Journal of Computational Methods in Molecular Design, 2012, 2 (3):107-121

Scholars Research Library
(http://scholarsresearchlibrary.com/archive.html)

ISSN : 2231- 3176
CODEN (USA): JCMMDA

2D and 3D QSAR using kNN-MFA method of pyrazolyl-thiazolinone derivatives as potential EGFR and HER-2 kinase inhibitors

Shraddha T. Thombare, Steffi I. Gonsalves and Anwar R. Shaikh*

Bhujbal Knowledge City, MET's Institute of Pharmacy, Adgaon, Nashik-422003, India

---

ABSTRACT

Quantitative structure–activity relationship (QSAR) analysis for recently synthesized pyrazolyl-thiazolinone derivatives was studied for their EGFR and HER-2 kinase inhibitory activities. The statistically significant 2D-QSAR models (r^2 = 0.9086; q^2 = 0.8370; F test = 49.6789; r^2 se = 0.1242; q^2 se = 0.1675; pred_r^2 = 0.8086; pred_r^2se = 0.1934 and r^2 = 0.9163; q^2= 0.8702; F test 54.7057; r^2 se= 0.0820; q^2 se=0.1020; pred_r^2 = 0.8249; pred_r^2se = 0.1195) were developed using molecular design suite (VLifeMDS 4.1). The study was performed with 36 compounds (data set) using sphere exclusion (SE) algorithm, random selection and manual selection methods used for the division of the data set into training and test set. Partial least square regression (PLSR) methodology with stepwise (SW) forward-backward variable selection method was used for building the QSAR models. The results of the 2D-QSAR models were further compared with 3D-QSAR models generated by kNN-MFA, (k-Nearest Neighbor Molecular Field Analysis) investigating the substitutional requirements for the favorable inhibitory activity for EGFR and HER-2 in tumor growth and providing useful information in the characterization and differentiation of their binding sites. The results derived may be useful in further designing novel EGFR and HER-2 kinase inhibitors prior to synthesis.

Keywords: pyrazolyl-thiazolinone, EGFR, HER-2, antitumor, quantitative structure-activity relationship, kNN-MFA

INTRODUCTION

The epidermal growth factor receptor (EGFR; ErbB-1; HER1 in humans) is the cell-surface receptor for members of the epidermal growth factor family (EGF-family) of extracellular protein ligands [1]. The epidermal growth factor receptor is a member of the ErbB family of receptors, a subfamily of four closely related receptor tyrosine kinases: EGFR (ErbB-1), HER2/c-neu (ErbB-2), Her 3 (ErbB-3) and Her 4 (ErbB-4). Mutations affecting EGFR expression or activity could result in cancer [2]. EGFR exists on the cell surface and is activated by binding of its specific ligands, including epidermal growth factor and transforming growth factor α (TGFα). ErbB2 has no known direct activating ligand, and may be in an activated state constitutively or become active upon heterodimerization with other family members such as EGFR. Upon activation by its growth factor ligands, EGFR undergoes a transition from an inactive monomeric form to an active homodimer [3].In addition to forming homodimers after ligand binding, EGFR may pair with another member of the ErbB receptor family, such as ErbB2/Her2/neu, to create an activated heterodimer. EGFR dimerization stimulates its intrinsic intracellular protein-tyrosine kinase activity. As a result, autophosphorylation of several tyrosine (Y) residues in the C-terminal domain of EGFR occurs. These include Y992, Y1045, Y1068, Y1148 and Y1173 [4]. This autophosphorylation elicits downstream activation and signaling by several other proteins that associate with the phosphorylated tyrosines through their own phosphotyrosine-binding SH2 domains. These downstream signaling proteins initiate several signal transduction cascades, principally the MAPK, Akt and JNK pathways, leading to DNA synthesis and cell proliferation [5]. EGFR overexpression (known as upregulation) or overactivity have been associated with a number of cancers, including head and neck, lung, breast, bladder, prostate, kidney and anal cancers and glioblastoma multiforme. Therefore,

EGFR tyrosin kinase represents the attractive target for the development of novel anticancer agents. EGFR and HER-2 are the hottest in currant cancer research and their overexpression or abnormal activation often cause cell malignant transformation [6]. Monoclonal antibodies such as panitumumab and cetuximab, erlotinib, gefitinib, and lapatinib are the representative drugs have been approved by US Food and Drug Administration for treatment of patient with non small cell lung cancer (NSCLC) [7].

Many pyrazole derivatives are acknowledged to possess a wide range of bioactivities. The pyrazole moiety makes up the core structure of numerous biologically active compounds. Thus representatives of this heterocycle exhibit antiviral [8-9], antitumor [10], antibacterial [11-12], antiinflamatory and analgesic [13-14], antidepressant and anticonvulsant [15], MAO-B inhibitors in alzheimer disease [16], anti-angiogenic activity [17]. Thiazolinone and their derivatives have attracted continuing interest over the years because of their varied biological activities, such as antimicrobial, antiproliferative, antiviral/anti-HIV, antidepressant, anticonvulsant, antifungal, and antibacterial [18], antiinflamatory [19], Hai-Liang Zhu et al.2012, reported that the pyrazole ring along with thiazolinone ring, exhibit synergistic anticancer effect that act as potential EGFR and HER-2 kinase inhibitors.

Traditional computer-assisted quantitative structure–activity relationship (QSAR) studies pioneered by C. Hansch et al.1962 [20] have been proved to be one of the useful approaches for accelerating the drug design process [21] which help to correlate the bioactivity of compounds with structural descriptors [22]. Recently synthesized pyrazolyl-thiazolinone derivatives were discovered with the EGFR and HER-2 kinase inhibitors which may have potential utility in treating cancer cells. To gain further insights into the structure–activity relationships of these derivatives and understand the mechanism of their substitutional specificity, we have performed 2D and 3D-QSAR on pyrazolyl-thiazolinone derivatives using Partial Least Squares Regression (PLSR) and k-Nearest Neighbor Molecular Field Analysis (kNN MFA), respectively. The significance of the QSAR models was evaluated using cross-validation tests, randomization tests and external test set prediction. The robust 2D/3D-models may be useful in further designing new candidates as potential EGFR and HER-2 kinase inhibitors prior to synthesis.

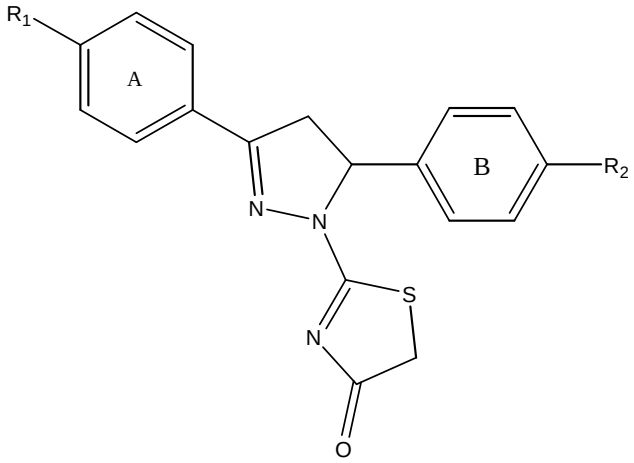

Figure 1: Basic structures of pyrazolyl-thiazolinone derivative

Table 1: structures of pyrazolyl-thiazolinone derivative

compounds	R_1	R_2	compounds	R_1	R_2
E1	H	H	E19	Br	H
E2	H	F	E20	Br	F
E3	H	Cl	E21	Br	Cl
E4	H	Br	E22	Br	Br
E5	H	Me	E23	Br	Me
E6	H	OMe	E24	Br	OMe
E7	F	H	E25	Me	H
E8	F	F	E26	Me	F
E9	F	Cl	E27	Me	Cl
E10	F	Br	E28	Me	Br
E11	F	Me	E29	Me	Me
E12	F	OMe	E30	Me	OMe
E13	Cl	H	E31	OMe	H
E14	Cl	F	E32	OMe	F
E15	Cl	Cl	E33	OMe	Cl
E16	Cl	Br	E34	OMe	Br
E17	Cl	Me	E35	OMe	Me
E18	Cl	OMe	E36	OMe	OMe

MATERIALS AND METHODS

2.1. Selection of molecules

Data set of 36 pyrazolyl-thiazolinone derivatives (Table 1) collected from published literature [23] were taken for the present study. The affinity data of inhibitory activities were converted into pKi values to get the linear relationship in equation using the following formula: pKi= -log Ki, where Ki value represents inhibitory activity in IC_{50} (µM) (Table 2). Molecules were rationally divided into the training set and test set based on the suggestions given by Alexander Tropsha et al. [24].

Table 2: Ki1 (µM) and Ki2 (µM) represent inhibition activities of compounds E1- E2 against EGFR and HER-2 respectively. All inhibitory activities are expressed as –log (Ki), which is pKi

Compounds	Ki1(µM)	Actual pKi1	Ki2(µM)	Actual pKi2
E1	3.38	0.529	5.12	0.709
E2	4.86	0.687	6.35	0.803
E3	3.49	0.543	4.83	0.684
E4	1.35	0.130	3.05	0.484
E5	3.03	0.481	4.64	0.667
E6	4.27	0.63	6.21	0.793
E7	8.14	0.911	10.53	1.022
E8	16.92	1.228	18.12	1.258
E9	10.92	1.038	13.16	1.119
E10	4.79	0.68	6.24	0.795
E11	8.36	0.922	10.26	1.011
E12	10.69	1.029	12.43	1.094
E13	5.34	0.728	7.05	0.848
E14	14.21	1.153	16.42	1.215
E15	8.16	0.912	9.96	0.998
E16	2.28	0.358	3.84	0.584
E17	6.67	0.824	8.13	0.91
E18	8.58	0.933	10.34	1.015
E19	3.20	0.505	4.87	0.688
E20	6.48	0.812	8.14	0.911
E21	4.12	0.615	5.87	0.769
E22	2.03	0.307	3.65	0.562
E23	5.58	0.747	7.04	0.848
E24	7.96	0.901	9.27	0.967
E25	1.08	0.033	2.24	0.35
E26	2.01	0.303	3.53	0.548
E27	1.66	0.22	3.11	0.493
E28	0.24	-0.62	1.07	0.029
E29	1.16	0.064	2.52	0.401
E30	4.24	0.627	6.23	0.794
E31	2.37	0.375	4.12	0.615
E32	5.95	0.775	8.04	0.905
E33	5.35	0.728	7.59	0.88
E34	1.26	0.1	2.75	0.439
E35	5.49	0.737	7.35	0.866
E36	8.89	0.949	10.48	1.02

2.2. Molecular modeling

All computational experiments were performed using on HCL computer having genuine Intel Pentium Dual Core Processor and Windows XP operating system using the software Molecular Design Suite (vlifeMDS 4.1) [25]. Structures were drawn using the 2D draw application and converted to 3D structures and subjected to an energy minimization and geometry optimization using Merck Molecular Force Field, force field and charges followed by Austin Model-1 with 10000 as maximum number of cycles, 0.01 as convergence criteria (root mean square gradient) and 1.0 as constant (medium's dielectric constant which is 1 for in vacuo) in dielectric properties. The default values of 30.0 and 10.0 Kcal/mol were used for electrostatic and steric energy cutoff.

2.3. 2D-QSAR analysis

2.3.1. Calculation of descriptors

Number of descriptors was calculated after optimization or minimization of the energy of the data set molecules. Various types of physicochemical descriptors were calculated: Individual (Molecular weight, H-Acceptor count, H-Donor count, XlogP, slogP, SMR, polarisablity, etc.), retention index (Chi), atomic valence connectivity index (ChiV), Path count, Chi chain, ChiV chain, Chain PathCount, Cluster, Pathcluster, Kappa, Element count (H, N, C, S count etc.), Distance based topological (DistTopo, ConnectivityIndex, WienerIndex, Balaban Index), Estate numbers (SsCH3count, SdCH2count, SssCH2count, StCHcount, etc.), Estate contribution (SsCH3-index., SdCH2-

index, SssCH2-index , StCH index), Information theory based (Ipc, Id etc.) and Polar surface area. More than 200 alignment independent descriptors were also calculated using the following attributes. A few examples are T_2_O_7, T_N_N_5, T_2_2_6, T_C_O_1, T_O_Cl_5 etc. The invariable descriptors (the descriptors that are constant for all the molecules) were removed, as they do not contribute to QSAR.

2.3.2. Generation of training and test sets:
In order to evaluate the QSAR model, data set was divided into training and test set using sphere exclusion, random selection and manual selection method. Training set is used to develop the QSAR model for which biological activity data are known. Test set is used to challenge the QSAR model developed based on the training set to assess the predictive power of the model which is not included in model generation.

Sphere Exclusion method: In this method initially data set were divided into training and test set using sphere exclusion method. In this method dissimilarity value provides an idea to handle training and test set size. It needs to be adjusted by trial and error until a desired division of training and test set is achieved. Increase in dissimilarity value results in increase in number of molecules in the test set.

Random Selection Method: In order to construct and validate the QSAR models, both internally and externally, the data sets were divided into training [90%-60% (90%, 85%, 80%, 75%, 70%, 65% and 60%) of total data set] and test sets [10%-40% (10%, 15%, 20%, 30%, 35% and 40%) of total data set] in a random manner. 10 trials were run in each case.

Manual data selection method: Data set is divided manually into training and test sets on the basis of the result obtained in sphere exclusion method and random selection method.

2.3.3. Generation of 2D-QSAR models:
PLSR was used for model generation. PLSR is an expansion of the multiple linear regression (MLR) models. In its simplest form, a linear model specifies the (linear) relationship between a dependent (response) variable and a set of predictor variables. PLSR extends MLR without imposing the restrictions employed by discriminant analysis, principal component regression (PCR) and canonical correlation. In PLSR, prediction functions are represented by factors extracted from the Y'XX'Y matrix. The number of such prediction functions that can be extracted typically will exceed the maximum of the number of Y and X variables. PLSR is probably the least restrictive of the various multivariate extensions of the multiple linear regression models. This flexibility allows it to be used in situations where the use of traditional multivariate methods is severely limited, such as when there are fewer observations than predictor variables. PLSR can be used as an exploratory analysis tool to select suitable predictor variables and to identify outliers before classical linear regression. All the calculated descriptors were considered as independent variable and biological activity as dependent variable.

2.4. 3D-QSAR analysis:
2.4.1. kNN-MFA
kNN-MFA is novel methodology, unlike conventional QSAR regression methods; this methodology can handle nonlinear relationships of molecular field descriptors with biological activity, thus making it a more accurate predictor of biological activity. Conventional correlation methods try to generate linear relationship with the activity, where as kNN is inherently non-linear method and is better able to explain activity trends. The kNN technique is a conceptually simple approach to pattern recognition problems. In this method, an unknown pattern is classified according to the majority of the class memberships of its k nearest neighbors in the training set. The nearness is measured by an appropriate distance metric (e.g. a molecular similarity measure, calculated using field interactions of molecular structures). The standard kNN method is implemented simply as follows: (i) calculate distances between an unknown object (u) and all the objects in the training set; (ii) select k objects from the training set most similar to object u, according to the calculated distances, (iii) classify object u with the group to which a majority of the k objects belong. An optimal k value is selected by the optimization through the classification of a test set of samples or by the leave-one out cross-validation. The variables and optimal k values are chosen using different variable selection methods as described below.

kNN-MFA with Simulated Annealing
Simulated Annealing (SA) is another stochastic method for function optimization employed in QSAR. Simulated annealing (SA) is the simulation of a physical process, 'annealing', which involves heating the system to a high temperature and then gradually cooling it down to a preset temperature (e.g., room temperature). During this process, the system samples possible configurations distributed according to the Boltzmann distribution so that at equilibrium, low energy states are the most populated.
kNN-MFA with Stepwise (SW) Variable Selection

This method employs a stepwise variable selection procedure combined with kNN to optimize the number of nearest neighbors (k) and the selection of variables from the original pool as described in simulated annealing.

kNN-MFA with Genetic Algorithm

Genetic algorithms (GA) first described by Holland mimic natural evolution by modeling a dynamic population of solutions. The members of the population, referred to as chromosomes, encode the selected features. The encoding usually takes form of bit strings with bits corresponding to selected features set and others cleared. Each chromosome leads to a model built using the encoded features. By using the training data, the error of the model is quantified and serves as a fitness function. During the course of evolution, the chromosomes are subjected to crossover and mutation. By allowing survival and reproduction of the fittest chromosomes, the algorithm effectively minimizes the error function in subsequent generations.

2.4.2. Alignment rules:

Molecular alignment was used to visualize the structural diversity in the given set of molecules. This was followed by generation of common rectangular grid around the molecules. The template structure, i.e. unsubstituted pyrazolyl-thiazolinone was used for alignment by considering the common elements of the series as shown in Figure 2. The reference molecule 28 is chosen high inhibitory effect which made it a valid lead molecule and therefore was chosen as a reference molecule. After optimizing, the template structure and the reference molecule were used to superimpose all molecules from the series using the template alignment method. kNN-MFA method requires suitable alignment of given set of molecules after optimization; alignment was carried out by template based alignment method. Stereoview of aligned molecules in training set and test set is shown in Figure 3.

Figure 2: Template molecule

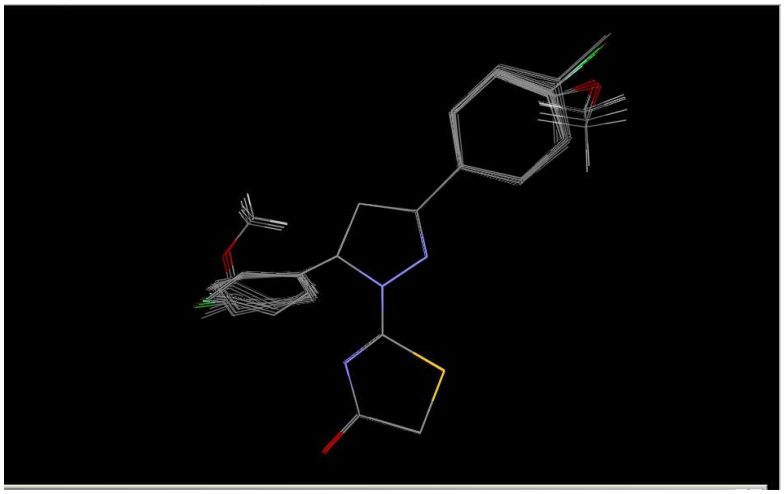

Figure 3: Stereoview of aligned molecules

2.4.3. Creation of interaction energies

Methyl probe with charge 1 and energy cut-off for electrostatic 10 Kcal/mol and for steric 30 Kcal/mol, dielectric constant 1 and charge type Gasteiger-marsili were used to calculate steric and electrostatic fields. The fields were computed at each lattice intersection of a regularly spaced grid of 2.0 A° within defined three-dimensional region.

2.4.4. Generation of training and test sets

In order to evaluate the QSAR model, data set was divided into training and test set using sphere exclusion, random selection and Manual selection method. Training set is used to develop the QSAR model for which biological activity data are known. Test set is used to challenge the QSAR model developed based on the training set to assess the predictive power of the model which is not included in model generation.

RESULTS AND DISCUSSION

3.1 2D-QSAR models

Different sets of 2D-QSAR models were generated using the PLSR analysis in conjunction with stepwise forward-backward variable selection method. Different training and test set were constructed using sphere exclusion, random and manual selection method. Training and test set were selected if they follow the unicolumn statistics, i.e., maximum of the test is less than maximum of training set and minimum of the test set is greater than of training set, which is prerequisite for further QSAR analysis. This result shows that the test is interpolative i.e., derived from the min-max range of training set. The mean and standard deviation of the training and test set provides insight to the relative difference of mean and point density distribution of the two sets. The statistical significant 2D-QSAR models for pKi1 and pKi2 are given in Table 3.

Table 3: Statistical evaluation of 2D-QSAR models for pKi1 and pKi2

Models	r^2	q2	r^2se	q^2se	pred_r^2	F test
pKi1						
1	0.9086	0.8336	0.1242	0.1675	0.8086	49.6789
2	0.8583	0.7138	0.1576	0.2188	0.8896	29.0050
3	0.8410	0.7522	0.1679	0.2096	0.7704	37.0280
4	0.8637	0.7861	0.1195	0.1497	0.7278	44.3586
5	0.8408	0.7601	0.1671	0.2052	0.7221	42.2478
pKi2						
1	0.9163	0.8702	0.0820	0.1020	0.8249	54.7057
2	0.7548	0.6471	0.1422	0.1706	0.4907	17.6988
3	0.7981	0.6997	0.1335	0.1628	0.7008	19.7688
4	0.8798	0.8345	0.1019	0.1196	-0.3342	49.3884
5	0.6201	0.5246	0.1751	0.1958	0.0718	14.6923

The selection of the best model is based on the values of r^2 (squared correlation coefficient), q^2 (cross-validated correlation coefficient), pred_r^2 (predicted correlation coefficient for the external test set), *F* (Fisher ratio) reflects the ratio of the variance explained by the model and the variance due to the error in the regression. High values of the F–test indicate that the model is statistically significant. r^2se, q^2se and pred_r^2se are the standard errors terms for r^2, q^2 and pred_r^2 respectively. The statistically significant 2D-QSAR model is shown as follows.

Model-1 (Test set: 5,6,7,8,9,12,14,19,30,35,36)
pKi1 = -0.5365 (T_2_C_6); + 0.5307 (H-AcceptorCount)
+0.1426 (T_T_Cl_7); + 0.1233 (T_2_C_7)
-0.0979 (T_2_T_7) + 22.3512

Statistics:
[Optimum Components= 4; n= 25; Degree of freedom= 20;
r^2 = 0.9086; q^2= 0.8336; F test= 49.6789; r^2se= 0.1242;
q^2se= 0.1675; pred_r^2= 0.8086; pred_r^2se = 0.1934]

Model-2 (Test size:1,6,11,13,18,19,22,28,30,31,33)
pKi2 = +0.3810 (H-AcceptorCount); -0.3070 (T_2_C_6)
+0.2479 (ChlorinesCount); + 0.0873 (T_2_C_7)
-0.0664 (T_2_T_7) +12.9440;

Statistics:
[Optimum Components= 4; n= 25; Degree of freedom = 20;

r^2= 0.9163; q^2= 0.8702, F test= 54.7057; r^2se= 0.0820;
q^2se= 0.1195; pred_r^2= 0.8249; pred_r^2se= 0.1195]

In the above QSAR equations, n is the number of molecules (Training set) used to derive the QSAR model, r^2 is the squared correlation coefficient, q^2 is the cross-validated correlation coefficient, pred_r^2 is the predicted correlation coefficient for the external test set, *F* is the Fisher ratio, reflects the ratio of the variance explained by the model and the variance due to the error in the regression. High values of the F-test indicate that the model is statistically significant. r^2se, q^2se and pred_r^2se are the standard errors terms for r^2, q^2 and pred_r^2 (smaller is better).

Interpretation of the Models:
Model-1
From equation, model 1 explains 90.86 % (r^2= 0.9086) of the total variance in the training set as well as it has internal (q^2) and external (pred_r^2) predictive ability of 83.36 % and 80.86 % respectively. The F test shows the statistical significance of 99.99 % of the model which means that probability of failure of the model is 1 in 10000. In addition, the randomization test shows confidence of 99.9999 (Alpha Rand Pred R^2 = 0.00000) that the generated model is not random and hence chosen as the QSAR model. From QSAR model 1, negative coefficient value of T_2_C_6 [count of number of double bonded atoms (i.e. any double bonded atom, T_2) separated from carbon atom by 6 bonds],T_2_T_7 [count of any bond separated from any atom by 7 bonds] on the biological activity indicated that lower values leads to good inhibitory activity while higher value leads to reduced inhibitory activity while positive coefficient value of H-AcceptorCount [number of hydrogen bond acceptor atoms], T_T_Cl_7 [count of any atom (represented as T) separated from Cl atom by 7 bonds], T_2_C_7 [count of number of double bonded atoms (i.e. any double bonded atom, T_2) separated from carbon atom by 7 bonds], on the inhibitory activity indicated that higher value leads to better inhibitory activity whereas lower value leads to decrease inhibitory activity.

Contribution chart for model 1 is represented in Figure 4 reveals that the descriptors H-AcceptorCount, T_T_Cl_7, T_2_C_7, contributing 53.07 %, 14.26 % and 12.33 % respectively. Two more descriptors T_2_C_6 and T_2_T_7 are contributing inversely 53.65 %, 9.79 % respectively to biological activity.

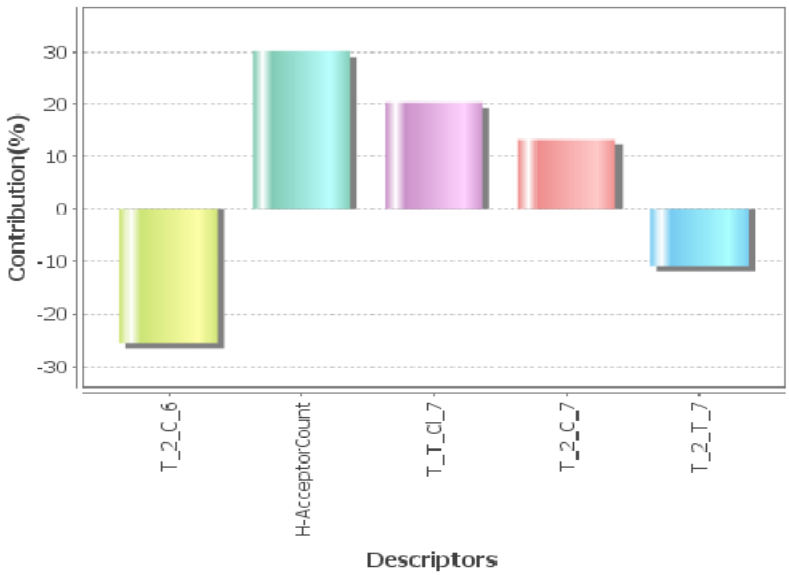

Figure 4: Contribution chart for model-1 showing contribution of different descriptors

Data fitness plot for model 1 is shown in Figure 5. The plot of observed vs predicted activity provides an idea about how well the model was trained and how well it predicts the activity of external test set.

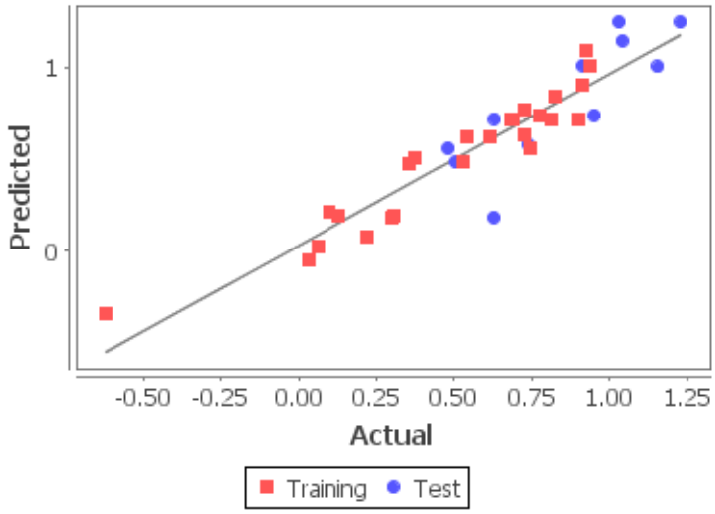

Figure 5: Data fitness plot for model-1

The graph of observed vs. predicted activity of training and test sets for model 1 is shown in Figure 6, it reveals that the model is able to predict the activity of training set quite well as well as external test set, providing confidence of model. Result of the observed and predicted inhibitory activity for the training and test compounds for the Model 1 is shown in Table 5.

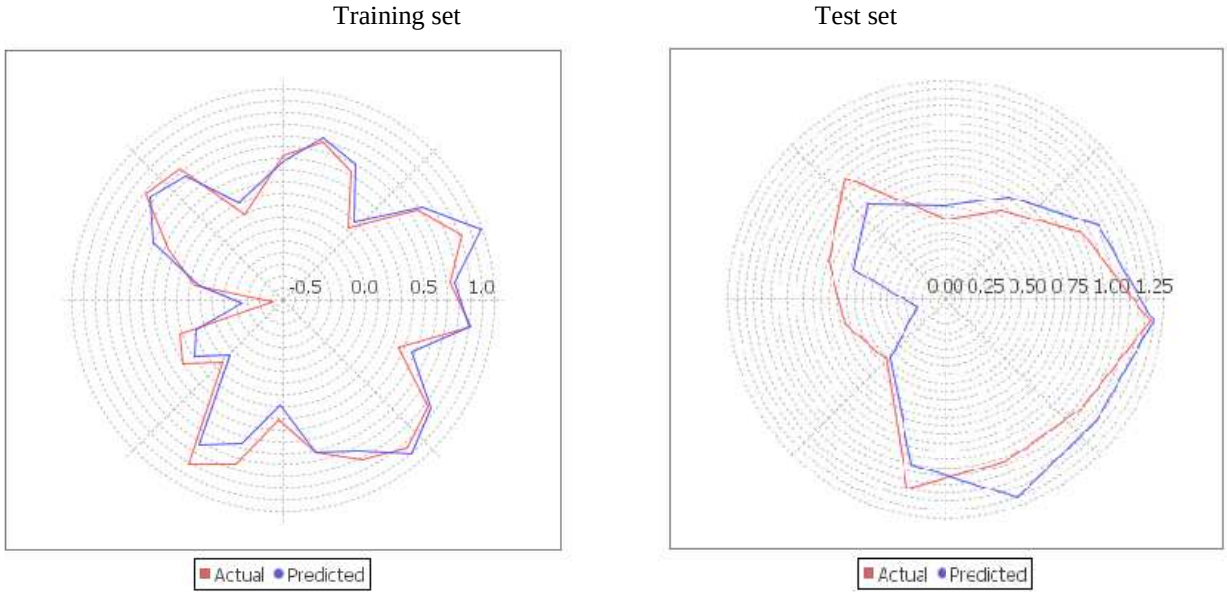

Figure 6: Graph between actual and predicted biological activity of training and test set for Model-1.

Model-2

From equation, Model 2 explains 91.63 % (r^2= 0.9163) of the total variance in the training set as well as it has internal (q^2) and external ($pred\_r^2$) predictive ability of 87.02% and 82.49% respectively. From QSAR model 2, it was observed that the positive coefficient value of H-AcceptorCount [number of hydrogen bond acceptor atoms], ChlorinesCount [number of chlorine atoms in compound], T_2_C_7 [count of number of double bonded atoms (i.e. any double bonded atom, T_2) separated from carbon atom by 7 bonds] on the biological activity indicated that higher value leads to better inhibitory activity whereas lower value leads to decrease inhibitory activity. Negative coefficient value of T_2_C_6 [count of number of double bonded atoms (i.e. any double bonded atom, T_2) separated from carbon atom by 6 bonds], T_2_T_7 [count of any bond separated from any atom by 7 bonds] on the biological activity indicated that lower values leads to good inhibitory activity while higher value leads to reduced inhibitory activity.

Contribution chart for model 2 is represented in Figure 7, it reveals that the descriptors H-AcceptorCount, ChlorinesCount, T_2_C_7 contributing 38.10%, 24.79% and 8.73 % respectively. Two more descriptors T_2_C_6 and T_2_T_7 are contributing inversely 30.70 %, 6.64 % respectively to biological activity.

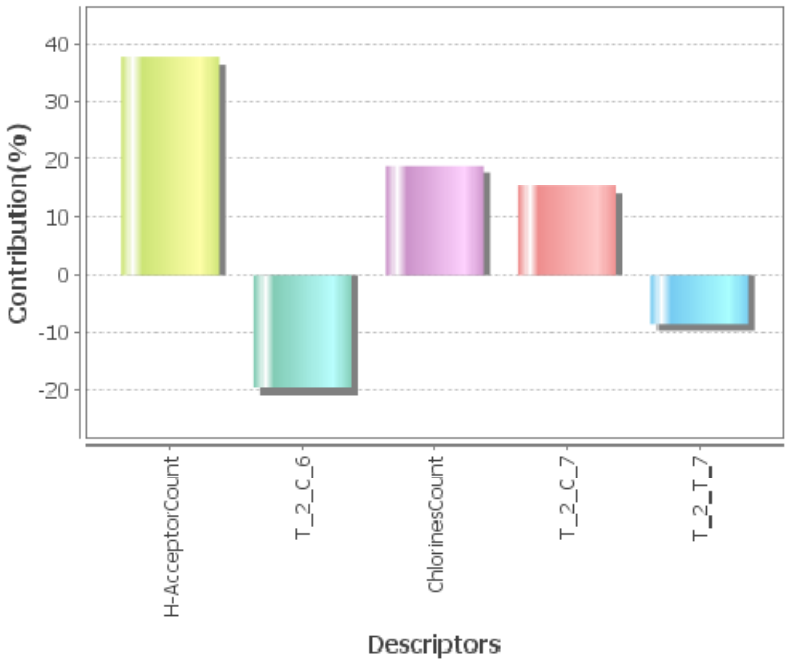

Figure 7: Contribution chart for model-2 showing contribution of different descriptors

Data fitness plot for model 2 is shown in Figure 8; the plot of observed vs predicted activity provides an idea about how well the model was trained and how well it predicts the activity of external test set.

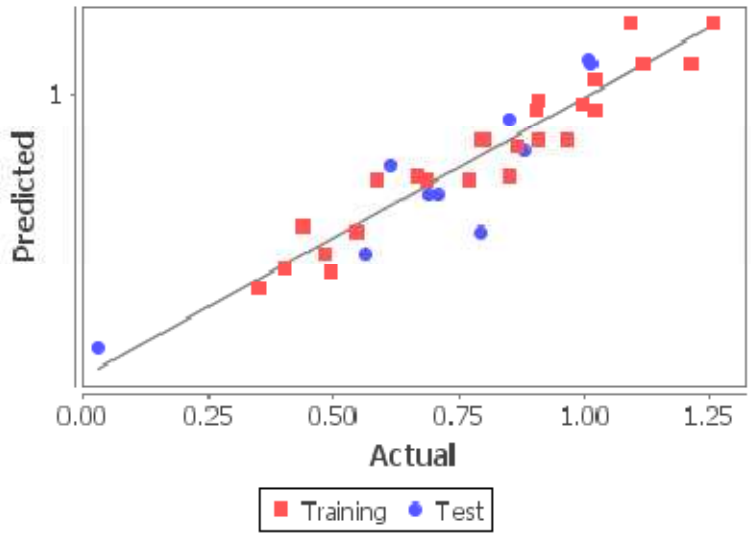

Figure 8: Data fitness plot for model-2

The graph of observed vs. predicted activity of training and test sets for model 2 is shown in Figure 9, reveals that the model is able to predict the activity of training set quite well as well as external test set, providing confidence of model. Result of the observed and predicted inhibitory activity for the training and test compounds for the Model 2 is shown in Table 5.

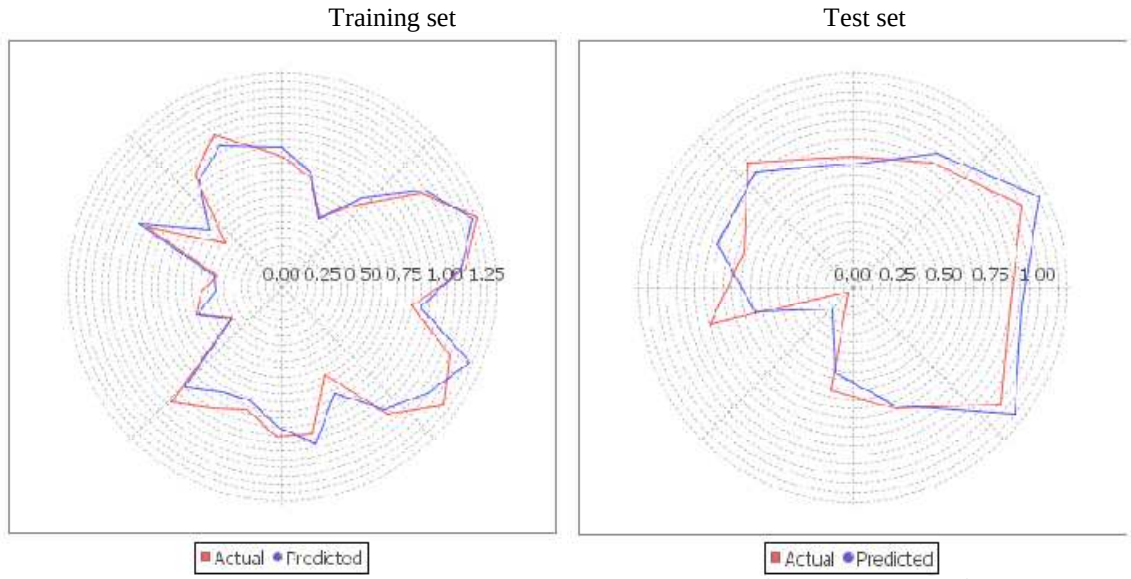

Figure 9: Graph between actual and predicted biological activity of training and test set for Model-2.

3.2. 3D-QSAR model

kNN-MFA samples the steric and electrostatic fields surrounding a set of ligands and constructs 3D-QSAR models by correlating these 3D fields with the corresponding biological activities. The statistical significant 3D-QSAR models for pKi1 and pKi2 are given in Table 4.

Table 4: Statistical evaluation of 3D-QSAR models for pKi1 and pKi2

Models	kNN	DOF	q^2	q2_se	pred_r^2	pred_r^2se
pKi1						
1	2	24	0.7194	0.1728	0.5587	0.3675
2	2	25	0.5148	0.2695	0.6807	0.2113
3	3	26	0.6285	0.2465	-0.1647	0.2983
4	2	18	0.6749	0.1904	0.3171	0.3591
5	2	29	0.6596	0.2301	-1.0269	0.2779
pKi2						
1	2	25	0.7034	0.1517	0.5080	0.1466
2	2	23	0.4922	0.1712	0.1636	0.2806
3	2	21	0.8281	0.1053	-0.2003	0.3223
4	2	21	0.6176	0.1486	0.1312	0.2860
5	2	23	0.7003	0.1378	-0.3363	0.3572

The selection of the best model is based on the values of q^2 (internal predictive ability of the model) and that of pred_r^2 (the ability of the model to predict the activity of external test set). The statistical significant 3D-QSAR models for pKi (model-3) and pKi2 (model-4) are given below.

Model-3

pKi1 = E_709 (0.0702 0.0745); S_391 (-0.1301 -0.1222)
S_784 (-0.0179 -0.0172)

Statistics:

[kNN= 2; n= 28; Degree of freedom= 24; q^2= 0.7194;
q^2_se= 0.1728; pred_r^2= 0.5587; pred_r^2se= 0.3675

The model 3 explains values of k (2), q^2 (0.7194), pred_r^2 (0.5587), q^2_se (0.1728), and pred_r^2 se (0.3675) prove that QSAR equation so obtained is statistically significant and shows the predictive power of the model is 71.94% (internal validation). Table 5 represents the predicted inhibitory activity by the model 3 for training and test set.

The data fitness plot for model 3 is shown in Figure 10. The plot of observed vs predicted activity provides an idea about how well the model was trained and how well it predicts the activity of the external test set.

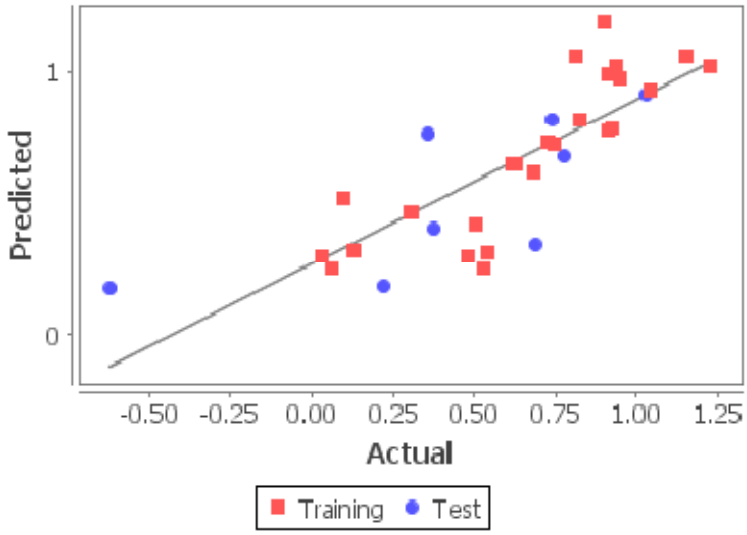

Figure 10: Data fitness plot for model-3

From figure 11 it can be seen that the model is able to predict the activity of the training set quiet well as well as external test set, providing confidence of the model.

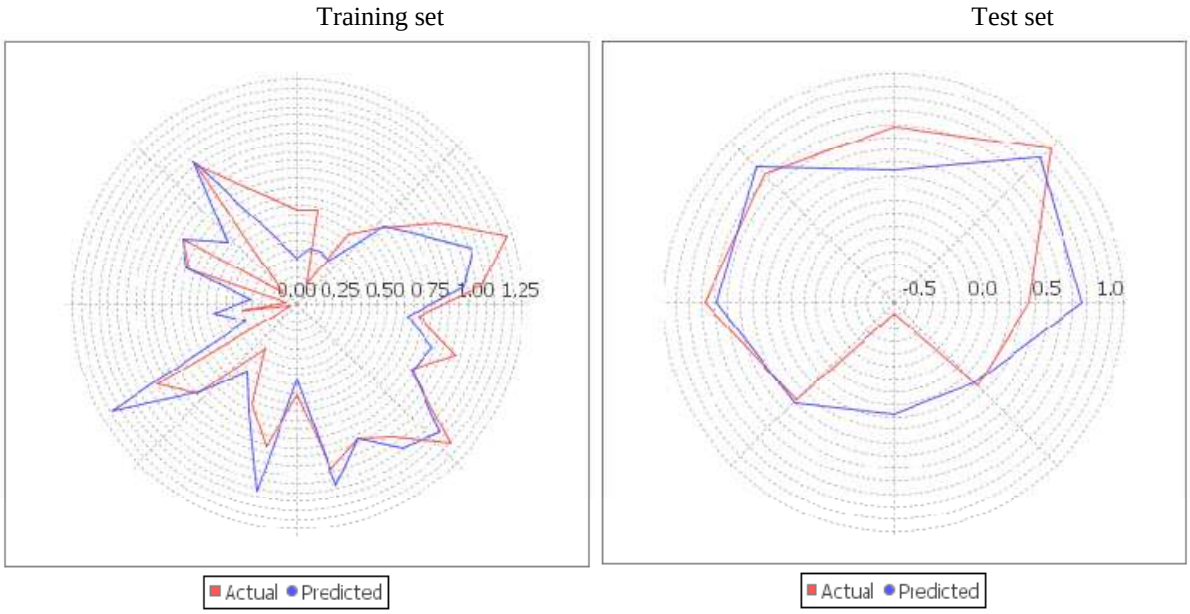

Figure 11: Graph between actual and predicted biological activity of training and test set for Model-3.

Result plot in which 3D-alignment of molecules with the important steric and electrostatic points contributing in the model-3 with ranges of values shown in the parenthesis represented in Figure 12. It shows the relative position and ranges of the corresponding important steric and electrostatic fields in the model provides guideline for new molecule design as follows-

(a) Electrostatic field, E_709 (0.0702 0.0745) has positive range indicates that positive electrostatic potential is favorable for increase in the activity and hence less electronegative substituent group is preferred in that region.
(b) Steric filed, S_391 (-0.1301 -0.1222) has negative range indicates that negative steric potential is favorable for increase in the activity and hence less bulky substituent group is preferred in that region.
(c) Steric filed, S_784 (-0.0179 -0.0172) also has negative range indicates that negative steric potential is favorable for increase in the activity and hence less bulky substituent group is preferred in that region.

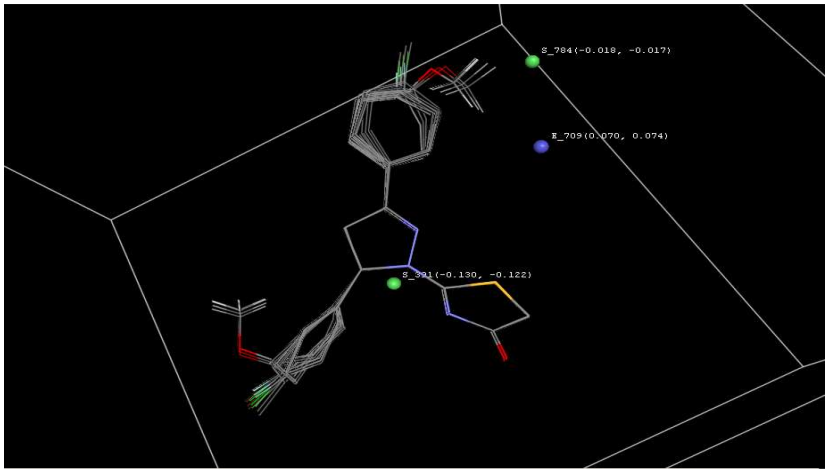

Figure 12: 3D-alignment of molecules with the important steric and electrostatic points contributing model-3 with ranges of values shown in parenthesis.

Table 5: Actual and predicted activities for 36 compounds based on the best 2D/3D-QSAR models

Compounds	Actual pKi1	2D-QSAR (model-1) Predicted	3D-QSAR (model-3) Predicted	Actual pKi2	2D-QSAR (model-2) Predicted	3D-QSAR (model-4) Predicted
E1	0.529	0.4834	0.2571	0.709	0.6680	0.6389
E2	0.687	0.7204	0.3443	0.803	0.8499	0.7332
E3	0.543	0.6176	0.3174	0.684	0.7168	0.6729
E4	0.130	0.1897	0.3216	0.484	0.4689	0.6270
E5	0.481	0.5596	0.2968	0.667	0.7307	0.5954
E6	0.63	0.7204	0.6474	0.793	0.8499	0.7369
E7	0.911		0.7761	1.022	1.0491	1.0142
E8	1.228	1.2512	1.0273	1.258	1.2310	1.0524
E9	1.038	1.1483	0.9304	1.119	1.0979	1.1167
E10	0.68	0.7204	0.6210	0.795	0.8499	0.8787
E11	0.922	1.0904	0.7895	1.011	1.1118	1.1750
E12	1.029	1.2512	0.9177	1.094	1.2310	1.1330
E13	0.728	0.7686	0.7375	0.848	0.9160	0.7156
E14	1.153	1.0057	1.0650	1.215	1.0979	1.0680
E15	0.912	0.9029	0.9937	0.998	0.9648	0.8392
E16	0.358	0.4749	0.7640	0.584	0.7168	0.8078
E17	0.824	0.8449	0.8198	0.91	0.9787	0.9635
E18	0.933	1.0057	1.0270	1.015	1.0979	0.9104
E19	0.505	0.4834	0.4234	0.688	0.6680	0.8769
E20	0.812	0.7204	1.0645	0.911	0.8499	0.9628
E21	0.615	0.6176	0.6535	0.769	0.7168	0.7120
E22	0.307	0.1897	0.4650	0.562	0.4689	0.7412
E23	0.747	0.5596	0.728	0.848	0.7307	0.9101
E24	0.901	0.7204	1.1905	0.967	0.8499	1.0670
E25	0.033	-0.0531	0.2965	0.35	0.3610	0.2548
E26	0.303	0.1839	0.4666	0.548	0.5429	0.6372
E27	0.22	0.0811	0.1834	0.493	0.4098	0.4004
E28	-0.62	-0.3468	0.1753	0.029	0.1619	0.4226
E29	0.064	0.0231	0.2568	0.401	0.4237	0.5343
E30	0.627	0.1839	0.6475	0.794	0.5429	0.6279
E31	0.375	0.5030	0.4032	0.615	0.7630	0.8434
E32	0.775	0.7401	0.6799	0.905	0.9448	0.9535
E33	0.728	0.6374	0.7374	0.88	0.8117	0.7216
E34	0.1	0.2093	0.5239	0.439	0.5638	0.7478
E35	0.737	0.5793	0.8209	0.866	0.8256	1.0157
E36	0.949	0.7401	0.9798	1.02	0.9448	0.8928

Model-4
pKi2 = E_591 (5.1470 5.1578); E_253 (-0.0888 -0.0734)
Statistics:
[kNN= 2; n= 28; Degree of freedom= 25; q^2= 0.7034;
q2_se= 0.1517; pred_r^2= 0.5080; pred_r^2se= 0.1466]

In model 4, values of k (2), q^2 (0.7034), pred_r^2 (0.5080), q^2_se (0.1517), and pred_r^2 se (0.1466) prove that QSAR equation so obtained is statistically significant and shows the predictive power of the model is 70.34% (internal validation). Table 5 represents the predicted inhibitory activity by the model-4 for training and test set.

The data fitness plot for model 4 is shown in Figure 13. The plot of observed vs predicted activity provides an idea about how well the model was trained and how well it predicts the activity of the external test set.

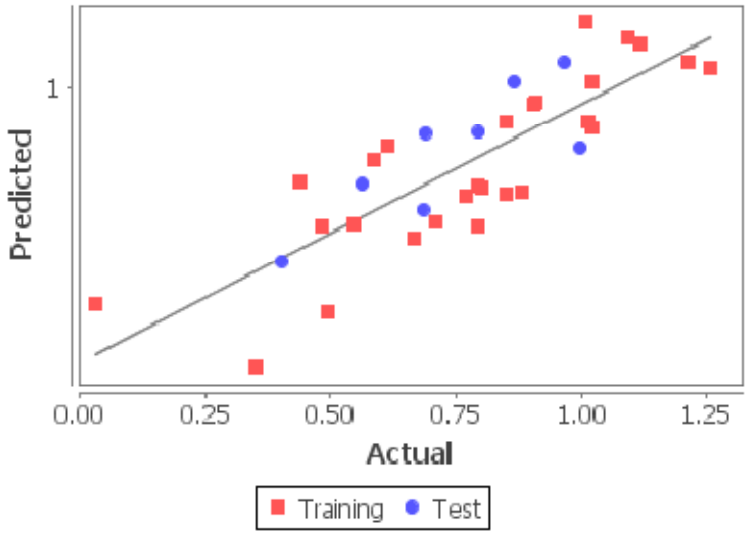

Figure 13: Data fitness plot for model-4

From figure 14 it can be seen that the model is able to predict the activity of the training set quiet well as well as external test set, providing confidence of the model.

Training set Test set

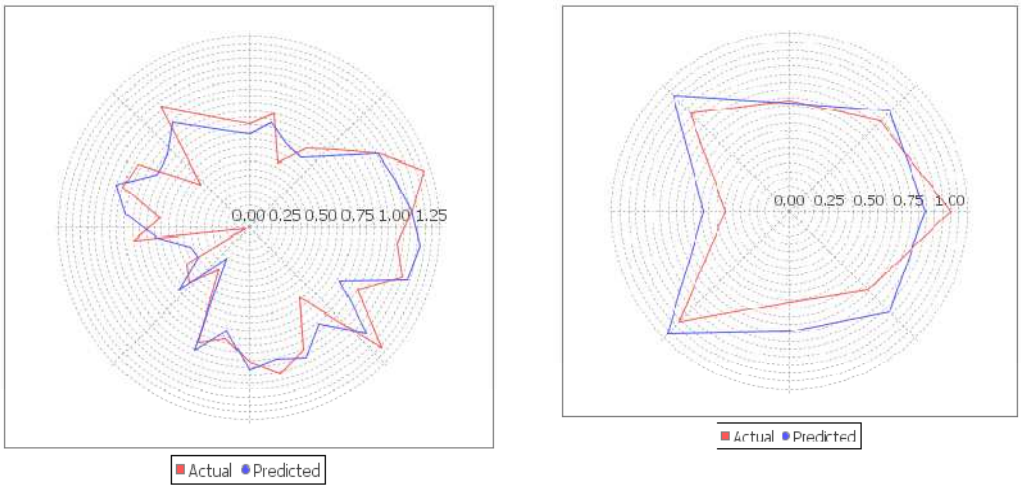

Figure 14: Graph between actual and predicted biological activity of training and test set for Model-4

Result plot in which 3D-alignment of molecules with the important steric and electrostatic points contributing in the model with ranges of values shown in the parenthesis represented in figure 15. It shows the relative position and ranges of the corresponding important steric and electrostatic fields in the model provides guideline for new molecule design as follows-

(a) Electrostatic field, E_591 (5.1470 5.1578) has positive range indicates that positive electrostatic potential is favorable for increase in the activity and hence less electronegative substituent group is preferred in that region.
(b) Electrostatic field, E_253 (-0.0888 -0.0734) has negative range indicates that negative electrostatic potential is favorable for increase in the activity and hence more electronegative substituent group is preferred in that region.

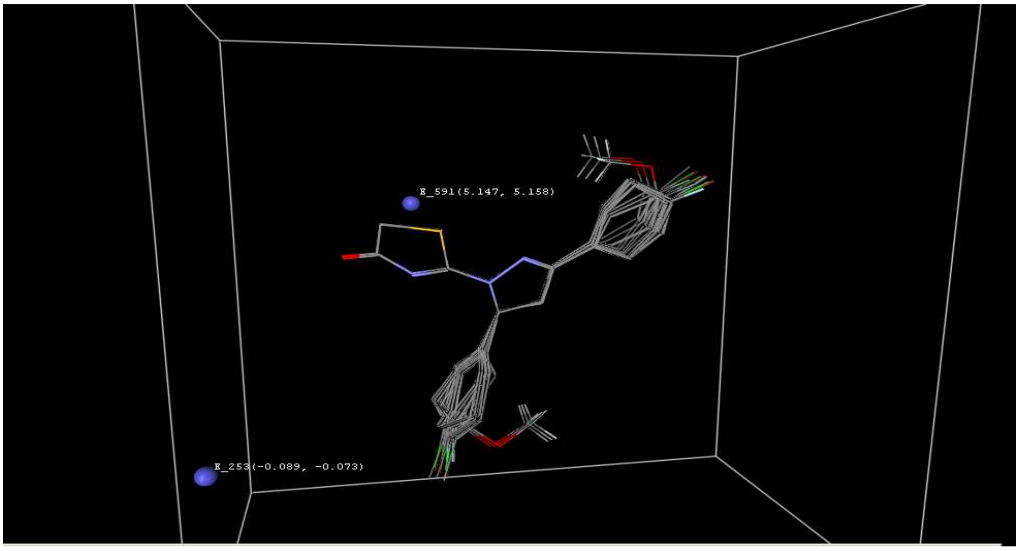

Figure 15: 3D-alignment of molecules with the important steric and electrostatic points contributing model-4 with ranges of values shown in parenthesis

CONCLUSION

Statistically significant 2D/3D-QSAR models were generated with the purpose of deriving structural requirements for the inhibitory activities of some pyrazolyl-thiazolinone derivatives against EGFR and HER-2 kinase. The validation of 2D-QSAR models was done by the cross-validation test, randomization tests and external test set prediction. The best 2D-QSAR models indicate that the descriptors of H-AcceptorCount, T_2_C_7, T_2_C_6, T_2_T_7 influenced the both EGFR and HER-2 inhibition activity while T_T_Cl_7 influenced only EGFR inhibition activity and ChlorinesCount controlled only HER-2 inhibition activity.

kNN-MFA investigated the substitutional requirements for the receptor-drug interaction and constructed the best 3D-QSAR models by PLSR method, providing useful information in characterization and differentiation of their binding sites. In conclusion, the information provided by the robust 2D/3D-QSAR models use for the design of new molecules and hence, this method is expected to provide a good alternative for the drug design.

REFERENCES

[1] R. Herbst, *Int. J. Radiat. Oncol. Biol. Phys*, **2004**, 59 (2), 21-26.
[2] H. Zhang, A. Berezov, Q. Wang, Z. Zhang, J. Drebin, R. Murali, M. Greene, *J. Clin. Invest*., **2007**, 117 (8),2051-2058.
[3] Y. Yarden, J. Schlessinger, *Biochemistry*, **1987**, 26 (5), 1443-1451.
[4] J. Downward, P. Parker, M. Waterfield, *Nature*, **1984**, 311 (5985), 483-485.
[5] K. Oda, Y. Matsuoka, A. Funahashi, H. Kitano, *Mol. Syst. Biol*.,**2005**, 1 (1), 2005-2010.
[6] F. Ciardiello, G. Tortora, *European Journal of Cancer*, **2003**, 39 (10), 1348-1354.
[7] R. Dienstmann, S. De Dosso, E. Felip, J. Tabernero, *Molecular oncology*, **2012** 6 (1), 15-26.
[8] G. Ouyang, Z. Chen, X. Cai, B. Song, P. Bhadury, S. Yang, et al., *Bioorganic & medicinal chemistry*, **2008**, 16 (22), 9699-9707.
[9] O. el-Sabbagh, M. Baraka, S. Ibrahim, C. Pannecouque, G. Andrei, R. Snoeck, J. Balzarini, et al. *European journal of medicinal chemistry*, **2009**, 44 (9), 3746-3753.
[10] G. Nitulescu, C. Draghici, A. Missir, *European journal of medicinal chemistry*, **2010,** 45 (11), 4914-4919.
[11] R. Mitchell, D. Greenwood, V. Sarojini, *Phytochemistry*, **2008,** 69 (15), 2704-2707.
[12] D. Castagnolo, A. De Logu, M. Radi, B. Bechi, F. Manetti, M. Magnani, S. Supino, et al., *Bioorganic & medicinal chemistry*, **2008,** 16 (18), 8587-8591.
[13] A. Bekhit, H. Ashour, Y. Abdel Ghany, A. Bekhit, et al., *European journal of medicinal chemistry*, **2008**, 43 (3), 456-463.
[14] A. Bekhit, T. Abdel-Aziem, *Bioorganic & medicinal chemistry*, **2004**, 12 (8), 1935-1945.
[15] M. Abdel-Aziz, G. Abuo-Rahma, A. Hassan, *European journal of medicinal chemistry*, **2009**, 44 (9), 3480-3487.
[16] N. Gökhan-Kelekçi, S. Yabanoğlu, E. Küpeli, U. Salgin, O. Ozgen, G. Uçar, E. Yeşilada, et al., *Bioorganic & medicinal chemistry*, **2007**, 15 (17), 5775-5786.

[17] M. Christodoulou, S. Liekens, K. Kasiotis, S. Haroutounian, *Bioorganic & medicinal chemistry*, **2010**, 18 (12), 4338-4350.
[18] A. Jain, A. Vaidya, V. Ravichandran, S. Kashaw, R. Agrawal, *Bioorganic & medicinal chemistry*, **2012**, 20 (11), 3378-3395.
[19] S. Barzen, C. Rödl, A. Lill, D. Steinhilber, H.Stark, B. Hofmann, *Bioorganic & medicinal chemistry*, **2012**, 20 (11), 3575-3583.
[20] C. Hansch, A. Kurup, R. Garg, H. Gao, Chem-Bioinformatics and QSAR: A Review of QSAR Lacking Positive Hydrophobic Terms. *Chem. Rev.*, **2001,** 101, 619–672
[21] M. Lill, (2007). Multi-dimensional QSAR in drug discovery, *Drug Discovery Today*, **2007**, 12, 1013–1017.
[22] G. Yang, X. Huang, *Curr. Pharm. Des.*, **2006**, 12, 4601–4611.
[23] K. Qiu, H. Wang, L. Wang, Y. Luo, X. Yang, X. Wang, H. Zhu, *Bioorganic & medicinal chemistry*, **2012**, 20 (6), 2010-2018.
[24] A. Golbraikh, M. Shen, Z. Xiao, Y. Xiao, K. Lee, A. Tropsha, *Journal of computer-aided molecular design*, **2003**, 17 (2-4), 241-253.
[25] VLifeMDS 4.1, Molecular Design Suite, Vlife Sciences Technologies Pvt. Ltd., Pune, India **2012**, www.vlifesciences.com.