Digest Journal of Nanomaterials and Biostructures Vol. 5, No 2, April-June 2010, p. 387 – 401

2D QSAR STUDIES ON A SERIES OF 4-ANILINO QUINAZOLINE DERIVATIVES AS TYROSINE KINASE (EGFR) INHIBITOR: AN APPROACH TO DESIGN ANTICANCER AGENTS

MALLESHAPPA N.NOOLVI*, HARUN M. PATEL, VARUN BHARDWAJ
Department of Pharmaceutical Chemistry, ASBASJSM College of Pharmacy, Bela (Ropar)-14011, Punjab, India

Epidermal growth factor receptor (EGFR) protein tyrosine kinases (PTKs) are known for its role in cancer. QSAR studies were performed on a set of 61 analogs of 4-anilino quinazoline using MDS vlife science QSAR plus module by using Multiple Linear Regression (MLR), Principal Component Regression (PCR) and Partial Least Squares (PLS) Regression methods. Among these methods, Multiple Linear Regression (MLR) method has shown very promising result as compare to other two methods. A QSAR model was generated by a training set of 42 molecules with correlation coefficient (r^2) of 0.912 , significant cross validated correlation coefficient (q^2) of 0.800 , F test of 60.5149 , r^2 for external test set (pred_r^2) 0.6042 , coefficient of correlation of predicted data set (pred_r^2se) 0.7438 and degree of freedom 38 by Multiple Linear Regression (MLR) method. Estate number, Electro- topological state indices, Bromine count, Chlorine count and alignment independent descriptors were found to be major contributing descriptors governing the activity.

(Received April 27, 2010; accepted May 5, 2010)

Keywords: - 4-Anilino Quinazoline, Tyrosine kinase (EGFR), Multiple Linear Regression (MLR), Principal Component Regression (PCR), Partial Least Squares (PLS) .

1. Introduction

Many of the tyrosine kinase enzymes are involved in cellular signaling pathways and regulate key cell functions such as proliferation, differentiation, anti-apoptotic signaling and neurite outgrowth. Unregulated activation of these enzymes, through mechanisms such as point mutations or over expression, can lead to a large percentage of clinical cancers [1, 2].The importance of tyrosine kinase enzymes in health and disease is further underscored by the existence of aberrations in tyrosine kinase enzymes signaling occurring in inflammatory diseases and diabetes. Inhibitors of tyrosine kinase as a new kind of effective anticancer drug are important mediators of cellular signal transduction that affects growth factors and oncogenes on cell proliferation [3, 4]. The development of tyrosine kinase inhibitors has therefore become an active area of research in pharmaceutical science. Epidermal growth factor receptor (EGFR) which plays a vital role as a regulator of cell growth is one of the intensely studied tyrosine kinase targets of inhibitors. EGFR is overexpressed in numerous tumors, including those derived from brain, lung, bladder, colon, breast, head and neck. EGFR hyper activation has also been implicated in other diseases including polycystic kidney disease, psoriasis and asthma [5-7]. Since the hyper activation of EGFR has been associated with these diseases, inhibitor of EGFR has potential therapeutic value and it has been extensively studied in the pharmaceutical industry.

One could not, however, confirm that the compounds designed would always possess good inhibitory activity to EGFR, while experimental assessments of inhibitory activity of these compounds are time-consuming and expensive. Consequently, it is of interest to develop a prediction method for biological activities before the synthesis. Quantitative structure activity relationship (QSAR) searches information relating chemical structure to biological and other

activities by developing a QSAR model. Using such an approach one could predict the activities of newly designed compounds before a decision is being made whether these compounds should be really synthesized and tested.

Anilinoquinazolines are the most developed class of drugs that inhibit EGFR kinase intracellularly [8].Structure–activity relationship (SAR) studies reveal the nature of desirable substituents on the Anilinoquinazolines moiety. Electron withdrawing, lipophilic substituents at the 3-position of aniline are favorable with Cl and Br being optimal. Similarly, electron donating groups at the 6- and 7-positions of quinazoline are preferred [9].Bulky substituents appear to be tolerated at the 6- and 7-positions [10].QSAR studies by Hou et al. have described the region around the 7-position as more electronegative than that near the 6-position [11].

With the above facts and in continuation of our research for newer anti-cancer agent in the present study, we reported 2D – QSAR studies on a series of EGFR kinase inhibitors to provide further insight into the key structural features required to design potential drug candidates of this class[12,13] . Here, we present our observations on the role of different substitution at the 4, 6- and 7-positions of quinazolines as EGFR inhibitor.

2. Computational methods

A] Chemical Data

A series of 61 molecules belonging to quinazoline derivatives as tyrosine kinase (EGFR) inhibitors were taken from the study by Bridges et al [14]. The 2D- QSAR models were generated using a training set of 42 molecules. The observed and predicted biological activities of the training and test set molecules are presented in Table 1. Predictive power of the resulting models was evaluated by a test set of 19 molecules with uniformly distributed biological activities. The observed selection of test set molecules was made by considering the fact that test set molecules represents a range of biological activity similar to the training set.

B] Data Set

All computational work was performed on Apple workstation (8-core processor) using Vlife MDS QSAR plus software developed by Vlife Sciences Technologies Pvt Ltd, Pune, India, on windows XP operating system . All the compounds were drawn in Chem DBS using fragment database and then subjected to energy minimization using batch energy minimization method. Conformational search were carried out by systemic conformational search method and all the compounds were aligned by template based method.

C] Biological Activities

The negative logarithm of the measured IC_{50} (μM) against tyrosine kinase (EGFR) as pIC_{50} [$pIC_{50} = -\log (IC_{50} \times 10^{-6})$] was used as dependent variable, thus correlating the data linear to the free energy change. Since some compounds exhibited insignificant/no inhibition, such compounds were excluded from the present study**.** All the IC_{50} values had been obtained using human A431 carcinoma cell vesicles by immunoaffinity chromatography [14]. The IC_{50} values of reference compounds were checked to ensure that no difference occurred between different groups. The pIC_{50} values of the molecules under study spanned a wide range from 5 to 11.

D] Molecular Descriptors

Various 2D descriptors (a total of 208) like element counts, molecular weight, molecular refractivity, log *P*, topological index, Baumann alignment independent topological descriptors *etc*., were calculated using VlifeMDS software**.** The preprocessing of the independent variables (i.e., descriptors) was done by removing invariable (constant column) and cross-correlated descriptors (with $r > 0.99$) which resulted in total 132, 121 and 116 descriptors for MLR,PCR and PLS respectively to be used for QSAR analysis.

E] Selection of Training and Test Set

The dataset of 61 molecules was divided into training and test set by Sphere Exclusion (SE) method for MLR, PCR and PCA model with pIC_{50} activity field as dependent variable and various 2D descriptors calculated for the molecules as independent variables.

F] Model Validation

This is done to test the internal stability and predictive ability of the QSAR models. Developed QSAR models were validated by the following procedure:

1] Internal Validation

Internal validation was carried out using leave-one-out (q^2, LOO) method. For calculating q^2, each molecule in the training set was eliminated once and the activity of the eliminated molecule was predicted by using the model developed by the remaining molecules. The q^2 was calculated using the equation which describes the internal stability of a model.

$$q^2 = 1 - \frac{\Sigma(y_i - \hat{y}_i)^2}{\Sigma(y_i - y_{mean})^2} \quad (1)$$

where y_i and $\hat{y}_i$ are the actual and predicted activity of the *i*th molecule in the training set, respectively, and y_{mean} is the average activity of all molecules in the training set.

2] External Validation

For external validation, the activity of each molecule in the test set was predicted using the model developed by the training set. The pred_r^2 value is calculated as follows.

$$pred\_r^2 = 1 - \frac{\Sigma(y_i - \hat{y}_i)^2}{\Sigma(y_i - y_{training\ set\ mean})^2} \quad (2)$$

where y_i and $\hat{y}_i$ are the actual and predicted activity of the *i*th molecule in the training set, respectively, and y_{mean} is the average activity of all molecules in the training set.
Both summations are over all molecules in the test set. Thus, the pred_r^2 value is indicative of the predictive power of the current model for external test set.

3. Randomization Test

To evaluate the statistical significance of the QSAR model for an actual dataset, one tail hypothesis testing was used [15, 16]. The robustness of the models for training sets was examined by comparing these models to those derived for random datasets. Random sets were generated by rearranging the activities of the molecules in the training set. The statistical model was derived using various randomly rearranged activities (random sets) with the selected descriptors and the corresponding q^2 were calculated. The significance of the models hence obtained was derived based on a calculated Z score [15, 16].

A Z score value is calculated by the following formula:

$$Z\text{score} = \frac{(h - \mu)}{\sigma} \quad (3)$$

where h is the q^2 value calculated for the actual dataset, μ the average q^2, and σ is its standard deviation calculated for various iterations using models build by different random datasets.

The probability (α) of significance of randomization test is derived by comparing Z score value with Z score critical value as reported in reference [17], if Z score value is less than 4.0; otherwise it is calculated by the formula as given in the literature. For example, a Z score value greater than 3.10 indicates that there is a probability (α) of less than 0.001 that the QSAR model constructed for the real dataset is random. The randomization test suggests that all the developed models have a probability of less than 1% that the model is generated by chance.

G] QSAR by Multiple Linear Regression (MLR) Analysis

Multiple regression is the standard method for multivariate data analysis. It is also called as ordinary least squares regression (OLS). This method of regression estimates the values of the regression coefficients by applying least squares curve fitting method. For getting reliable results, dataset having typically 5 times as many data points (molecules) as independent variables (descriptors) is required.

The regression equation takes the form

$$Y = b1*x1 + b2*x2 + b3*x3 + c, \quad (4)$$

where Y is the dependent variable, the 'b's are regression coefficients for corresponding 'x's (independent variable), 'c' is a regression constant or intercept.

In the present study QSAR model was developed using multiple regression by forward-backward variable selection method with pIC_{50} activity field as dependent variable and 132 physico-chemical descriptors as independent variable having cross-correlation limit of 1. Selection of test and training set was done by sphere exclusion method.

H] QSAR by Principal Component Regression (PCR) Method

Principal components analysis rotates the data into a new set of axes such that the first few axes reflect most of the variations within the data. By plotting the data on these axes, we can spot major underlying structures automatically. The value of each point, when rotated to a given axis, is called the principal component value. Principal Components Analysis selects a new set of axes for the data. These are selected in decreasing order of variance within the data. They are also perpendicular to each other. Hence the principal components are uncorrelated. Some components may be constant, but these will be among the last selected. The problem noted with MLR was that correlated variables cause instability. So, how about calculating principal components, throwing away the ones which only appear to contribute noise (or constants), and using MLR on these? This process gives the modeling method known as Principal Components Regression. Rather than forming a single model, as with MLR, a model can be formed using 1, 2,… components and a decision can be made as to how many components are optimal. If the original variables contained collinearity, then some of the components will contribute only noise. So long as these are dropped, the models can be guarantee that our model will be stable**.**

The QSAR model was developed using principal component regression by forward-backward variable selection method with pIC_{50} activity field as dependent variable and 121 physico-chemical descriptors as independent variable having cross-correlation limit of 0.5. Selection of test and training set was done by sphere exclusion method.

I] QSAR by Partial Least Squares (PLS) Regression Method

PLS is an effective technique for finding the relationship between the properties of a molecule and its structure. In mathematical terms, PLS relates a matrix Y of dependent variables to a matrix X of molecular structure descriptors, i.e., a latent variable approach to modeling the covariance structures in these two spaces. PLS have two objectives: to approximate the X and Y data matrices, and to maximize the correlation between them. Whereas the extraction of PLS components is performed stepwise and the importance of a single component is assessed independently, a regression equation relating each Y variable with the X matrix is created. PLS decomposes the matrix X into several latent variables that correlate best with the activity of the molecules. PLS can be done using NIPALS or SIMPLS iterative algorithm, with consecutive estimates obtained using the residuals from previous iterations as the new dependent variable

The QSAR model was developed using partial least squares by forward-backward variable selection method with pIC_{50} activity field as dependent variable and 116 physico-chemical descriptors as independent variable having cross-correlation limit of 0.8. Selection of test and training set was done by sphere exclusion method.

Despite its wide acceptance, a high value of q^2 alone is an insufficient criterion for a QSAR model to be highly predictive. Use of greater number of descriptors particularly requires the model to be validated by external predictive power (r^2 predictive). Hence a set of 19 molecules covering different quinazoline derivatives was employed as test to evaluate the predictivity of training set.

4. Results and discussion

Training set of 42 and test set 19 quinazoline derivatives having different substitution, were employed. Following statistical measure was used to correlate biological activity and molecular descriptors; n ,number of molecules; k ,number of descriptors in a model; df ,degree of freedom; r^2 ,coefficient of determination; q^2 , cross validated r^2; pred_r^2 , r^2 for external test set; pred_r^2se , coefficient of correlation of predicted data set; Z score, Z score calculated by the randomization test; best_ran_r^2; best_ran_q^2,highest q^2 value in the randomization test; α, statistical significance parameter obtained by the randomization test. Selecting training and test set by sphere exclusion method, Unicolumn statics shows that the max of the test is less than max of train set and the min of the test set is greater than of train set shown in Table 2**,** which is prerequisite analysis for further QSAR study. The above result shows that the test is interpolative i.e. derived within the min-max range of the train set. The mean and standard deviation of the train and test provides insight to the relative difference of mean and point density distribution of the two sets. In this case the mean in the test set higher than the train set shows the presence of relatively more active molecules as compared to the inactive ones. Also the similar standard deviation in both set indicates that the spread in both the set with their respective mean is comparable.

Table 2. Unicolumn Statics of Training and Test Set.

Unicolumn statics	Average	Max	Min	Stand. Deviation
Training set	6.7023	9.6819	4.6985	1.1759
Test set	6.8064	8.6989	4.6989	1.1572

5. Generation of qsar models

The dataset of 61 molecules were used for the present study. The common structure of 4-anilino quinazoline ring is shown in Fig. (1).

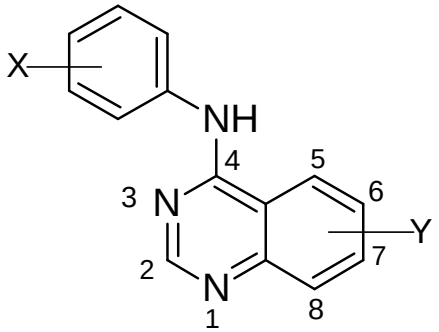

Fig. 1. Structure of 4-(X-bromoanilino)-Y-quinazolines

MODEL – 1 (Multiple Linear Regression (MLR) Analysis)

After 2D QSAR study by Multiple Linear Regression method using forward-backward stepwise variable selection method, the final QSAR equation developed was as follows.

pIC50 = 0.4412 (bromine count) – 1.4112 (T_2_F_1) – 1.3010 (SsssCHcount) + 0.7216 (T_N_O_5) – 1.6538 (T_O_O_3) +8.4330.

Model – 1 developed has a correlation coefficient (r^2) of 0.912 , significant cross validated

Table 1. Structure, Experimental and Predicted Activity of Quinazolines used in Training and Test Set using (MLR-method model 1.

No.	Substituent		IC_{50}[a] (μM)	pIC_{50}[b]		Residual
	X	Y		Exp.	Pred.	
1	H	H	0.344	6.463	6.167	0.296
2[T]	3-F	H	0.056	7.251	7.467	-0.216
3	3-Cl	H	0.023	7.638	7.543	0.095
4[T]	3-Br	H	0.027	7.568	7.441	0.127
5	3-I	H	0.08	7.096	7.112	-0.016
6[T]	3-CF3	H	0.577	6.238	6.447	-0.209
7	H	6-OMe	0.055	7.259	7.063	0.196
8	3-Br	6-OMe	0.03	7.522	8.363	-0.841
9[T]	H	6-NH2	0.77	6.113	7.042	-0.929
10	3-CF3	6-NH2	0.574	6.241	6.321	-0.08
11	3-Br	6-NH2	0.00078	9.107	8.342	0.765
12	H	6-NO2	5.00	5.301	4.836	0.465
13[T]	3-Br	6-NO2	0.9	6.045	6.136	-0.091
14	H	7-OMe	0.12	6.920	7.063	-0.143
15	3-Br	7-OMe	0.01	8	8.363	-0.363
16	H	7-NH2	0.1	7	7.254	-0.254
17	3-F	7-NH2	0.002	8.698	8.554	0.144
18[T]	3-Cl	7-NH2	0.00025	9.602	9.554	0.048
19	3-Br	7-NH2	0.0001	10	10.554	0.554
20	3-I	7-NH2	0.00035	9.455	9.554	-0.099
21[T]	3-CF3	7-NH2	0.0033	8.481	7.534	0.947
22	H	7-NO2	12.00	4.920	5.049	-0.129
23	3-F	7-NO2	6.100	5.214	5.348	-0.134
24[T]	3-Cl	7-NO2	0.81	6.091	6.348	-0.257
25	3-Br	7-NO2	1.000	6	6.348	-0.348
26	3-I	7-NO2	0.54	6.267	6.348	-0.081
27[T]	H	6,7-Di-OMe	0.029	7.537	7.259	0.278
28[T]	3-F	6,7-Di-OMe	0.0038	8.420	7.959	0.461
29	3-Cl	6,7-Di-OMe	0.00031	9.508	9.259	0.249
30	3-Br	6,7-Di-OMe	0.000025	10.602	10.259	0.343
31[T]	3-I	6,7-Di-OMe	0.00089	9.050	9.259	-0.209
32	3-CF3	6,7-Di-OMe	0.00024	9.619	9.239	0.38
33	3-Br	6-NHMe	0.004	8.397	7.938	0.459
34	3-Br	6-NMe2	0.084	7.057	7.130	-0.073
35	3-Br	6-NHCOOMe	0.012	7.920	7.534	0.386
36[T]	3-Br	7-OH	0.0047	8.327	8.767	-0.44
37	3-Br	7-NHCOMe	0.04	7.397	8.150	-0.753
38	3-Br	7-NHMe	0.007	8.154	8.150	0.004
39	3-Br	7-NHC2H5	0.012	7.920	8.150	-0.23
40[T]	3-Br	7-NMe2	0.011	7.958	7.343	0.615
41	3-Br	6,7-Di-NH2	0.00012	9.920	9.429	0.491
42	3-Br	6-NH2,7-NHMe	0.00069	9.161	8.812	0.349

43[T]	3-Br	6-NH2,7-NMe2	0.159	6.798	7.792	-0.994
44	3-Br	6-NH2,7-OMe	0.0038	8.420	9.025	-0.605
45	3-Br	6-NH2,7-Cl	0.0065	8.187	8.641	-0.454
46[T]	3-Br	6-NO2,7-NH2	0.053	7.275	7.223	0.052
47	3-Br	6-NO2,7-NHMe	0.068	7.167	6.607	0.56
48	3-Br	6-NO2,7-NMe2	2.000	5.698	5.586	0.112
49	3-Br	6-NO2,7-NHCOMe	0.028	7.552	6.607	0.945
50[T]	3-Br	6-NO2,7-OMe	0.015	7.823	7.819	0.004
51	3-Br	6-NO2,7-Cl	0.025	7.6020	7.436	0.166
52	3-Br	6,7 Di-OH	0.00017	9.769	10.067	-0.298
53	3-Br	6,7 Di-OC2H5	0.000006	11.211	10.859	0.352
54[T]	3-Br	6,7 Di-OC3H7	0.00017	9.769	9.451	0.318
55	3-Br	6,7 Di-OC4H9	0.105	6.978	7.644	-0.666
56[T]	3-Br	2-NH2	0.463	6.334	6.259	0.075
57	3-Br	5,6,7-Tri-OMe	0.00067	9.245	10.155	-0.91
58	2-Br	6,7-Di-OMe	0.128	6.892	6.259	0.633
59	4-Br	6,7-Di-OMe	0.00096	9.017	9.259	-0.242
60[T]	3,5-di-Br	6,7-Di-OMe	0.000072	10.142	10.559	-0.417
61	3,5-di-Br	6,7-Di-OMe	0.113	6.946	6.998	-0.052

Expt. = Experimental activity, Pred. = Predicted activity

a = concentration of drug (μM) to inhibit the phosphorylation of a 14-residue fragment of phospholipase by
EGFR (prepared from human A431 carcinoma cell vesicles by immunoaffinity chromatography).

b = -Log (IC_{50} × 10^{-6}): Training data set developed using model 1

[T] Test Set

coefficient of correlation of predicted data set (pred_r^2se) 0.7438 and degree of freedom 38.The model is validated by α_ran_r^2 = 0.00000, α_ran_q^2 = 0.00000, α_ran_pred_r^2 = 0.001, best_ran_r^2 = 0.37610, best_ran_q^2 = 0.03060, Z score_ran_r^2 = 8.47533 and Z score_ran_q^2 = 7.40448.The randomization test suggests that the developed model have a probability of less than 1% that the model is generated by chance. Statistical data is shown in Table 3.The observed and predicted pIC_{50} along with residual values are shown in Table 1. The plot of observed vs. predicted activity is shown in Fig. (2). From the plot it can be seen that MLR model is able to predict the activity of training set quite well (all points are close to regression line) as well as external.

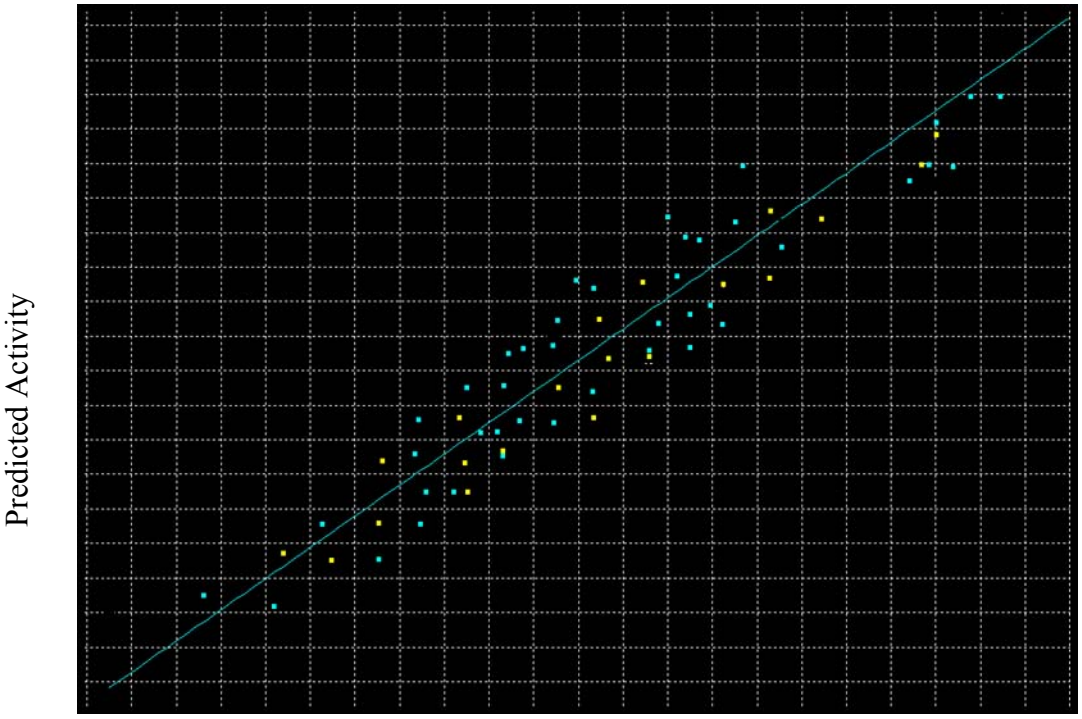

Fig.(2). Graph of Actual vs. Predicted activities for training and test set molecules from the Multiple Linear Regression model. A) Training set (Blue dots) B) Test Set (Yellow dots).

The descriptors which contribute for the pharmacological action are shown in Fig. (3).The major group of descriptors involved sub groups like bromine count, SsOHcount and alignment independent descriptors, help in understanding the effect of substituent at different position of quinazolines.

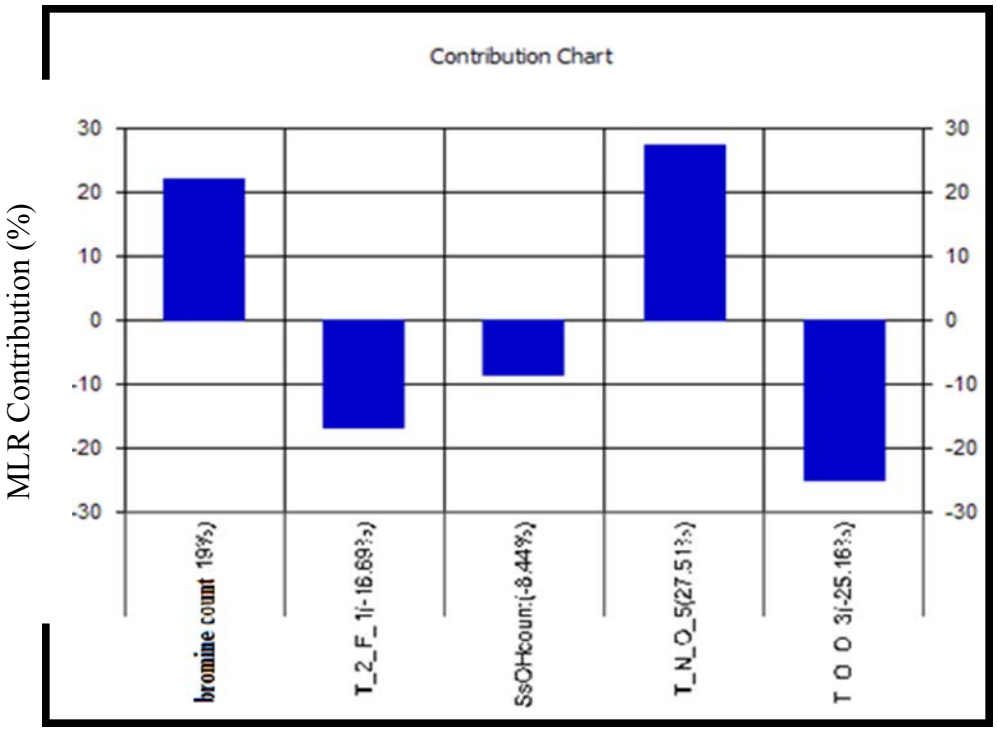

Fig.(3). Plot of percentage contribution of each descriptor in developed MLR model explaining

variation in the activity.

The above study leads to the development of statistically significant QSAR model, which allows understanding of the molecular properties/features that play an important role in governing the variation in the activities. In addition, this QSAR study allowed investigating influence of very simple and easy-to-compute descriptors in determining biological activities, which could shed light on the key factors that may aid in design of novel potent molecules.

The present QSAR model reveals that Baumann's alignment independent topological descriptor has a major contribution in explaining variation in activities. In general, a descriptor T_X_Y_Z can be defined as a count of fragments formed with atom types X and Y separated by topological distance of Z bonds. The definition for the descriptors that were found to be dominating in the developed QSAR models is given below.

T_2_F_1: This descriptor means the count of number of double bonded atoms (single, double or triple bonded) seprated from any fluorine atom (single or double bonded) by one bond distance, e.g., □C_F.

T_N_O_5: This descriptor means the count of number of nitrogen atoms (single, double or triple bonded) seprated from any oxygen atom (single or double bonded) by five bond distance,e.g., N_C_C_C_C_O.

T_O_O_3: This descriptor means the count of number of oxygen atoms (single, double or triple bonded) seprated from any oxygen atom (single or double bonded) by three bond distance, e.g., O_C_C_O.

The careful examination of the descriptors in the model suggests that descriptor T_2_F_1 (-16.69%) is inversly proportional to the activity and shows that any increase in carbon chain between the fluorine atom and aniline moiety which is present at C-4 will increase the activity. In other language we can say that Fluorine atom should not be directly attached with aniline ring as in case of compounds 10, 21 and 32.

The presence of descriptor T_N_O_5 (having positive MLR contribution of 27.51%) in the QSAR model reveals that the presence of -O-CH_3, O-C_2H_5 or aryloxy at C-6 and C-7 position of 4-anilino quinazoline is favorable for the activity as in case of compounds 7, 8, 27 and structurally similar other analogs in the series, in which -N- atom at 3rd position of quinazoline ring is separated from - O - atom at 6th position by five bond.

The negative contribution of T_O_O_3 descriptor (-25.16%) shows that there should be maximal distance (more than 3 bond) between the two Methoxy group which is present at C-6 and C-7 of 4-anilino quinazoline ring as in case of compound 57.

An estate number descriptor SsOH count (-8.44%), which represents total number of hydroxy group connected with one single bond is inversely proportional to the activity. It reveals that hydroxy group should not be directly attached with aniline ring for maximal activity.

The next influential descriptor is bromine count (22.19%) directly proportional to the activity and shows the role of the total number of bromine atom in a molecule. It reveals that presence of electron withdrawing groups over the 4-anilino quinazoline is favourable for the activity as in case of compound 11,13 and similar analogues.

Table 3. Statistical parameters of MLR, PCR and PLS

Parameters	**MLR**	**PCR**	**PLS**
N	42	48	46
Df	38	46	42
r^2	0.912	0.7998	0.8249
q^2	0.8008	0.713	0.7513
F test	60.5149	45.5477	68.5489

r2 se	0.6675	0.555	0.5912
q2 se	0.5732	0.6369	0.6786
pred_r2	0.6042	0.6399	0.6462
pred_r2se	0.7438	0.6973	0.6911
best_ran_r2	0.37561	0.28217	0.22974
best_ran_q2	0.30601	0.12257	0.10927
Z score_ran_r2	8.47533	14.18812	12.05239
Z score_ran_q2	7.4044	10.25775	9.23195
α_ran_r2	0.00000	0.00000	0.00000
α_ran_q2	0.00000	0.00000	0.00000
α_ran_pred_r2	0.001	0.00004	0.00005

MLR = Multiple Linear Regression, PCR = Principal Component Regression, PLS = Partial Least Squares, n = number of molecules of training set, df =degree of freedom, r2 = coefficient of determination, q2 = cross validated r2, pred_r2 = r2 for external test set, pred_r2se = coefficient of correlation of predicted data set.

MODEL – 2 (Principal Component Regression (PCR) Analysis)

Model - 2 is having following QSAR equation.

pIC50 = 0.8868 (bromine count) – 1.3601 (SsCIE-index) + 0.2012 (T_2_N_5) – 1.8114 (T_2_F_1) + 6.6729.

The model -2 gave correlation coefficient (r^2) of 0.799, cross validated correlation coefficient (q^2) of 0.713, F test of 45.5477 , r^2 for external test set (pred_r^2) 0.6399 , coefficient of correlation of predicted data set (pred_r^2se) 0.6973 and degree of freedom 46. The model is validated by α_ran_r^2 = 0.00000, α_ran_q^2 = 0.00000, α_ran_pred_r^2 = 0.00004, best_ran_r^2 = 0.28217, best_ran_q^2 = 0.12257, Z score_ran_r^2 = 14.18812 and Z score_ran_q^2 = 10.25775.The randomization test suggests that the developed model have a probability of less than 1% that the model is generated by chance. Statistical data is shown in Table 3. The plot of observed vs. predicted activity is shown in Fig.(4) .

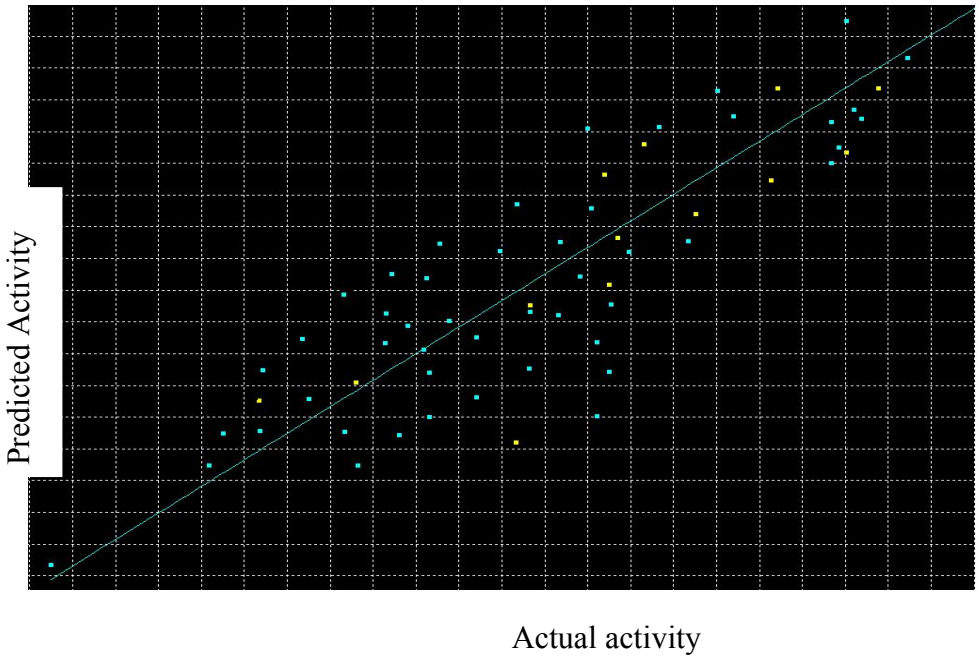

Fig.(4). Graph of Actual vs. Predicted activities for training and test set molecules by Principal Component Regression model. A) Training set (Blue dots) B) Test Set (Yellow dots).

The descriptors which contribute for the pharmacological action are shown in Fig.(5) .

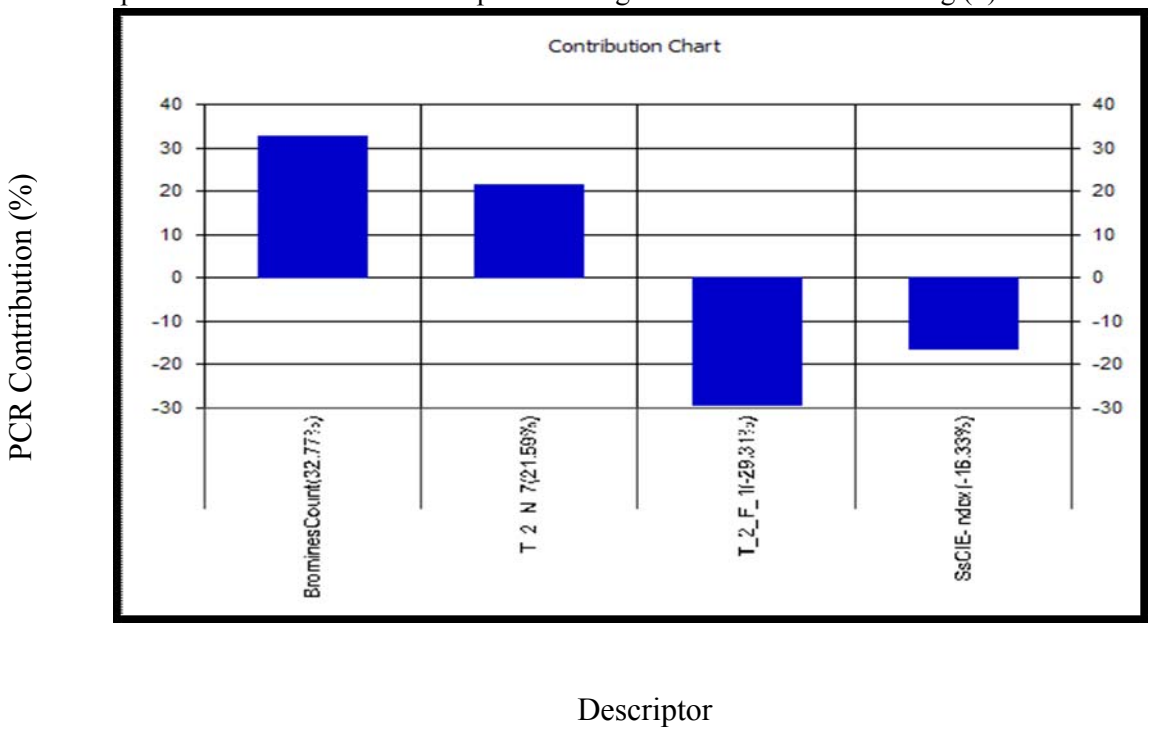

Fig.(5). Plot of percentage contribution of each descriptor in developed PCR model explaining variation in the activity.

Baumann's alignment independent topological descriptors T_2_F_1 (-29.31%) and element count descriptor for bromine atom (32.77%) are common between PCR and MLR; only differs from each other in their percentage of contribution. The definition of the remaining descriptors that were found to be dominating in the developed QSAR models is given below.

T_2_N_7: This descriptor means the count of number of double bonded atoms (single, double or triple bonded) seprated from any nitrogen atom (single or double bonded) by seven bond distance, e.g., □C_C_C_C_C_C_C_N.

The presence of descriptor T_2_N_7 (having positive PCR coefficient of 21.59%) in the QSAR model reveals that the presence of amino group at C-4 position of quinazoline is essential for interaction with tyrosine kinase (EGFR).

The inverse relationship of Electrotopological state indices descriptor, SsCIE-index (-16.33%) defines the number of bromine atom connected with one single bond, indicates that bromine atom should not be directly attached to the anilino ring at C-4 position of quinazoline ring for maximal activity.

MODEL – 3 (Partial Least Squares (PLS) Analysis)

Model - 3 is having following QSAR equation.

pIC50 = + 0.8867 (T_2_N_7) – 1.8572 (T_2_F_1) + 0.8054 (SaasCE-index) - 1.4965(T_C_N_7) + 0.4098 (Chlorine count) + 7.6315.

The model -3 gave 67% variance of prediction with correlation coefficient (r^2) of 0.8249, cross validated correlation coefficient (q^2) of 0.751,F test of 68.54 , r^2 for external test set (pred_r^2) 0.6462 , coefficient of correlation of predicted data set (pred_r^2se) 0.6911 and degree of freedom 42. The model is validated by α_ran_r^2 = 0.00000, α_ran_q^2 = 0.00000, α_ran_pred_r^2 = 0.00005, best_ran_r^2 = 0.22974, best_ran_q^2 = 0.10927, Z score_ran_r^2 = 12.05239 and Z score_ran_q^2 = 9.23195.The randomization test suggests that the developed model have a probability of less than 1% that the model is generated by chance. Statistical data is shown in

Table 4. The plot of observed vs. predicted activity is shown in Fig.(6) .

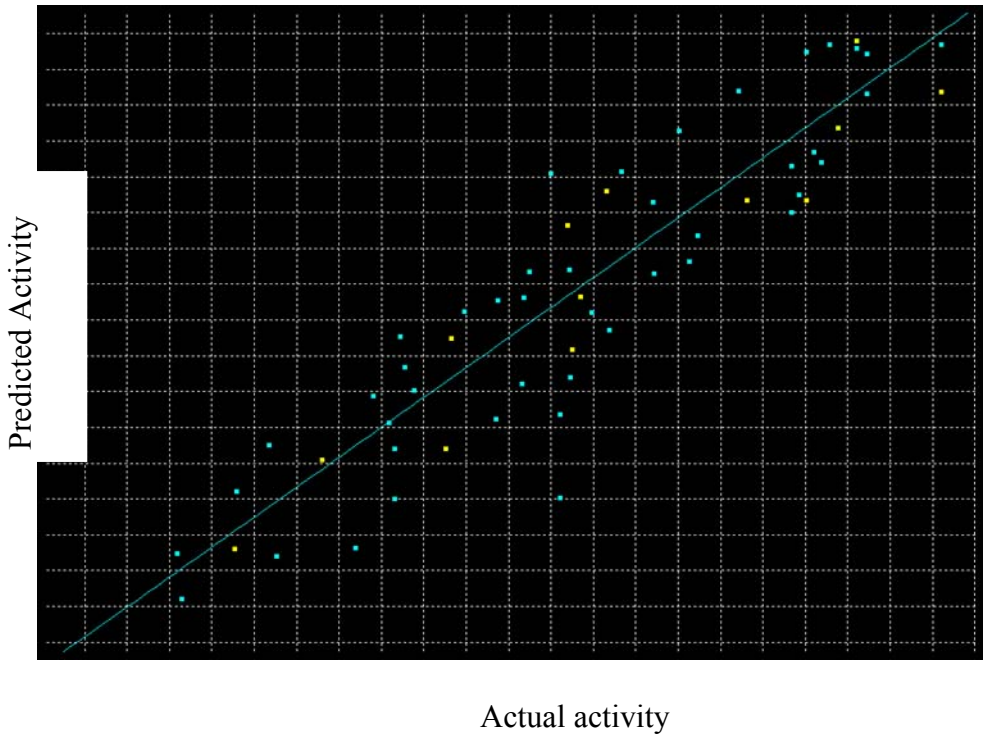

Fig.(6). Graph of Actual vs. Predicted activities for training and test set molecules by Partial Least Square model. A) Training set (Blue dots) B) Test Set (Yellow dots).

The descriptors which contribute for the pharmacological action are shown in Fig.(7). The QSAR model by PLS reveals that Baumann's alignment independent topological descriptor has a major contribution in explaining variation in activities. Descriptors T_2_F_1 (-21.93%) is common among all the three methods and T_2_N_7 (25.61%) is common between PLS and PCR, only the difference is of percentage contribution. The definition of the remaining descriptors that were found to be dominating in the developed QSAR models is given below.

The chlorine count descriptor (11.82%) is directly proportional to the activity and shows the role of the total number of chlorine atom in a molecule. It reveals that presence of electron withdrawing groups over the 4-anilino quinazoline is favourable for the activity as in case of compounds 18,24 and 29.

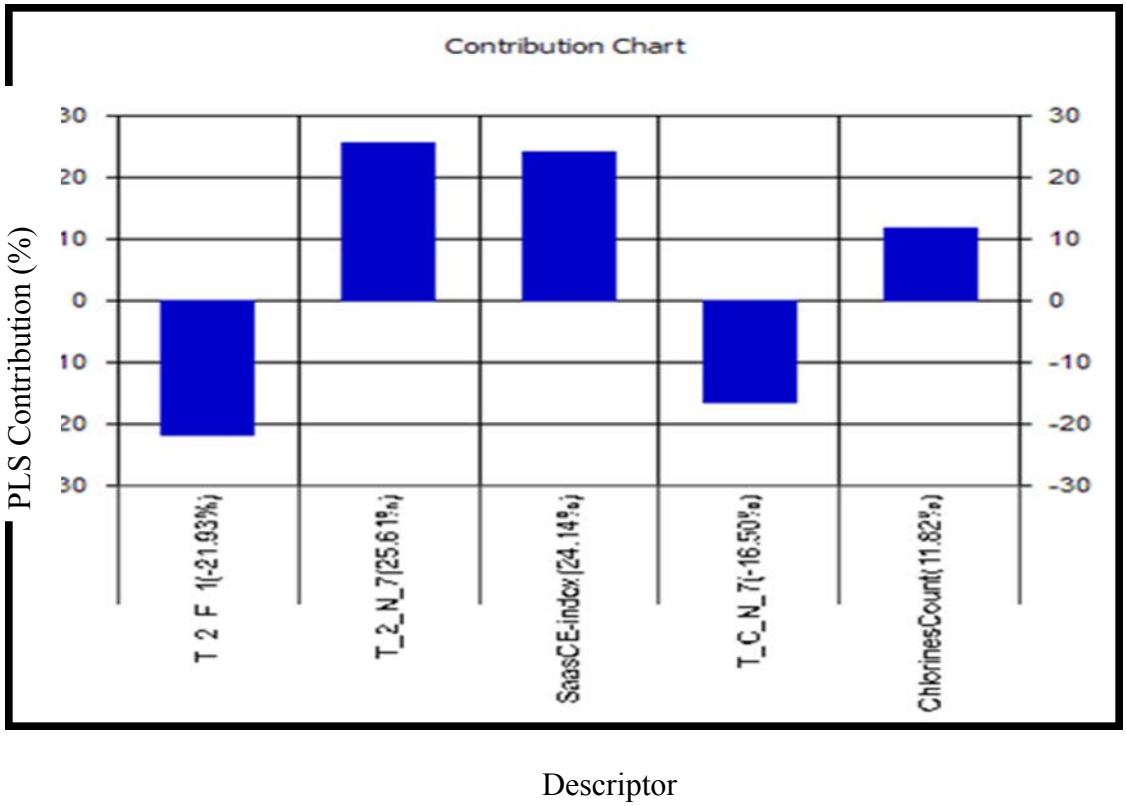

Fig.(7). Plot of percentage contribution of each descriptor in developed PLS model explaining variation in the activity.

The next important Electrotopological indices descriptor is SaasCE-index (24.14%) directly contributing to the activity and shows the importance of number of carbon atom connected with one single bond along with two aromatic bonds.

QUINAZOLINE – EGFR (1M17.pdb) INTERACTION (HYPOTHETICAL MODEL)

The values obtained from the descriptors calculations explain the structural parameters and the possible interaction with the binding site of enzyme. Quinazoline act primarily by binding to ATP binding site of protein kinase. Though ATP binding site is highly conserved among the protein kinase, architecture in the regions proximal to ATP binding site does afford key diversity [18]. The binding interactions of quinazolines with nucleotide are of lipophilic/van der Waals nature. Nitrogen atoms of aniline group and quinazoline ring are involved in hydrogen bond formation with the hinge region of protein kinase.

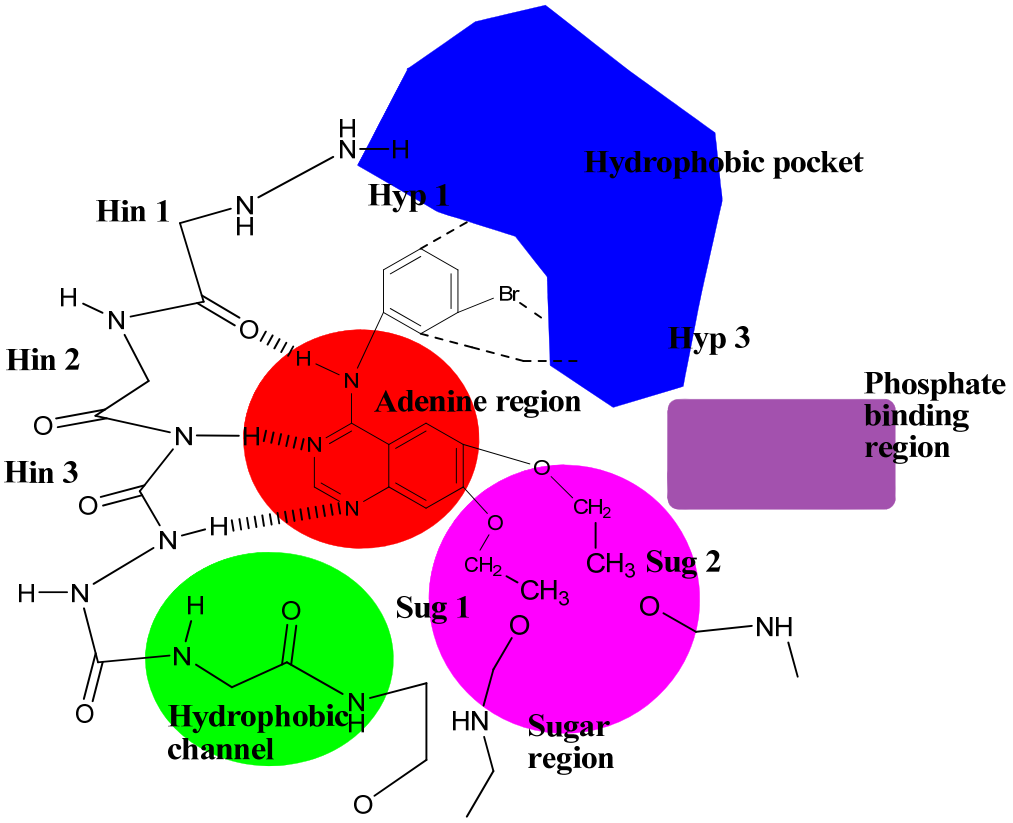

Fig. (8). Proposed hypothetical model of the 4-Anilino Quinazoline derivatives (Compound 53) bound to ATP binding site of EGFR protein tyrosine kinase

Hydrophobic pockets of protein kinase though not used by ATP, but is exploited by most of kinase inhibitors and plays an important role in selecting inhibitors. Hydrophobic channel used to gain binding site of protein kinase is important in improving the selectivity of inhibitors. Methoxy group is going to interact with sugar region. The backbone carbonyl of the residue corresponding to valine serves as a hydrogen bond acceptor for inhibitor binding. Based on these observations, a proposed hypothetical model to explain the drug - protein interaction is depicted in Fig.(8).

6. Conclusions

In conclusion, the model developed to predict the structural features of quinazoline to inhibit EGFR tyrosine kinase, reveals useful information about the structural features requirement for the molecule. In all three optimized models, Multiple Linear Regression method is giving very significant results .The alignment independent descriptors, bromine count, chlorine count, Sss OH-count, SsCIE-index and SaaCHE-index were major contributing descriptors. Descriptor values obtained helps us to understand the structural features required by ATP binding site of EGFR tyrosine kinase. The QSAR results obtained are in agreement with the observed SAR of quinazoline studied. Hence the model proposed in this work is useful in describing QSAR of quinazoline derivatives as EGFR tyrosine kinase inhibitor and can be employed to design new derivatives of quinazoline with specific inhibitory activity.

Acknowledgements

The authors would like to thank Director General, Department of Science and Technology, New Delhi for funding the project (Grant.No.SR/FT/LS-0083/2008) and Sardar Sangat Singh Longia, Secretary ASBASJSM College of Pharmacy for providing the necessary facilities.

References

[1] R. Kurup, C. Garg, Hansch, Chem. Rev., **101**, 2573-2600 (2001).
[2] B.D. Palmer, A.J. Kraker, B.G. Hartl, A.D. Panopoulos, R.L. Panek, B.L. Batley, G.H. Lu, T.K Susanne, H.D.H Showalter, W.A. Denny, J. Med. Chem., **42**, 2373-2382.3(1999).
[3] M. Oblak, M. Randic, T. Solmajer, J. Chem. Inf. Comput. Sci.,**40**, 994-1001 (2000).
[4] T. Naumann, T. Matter, J. Med. Chem., **45**, 2366-2378 (2002).
[5] A.J. Bridges, H. Zhou, D.R. Cody, G.W. Rewcastle, A. McMichael, H.D.H Showalter, D.W. Fry, A.J Kraker, W.A. Denny, J. Med. Chem., **39**, 267-276(1996).
[6] K.A. Ma, H. Bower, G. Lin, C. Chen, X. Huang, J. Shi, Biochem. Pharmacol.,69,1785-1794 (2005).
[7] R. Albuschat, W. Lowe, M. Weber, P. Luger, V. Jendrossek, Eur. J. Med. Chem. **39**, 1001- 1011(2004).
[8] A. Bridges, J. Chem. Rev.,**101**, 2541(2001).
[9] G.W. Rewcastle, W.A Denny, A. J. Bridges, H. Zhou, D.R Cody, J. Med. Chem. **38**, 3482 (1995).
[10] A.J. Bridges, H. Zhou, D.R. Cody, G.W. Rewcastle, A. McMichael, H.D.H. Showalter, D.W. Fry, A.J. Kraker, W.A. Denny, J. Med.Chem., **39**, 267 (1996).
[11] T. Hou, L. Zhu, L.; Chen, X Xu, J. Chem. Inf. Comput. Sci., **43**, 273(2003).
[12] S.N. Manjula, M.N. Noolvi, K.V. Parihar, S.A. Manohara Reddy, V. Ramani, A.K. Gadad, G Singh, N.G Kutty, M. Rao,Synthesis and anti-tumour activity of optically active thiourea and their 2-aminobenzothiazole derivatives: A novel class of anticancer agents. Eur J. Med. Chem.,1-7(2009).
[13] A.M. Badiger, M.N. Noolvi, P.V Nayak, QSAR study of Benzthiazole derivatives as p56 lck inhibitors. Lett. Drug. Des. Discov., **3**,550-560(2006).
[14] A.J. Bridges, H. Zhou, D.R. Cody, G.W. Rewcastle, A. McMichael, H.D.H Showalter, D.W. Fry, A.J. Kraker, W.A. Denny, J. Med.Chem., **39**, 267 (1996).
[15] W. Zheng, A. Tropsha, J. Chem. Inf. Comput. Sci., **40**,185 – 194 (2000).
[16] N. Gilbert, W.B. Statistics, Co Saunders, P.A .Philadelphia (1976).
[17] M. Shen, Y. Xiao, A. Golbraikh, V.K. Gombar, A. Tropsha , J. Med .Chem., **46**, 3013 – 3020 (2003).
[18] D. Fabbro, S. Ruetz, E. Buchdnger, S.W. Cowan Jacob, G. Fendrich, J. Liebetanz, J. Mestan, P. Traxler, B. Chaudhari, H. Fertz, J. Zimmermann, T. Mayer, G. Caravatti, P. Furret, P.W. Manley, Protein kinases as targets for anti-cancer agents: from inhibitors to useful drugs. Pharmacol. Ther., **93**, 79-98 (2002).