Mutat Res Fund Mol Mech Mutagen 806 (2017) 19–26

Contents lists available at ScienceDirect

Mutat Res Fund Mol Mech Mutagen

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

Investigating mutation-specific biological activities of small molecules using quantitative structure-activity relationship for epidermal growth factor receptor in cancer

P. Anoosha, R. Sakthivel, M. Michael Gromiha*

Department of Biotechnology, Bhupat and Jyoti Mehta School of BioSciences, Indian Institute of Technology Madras, Chennai 600 036, Tamilnadu, India

ARTICLE INFO

Keywords:
EGFR
Cancer
Driver mutation
Quantitative structure-activity relationship (QSAR)
Regression
Docking

ABSTRACT

Epidermal Growth Factor Receptor (EGFR) is a potential drug target in cancer therapy. Missense mutations play major roles in influencing the protein function, leading to abnormal cell proliferation and tumorigenesis. A number of EGFR inhibitor molecules targeting ATP binding domain were developed for the past two decades. Unfortunately, they become inactive due to resistance caused by new mutations in patients, and previous studies have also reported noticeable differences in inhibitor binding to distinct known driver mutants as well. Hence, there is a high demand for identification of EGFR mutation-specific inhibitors. In our present study, we derived a set of anti-cancer compounds with biological activities against eight typical EGFR known driver mutations and developed quantitative structure-activity relationship (QSAR) models for each separately. The compounds are grouped based on their functional scaffolds, which enhanced the correlation between compound features and respective biological activities. The models for different mutants performed well with a correlation coefficient, (r) in the range of 0.72–0.91 on jack-knife test. Further, we analyzed the selected features in different models and observed that hydrogen bond and aromaticity-related features play important roles in predicting the biological activity of a compound. This analysis is complimented with docking studies, which showed the binding patterns and interactions of ligands with EGFR mutants that could influence their activities.

1. Introduction

The Epidermal Growth Factor Receptor (EGFR) is one of the well known drug targets for cancer treatment. It is the second most genes with high frequency of point mutations observed in different cancer types [1]. Mutations result in abnormal activation of EGFR kinase and are implicated in the development and progression of several cancer types [2]. The activating mutations such as L858R and T790M have impact on protein function leading to its over-expression which results in increased cell growth, proliferation and metastases [3]. Although a number of first- and second-generation small molecule tyrosine kinase inhibitors were developed to treat mutated EGFR, they suffer from drug resistance after long-term drug administration and dose-limiting adverse effects, respectively [4–8]. Currently, research is mainly focused on the development of third generation inhibitors which could display high mutant selectivity while minimizing adverse effects by sparing wild-type EGFR activity [9,10].

Earlier studies reported that distinct EGFR mutations differ markedly in their inhibitor susceptibilities [11]. For example, the well known EGFR inhibitor gefitinib showed 50-fold weaker binding affinity towards G719S mutant and is significantly less sensitive compared to L858R mutant [12]. This shows that intrinsic differences in inhibitor binding of the altered kinases can explain the differential sensitivity of cell lines bearing these mutations and another possibility might be differences in their signaling pathways. These observations trigger the importance of developing mutation specific inhibitors for efficient treatment of tumors bearing distinct mutations. Several studies have been reported in the literature to understand the differences in altered drug sensitivities by investigating mutant structures with respect to their binding patterns to small molecules by using computational [13] and biochemical analyses [12,14,15]. Among them, 2D and 3D QSAR models are key tools for predicting the biological activity of new inhibitor compounds [16]. In the past, QSAR models have been developed for EGFR using single scaffold based analogues such as anilinoquinolines and quinazoline derivatives with experimental data generated from a single bioassay system [17,18]. The predictive coverage is minimal as the methods are based on a limited set of compounds with a particular scaffold. Further, QSAR based analysis of EGFR inhibitors with different functional scaffolds have been performed against wild-type EGFR protein using a large set of molecules to

* corresponding author.
E-mail address: gromiha@iitm.ac.in (M.M. Gromiha).

http://dx.doi.org/10.1016/j.mrfmmm.2017.08.003
Received 14 July 2017; Received in revised form 21 August 2017; Accepted 22 August 2017
Available online 26 August 2017
0027-5107/

identify or predict novel compounds [19,20]. Apart from these studies, it is very important to analyze the structure – activity relationships of the compounds against specific EGFR driver mutants which would be beneficial in understanding their altered responses to drugs and also to predict the activity of new compounds.

In the present study, we have collected biological activities (IC_{50}) of diverse set of compounds against eight typical EGFR mutants which are known to be driver mutations [21] and are highly sensitive to tyrosine kinase inhibitors. We developed individual QSAR models for each mutant separately by different combinations of features and achieved a correlation, (r) in the range of 0.78–0.92. We have examined the models using n-fold cross validation and jack-knife test. Further, for evaluating the performance of our models, we trained 90% of the randomly chosen data using the same set of selected features and tested the remaining 10% of the data and repeated this for five times by shuffling the data. Interestingly, the performance of the model on different test sets is consistent in all iterations and the average correlation coefficient (r) is > 0.90 in all the cases. We have also analyzed the binding patterns of the compounds in our dataset with mutants using docking studies and observed that hydrophobic interactions play major role in binding, which is in strong agreement with the previous analyses. Further, we proposed few compounds (known anti-cancer drugs but new to EGFR) with IC_{50} in nano-molar range against different EGFR mutants, which could be a basis and promising step towards the development of mutation specific inhibitors.

2. Materials and methods

2.1. Dataset

We have collected diverse sets of anti-cancer compounds with biological activity (IC_{50}) against eight typical EGFR mutants viz. A289V, G598V, G719S, T751I, P753S, S768I, R832C and double mutant L858R/T790M, which are known to be driver mutations. The IC_{50} values of all the compounds vary widely in the range of 37 mM–0.2 nM and are obtained under same experimental conditions from COSMIC database [22], which is extremely suitable for the development of high quality and reliable QSAR models. For example, distribution of IC_{50} values of compounds against L858R/T790M double mutant is represented in Fig. 1. The activity values of 130 compounds range widely from 0.002 μM–2750 μM. IC_{50} values of the compounds are converted to pIC50 (Eq. (1)) for better interpretation and to avoid over fitting.

$$pIC_{50} = -\log_{10}(IC_{50}) \quad (1)$$

2.2. Grouping of the compounds

Further, the compounds of each mutant are grouped into different datasets based on their functional scaffolds [Table 1]. For the mutants A289V, G719S, T751I, S768I, R832C and L858R/T790M, compounds are grouped into three datasets with analogous functional groups (i) Azole, Quinazoline, Indole, azine and Aniline; (ii) Imidazole, Thiophene; (iii) Peptide, Carbohydrate, Alkaloids and Lactone; respectively. Compounds of P753S mutant are grouped into two datasets in which proteins, peptides, carbohydrate and lactones are grouped into dataset2 and rest of the compounds in dataset1 and for G598V mutant, compounds are divided into four datasets belonging to different functional groups. To validate the performance of developed QSAR models, 10% of each dataset for all the mutants has been randomly chosen for test set.

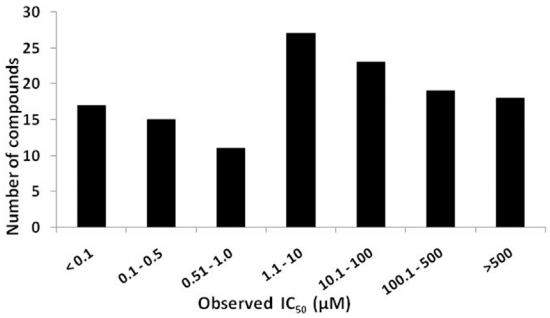

Fig. 1. Distribution of experimental IC_{50} values of L858R/T790M double mutant.

Table 1
Dataset and classification of compounds based on functional groups for eight mutants.

Mutant	Dataset1		Dataset2	Dataset3	Total compounds
P753S		60		26	86
A289V	67		28	37	132
G719S	37		29	26	92
T751I	28		28	35	91
S768I	40		26	26	92
R832C	66		28	37	131
L858R/T790M	66		27	37	130
G598V	39	28	28	37	132

2.3. Molecular descriptors of chemical compounds

We derived a set of 590 features representing 1-D, 2-D and 3-D molecular descriptors encoding chemical composition, topology and geometry, respectively using PaDEL Descriptor server [23]. The total number of features has been reduced to 154 by employing 'dimensionality reduction by correlation' criteria as discussed in previous reports [24] to remove the redundancy with a correlation cut-off of 0.75 between any two features.

2.4. Multiple linear regression and feature selection

We have developed independent QSAR models for all the datasets of eight mutants using multiple linear regression technique [25] by combining more than one feature (Eq. (2)).

$$Y = \beta_1.x_1\beta_1.x_1 + \beta_2.x_2 + \beta_3.x_3 + \ldots. + \beta_n.x_n + C \quad (2)$$

Where, Y is dependent or response variable representing biological activity of the compound (IC_{50}) in this study and x_i is independent variable which represents each feature. The intercept C and slopes β_1, $\beta_2 \ldots \beta_n$ are partial regression coefficients. Best set of features for each dataset are identified by using step-wise least square fit [26] and regression technique to develop QSAR regression models.

2.5. Model performance and validation

The performance of the developed QSAR regression models is measured using absolute correlation-coefficient (r) between the experimental and predicted IC_{50} values. The statistical significance of the models is evaluated using p-value and mean absolute error, MAE [27]. The developed models are validated using (i) n-fold cross validation: entire dataset is divided into n equal subsets and single subset is used for validation and n-1 subsets as training. This process is repeated for n times such that each subset is used for testing at least once and (ii) jack-knife (leave-one-out cross validation) test: regression model is developed using n-1 compounds (n: total number of compounds in dataset) and tested on the remaining compound to predict its IC_{50} value and this process is iterated n times to predict IC_{50} of each compound in the dataset.

2.6. Contribution of each feature in the regression models

We assessed the significance of each feature in different mutant regression models using proportional reduction of error (PRE) measure