studies [8]. As alternatives, *insilico* methodologies work effectively in reducing the cost incurred by experimental design. Of lately, techniques like pharmacophore mapping, molecular docking and quantitative structure activity relationship (QSAR) models are implemented for reducing the time scale in drug discovery

pipeline [9]. These approaches aid in detecting probable lead molecules based on physicochemical characteristics which in turn reduces the cost of chemical synthesis of drugs. QSAR models are constructed either based on targeted single-mechanisms based "local" models or by deploying multi-mechanism based "global" models. Local models are confined to compounds having a single mechanism of action, while global QSAR models are developed to cover compounds having a wide range of mechanism of actions [10]. Previous literature suggests that global models are dynamic and accurate when compared to local models. This study focuses on developing global QSAR models for compounds having different modes of action against the TG2 protein. The QSAR models are developed by determining the predicted pIC50 values of these compounds. The values generated from the models are compared with the experimental pIC50 values to screen drug molecules against the TG2 protein.

2. MATERIALS AND METHODS

2.1Preparation of ligand dataset

It is important to design an inhibitor that blocks specific functionality of target protein, rather than a generalized approach for all protein targets. It is commonly referred to as target-based drug discovery [11]. Keeping this in mind, the inhibitor dataset for TG2 protein is designed manually to avoid computational inaccuracies. Previous literature highlights three classes of inhibitors against the protein based on a different mechanism of actions namely competitive amine inhibitors, reversible and irreversible inhibitors. It is relevant to consider all the three categories of inhibitors, generating multiple inhibitor datasets [12]. A random sample of ten inhibitors is selected from each category of inhibitors resulting in thirty TG2 inhibitors to eliminate the probability of bias in data. The random sample is selected based on minimum IC50 value, as it is known fact that smaller amount of drug must inhibit the target protein effectively. The information from experimentally known inhibitors is obtained for each of the three subsets of inhibitors resulting in thirty inhibitors [13] . Such a dataset having multiple classes of inhibitors are generally referred to as multiple class attributes in machine learning. Global QSAR models are constructed to predict the multi-class attributes by partitioning the dataset into instances having the same class variable [14]. PubChem sketcher, online software is used for drawing the 2D structures of ligand molecules to eliminate the distorted side chains and functional groups [15]. The concentration of each of the inhibitors required in diminishing the activity of TG2 is calculated based on their pIC50 values obtained from the literature. As an extent, to standardize the units, the online measurement tool from Sanjeev's lab is utilized to convert all the pIC50 values in nanomolar (nM) concentration [16]. The dataset is prepared for each of the three subcategories of TG2 inhibitors in the same fashion.

2.2 Generation of molecular descriptors

Molecular descriptors are generated for the ligand dataset using PaDEL-descriptor software [17]. For the ligand dataset, 1D along with 2D molecular descriptors followed by fingerprints is generated.

The descriptors are generated based on different molecular characteristics such as Extended Topochemical Atoms (ETA), hydrogen bond information, functional groups, constitutional indicators, kappa shape indices, among many others.

2.3 Reduction of molecular descriptors

Some of the descriptors generate sparse entries for the dataset which are highly interrelated. It is essential to eliminate unassigned attributes from dataset to aid towards accurate model development. This procedure is often called as the curse of dimensionality in machine learning community [18]. As a measure, highly correlated attributes above a threshold of 0.75 are removed from the dataset. Further, attributes are reduced by employing principal component analysis (PCA) algorithm which minimizes the size of the dataset. Relevant features are selected based on the relative importance of each attribute in the dataset [19]. The dependency in R programming language called 'Boruta' is invoked to identify significant features. It is a wrapper based feature selection implementation which explores the importance of original attributes based on permutations [20,21].

2.4 Development of mathematical models

2D-QSAR modeling is performed to predict the interrelation between molecular descriptors and the pIC50 value for the ligand dataset. Mathematical models are developed in this study using linear regression and penalized techniques. Linear regression model identifies the relationship among the multiple independent variables (i.e. molecular descriptors) and the dependent variable (i.e. pIC50) by fitting a linear equation equivalent to the data items along with error value. Similarly, penalized regression models are applied to minimize the residual sum of squares (RSS) by introducing penalty function which maximizes likelihood and minimizes information loss. Three penalized regression techniques namely ridge, lasso and elastic net regression is applied to the ligand dataset. The models are developed in R programming language by invoking the 'glmnet' dependency [22].

2.5 Evaluation of models

The predictive performance of QSAR models is evaluated using ten-fold cross-validation metric. The original dataset is partitioned randomly into ten equal samples. Out of ten samples, nine subsamples are used for training the model, and one sample is used for the testing purpose. Cross-validation is re-iterated ten times by repeated resampling. The regression model developed after cross-validation is validated using statistical significance and goodness of fit measures defined below.

$$Residualsumofsquares(RSS) = \sum_{i=1}^{n}\left(y_i - f(x_i)\right)^2 \rightarrow (1)$$

$$MeanSquaredError(MSE) = \frac{1}{N}\sum_{i=1}^{n}\left(y_i - f(x_i)\right)^2 \rightarrow (2)$$

$$R = \frac{\sum x_i y_i - \frac{\sum x_i y_i}{N}}{\sqrt{\left(\sum {x_i}^2 - \left(\frac{\sum x_i}{N}\right)^2\right)\left(\sum {y_i}^2 - \frac{(\sum y_i)^2}{N}\right)}} \rightarrow (3)$$

Here, x_i and y_i represent the actual and predicted value of *pIC50* for the i^{th} ligand and N represents the total number of ligands in the dataset. The performance of QSAR models is evaluated on the test set to determine the accuracy of prediction [23].

3. RESULTS AND DISCUSSION

3.1 TG2 ligand dataset

The dataset of TG2 ligands is prepared for each of the three sub-categories of inhibitors namely competitive amine inhibitors, reversible inhibitors and irreversible inhibitors. Ten ligands are chosen in each category, resulting in 30 known inhibitors of TG2. Biological activities of the ligands are determined based on their pIC50 values available in the literature [24,25]. The ligands along with their pIC50 values are shown in Table 1. In order to maintain a balance between positive and negative groups, 30 ligands are chosen from literature which is known to have no inhibitory effect against TG2. The pIC50 values for these compounds were assumed to be zero for model building purpose. The dataset comprises both inhibitors and non-inhibitors of TG2. A flowchart depicting the steps performed in this study is shown in Figure 1.

Table 1: The inhibitors of TG2 identified by literature

Sl. No	Inhibitor name	Type of TG2 inhibitor	pIC50 value (nM)
1.	Putrescine	Competitive amine inhibitors	4.87
2.	Monodansylcadaverine		5.05
3.	5-(biotinamido) pentylamine		3.56
4.	6-diazo-5-oxo-*norleucine*		6.39
5.	Cystamine		8.53
6.	Dermatan sulphate		5.71
7.	Spermidine		6.32
8.	Fluorescein cadaverine		5.92
9.	Histamine		4.23
10.	Spermine		5.96
11.	GDP	Reversible inhibitors	4.49
12.	GTP		5.67
13.	GMP-PCP		4.04
14.	GTPγS		3.89
15.	LDN-27219		6.73
16.	Tyrphostin 47		7.23
17.	ZM39923		4.56
18.	ZM449829		4.97

19.	CP4d		5.05
20.	Naphthoquinone		6.33
21.	Iodoacetamide	Irreversible inhibitors	7.84
22.	3-halo-4,5-dihydroisoxazole		4.52
23.	NC9		3.67
24.	4-aminopiperidine		5.09
25.	Doxorubicin		3.76
26.	KCA075		6.05
27.	KCC009		5.66
28.	Cbz-gln(epoxide)		5.03
29.	3-bromo-4,5-dihydroisoxazole		4.98
30.	Chloroacetamide		4.03

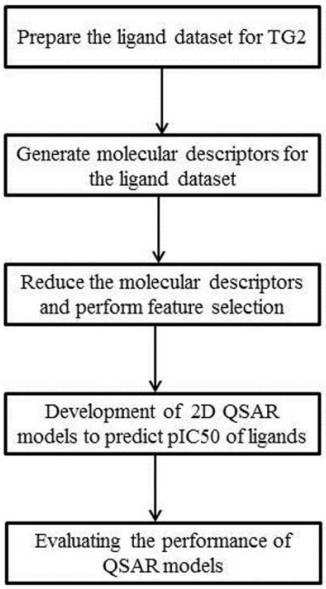

Figure 1: Flowchart describing the steps implemented in study

3.2 Molecular descriptors generation

The SDF file format is created for every compound using PubChem Sketcher. The.sdf file created is given as input to PaDEL-descriptor for generating the molecular descriptors. Based on the properties of ligand molecule, 2325 molecular descriptors are obtained. The molecular descriptors are consideredas an independent variable for building QSAR models.

3.3 Selection of relevant descriptors

It is not feasible to develop QSAR models having a large number of independent attributes. Hence, feature selection techniques are applied to the dataset. Initially, features having more than 75% correlation are eliminated from the dataset. It is followed by implementing PCA algorithm to reduce the high dimensionality of features. The algorithm resulted in 624 features. However, the features

obtained after PCA algorithm were still large for generating an accurate representation of data instances. Hence, Boruta algorithm available in R programming language is used for feature selection which based on wrapper method. The algorithm is iterated in ten folds to yield informative features with respect to pIC50 attribute. After the iterations, the algorithm yields seven relevant attributes for model development. The molecular descriptors obtained after feature selection is shown in Table 2.

Table 2: Relevant molecular descriptors generated as a function of pIC50 value

Sl. No	Molecular descriptor	Meaning
1.	AlogP	Ghose-Crippen water-octanol partition coefficient
2.	ATS6m	Broto-Moreau autocorrelation of lag 6 coefficient, weight by mass
3.	ATS6e	Broto-Moreau autocorrelation of lag 6 coefficient, weight by Sanderson electronegativity
4.	ATS7e	Broto-Moreau autocorrelation of lag 7 coefficient, weight by Sanderson electronegativity
5.	GATS6c	Geary autocorrelation lag 6 coefficient
6.	SpMax_Dzm	Barysz matrix coefficient
7.	Kier1	Kappa shape indices

3.4 QSAR modeling

Prior to data modeling, the ligand dataset is divided into training (75%) and test (25%) datasets. The model is developed on training dataset, while the predictive performance of the model is assessed using the independent test set. QSAR models are generated to find the interrelation between pIC50 and the seven molecular descriptors. Four different regression techniques are implemented on the dataset [26].

i. Linear regression
ii. Ridge regression
iii. Lasso regression
iv. Elastic-net regularization

3.4.1 Linear regression

Initially, a linear regression model was applied to the dataset to predict the dependency between the variables. A linear equation is generated along with error term defined as:

$$pIC50 = -0.0479(AlogP) - 0.0005(ATS7e) - 0.0014(SpMax\_Dzm) + 0.0585(Kier1) - 0.0767 \rightarrow (4)$$

The equation (4) describes pIC50 as a function of molecular descriptors AlogP, ATS7e and SpMax_Dzm which are negatively correlated while Kier1 is positively correlated. The equation

concludes with a negative error estimate of 0.0767. The accuracy of the model is found to be 78.4% with RSS value of 0.632.

To get better predictive performance, penalized regression methods are applied to the ligand dataset.

3.4.2 Ridge regression

This method minimizes the residual sum of squares (RSS) by introducing a penalty parameter λ, which initializes the weights to zero by shrinking to low variance. Shrinkage avoids over fitting ensuring that the estimator provides no solution, as seen in the case of multi collinearity [27]. Ridge regression aims to minimize the penalty factor$\|\beta\|_2^2$ by shrinking λ upto zero.

$$min_{\beta \epsilon \mathcal{R}^p} \|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2 \rightarrow (5)$$

Ridge regression is applied in bioinformatics applications when a number of attributes (p) is greater than the data instances (n) (i.e. p>n). The generalized ridge regression model is denoted as:

$$Y = XB + e \rightarrow (6)$$

Here, Y denotes independent variable; X denotes independent variables, B is the regression coefficient, and e is the error estimate. The dependency in R programming language, 'glmnet' is invoked to perform ridge regression by defining α=0 (Jerome et al., 2010). The glmnet function for ridge regression model represented as:

$$pIC50 = -6.9236(AlogP) + 2.3273(ATS6m) + 2.2651(ATS6e) + 2.4839(ATS7e) + 2.7070(GATS6e) + 8.2010(SpMax_{Dzm}) + 1.3116(Kier1) - 8.0316 \rightarrow (7)$$

The ridge model obtained after cross-validation, describes pIC50 value as a function of all the seven molecular descriptors associated with the error estimate, while none of the variables is reduced to zero. The plot describing coefficients of ridge regression versus penalty function log lambda (λ) is shown in Figure 2(a). As observed in the plot, when λ approaches zero, ridge model behaves similarly to ordinary least squares model. While λ value increases, ridge model approaches to zero. The mean square error is found to be 2.4014 for the ridge regression model. The plot representing mean square error as a function of log lambda is shown in Figure 3(a).

3.4.3 LASSO regression

Abbreviated as Least Absolute Shrinkage and Selection Operator, it is penalized regression model which introduces shrinkage procedure for reducing the data instances towards a central measure. The algorithm performs L1 regularization, which defines a penalty function that equates the absolute value of the coefficients to zero [28]. This method is suitable for data with high multi collinearity, larger the penalty value; greater is the shrinkage towards zero. The generalized equation of LASSO is defined as:

$$min_{\beta \epsilon \mathcal{R}^p} \|y - X\beta\|_2^2 + \lambda\|\beta\|_1 \rightarrow$$

Here, $\|\beta\|_1$ is a penalty function that shrinks the penalty parameter λ to zero. Based on the parameters in the dataset, lasso model is developed using the 'glmnet' dependency in R programming language [29]. The model is described as:

$$pIC50 = -0.1367(AlogP) + 0.0010(ATS7e) + 0.6073(GATS6c) + 0.004(Kier1) - 1.096 \rightarrow (9)$$

Lasso model defines pIC50 value as a function of three molecular descriptors namely AlogP, GATS6c and Kier1 associated with the error estimate. As observed, the model eliminates other descriptors which are highly correlated. The coefficients derived in the lasso equation are shown in the plot obtained by equating α=1 in the glmnet fit function as shown in Figure 2(b)A plot representing the mean square error observed in lasso regression model as a function of log lambda is shown in Figure 3(b). The plot is obtained after ten-fold cross-validation represents the range of mean square error obtained for different models. The best performing model results in terms of the mean square value of 3.4075.

3.4.4 Elastic Net regularization

It is a hybrid technique based on ridge and lasso regression methods [30]. The model is trained based on L_1 and L_2 regularization exhibiting a grouping effect using the 'glmnet' dependency in R programming language. The value of α is chosen to be intermediate to that of ridge (α=0) and lasso (α=1), which is 0.5. The generalized equation of elastic net is defined as:

$$\hat{\beta} = argmin_{\beta}\|y - X\beta\|^2 + \lambda_2\|\beta\|^2 \quad + \lambda_1\|\beta\|_1 \rightarrow (10)$$

Based on the parameters of the ligand dataset, the elastic net model is obtained given as:

$$pIC50 = \quad -1.0005(AlogP) + 2.7188(ATS6m) + 1.2507(ATS6e) + 4.3033(ATS7e) + 4.2718(GATS6c) + 2.1118(Kier1) - 9.6997 \rightarrow (11)$$

The coefficients of the elastic net model are plotted as a function of log lambda as shown in Figure 2(c). After cross-validation on the test set the model is being evaluated for mean square error which is found to be 0.36656. The plot referring to mean square error plotted against log lambda is shown in Figure 3(c).

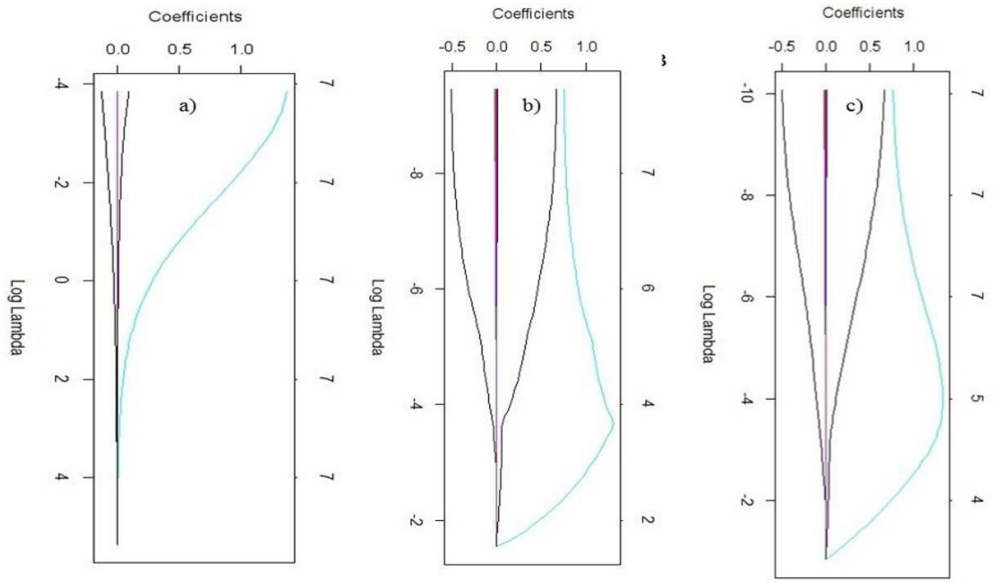

Figure 2: The figure representing plots of coefficients of QSAR models as a function of penalty function log lambda; 2(a) represents ridge regression coefficients; 2(b) represents lasso regression coefficients and 2(c) represents elastic net coefficients respectively.

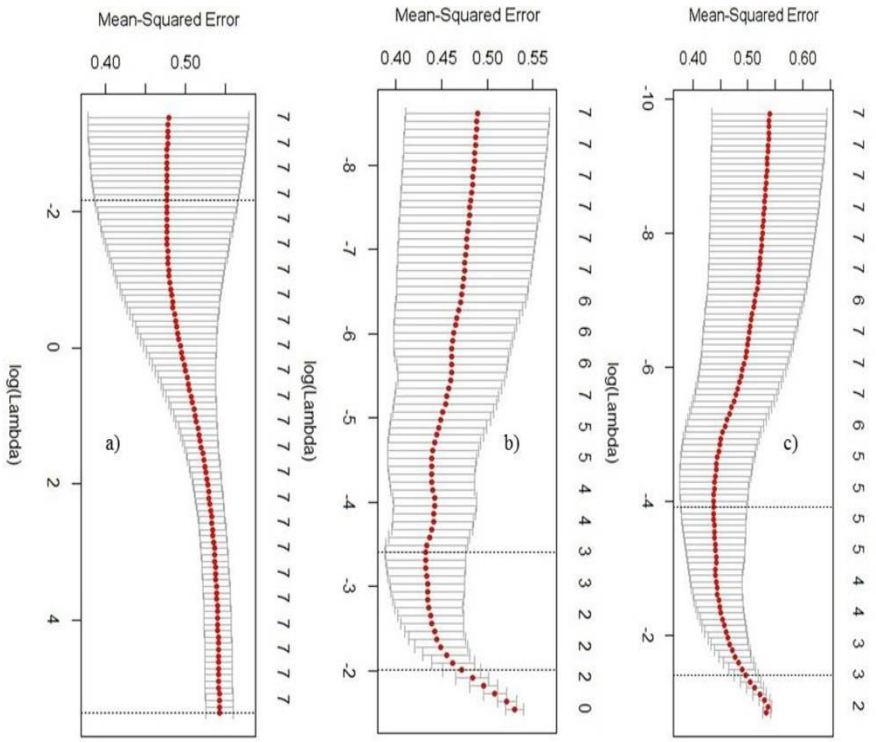

Figure 3: Mean square error plotted as a function of log lambda. Two lines seen in each figure indicates the range of mean square error value. Numbers seen on the right side of the figures represents the molecular descriptor attribute. 3(a) represents mean square error for ridge regression while 3(b) and 3(c) represent MSE's for lasso regression and elastic net regression respectively.

3.5 Evaluation of model performance

Based on the QSAR equations derived from the regression methods, the predicted pIC50 values were calculated for each model. The error is measured as a function of the difference in values of actual pIC50 compared to the predicted ones for each regression model. It is found that linear regression performs poorly for the dataset, while elastic net model outperforms other models. The detailed summary of evaluation metrics is shown in Table 3.

Table 3: Evaluation metrics for regression models

Sl. No	QSAR model	Residual sum of squares	Mean square error	R value
1.	Linear regression	0.632	-	0.58
2.	Ridge regression	-	2.4014	0.79
3.	Lasso regression	-	3.4075	0.68
4.	Elastic net regularization	-	0.36656	0.88

The values in the table indicate elastic net model performs better in the prediction of pIC50 metric, resulting in the least mean square error of 0.36656. The predictive performance of all the regression models is shown in the plot depicted as Figure 4. The plot 4(d) shows minimal distortion with

original values as compared to other models. Hence, elastic net model outperforms having a better predictive performance for TG2 ligand dataset.

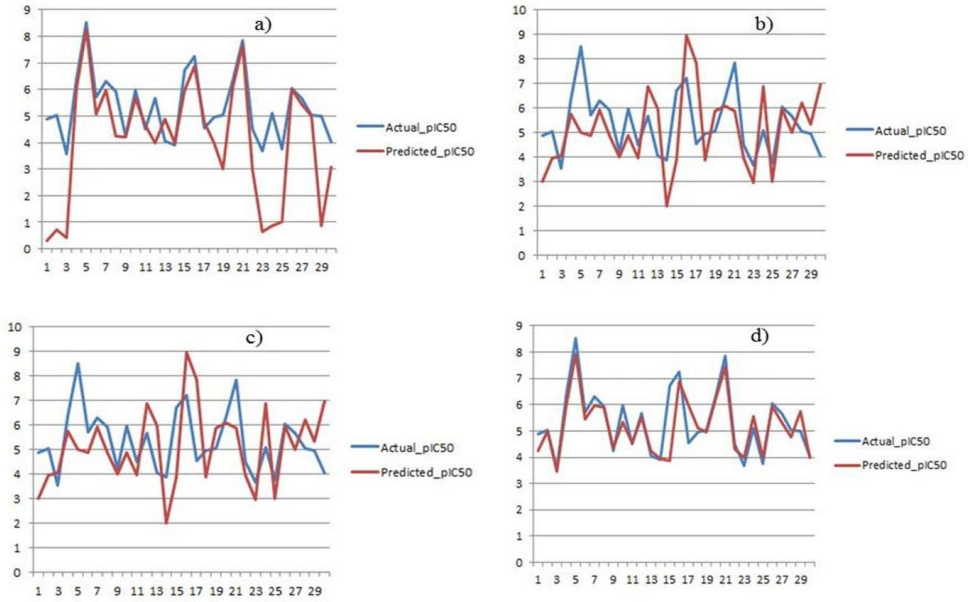

Figure 4: Analysis of predictive performance of regression models. Here X-axis represents the actual pIC50 values and Y-axis represents the predicted pIC50 values from each model. 4(a) represents linear regression; 4(b) represents ridge regression; 4(c) represents lasso regression and 4(d) represents elastic net regression plot respectively.

4. CONCLUSION

The present study will help to identify the set of potent ligand molecules in inhibiting the expression of TG2 and also it can be used as a new strategy to control over the lung carcinoma.The focuses on an approach to predict novel drug molecules based on the outcome of mathematical models. By implementing this methodology, encouraging insights are obtained for predicting the biological activity of drug molecules against TG2. These models will aid in filtering out the drug-like molecules and development of better anti-tumor drugs against transglutaminase 2 (TG2).

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

REFERENCES

1. Kris MG, Gaspar LE, Chaft JE, Kennedy EB, Azzoli CG, Ellis PM, Lin SH, Pass HI, Seth R, Shepherd FA, Spigel DR. Adjuvant systemic therapy and adjuvant radiation therapy for stage I to IIIA completely resected non-small-cell lung cancers: American Society of Clinical Oncology/Cancer Care Ontario clinical practice guideline update. J Clin Oncol. 2017 Apr 24;35(25):2960-74.
2. Behera D. SC17. 03 Lung Cancer in India: Challenges and Perspectives. Journal of Thoracic Oncology. 2017 Jan 1;12(1): S114-5.
3. Chihong Z, Yutian L, Danying W, Ruibin J, Huaying S, Linhui G, Jianguo F. Prognostic value

of transglutaminase 2 in non-small cell lung cancer patients. Oncotarget. 2017 Jul 11;8(28):45577.

4. Agostinelli E. Polyamines and transglutaminases: biological, clinical, and biotechnological perspectives.
5. Schaertl S, Prime M, Wityak J, Dominguez C, Munoz-Sanjuan I, Pacifici RE, Courtney S, Scheel A, Macdonald D. A profiling platform for the characterization of transglutaminase 2 (TG2) inhibitors. Journal of biomolecular screening. 2010 Jun;15(5):478-87.
6. Griffin M, Casadio R, Bergamini CM. Transglutaminases: nature's biological glues. Biochemical Journal. 2002 Dec 1;368(2):377-96.
7. Verma A, Mehta K. Tissue transglutaminase-mediated chemoresistance in cancer cells. Drug resistance updates. 2007 Aug 1;10(4-5):144-51.
8. Park KS, Kim HK, Lee JH, Choi YB, Park SY, Yang SH, Kim SY, Hong KM. Transglutaminase 2 as a cisplatin resistance marker in non-small cell lung cancer. Journal of cancer research and clinical oncology. 2010 Apr 1;136(4):493-502.
9. Song M, Hwang H, Im CY, Kim SY. Recent Progress in the Development of Transglutaminase 2 (TGase2) Inhibitors: Miniperspective. Journal of medicinal chemistry. 2016 Nov 21;60(2):554-67.
10. Schneider G, Fechner U. Computer-based de novo design of drug-like molecules. Nature Reviews Drug Discovery. 2005 Aug;4(8):649.
11. Han I, Park HJ, Seong SC, Lee S, Kim IG, Lee MC. Role of transglutaminase 2 in apoptosis induced by hydrogen peroxide in human chondrocytes. Journal of Orthopaedic Research. 2011 Feb;29(2):252-7.
12. Dimitrov SD, Mekenyan OG, Sinks GD, Schultz TW. Global modeling of narcotic chemicals: ciliate and fish toxicity. Journal of Molecular Structure: THEOCHEM. 2003 Mar 7;622(1-2):63-70.
13. Enoch SJ, Cronin MT, Schultz TW, Madden JC. An evaluation of global QSAR models for the prediction of the toxicity of phenols to Tetrahymena pyriformis. Chemosphere. 2008 Apr 1;71(7):1225-32.
14. Narayanan S, Velmurugan D. Modeling, ADME and QSAR studies on Dihydroisoxazole derivatives as Transglutaminase-2 Inhibitors against Neurodegenerative Disorders.
15. Siegel M, Khosla C. Transglutaminase 2 inhibitors and their therapeutic role in disease states. Pharmacology & therapeutics. 2007 Aug 1;115(2):232-45.
16. Zhang S, Golbraikh A, Oloff S, Kohn H, Tropsha A. A novel automated lazy learning QSAR (ALL-QSAR) approach: method development, applications, and virtual screening of chemical databases using validated ALL-QSAR models. Journal of chemical information and modeling. 2006 Sep 25;46(5):1984-95.

17. Wisowski G, Koźma EM, Bielecki T, Pudełko A, Olczyk K. Dermatan sulfate is a player in the transglutaminase 2 interaction network. PloS one. 2017 Feb 15;12(2):e0172263.
18. Selvaraj C, Tripathi SK, Reddy KK, Singh SK. Tool development for Prediction of pIC 50 values from the IC 50 values-A pIC 50 value calculator. Current Trends in Biotechnology & Pharmacy. 2011 Apr 1;5(2).
19. Eriksson L, Andersson PL, Johansson E, Tysklind M. Megavariate analysis of environmental QSAR data. Part I–A basic framework founded on principal component analysis (PCA), partial least squares (PLS), and statistical molecular design (SMD). Molecular diversity. 2006 May 1;10(2):169-86.
20. Yap CW. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. Journal of computational chemistry. 2011 May;32(7):1466-74.
21. Gentleman R. R programming for bioinformatics. Chapman and Hall/CRC; 2008 Jul 14.
22. Kursa MB, Rudnicki WR. Feature selection with the Boruta package. J Stat Softw. 2010 Sep 16;36(11):1-3.
23. Zhang S, Golbraikh A, Oloff S, Kohn H, Tropsha A. A novel automated lazy learning QSAR (ALL-QSAR) approach: method development, applications, and virtual screening of chemical databases using validated ALL-QSAR models. Journal of chemical information and modeling. 2006 Sep 25;46(5):1984-95.
24. Badarau E, Collighan RJ, Griffin M. Recent advances in the development of tissue transglutaminase (TG2) inhibitors. Amino Acids. 2013 Jan 1;44(1):119-27.
25. Duval E, Case A, Stein RL, Cuny GD. Structure–activity relationship study of novel tissue transglutaminase inhibitors. Bioorganic & medicinal chemistry letters. 2005 Apr 1;15(7):1885-9.
26. Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2005 Apr 1;67(2):301-20.
27. Hoerl AE, Kennard RW. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics. 1970 Feb 1;12(1):55-67.
28. Jerome, F., Trevor, H., Noah, S., & Rob, T.- Package “glmnet.” Retrieved from 2010.
29. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological). 1996 Jan 1:267-88.
30. Cherkasov A, Muratov EN, Fourches D, Varnek A, Baskin II, Cronin M, Dearden J, Gramatica P, Martin YC, Todeschini R, Consonni V. QSAR modeling: where have you been? Where are you going to? Journal of medicinal chemistry. 2014 Jan 6;57(12):4977-5010.