Journal of Computational Methods in Molecular Design, 2013, 3 (2):1-9

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

ISSN : 2231- 3176
CODEN (USA): JCMMDA

2D and 3D QSAR using kNN-MFA method of N-[3-(4-benzylpiperidin-1-yl)propyl]-*N,N'*-diphenylureas as CCR5 antagonists as anti-HIV-1 agents

Ajay B. Bedadurge, Anwar R. Shaikh*

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

---

ABSTRACT

Quantitative structure–activity relationship (QSAR) analysis for recently synthesized N-[3-(4-benzylpiperidin-1-yl)propyl]-N,N'-diphenylureas derivatives was studied for their CCR5 antagonists as anti-HIV-1 agents [1]. The statistically significant 2D-QSAR model (r^2 = 0.9493; q^2 = 0.7653; F test = 42.09; r^2 se = 0.1672; q^2 se = 0.3597; pred_r^2 = 0.5311; pred_r^2se = 0.5001) were developed using molecular design suite (VLifeMDS 4.2). The study was performed with 20 compounds (data set) using random selection and manual selection methods used for the division of the data set into training and test set. Multiple linear regression (MLR) 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). The statistical significant model (q^2 = 0.4644; q^2se = 04751; pred_r^2 = 0.4332; pred_r^2se = 0.4890) were developed using molecular design suite (VLifeMDS 4.2) these investigating the substitutional requirements for the favorable anti-HIV-1 agents. The results derived may be useful in further designing novel N,N'-diphenylurea derivatives as CCR5 antagonists prior to synthesis.

Keywords: *N,N'*-diphenylurea derivatives , CCR5 antagonist, AntiHIV-1, Quantitative structure-activity relationship, kNN-MFA

INTRODUCTION

Treatment with highly active antiretroviral therapy (HAART) has successfully suppressed viral replication, recovered immune function, and improved quality of life in human immunodeficiency virus type 1 (HIV-1)-infected individuals. However, the effectiveness of currently available HAART is limited by the development of viral resistance as well as the toxicity and complexity of drug regimens. Therefore, there remains a need to develop new anti-HIV-1 drugs with improved efficacy and less toxicity. The discovery of chemokine receptors as HIV-1 coreceptors has provided a greater understanding of how HIV-1 enters human cells and has led to a novel approach for controlling HIV-1.[1] HIV-1 strains that cause the initial infection predominantly utilize CC chemokine receptor 5 (CCR5),[2] and CCR5-using (R5) HIV-1 is isolated exclusively during the asymptomatic stage of the infection, which usually persists 5–10 years.[3] CCR5 is a member of the seven-transmembrane G protein-coupled receptor superfamily, and its natural ligands include the CC chemokines [regulated on activation, normal T cell expressed and secreted (RANTES), macrophage inflammatory protein 1a (MIP-1a), and MIP-1b], which have been reported to inhibit R5 HIV-1 infection in vitro.[4] Individuals homozygous for a defect in CCR5 expression are highly resistant to HIV-1 infection, while this defect does not represent a significant health problem.[5–7] In addition, infected individuals heterozygous for the defective CCR5 gene appear to have delayed disease progression.[8] These observations suggest that CCR5 antagonists functioning as HIV-1 entry inhibitors could be promising anti-HIV-1 therapeutic agents.
Traditional computer-assisted quantitative structure–activity relationship (QSAR) studies pioneered by C. Hansch et al.1962[10] have been proved to be one of the useful approaches for accelerating the drug design process[11] which help to correlate the bioactivity of compounds with structural descriptors[15].

MATERIALS AND METHODS

2.1. Selection of molecules

Data set of 20 *N,N'*-diphenylureas derivatives (Table 1) collected from published literature were taken for the present study.[9] The affinity data of inhibitory activities were converted into IC_{50} values to get the linear relationship in equation using the following formula: pIC_{50}= -logIC_{50}, where IC_{50} value represents inhibitory activity in IC_{50} (µM) (Table 1). Molecules were rationally divided into the training set and test set based on the suggestions given by Alexander Tropsha et al.[12]

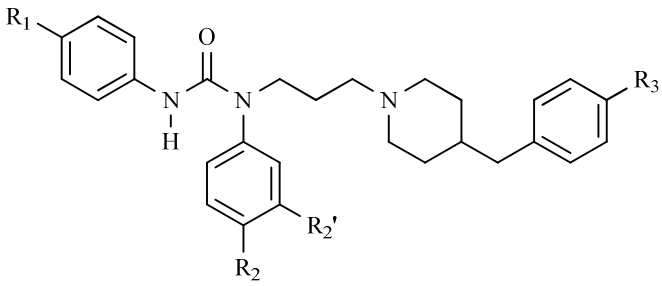


Figure 1: Basic structures of *N,N'*-diphenylureas derivatives.

Table 1: structures of derivative

Compd.	R_1	R_2	R_2'	R_3	CCR5 IC_{50}(nM)	pIC_{50}
1	H	H	H	H	18	1.255
2	Cl	H	H	H	5.9	0.771
3	Cl	H	H	F	7.8	0.892
4	F	H	H	F	13	1.114
5	Br	H	H	F	18	1.255
6	CH_3	H	H	F	6.6	0.820
7	i-Pr	H	H	F	76	1.881
8	CF_3	H	H	F	14	1.146
9	CN	H	H	F	15	1.176
10	EtOCO	H	H	F	280	2.447
11	HOCO	H	H	F	350	2.544
12	H_2NCO	H	H	F	29	1.462
13	MeO	H	H	F	30	1.477
14	MeS	H	H	F	15	1.176
15	$MeSO_2$	H	H	F	26	1.415
16	Cl	CH_3	H	H	11	1.041
17	Cl	H	Cl	H	19	1.279
18	Cl	Cl	Cl	H	62	1.792
19	Cl	H	H	SO_2Me	1.2	0.079
20	Cl	H	H	SO_2(morpholino)	1.0	0.000

2.2. Molecular modeling

All computational experiments were performed using on Lenovo computer having genuine Intel Pentium i3Core Processor and Windows XP operating system using the software Molecular Design Suite (vlifeMDS 4.2).[13] 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.

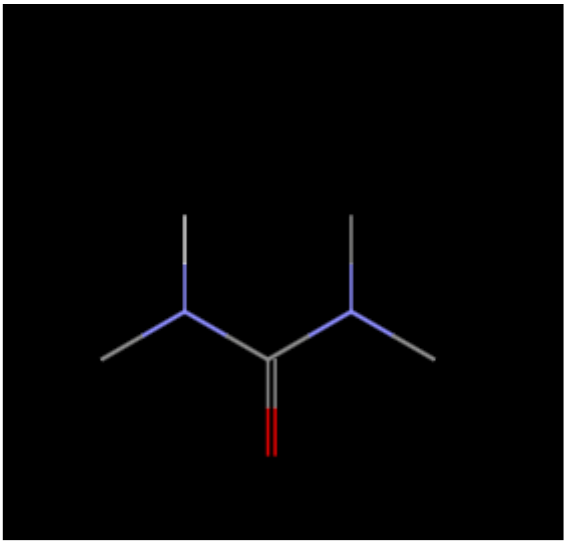


Figure 2: Template molecule

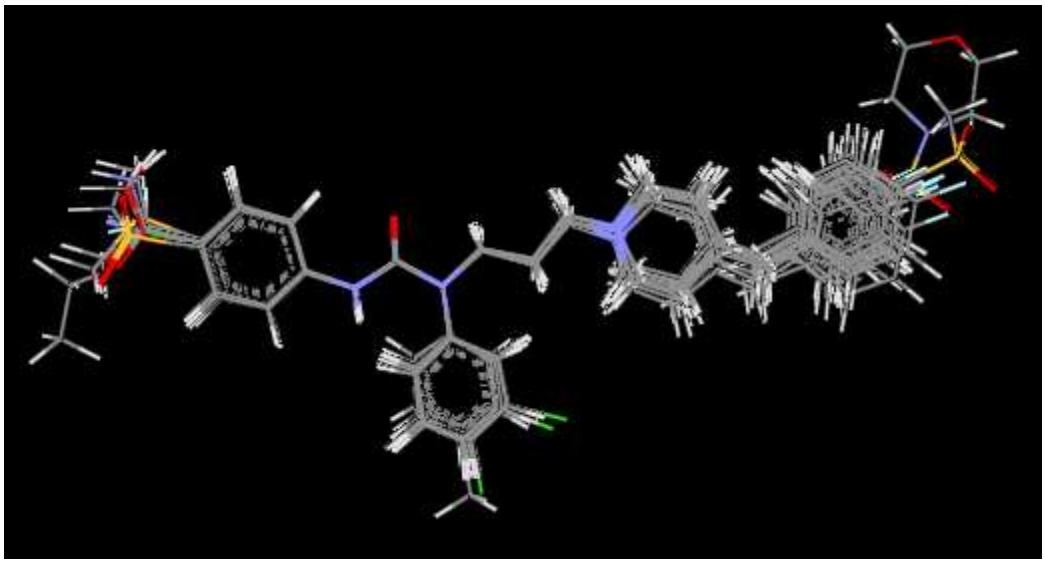


Figure 3: Stereoview of aligned molecules

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 was used for alignment by considering the common elements of the series as shown in Figure 2. The reference molecule 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.

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.[14] 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 Multiple Linear Regression analysis in conjunction with stepwise forward-backward variable selection method. Different training and test set were constructed using 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 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:11,15,3,6,7,8)

***p*IC$_{50}$(column)** = 0.2525(SddssS(sulfate)E-index + 0.4719 (T_O_N_6) + 0.6602 (T_Cl_Cl_3) + 0.3005(SssOcount) + 0.6598

Statistics:

[n= 14; Degree of freedom= 9; r^2 =0.9493; q^2=0.7653; F test=42.09; r^2se=0.1672;
q^2se= 0.3597; pred_r^2= 0.5311; pred_r^2se = 0.5001]

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 94.93% (r^2=0.9493) of the total variance in the training set as well as it has internal (q^2) and external (pred_r^2) predictive ability of 76.53 % and 53.11 % 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,positive coefficient value of SddssS(Sulfate) E-index [the total number of sulphate group connected with two single and two double bonds.] on the biological activity indicated that higher values leads to good inhibitory activity, positive coefficient value of T_O_N_6 [single or double bonded oxygen atom separated from nitrogrn atom by six bonds in molecule] on the biological activity indicated that higher value leads to better inhibitory activity whereas lower value leads to decrease inhibitory activity, positive coefficient value of T_Cl_Cl_3 [single or double bonded chlorine atom separated from any other chlorine atom by three bonds in molecule] on the biological activity indicated that higher

value leads to better inhibitory activity whereas lower value leads to decrease inhibitory activity and positive coefficient value of SssOcount[the total number of oxygen connected with two single bonds.]
Contribution chart for model 1 is represented in Figure 4 reveals that the descriptors SddssS(Sulfate) E-index descriptor is contributing directly 43% respectively, T_O_N_6 descriptors is contributing directly 26% respectively, T_Cl_Cl_3 descriptor is contributing directly 16% respectively and SssOcount descriptor is contributing directly 12% respectively to biological activity.

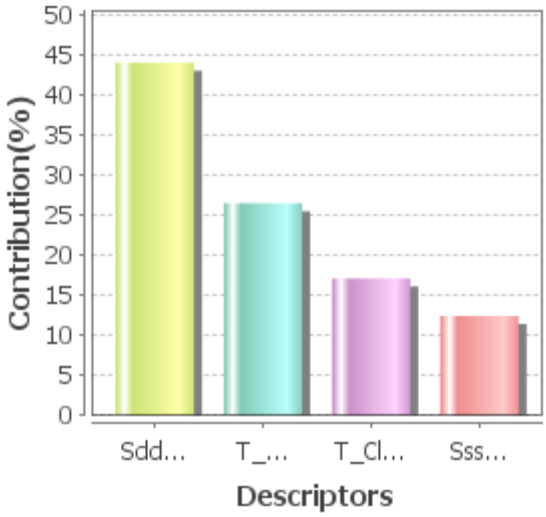


Figure 4: Contribution chart for model-1 showing contribution of different descriptors

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.

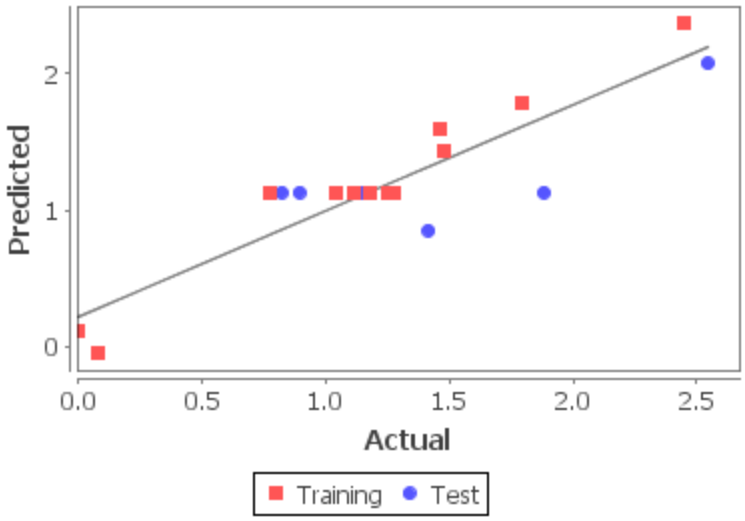
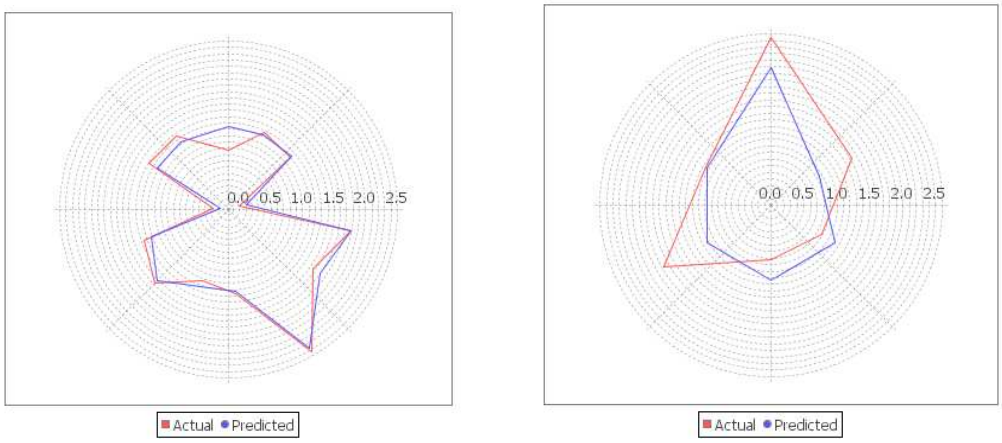


Figure 5: Data fitness plot for model-1

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 2.



Training set **Test set**

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

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 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 model for pIC_{50} (model-1) is given below.

Model-1(Test set:10,15,16,5,9)
***p*IC$_{50}$**(column) = -S_1325(-0.0011 -0.0010)

Statistics:
[kNN= 2; n= 15; Degree of freedom= 13; q^2= 0.4644; q^2_se= 0.4751; pred_r^2= 0.4332; pred_r^2se= 0.4890
The model 1 explains values of k (2), q^2 (0.4644), pred_r^2 (0.4332), q^2_se (0.4751), and pred_r^2 se (0.4890) prove that QSAR equation so obtained is statistically significant and shows the predictive power of the model is 46.44% (internal validation) and 43.32%(external validation) . Table 2 represents the predicted inhibitory activity by the model 1 for training and test set.

The data fitness plot for model 1 is shown in Figure 7. 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.

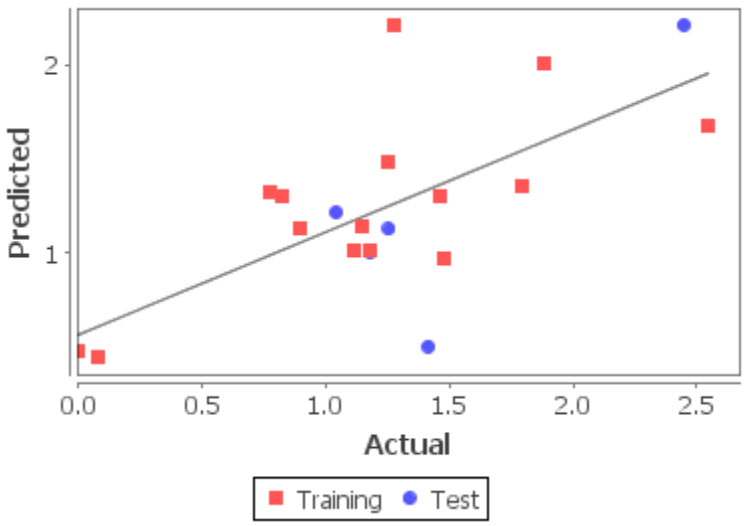
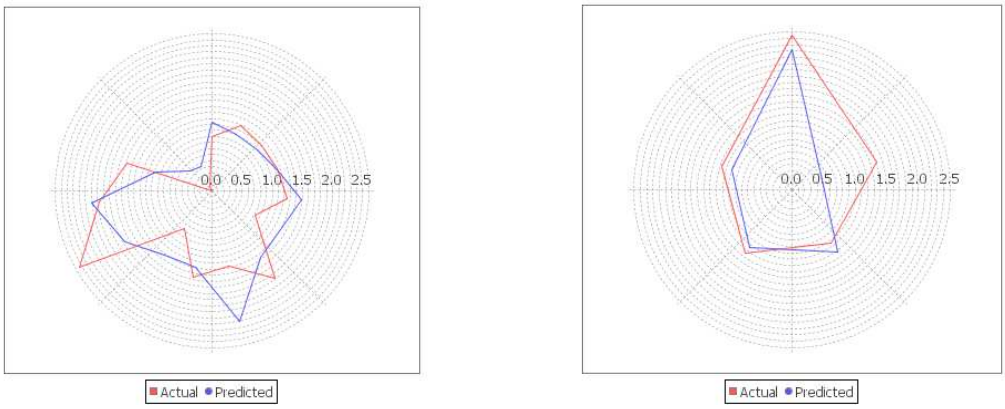


Figure 7: Data fitness plot for model-1(3D)

From Figure 8 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 8: Graph between actual and predicted biological activity of training and test set for Model-1(3D).

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 9. 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) Steric filed, -S_1325(-0.0011 -0.0010) 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.

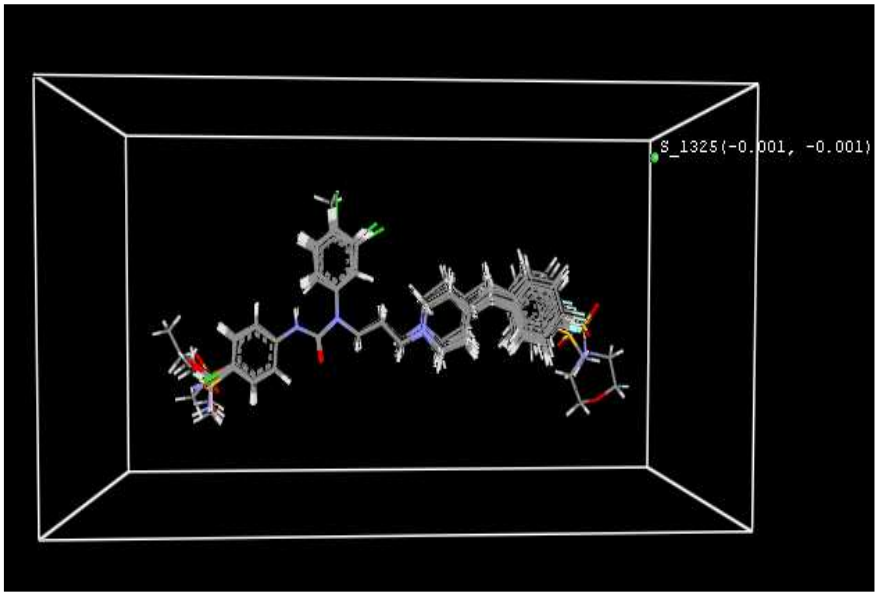


Figure 9: 3D-alignment of molecules with the important steric and electrostatic points contributing model-1(3D) with ranges of values shown in parenthesis.

CONCLUSION

Statistically significant 2D/3D-QSAR models were generated with the purpose of deriving structural requirements for the inhibitory activities of some *N*,*N*'-diphenylureas derivatives as CCR5 antagonist anti-HIV agents. 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 SddssS(sulfate)E-index, T_O_N_6, T_Cl_Cl_3, SssOcount influenced the inhibition activity.

Table 2: Actual and predicted activities for 20 compounds based on the best 2D/3D-QSAR models

Compd.	pIC_{50}	2D-QSAR model 1 Predicted activity	3D-QSAR model Predicted activity
1	1.255	1.1317	1.8869
2	0.771	1.1317	1.9115
3	0.892	1.1317	1.9115
4	1.114	1.1317	1.9115
5	1.255	1.1317	1.9115
6	0.820	1.1317	1.9115
7	1.881	1.1317	1.9115
8	1.146	1.1317	1.9115
9	1.176	1.1317	1.9115
10	2.447	2.3761	1.9115
11	2.544	2.0756	1.9115
12	1.462	1.6037	1.9115
13	1.477	1.4322	1.9115
14	1.176	1.1317	1.9115
15	1.415	0.8537	1.9115
16	1.041	1.1317	1.9115
17	1.279	1.1317	1.9115
18	1.792	1.7921	1.9115
19	0.079	-0.0497	1.9115
20	0.000	0.1155	1.9115

kNN-MFA investigated the substitutional requirements for the receptor-drug interaction and constructed the best 3D-QSAR models by Multiple Linear Regression 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 used for design of new molecules. The newly designed molecules have increased activity than reported biological activity (Table 3).

Table 3: Newly designed molecules

Compd.	R_1	R_2	R_2'	R_3	Predicted activity (pIC_{50})	Antilog of pIC_{50}
1	C2H5OCO	Cl	Cl	H	3.0363	1087.17
2	C2H5OCO	H	H	OCH3	2.6766	474.89
3	C2H5OCO	H	H	Cl	2.3761	237.73
4	C2H5OCO	H	H	Br	2.3761	237.73
5	HOOC	H	H	OCH3	2.3761	237.73

REFERENCES

[1] E A Berger; P M Murphy; Farber J M. *Annu. Rev.Immunol*. **1999**, 17, 657.
[2] Fauci AS. *Nature* **1996**, 384, 529.
[3] RI Connor; KE Sheridan; D Ceradini; S Choe; Landau NR. *J. Exp. Med.*, **1997**, 185, 621.
[4] F Cocchi; AL DeVico; A Garzino-Demo; *et al*. *Science*, **1995**, 270, 1811.
[5] M Dean; M Carrington; *et al*; O_Brien S J. *Science*, **1996**, 273, 1856.
[6] R Liu; WA Paxton; S Choe; D Ceradini; *et al*. *Cell*, **1996**, 86, 367.
[7] M Samson; M Parmentier; *et al*. *Nature*, **1996**, 382, 722.
[8] NL Michael; G Chang; LG Louie; *et al*. *Nat. Med*.,**1997**, 3, 338.
[9] Shinichi Imamura, Osamu Kurasawa; *et al*. *Bio. Med. Chem*., **2004**, 12, 2295.
[10]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.
[11]M. Lill, (2007). Multi-dimensional QSAR in drug discovery, *Drug Discovery Today*, **2007**, 12, 1013
[12]A. Golbraikh, M. Shen, Z. Xiao, Y. Xiao, K. Lee, A. Tropsha, *Journal of computer-aided molecular design*, **2003**, 17 (2-4), 241
[13]VLifeMDS 4.2, Molecular Design Suite, Vlife Sciences Technologies Pvt. Ltd., Pune, India **2012**, www.vlifesciences.com.
[14]J Gasteiger; Marsili M. *Tetrahedron*. **1980**, 36, 3219.
[15]G Yang, X Huang, Development of Quantitative Structure-Activity Relationships and Its Application in Rational Drug Design, *Curr. Pharm. Des*., **2006**, 12, 4601.