Journal of Enzyme Inhibition and Medicinal Chemistry, December 2008; 23(6): 980–995

informa
healthcare

Development of linear and nonlinear predictive QSAR models and their external validation using molecular similarity principle for anti-HIV indolyl aryl sulfones

KUNAL ROY & ASIM SATTWA MANDAL

Division of Medicinal and Pharmaceutical Chemistry, Drug Theoretics and Cheminformatics Lab, Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700 032, India

(*Received 19 September 2007; accepted 29 October 2007*)

Abstract

Quantitative structure–activity relationship (QSAR) studies have been carried out on indolyl aryl sulfones, a class of novel HIV-1 non-nucleoside reverse transcriptase inhibitors, using physicochemical, topological and structural parameters along with appropriate indicator variables. The statistical tools used were linear methods (e.g., stepwise regression analysis, partial least squares (PLS), factor analysis followed by multiple regression (FA-MLR), genetic function approximation combined with multiple linear regression (GFA-MLR) and GFA followed by PLS or G/PLS and nonlinear method (artificial neural network or ANN). In case of physicochemical parameters, GFA-MLR generated the best Equation ($n = 97$, $R^2 = 0.862$, $Q^2 = 0.821$). Using topological parameters, the best Equation (based on leave-one-out Q^2) was obtained with stepwise regression technique ($n = 97$, $R^2 = 0.867$, $Q^2 = 0.811$). When topological and physicochemical parameters were used in combination, statistical quality increased to a great extent ($n = 97$, $R^2 = 0.891$, $Q^2 = 0.849$ from stepwise regression). Furthermore, the whole dataset had been divided into test (25% of whole dataset) and training (remaining 75%) sets. Models were developed based on the training set and predictive potential of such models was checked from the test set. The selection of the training set was based on *K*-means clustering of the standardized descriptors (topological and physicochemical). In this case also the best results were obtained with stepwise regression ($n = 72$, $R^2 = 0.906$, $Q^2 = 0.853$) but external predictive capacity of this model ($R^2_{\text{pred}} = 0.738$) was inferior to the model developed from GFA-MLR technique ($R^2 = 0.883$, $Q^2 = 0.823$, $R^2_{\text{pred}} = 0.760$). However, the squared regression coefficient between observed activity and predicted activity values of the test set compounds for the best linear model, i.e., GFA-MLR ($r^2 = 0.736$) was lower in comparison to the best nonlinear model developed using artificial neural network ($r^2 = 0.781$). Thus, based on external validation, the ANN models were superior to the linear models. The predictive potential of the best linear Equation (stepwise regression model) was superior to that of the previously published CoMFA ($Q^2 = 0.81$, $\text{SDEP}_{\text{Test}} = 0.89$) on the same data set (Ragno R. et al., *J Med Chem* 2006, *49*, 3172–3184). Furthermore, the physicochemical parameter based models also supported the previous observations based on docking (Ragno R. et al., *J Med Chem* 2005, 48, 213–223).

Keywords: *QSAR, indolyl aryl sulfones, validation, anti-HIV-1 activity*

Introduction

Acquired immunodeficiency syndrome (AIDS), characterized by opportunistic infections (T4 cell falls below 200/μL) and opportunistic neoplasms, is one of the leading causes of death worldwide [1]. About 39.5 million people are living with HIV positive till 2006. Nearly 4.3 million people have newly infected with HIV, and AIDS have claimed 2.9 million people including 3,80,000 children under 15 years in the year of 2006 [2].

There are generally two serotypes of HIV virus, which can be distinguished genetically and antigenetically. HIV-1 causes more fatal and rapid infection

Correspondence: K. Roy, Division of Medicinal and Pharmaceutical Chemistry, Drug Theoretics and Cheminformatics Lab, Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700 032, India. E-mail: kunalroy_in@yahoo.com; http://www.geocities.com/kunalroy_in

ISSN 1475-6366 print/ISSN 1475-6374 online © 2008 Informa UK Ltd.
DOI: 10.1080/14756360701811379

than HIV-2. There are three subgroups of HIV-1 including M (major or main), N (new) and O (outlier) [3]. HIV is a special type of retrovirus of lentivirus family. There are at least nine recognizable genes in the HIV virus, but the major structures are composed of *gag*, *pol* and *env*. The other six genes are involved in the infection process as well as regulatory production in *gag*, *pol* and *env* genes. The *gag* gene is "group specific antigen" composed of viral nucleocapsid. It is responsible for development of virus in the absence of *pol* and *env* genes. The *pol* gene codes for HIV enzymes—reverse transcriptase, protease and integrase. Finally the *env* gene codes for the two major envelope's glycoproteins (gp120 and gp4) [4]. When HIV enters into the blood stream, it binds its glycoprotein (gp120) to a T4 cell's or macrophage's, CD4 receptor and the coreceptor CCR5 and/or CXCR4. Then it fuses with the cell membrane and penetrate through it. Inside the cell virion sheds off its coat and leaves its envelope. Single stranded RNA is converted to single stranded DNA using reverse transcriptase from which DNA synthesis of a second strand occurs to form double stranded DNA. This migrates to the nucleus of the cell and integrates into host nucleus by integrase. This provirus transfers its genetic codes to that of the host which becomes a virus factory [5]. Newly formed HIV core proteins, enzymes and genomic RNA gather inside the cell and an immature viral particles composed of long chain proteins are not infectious. These are divided into small fragments to make them infectious using protease enzyme [6]. Thus, the inhibition of reverse transcriptase and protease are the most effective target for anti-HIV drug development.

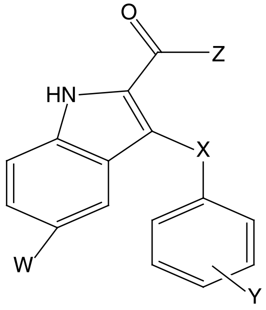
Because of emergence of resistance to present antiviral therapy, development of new anti-HIV drugs is necessary. This necessitates QSAR studies for developing good predictive models involving ligands with different functionalities acting on different anti-HIV targets. Villar et al. have developed a theoretical model using probabilistic neural network that discriminates between active and non-active drugs against HIV-1 with four different mechanisms of action of active drugs [7]. QSAR of HIV-1 reverse transcriptase inhibitory activities of 2-(2,6-dihalophenyl-2-yl)-thiazolidine-4-ones have been studied by Probhakar et al. using topological descriptors obtained from DRAGON software [8]. Senese and Hopfinger have used HIV-1 protease inhibitors derived from norstatine containing 3*S*-amino-2*S*-hydroxy-4-phenylbutanoic acid core to construct 4-D QSAR model [9]. 3D-QSAR studies have been performed on a series of inhibitors of HIV-1 integrase with respect to their inhibition of $3'$-processing and $3'$-end joining steps *in vitro* by Makhija and Kulkarni [10]. CoMFA and CoMSIA and docking studies have been performed by Buolamwini et al. on conformationally restrained cinnamoyl HIV-1 integrase inhibitors to explore binding mode at the active site [11]. Niwa has predicted responses for biological targets including HIV-1 protease for diverse molecules using probabilistic neural network and atom type descriptors from their chemical structure for generating focused libraries, selecting compounds for screening and annotating biological activity for those compounds whose activities are unknown [12]. Weekes and Fogel have used evolutionary optimization, back propagation and data preparation issues in QSAR modeling of HIV inhibition by HEPT derivatives. Evolutionary computation gives appropriate set of weights and bias terms associated with artificial neural network that minimize selected functions of the error between the actual and desired outputs [13]. Hopfinger and Senese have used simple clustering technique to facilitate and improve model selection and test set prediction (training set of 50 tetrahydropyrimidine 2-one based inhibitors of HIV-1 protease) [14]. A 3D-QSAR was applied to a set of dipyridodiazepinone derivatives which is active against wild and mutant type HIV-1 reverse transcriptase by Pungpo et al. [15]. Ragno et al. have also performed docking and 3D-QSAR (CoMFA) on indolyl aryl sulfones to explore binding mode at HIV-1 reverse transcriptase binding site [16].

The present group of authors [1,17–25] has developed a few anti-HIV QSARs involving different series of chemicals acting on different targets. In continuation of such efforts, the present paper deals with modeling of anti-HIV-1 activity of reverse transcriptase inhibitor indolyl aryl sulfones reported by Ragno et al. [16,26]. Some compounds were excluded from our study due to lack of quantitative activity data. We have modeled the data set using linear techniques like multiple linear regression (with stepwise regression, factor analysis and genetic function approximation as variable selection strategy) and partial least squares and compared the results with those obtained from nonlinear method (feed-forward back propagation artificial neural network). The objectives of the present study include development of QSAR models with physicochemical significance in one hand and development of predictive model with good validation characteristics on the other hand for anti-HIV-1 activity of reverse transcriptase inhibitor indolyl aryl sulfones.

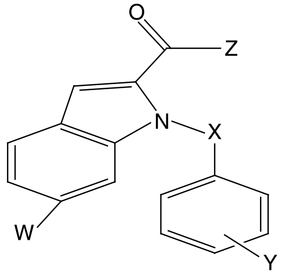
Materials and methods

The anti-HIV data (EC_{50}) of indolyl aryl sulfone derivatives [16,26] had been converted to logarithmic scale [$\mathrm{p}C = -\log EC_{50}$ (M)] and then used for the QSAR study. Though the original paper reported 117 compounds in total, twenty of the reported compounds do not have exact biological activity values. Thus, 97 compounds were considered for the present QSAR study (Table I). There are four different

Table I. Structural features, observed and calculated data of HIV-1 reverse transcriptase inhibitory activity of indole aryl sulfones



A



B

.

		Structural features				Anti-HIV activity (−logEC50(M))			
Sl. No.	Structure	X	Y	Z	W	obs	cal[a]	cal[b]	cal[c]
1.	A	SO_2	H	NH_2	Cl	9.00	8.16	7.99	8.02
2.*	A	S	H	OEt	H	5.85	4.73	5.52	4.98
3.	A	S	2-NH_2	OEt	Cl	5.64	5.07	5.36	5.21
4.	A	S	2-NH_2-5-Cl	OEt	Cl	5.60	5.14	5.49	5.09
5.	A	SO_2	H	OEt	H	5.43	5.86	6.37	6.24
6.	A	SO_2	2-NH_2-5-Cl	OEt	H	5.60	5.65	5.88	5.82
7.	A	SO_2	2-NH_2-5-Cl	OEt	Cl	5.72	6.33	6.68	6.35
8.	A	S	H	NH_2	H	5.85	6.31	6.35	6.25
9.*	A	S	2-NH_2-5-Cl	NH_2	H	5.05	6.00	5.59	5.85
10.	A	S	H	NH_2	Cl	7.69	6.98	7.07	6.79
11.	A	S	2-Me	NH_2	Cl	6.52	6.93	6.96	6.77
12.	A	S	4-F	NH_2	Cl	5.85	6.80	6.57	6.54
13.	A	S	4-Cl	NH_2	Cl	5.51	6.18	6.23	6.21
14.	A	S	4-i-Pr	NH_2	Cl	5.72	5.63	5.53	5.94
15.*	A	S	4-t-Bu	NH_2	Cl	5.10	6.31	7.26	6.35
16.	A	S	3,5-Me_2	NH_2	Cl	8.22	7.64	7.69	7.72
17.	A	S	2,6-Cl_2	NH_2	Cl	5.92	6.17	6.58	6.27
18.	A	S	2-NH_2-5-Cl	NH_2	Cl	5.79	6.67	6.33	6.39
19.*	A	SO_2	H	NH_2	H	6.74	7.51	7.21	7.51
20.*	A	SO_2	2-NH_2-5-Cl	NH_2	H	6.52	7.26	6.74	7.12
21.	A	SO_2	2-Me	NH_2	Cl	9.00	8.14	8.06	8.01
22.	A	SO_2	3-Me	NH_2	Cl	9.00	8.48	8.38	8.45
23.	A	SO_2	4-Me	NH_2	Cl	8.52	7.81	7.61	7.77
24.	A	SO_2	4-F	NH_2	Cl	7.85	7.98	7.55	7.77
25.	A	SO_2	4-Cl	NH_2	Cl	7.96	7.36	7.21	7.47
26.	A	SO_2	4-i-Pr	NH_2	Cl	7.10	6.80	6.52	7.22
27.	A	SO_2	4-t-Bu	NH_2	Cl	6.87	7.47	8.24	7.24
28.	A	SO_2	2,4-Me_2	NH_2	Cl	8.40	7.82	7.70	7.76

Table I – *continued*

		Structural features				Anti-HIV activity (−logEC50(M))			
Sl. No.	Structure	X	Y	Z	W	obs	cal[a]	cal[b]	cal[c]
29.	A	SO_2	3,5-Me_2	NH_2	Cl	8.40	8.83	8.79	8.87
30.	A	SO_2	2,6-Cl_2	NH_2	Cl	7.00	7.39	7.86	7.53
31.	A	SO_2	2-NH_2-5-Cl	NH_2	Cl	7.40	7.91	7.54	7.64
32.	A	SO_2	3,5-Me_2	NH_2	Br	8.70	8.28	8.30	8.50
33.	A	SO_2	3,5-Me_2	NH_2	COMe	7.82	8.56	8.20	8.49
34.*	A	SO_2	3,5-Me_2	NH_2	CH(OH)Me	7.60	5.61	6.51	7.37
35.*	A	S	H	$NHNH_2$	Cl	6.26	5.80	5.97	5.67
36.	A	S	4-Me	$NHNH_2$	Cl	5.82	5.47	5.54	5.41
37.*	A	S	4-F	$NHNH_2$	Cl	5.30	5.64	5.47	5.42
38.	A	S	4-Cl	$NHNH_2$	Cl	5.00	5.02	5.14	5.09
39.	A	SO_2	H	$NHNH_2$	H	6.28	6.29	6.10	6.42
40.*	A	SO_2	H	$NHNH_2$	Cl	8.00	6.96	6.89	6.95
41.*	A	SO_2	4-Me	$NHNH_2$	Cl	7.30	6.63	6.51	6.68
42.	A	SO_2	4-F	$NHNH_2$	Cl	6.49	6.80	6.45	6.69
43.	A	SO_2	4-Cl	$NHNH_2$	Cl	6.72	6.18	6.11	6.37
44.	A	SO_2	3,5-Me2	$NHNH_2$	Cl	6.89	7.66	7.69	7.87
45.	A	SO_2	2-NH_2-5-Cl	$NHNH_2$	Cl	6.52	6.73	6.44	6.54
46.	B	SO_2	2-NO_2	COOEt	H	5.74	5.58	5.80	5.63
47.	B	SO_2	2-NH_2-5-Cl	COOEt	H	5.74	5.16	5.69	5.24
48.*	B	SO_2	2-NH_2-5-Cl	COOEt	5-Cl	5.08	5.81	6.46	5.78
49.*	B	SO_2	H	$CONH_2$	H	4.82	4.77	4.92	5.37
50.	B	SO_2	H	$CONH_2$	5-Cl	4.18	5.38	5.68	5.90
51.	B	SO_2	2-NO_2	H	H	5.30	5.56	5.13	5.19
52.	B	SO_2	2-NH_2	H	H	4.96	5.12	4.87	4.93
53.*	B	SO_2	2-NO_2-5-Cl	H	H	5.40	5.67	5.36	5.09
54.	B	SO_2	2-NH_2-5-Cl	H	H	6.00	5.18	5.04	4.83
55.	B	SO_2	2-NO_2-4-Cl	H	H	4.80	4.86	4.48	4.65
56.	B	SO_2	2-NH_2-4-Cl	H	H	4.13	4.35	4.14	4.39
57.	B	SO_2	2-Cl-5-NO_2	H	H	5.22	4.96	5.27	5.00
58.*	B	SO_2	2-Cl -5- NH_2	H	H	3.82	4.09	5.06	5.25
59.	A	S	H	$NHCH_2CH_2OH$	5-Cl	4.40	4.11	4.23	4.11
60.	A	S	2-Me	$NHCH_2CH_2OH$	5-Cl	4.30	4.09	4.12	4.08
61.	A	S	3-Me	$NHCH_2CH_2OH$	5-Cl	4.10	4.47	4.53	4.53
62.	A	S	4-Me	$NHCH_2CH_2OH$	5-Cl	3.92	3.82	3.79	3.88
63.*	A	S	2,3-Me_2	$NHCH_2CH_2OH$	5-Cl	4.00	4.35	3.95	4.51
64.	A	S	3,5-Me_2	$NHCH_2CH_2OH$	5-Cl	4.92	4.84	4.85	4.97
65.*	A	SO_2	H	$NHCH_2CH_2OH$	5-Cl	6.00	5.22	5.14	5.30
66.*	A	SO_2	2-Me	$NHCH_2CH_2OH$	5-Cl	4.50	5.22	5.21	5.26
67.	A	SO_2	3-Me	$NHCH_2CH_2OH$	5-Cl	5.52	5.58	5.53	5.76
68.	A	SO_2	4-Me	$NHCH_2CH_2OH$	5-Cl	4.37	4.93	4.76	5.04
69.	A	SO_2	2,4-Me_2	$NHCH_2CH_2OH$	5-Cl	5.10	4.95	4.85	5.00
70.	A	SO_2	3,5-Me_2	$NHCH_2CH_2OH$	5-Cl	6.10	5.96	5.94	6.23

Table I – *continued*

		Structural features				Anti-HIV activity ($-$logEC50(M))			
Sl. No.	Structure	X	Y	Z	W	obs	cal[a]	cal[b]	cal[c]
71.	A	S	H	$NHNHCH_2CH_2OH$	5-Cl	3.70	3.69	3.86	3.81
72.⋆	A	S	2-Me	$NHNHCH_2CH_2OH$	5-Cl	4.00	3.68	3.75	3.79
73.	A	S	4-Me	$NHNHCH_2CH_2OH$	5-Cl	3.22	3.42	3.42	3.59
74.⋆	A	S	2,4-Me_2	$NHNHCH_2CH_2OH$	5-Cl	2.92	3.43	3.33	3.57
75.	A	S	3,5-Me_2	$NHNHCH_2CH_2OH$	5-Cl	4.52	4.45	4.48	4.65
76.	A	SO_2	H	$NHNHCH_2CH_2OH$	5-Cl	5.00	4.78	4.77	4.98
77.	A	SO_2	2-Me	$NHNHCH_2CH_2OH$	5-Cl	4.30	4.79	4.84	4.95
78.⋆	A	SO_2	3-Me	$NHNHCH_2CH_2OH$	5-Cl	5.15	5.16	5.16	5.43
79.	A	SO_2	4-Me	$NHNHCH_2CH_2OH$	5-Cl	4.39	4.51	4.39	4.73
80.⋆	A	SO_2	2,4-Me_2	$NHNHCH_2CH_2OH$	5-Cl	4.10	4.54	4.48	4.69
81.	A	SO_2	3,5-Me_2	$NHNHCH_2CH_2OH$	5-Cl	6.00	5.55	5.57	5.90
82.	A	SO_2	3,5-Me_2	$NHCH_2CONH_2$	5-Cl	8.22	7.89	8.44	7.77
83.	A	SO_2	3,5-Me_2	$NHCH_2CONHNH_2$	5-Cl	8.00	7.36	7.26	7.57
84.⋆	A	SO_2	3,5-Me_2	$NHCH_2CONHCH_2CONH_2$	5-Cl	9.16	7.46	8.00	7.45
85.	A	SO_2	3,5-Me_2	$NHCH_2CONHCH_2CONHNH_2$	5-Cl	7.22	6.93	6.82	7.24
86.	A	SO_2	3,5-Me_2	$NHCH_2CONHCH(Me)CONHNH_2$	5-Cl	7.10	7.11	7.15	7.38
87.	A	SO_2	3,5-Me_2	$NHCH_2CONHCH(Me)CONH_2$	5-Cl	7.85	7.58	8.26	7.59
88.⋆	A	SO_2	3,5-Me_2	$NHCH_2CONHCH_2CONHCH_2CONH_2$	5-Cl	6.92	7.07	7.57	7.11
89.	A	SO_2	3,5-Me_2	$NHCH_2CONHCH_2CONHCH_2CONHNH_2$	5-Cl	6.75	6.56	6.39	6.90
90.	A	SO_2	H	2-oxazolidone-3-yl amino	5-Cl	7.82	8.07	7.84	7.59
91.	A	SO_2	3,5-Me_2	2-oxazolidone-3-yl amino	5-Cl	9.05	8.83	8.66	8.48
92.⋆	A	SO_2	3,5-Me_2	Me	5-Cl	7.70	8.72	7.95	7.80
93.	A	SO_2	3,5-Me_2	i-Pr	5-Cl	9.16	8.99	9.04	7.82
94.	A	SO_2	3,5-Me_2	c-He	5-Cl	7.30	7.62	7.62	7.66
95.	A	SO_2	3,5-Me_2	COMe	5-Cl	8.30	9.01	8.95	7.83
96.	A	SO_2	3,5-Me_2	COOEt	5-Cl	7.70	8.31	7.45	7.79
97.⋆	A	SO_2	3,5-Me_2	$CONHNH_2$	5-Cl	9.00	7.36	7.24	7.55

⋆stands for a member of the test set. [a]from Equation 16. [b]from Equation 19. [c]from ANN (Model 4).

Table II. Values of physicochemical parameters (substituent constants)[#].

Ring substitution	π	MR[a]	σ_m	σ_p	L	B1	B2	B3	B4
NH_2	− 1.23	0.542	− 0.16	− 0.66	2.93	1.50	1.50	1.84	1.84
Cl	0.71	0.603	0.37	0.23	3.52	1.80	1.80	1.80	1.80
CH_3	0.56	0.565	− 0.07	− 0.17	3.00	1.52	2.04	1.90	1.90
F	0.14	0.092	0.34	0.06	2.65	1.35	1.35	1.35	1.35
i-Propyl	1.53	1.496	− 0.07	− 0.15	4.11	2.04	2.76	3.16	3.16
t-Butyl	1.98	1.962	− 0.10s	− 0.20	4.11	2.59	2.97	2.86	2.86
NO_2	− 0.28	0.736	0.71	0.78	3.44	1.70	1.70	2.44	2.44
H	0.00	0.103	0.00	0.00	2.06	1.00	1.00	1.00	1.00

[#] Obtained from reference [27]. [a] MR values are scaled with 0.1 as usual.

positions of substitutions: one is the fragment flanked between the indole nucleus and phenyl ring (X) and the second one is the substitution on the phenyl ring (Table I). The other two positions are second and fifth positions of the indole nucleus.

For the construction of the linear models, multiple regression (with stepwise regression, factor analysis and genetic function approximation as variable selection tools) and partial least squares were used.

Descriptors

Physicochemical parameters like hydrophobicity (π), electronic (Hammett σ), steric (molar refractivity MR and STERIMOL L, B1 to B4) substituent constants (Table II) of phenyl ring substituents were taken from reference [27]. The topological descriptors including Balaban index (Jx), connectivity indices ($^1\chi, ^2\chi, ^1\chi^v, ^2\chi^v$ etc), kappa shape indices ($^1\kappa, ^2\kappa, ^3\kappa$ etc), molecular flexibility index (Φ), Wiener index, Zagreb index, E-state indices ($S_{\_sCH_3}, S_{\_ssCH_3}$ etc) were calculated using Cerius2 version 10 [28] on a silicon graphics computer. Besides these, structural descriptors like number of chiral centres, molecular weight (MW), number of rotatable bond (Rotlbonds) and indicator parameters as defined in Table III were also used. For the development of the QSAR models we initially considered physicochemical and topological parameters separately and then combination of both types of parameters along with indicators variables were used. In the total pool of descriptors, there were 23 physicochemical, 52 topological and 9 indicator parameters from which variable selection was made using different strategies as detailed below.

Cluster analysis and validation

Initially QSAR models were developed on the whole data set. The models were crossvalidated using leave-one-out method. However, internal validation does not ascertain that the model will perform well on a new set of data. Thus, the whole data set was divided into training and test sets and the models developed from training set were externally validated using the test set. Predictive capacity of a model for new chemical entities is influenced by chemical nature of the training set molecules used for development of the model [29–31]. In actual case, the test set molecules will be predicted well when these molecules are structurally very similar to the training set molecules. The reason is that the model has captured all features common to the training set molecules.

Any QSAR modeling should ultimately lead to statistically robust models capable of making accurate and reliable predictions of biological activities of compounds. When QSAR models are developed, it is important to validate any fitted models to check that it is plausible that its predictions will carry over to fresh data not used in the model fitting exercise. The validation strategies check the reliability of the developed models for their possible application on a new set of data, and confidence of prediction can thus be judged. Often, truly external data points being unavailable for prediction purpose, original data set compounds are divided into training and test sets. A QSAR model's ability to predict the properties of unknown chemicals depends largely on the nature of the training set and the algorithm used to establish the structure–activity relationship [29]. The process

Table III. Definition of indicator variables.

Used indicators	Meaning of indicators
I_{Z_amino}	Indicator having value 1 if amino group is present at Z position otherwise value 0
I_{Z_ethoxy}	Indicator having value 1 if ethoxy group is present at Z position otherwise value 0
$I_{Z\_hydrazine}$	Indicator having value 1 if hydrazine group is present at Z position otherwise value 0
$I_{Z\_ethcarb}$	Indicator having value 1 if ethyl carboxylate group is present at Z position otherwise value 0
I_{Z_hetami}	Indicator having value 1 if hydroxyethylamino group is present at Z position otherwise value 0
$I_{Z\_hethydr}$	Indicator having value 1 if hydroxyethylhydrazine group is present at Z position otherwise value 0
I_{W_Cl}	Indicator having value 1 if chlorine atom is present at W position otherwise value 0
I_X	Indicator having value 1 if sulfur dioxide group is present at X position otherwise value 0
I_{NH}	Indicator having value 1 if unsubstituted nitrogen atom is present in the indole nucleus otherwise value 0

is based on the assumption that a molecule that is structurally very similar to the training set molecules will be predicted well because the model has captured features that are common to the training set molecules and is able to find them in the new molecule. On the other hand, a new molecule which has very little in common with the training set data should not be predicted very well, i.e., the confidence in its prediction should be low [30]. A model's predictive accuracy and confidence for different unknown chemicals varies according to how well the training set represents the unknown chemicals and how robust the model is in extrapolating beyond the chemistry space defined by the training set. Therefore, assessing a model's prediction accuracy outside the training domain is a vital step toward defining the application domain of a model for the regulatory acceptance of QSARs. The selection of the training set is significantly important in QSAR analysis. In this paper, the data set was divided into training and test sets (75% and 25% respectively of the total number of compounds) based on clusters obtained from *K*-means clustering applied on standardized descriptor matrix. All the parameters were standardized to values between 0 and 1 and the whole dataset was clustered into seven subgroups from each of which 25% of compounds were selected as members of the test set. Cluster analysis is a method of arrangement of objects into groups [32–34]. It classifies different objects into groups in such a way that the degree of association between two objects is maximum if they possess same group and otherwise minimum. Most clustering techniques are hierarchical, i.e, the resultant classification has an increasing number of nested classes [32]. There are some non-hierarchical methods e.g., *K*-means clustering [32–34]. In this method, number of groups or clusters (*K*) generated is specified by the user. At the end of the analysis the data are split between *K* clusters. From the results of *K*-means clustering analysis, one can examine the means for each cluster on each dimension to assess how distinct the *K* clusters are. After clustering, the test set compounds are selected from each cluster by taking approximately 25% of the compounds from each cluster so that test and training set can represent all clusters and the whole dataset.

For the development of equations different chemometric tools were utilized.

Stepwise regression

In stepwise regression, a multiple-term linear equation is built step-by-step. First, an initial model is recognized and then the model is altered repeatedly at the previous step by adding or removing a predictor variable according to the "stepping criteria" [35]. At the last step the search is terminated when stepping is no longer possible or when a specified maximum number of steps has been reached. Specifically, at each step all variables are reviewed and evaluated to determine which one will contribute most to the equation. The method selected for stepwise regression is forward selection and backward elimination. The criteria "F to Enter" and "F to Remove" determine how significant or insignificant the contribution of a variable in the regression equation respectively for adding to the equation and removing from the equation. A limitation of the stepwise regression search approach is that it presumes that there is a single "best" subset of *X* variables and seeks to identify it.

PLS

For PLS [36,37], "leave-one-out" method was used for crossvalidation to obtain the optimum number of components. PLS is a useful technique when number of factors is large and they are highly collinear. This technique generalizes and combines features from principal component and multiple regression. In case of PLS analysis on the present data set, based on the standardized regression coefficients, the variables with smaller coefficients were removed from the PLS regression, until there was no further improvement in Q^2 value, irrespective of the components. It gives statistically more robust solution than MLR. To avoid overfitting, a strict test for the significance of each consecutive PLS component is necessary and then stopping when the components are non-significant. This ensures that the QSAR equations are selected based on their ability to predict the data rather than to fit the data.

FA-MLR

Factor analysis [38,39] has been performed to find out the relationship among variables. It reduces the large numbers of variables to few factors from which important variables for multiple linear regression can be identified. It is a data processing step to identify the variables contributing to the response variable. The whole dataset containing biological activity and all descriptor variables is extracted by principle component method and rotated by VARIMAX rotation to obtain Thurston's simple structure. The effective variables are selected from rotated component matrix obtained from the previous operation. Linear regression is performed using these variables.

GFA-MLR

For the development of genetic function approximation (GFA) model Cerius2 4.10 version has been used. GFA provides a new approach to the problem of developing QSAR models. Genetic algorithms are derived with the spread of mutations in a population. It was initially conceived from two seemingly disparate algorithms - i) Holland's genetic algorithm and ii) Friedman's multivariate adaptive regression splines algorithm [40]. GFA allows construction of models superior to standard techniques and gives additional information not provided by other techniques. A distinctive feature of GFA is that it produces a population of models (e.g., 100), instead of generating a single model, as do most other statistical methods. These multiple models are generated from random initial models using a genetic algorithm. A "fitness function" (lack-of-fit or LOF) is used to measure the quality of each model, so that the best model receives the best fitness score. Features, which are necessary for construction of the model, are automatically selected by GFA. GFA can build not only linear models but also higher order polynomials, splines and Gaussian, i.e., the type of model is user-defined. It can automatically remove outlier and perform classification using spline-based terms. Its random search procedure for building of the model leads to the discovery of highly predictive models.

G/PLS

Genetic partial least squares (G/PLS) model is derived from two methods: i) genetic function approximation (GFA) and ii) partial least squares (PLS) [36,37]. Both of these methods are valuable analytical tools for QSAR modeling where numbers of descriptors are more than samples. Genetic function approximation is used to select the appropriate variables to be used in the development of model. It is followed by PLS regression as fitting technique to weigh the relative contribution of the selected variables in the final model.

ANN

For the purpose of development of nonlinear model, multilayer perceptron (MLP) of "Custom Network Designer" had been selected to design the network. We selected the back propagation method of MLP followed by conjugate gradient descent to train the network. The back propagation method is the most popular method of developing nonlinear model. There are at least three layers including one input layer, one hidden layer and one output layer. Each layer is interconnected with each other. In this method difference between output of the network and the desired output is calculated. This error value is back propagated to the transfer function for adjustment of weight. Through transfer function (sigmoid function), the output is obtained as [41]

$$O_j = f(i_j) = \frac{1}{1 + \exp(-\beta i_j)}$$

where O_j is the output of node j and β is a gain, being able to adjust the form of the function. Usually β is taken as 1. Using the error signal to adjust the connected weights, the following adjusted weights are obtained for the output layer.

$$W_{i_j}(new) = W_{i_j}(old) + \eta\delta_i O_j + \alpha[\Delta W_{i_j}(old)]$$

In back propagation, the gradient vector of the error surface is calculated. This vector points in the direction of steepest descent from the current point, so one knows that if one moves along it a "short" distance, one will decrease the error. A sequence of such moves will eventually find a minimum of some sort.

Conjugate gradient descent method is a good secondary and advanced method of training multilayer perceptron. It is generally used for the network of large numbers of weights and/or multiple output units. It is a batch update algorithm whereas back propagation adjusts the weights of the network. Learning rate and momentum of each epoch are adjusted and weight decay is regularized.

Most work on assessing performance in neural modeling concentrates on approaches to resampling. A neural network is optimized using a training subset. Often, a separate subset (the selection subset) is used to halt training to mitigate over-learning, or to select from a number of models trained with different parameters. Then, a third subset (the test subset) is used to perform an unbiased estimation of the network's likely performance.

Although the use of a test subset set allows us to generate unbiased performance estimates, these estimates may exhibit high variance. Ideally, one would like to repeat the training procedure a number of different times, each time using new training, selection and test cases drawn from the population - then, one could average the performance prediction over the different test subsets, to get a more reliable indicator of generalization performance.

In reality, one seldom has enough data to perform a number of training runs with entirely separate training, selection and test subsets. Crossvalidation is the simplest resampling technique. We have cross-validated the network using 15 resampling. Using different numbers of hidden layers and different numbers of units per layer it was shown that the one hidden layer of 39 units had good predictive capacity on this dataset.

Model quality

The statistical qualities of the multiple regression equations [42] were judged by the parameters like *explained variance* (R_a^2), *correlation coefficient* (R), *standard error of estimate* (s) and *variance ratio* (F) at specified *degrees of freedom* (*df*). All accepted MLR equations have regression coefficients and F ratios significant at 95% and 99% levels respectively, if not stated otherwise. The generated QSAR equations were validated by *leave-one-out or LOO statistics* [43,44] and *cross-validation* R^2 (Q^2) and *predicted residual sum* *of squares* (*PRESS*) values were reported. In case of external validation, predictive capacity of a model was judged by its application for prediction of test set activity values and calculation of predictive R^2 (R^2_{pred}) value.

Softwares

MINITAB [45] was used for stepwise regression and PLS whereas SPSS [46] and STATISTICA [47] were used for FA-MLR and ANN respectively. Cerius2 version 4.10 [28] was used for GFA and G/PLS analyses.

Results and discussions

QSAR of the whole data set using physicochemical descriptors and indicator variables

Stepwise regression. The following best equation was obtained with ten variables using F criterion (F = 4 for inclusion; F = 3.9 for exclusion).

$$\begin{aligned} pC = {} & 2.327 + 0.43(\pm 0.304)\pi_m + 1.00(\pm 0.354)I_{Z\_amino} \\ & + 1.32(\pm 0.306)I_X - 2.18(\pm 0.455)I_{Z\_hethydr} \\ & + 1.23(\pm 0.569)I_{NH} - 1.76(\pm 0.441)I_{z\_hetami} \\ & + 0.89(\pm 0.400)I_{W\_Cl} - 0.99(\pm 0.447)B1_{\_p} \\ & - 15.80(\pm 9.302)B1_{\_o} + 8.30(\pm 5.117)L_{\_o} \\ & n = 97,\ R^2 = 0.854,\ R_\alpha{}^2 = 0.837,\ Q^2 = 0.816, \\ & F = 50.324,\ \mathrm{s} = 0.632,\ PRESS = 43.355 \qquad (1) \end{aligned}$$

All regression coefficients were significant at 95% confidence level and the corresponding confidence intervals were mentioned within parentheses. Equation 1 could explain 83.7% of the variance of the anti-HIV activity while the leave-one-out predicted variance was 81.6%. Equation 1 contains 10 number of independent variables, which is considerably large for 97 compounds, though it maintains the recommended 1:5 ratio for the number of descriptors and number of observations [29]. According to Livingstone and Salt [48,49], when multiple linear regression models are constructed from a large pool of potential independent variables, they suffer from an effect known as "selection bias". The effect of selection bias is to make the resulting models appear more significant than they really are. According to them, a critical F 5% values should be used to judge the significance of MLR models constructed by best subset selection and the critical value (F_{max}) is calculated as follows [49]:

$$F_{\max} = \frac{29.96 n^{3.18} N^{0.21}}{p^{0.82}} e^{\ln(v2)[1.06\ln(v2) - 0.97\ln(n) - 3.97]}$$

In the above equation, p is the number of predictor variables used in a MLR equation, k is the total number of variables from which the p variables have been chosen and n is the number of compounds. For Equation 1, the values of p, k and n are 10, 32 and 97 respectively. N is defined as k!/(p!(k − p)!) and v2 is the second degree of freedom of the F-statistics, i.e., n − p − 1. For Equation 1, F_{max} is calculated to be 27.222 whereas the F value of the equation is 50.324. Thus, Equation 1 passes the criterion of critical F value as prescribed by Livingstone and Salt [49].

The positive value of hydrophobic substituent constant (π_m) indicates that the anti-HIV-1 activity increases with increase in lipophilicity of *meta* substituents of the phenyl ring. The previously reported docking study [16] indicates that the phenyl ring of indolyl aryl sulfones occupies a hydrophobic aromatic-rich pocket formed mainly by the side chain of Tyr181, Tyr188, Phe227 and Tryp229. This explains why compound **29** (containing 3,5-dimethylphenyl moiety) has better anti-HIV-1 activity than compound **18** (containing 2-amino-5-chlorophenyl moiety). The negative coefficients of the STERIMOL width parameters (B1) at the *ortho* and *para* positions of the phenyl ring also indicate that the anti-HIV-1 activity decreases with increase in width of the *ortho* and *para* substituents. This indicates that the pocket within which the phenyl ring fits is of restricted size. The chlorine atom at the W position of the indole moiety also increases the anti-HIV-1 activity as evidenced from the positive coefficient of I_{W_Cl}. This is supported by the observation of the previously reported docking study that this chlorine atom interacts with Pro236. The positive coefficient of the parameter I_{Z_amino} indicates that a carboxamido substituent at 2 position of the indole nucleus is conducive to the anti-HIV-1 activity while the negative coefficients of $I_{Z\_hethydr}$ and I_{Z_hetami} indicate that presence of hydroxyethylamino or hydroxyethylhydrazine group at Z position contributes negatively to the anti-HIV-1 activity. The positive coefficient of I_X indicates that $-SO_2-$ group at X position increases the activity. The sulfonyl group fits in a little hydrophobic pocket made by the side chains of Val106, Lys103 (only α and β CH_2) and Val179 as evidenced from the results of the docking study[16]. The positive coefficient of I_{NH} indicates that presence of the unsubstituted indole NH is conducive for the anti-HIV-1 activity as evidenced from the interaction of indole NH with the Lys101 carbonyl by hydrogen bond[16].

PLS. In case of PLS, the following equation with two components was developed.

$$\begin{aligned} pC = {} & 4.514 + 0.746\pi_m - 1.468\sigma_m - 0.467B1_{\_p} \\ & + 1.271I_X + 1.120I_{Z\_amino} - 1.584I_{Z\_hetami} \\ & - 1.989I_{Z\_hethydr} + 0.460I_{W\_cl} + 0.941I_{NH} \\ & n = 97,\ R^2 = 0.826,\ R_\alpha{}^2 = 0.822, \\ & Q^2 = 0.792,\ \mathrm{F} = 222.460,\ \mathrm{s} = 0.410, \\ & PRESS = 48.914 \qquad (2) \end{aligned}$$

Equation 2 could explain and predict 82.2% and 79.2% respectively of the variance of the anti-HIV activity, the values being slightly inferior to the corresponding values in case of stepwise regression equation. The negative coefficient of the electronic parameter (σ_m) indicates that presence of electron withdrawing *meta* substituent on the phenyl ring decreases the anti-HIV-1 activity.

FA-MLR. From the factor analysis on the data matrix consisting of anti-HIV activity of indolyl aryl sulfones and physicochemical parameters with indicator variables, it was observed that 8 factors could explain the data matrix to the extent of 96.531%. The anti-HIV activity was highly loaded with factor 6 (loaded in I_{Z_hethy} and I_{Z_amino}) and moderately loaded in factor 1 (loaded in π_m, MR_m, $L\__m$, $B1\__m$, $B2\__m$, $B3\__m$ and $B4\__m$), factor 7 (loaded in I_{Z_hetami}) and factor 9 (loaded in Ix). Based on factor analysis the following variables were selected for multiple linear regression. The best equation evolved was as follows:

$$\begin{aligned} pC = {} & 3.97 + 0.567(\pm 0.325)\pi_m - 1.43(\pm 1.017)\sigma_m \\ & + 1.98(\pm 0.529)I_{NH} - 2.50(\pm 0.511)I_{Z\_hethydr} \\ & - 2.03(\pm 0.493)I_{Z\_hetami} + 1.21(\pm 0.372)I_X \\ & n = 97,\ R^2 = 0.772,\ R_\alpha{}^2 = 0.757, \\ & Q^2 = 0.739,\ \mathrm{F} = 50.890,\ \mathrm{s} = 0.771, \\ & PRESS = 61.309 \qquad (3) \end{aligned}$$

Equation 3 involved six descriptors explaining and predicting 75.7% and 63.7% respectively of the variance of the activity. The statistical quality of Equation 3 was inferior to both Equations 1 and 2. The critical F_{max} value for Equation 3 calculated according to Livingstone and Salt [49] was 17.958. The F value of Equation 3 being 50.890, this equation passed the critical F value test.

GFA. In case of GFA (100,000 iterations), the following equation with 11 variables appeared as the best equation.

$$\begin{aligned} pC = {} & 0.755 - 2.27I_{Z\_hethydro} - 1.85I_{Z\_hetami} \\ & + 0.886I_{W\_Cl} + 1.57B3\_m + 1.03I_{Z\_a\min o} \\ & + 1.32I_{NH} + 1.30I_X - 0.598L_{\_p} - 6.80MR_o \\ & + 3.39B4_{\_o} - 2.15B1_{\_m} \\ & n = 97,\ R^2 = 0.862,\ R_\alpha{}^2 = 0.844, \\ & Q^2 = 0.821,\ \mathrm{F} = 48.280,\ \mathrm{s} = 0.618, \\ & PRESS = 42.022 \qquad (4) \end{aligned}$$

Equation 4 was comparable in statistical quality to that of Equation 1. The critical F_{max} value for Equation 4 calculated according to Livingstone and Salt [49] was 28.723. The F value of Equation 4 being 48.280, this equation passed the critical F value test. Equation 4 showed negative coefficient of B1 for *meta* substituent on the phenyl ring while positive coefficient of B3 for the same substituent and this suggested that the *meta* substituents on the phenyl ring should be of optimum shape for interaction with the enzyme cavity.

G/PLS. In G/PLS (5,000 iterations) the best equation was obtained with nine variables and three components.

$$\begin{aligned} pC = {} & 4.921 - 1.616I_{Z\_hetami} - 2.064I_{Z\_hethydr} \\ & + 1.014I_{Z\_a\min o} + 1.449I_X + 0.423B1_{\_m} \\ & + 0.865I_{W\_Cl} - 0.572B3_{\_p} - 0.702B1_{\_o} \\ & + 1.138I_{NH} \\ & n = 97,\ R^2 = 0.825,\ R\alpha^2 = 0.819, \\ & Q^2 = 0.788,\ \mathrm{F} = 145.810,\ \mathrm{s} = 0.412, \\ & PRESS = 49.924 \qquad (5) \end{aligned}$$

Equation 5 was comparable in statistical quality to Equation 2. Equation 5 could explain and predict 81.9% and 78.8% respectively of the variance of the anti-HIV-1 activity.

QSAR of the whole data set using topological descriptors and indicator variables. In search of models of better statistical quality, equations using topological parameters and indicator variables were developed which are shown in Table IV. From stepwise regression, Equation 6 with twelve predictor variables was obtained. This equation involves five E-state parameters, two connectivity parameters and one flexibility index. Equation 6 was comparable in statistical quality to Equation 1 obtained from physicochemical descriptors. In case of PLS regression, Equation 7 with nineteen variables and six components (optimized with crossvalidation) was obtained. Equation 7 involves five E-state parameters, two molecular connectivity parameters, four indicator parameters apart from Balaban index, hydrogen bonding parameter and number of rotatable bonds. The predicted variance (Q^2) value of Equation 7 was lower than that of the stepwise regression model Equation 6 but was comparable to that of Equation 2, the PLS model obtained from physicochemical and indicator parameters. From FA-MLR, Equation 8 was obtained which showed statistical quality and prediction potential less than those of both stepwise regression and PLS models. Equation 8 was slightly inferior to Equation 3 obtained from the physicochemical descriptors. The GFA model of 100,000 iterations produced the best equation with eleven variables [Equation 9]. The predicted variance of this equation was 80.4% while the explained variance was 84.1%. In case of G/PLS with 5,000 iterations, Equation 10 (five components) with 73.6% predicted variance and

Table IV. Different equations obtained from topological descriptors using different statistical methods (whole set).

Parameters	Statistical method	Equation No.	
Topological and indicator	Stepwise regression	(6)	$pC = -0.334 + 6.67(\pm 1.215)^{3}\chi_{c} - 0.192(\pm 0.048)S_{\_sOH} + 9.00(\pm 4.114)I_{NH}$ $- 7.24(\pm 1.712)^{3}\chi_{c}^{v} - 0.118(\pm 0.062)S_{\_sNH_2} - 0.135(\pm 0.054)S_{\_sF}$ $+ 1.38(\pm 0.533)I_{w\_Cl} - 0.183(\pm 0.080)S_{\_sCl} + 2.39(\pm 1.175)I_{Z\_ethcarb}$ $+ 5.60(\pm 3.675)S_{\_aasN} - 0.779(\pm 0.272)^{1}\kappa + 0.022(\pm 0.018)MW$ $n = 97,\ R^{2} = 0.867,\ R_{a}^{2} = 0.849,\ Q^{2} = 0.811,\ \mathrm{F} = 45.799,\ \mathrm{s} = 0.609,\ PRESS = 44.385$
	PLS	(7)	$pC = -2.032 + 2.898J_{X} + 0.166^{3}\kappa - 0.182\phi + 0.157^{2}\chi + 0.391^{3}\chi_{c} - 0.695^{3}\chi_{c}^{v}$ $- 0.121S_{\_sNH2} + 0.211S_{\_ssNH} + 0.285S_{\_aaNH} + 1.200S_{\_sssN} + 0.042S_{\_dO}$ $- 0.136Rotlbonds - 0.464Hbondacceptor + 0.635I_{Z\_amino} + 0.925I_{Z\_ethcarb}$ $- 2.164I_{Z\_hetami} - 2.518I_{Z\_hethydr} + 0.886I_{W\_cl} + 1.040I_{NH}$ $n = 97, R^{2} = 0.847, R_{a}^{2} = 0.837, Q^{2} = 0.782, \mathrm{F} = 83.120, \mathrm{s} = 0.359, PRESS = 51.130$
	FA-MLR	(8)	$pC = 8.391 - 1.489(\pm 0.404)S_{\_aaaC} - 1.528(\pm 0.528)S_{\_aasN} - 1.70(\pm 0.585)I_{Z\_hetami}$ $- 2.264(\pm 0.598)I_{Z\_hethydr} + 0.523(\pm 0.442)I_{Z\_a\min o}$ $n = 97, R^{2} = 0.721, R_{a}^{2} = 0.706, Q^{2} = 0.688, \mathrm{F} = 47.010, \mathrm{s} = 0.656, PRESS = 73.354$
	GFA-MLR	(9)	$pC = 2.52 - 1.883^{3}\chi_{c}^{v} + 0.638I_{W\_Cl} - 1.167S_{\_dssC} + 4.301I_{X} + 0.298S_{\_sCH_3} - 0.432SC_{3C}$ $- 1.013Rotlbonds - 0.586SC_{3P} + 0.982S_{\_aaNH} + 0.298Zagreb + 1.971I_{Z\_ethcarb}$ $n = 97, R^{2} = 0.859, R_{a}^{2} = 0.841, Q^{2} = 0.804, \mathrm{F} = 47.030, \mathrm{s} = 0.625, PRESS = 46.199$
	G/PLS	(10)	$pC = 4.377 + 0.072S_{\_dO} - 0.422Rotlbonds + 0.305S_{\_sCH_3} + 2.489I_{NH}$ $+ 0.546I_{W\_Cl} - 0.108S_{\_sOH} - 1.020^{3}\chi_{c}^{v}$ $n = 97, R^{2} = 0.782, R_{a}^{2} = 0.770, Q^{2} = 0.736, \mathrm{F} = 65.290, \mathrm{s} = 0.512, PRESS = 62.105$

77.0% explained variance was obtained. Because of large pool of descriptors used (from which variables were chosen), Equations 6 and 9 did not pass the criterion of critical $\mathrm{F}_{\max}$ value as recommended by Livingstone and Salt [49]. However, we selected the best equation based on internal validation statistics (*see* the Overview section below).

QSAR of the whole data set using combination of physicochemical and topological descriptors and indicator variables. Attempt was also made to develop models using combined pool of descriptors and the best equations are shown in Table V. Equation 11 was developed from stepwise regression. This equation consisted of eleven predictors which included one flexibility index, three E-state indices, two sterimol parameters and one kappa shape index. The values of explained variance and predicted variance had been improved to some extent on using physicochemical and topological descriptors in combination. Equation 12 involving 18 descriptors (five components) was a PLS model obtained from combined pool of descriptors and statistical quality of this model was better than that of the PLS models obtained from individual group of descriptors [Equations 2 and 7]. From FA-MLR, Equation 13 was developed using six variables. This equation had more predictive power than Equation 3

Table V. Different equations obtained from combined set of descriptors using different statistical methods (whole set).

Physicochemical, topological and indicator	Stepwise regression	(11)	$pC = 2.222 + 2.55(\pm 0.424)^{3}\chi_{c} - 0.146(\pm 0.046)S_{\_sOH} + 2.94(\pm 0.507)I_{NH}$ $- 7.00(\pm 2.080)B1_{\_p} - 2.77(\pm 0.843)\phi + 2.17(\pm 0.837)I_{Z\_ethcarb} - 0.117(\pm 0.052)S_{\_sNH2}$ $+ 2.23(\pm 0.982)^{3}\kappa_{AM} + 3.18(\pm 1.217)L_{\_p} + 0.70(\pm 0.396)I_{W\_cl} + 0.152(\pm 0.117)S_{\_aaCH}$ $n = 97,\ R^{2} = 0.891,\ R_{a}^{2} = 0.877,\ Q^{2} = 0.849, \mathrm{F} = 63.403,\ \mathrm{s} = 0.548,\ PRESS = 35.387$
	PLS	(12)	$pC = 8.698 + 0.398^{3}\chi_{c} + 0.139S_{\_sCH3} - 0.578S_{\_aaaC} + 0.162S_{\_aaNH} - 0.480S_{\_aasN}$ $- 0.094S_{\_sOH} + 0.014S_{\_dO} - 0.148Rotlbonds - 0.206Hbonddonor + 0.665Iz_{\_amino}$ $- 0.892I_{Z\_ethoxy} - 0.750I_{z\_hetami} - 0.920I_{Z\_hethydro} + 0.512I_{W\_cl} - 0.937B1_{\_o} - 0.612\pi_{p}$ $- 0.767B1_{\_p} + 0.634I_{NH}$ $n = 97, R^{2} = 0.873,\ R_{a}^{2} = 0.866, Q^{2} = 0.831, \mathrm{F} = 125.00, \mathrm{s} = 0.299, PRESS = 39.759$
	FA-MLR	(13)	$pC = 6.427 + 0.972(\pm 0.292)^{3}\chi_{c} - 0.632(\pm 0.344)B2_{\_p} - 0.414(\pm 0.392)B3_{\_o}$ $- 1.46(\pm 0.537)S_{\_aasN} - 1.86(\pm 0.536)I_{Z\_a\min o} - 2.289(\pm 0.555)I_{Z\_hethydr}$ $n = 97, R^{2} = 0.735, R_{a}^{2} = 0.718, Q^{2} = 0.701, \mathrm{F} = 41.680, \mathrm{s} = 0.831, PRESS = 70.320$
	GFA-MLR	(14)	$pC = 5.835 - 0.985\pi_{p} - 0.106S_{\_sOH} + 0.230S_{\_sCH_3} - 0.358Rotlbonds + 0.819I_{W\_Cl}$ $+ 0.056S_{\_dO} + 0.483S_{\_aaNH} - 0.272B1_{-p} - 0.613B3_{\_o} + 0.670I_{Z\_amino}$ $n = 97, \mathrm{R}^{2} = 0.869, R_{a}^{2} = 0.853, Q^{2} = 0.828, \mathrm{F} = 56.870, \mathrm{s} = 0.599, PRESS = 40.365$
	G/PLS	(15)	$pC = 7.906 - 0.969Rotlbonds + 0.165S_{\_sCH_3} - 0.850B4_{\_p} + 1.517I_{NH} + 1.439^{3}\kappa$ $- 1.593S_{\_aaaC} - 0.654B3_{\_o}$ $n = 97, R^{2} = 0.818, R_{a}^{2} = 0.808, Q^{2} = 0.785, \mathrm{F} = 81.890, \mathrm{s} = 0.427, PRESS = 50.440$

Table VI. Comparative study of statistical parameters of models using different descriptors.

Type of descriptors	Statistical method	Q^2	R^2	Ra^2	F	s
Physicochemical + indicators	FAMLR	0.739	0.772	0.757	50.890	0.771
	Stepwise	0.816	0.854	0.837	50.324	0.632
	PLS	0.792	0.826	0.822	222.460	0.410
	GFA-MLR	**0.821**	**0.862**	0.844	48.280	0.618
	G/PLS	0.788	0.825	0.819	145.810	0.412
Topological + indicators	FAMLR	0.688	0.721	0.706	47.010	0.849
	Stepwise	**0.811**	**0.867**	0.849	45.799	0.609
	PLS	0.782	0.847	0.837	83.120	0.359
	GFA-MLR	0.804	0.859	0.841	47.030	0.625
	G/PLS	0.736	0.782	0.770	65.290	0.512
Combined (whole)	FAMLR	0.701	0.735	0.718	41.680	0.831
	Stepwise	**0.849**	**0.891**	0.877	63.403	0.548
	PLS	0.831	0.873	0.866	125.000	0.299
	GFA-MLR	0.828	0.869	0.853	56.870	0.599
	G/PLS	0.785	0.818	0.808	81.890	0.427
Combined (training)	FAMLR	0.677	0.723	0.702	34.510	0.845
	Stepwise	**0.853**	**0.906**	0.887	47.243	0.522
	PLS	0.782	0.842	0.831	57.650	0.269
	GFA-MLR	0.823	0.883	0.866	52.110	0.566
	G/PLS	0.806	0.844	0.604	71.610	0.265

and Equation 8 obtained from individual group of descriptors. The GFA model obtained from 100,000 iterations (Equation 14) consisted of ten predictors out of which five are topological three are physicochemical and two are indicator variables. In case of G/PLS obtained from 5000 iterations, the best equation (five components) showed 80.8% explained variance and 78.5% predicted variance. Table VI shows a comparison of statistical parameters of different models applied on the whole data set. For QSAR using the combined set of descriptors (total 84 in number as a pool from which variables were chosen), we did not check the F_{max} criterion of Livingstone and Salt [49]. However, we selected equations based on validation criteria (internal, or external for the next section).

Splitting of the data set into training and test sets. Not even a robust and validated QSAR model can be expected to reliably predict the response for the entire universe of chemicals [50]. Only the predictions for chemicals falling within the applicability domain can be considered reliable. The selection of training and test sets should be based on the proximity of the representative points of the test set to representative points of the training set in the multidimensional descriptor space.There are different methods of rational division of QSAR data set into training and test sets [51] out of which a clustering technique has been used in the present case.

At first all compounds were clustered according to their structural similarities (using *K*-means clustering technique) as shown in Table VII. From each of seven clusters, approximately 25% of the compounds were taken in the test set as shown in Table I by superscript "$\star$". Using 72 compounds as the training set, equations were developed which are shown in Table VIII. The test set compounds were used to examine predictive potential of the compounds and predictive R^2 values (R^2_{pred}) were noted. While developing the models, both physicochemical and topological descriptors along with indicator variables were used. In this case, stepwise regression gave the best Equation (Equation 16) with twelve predictors including E-state parameters, connectivity index, Balaban index, hydrophobicity parameters etc. In case of PLS, Equation 17 was obtained with eleven variables (six components). Although the value of explained variance was low but predictive power of the model was good. In case of FA-MLR, Equation 18 had shown that except connectivity index all other four variables had negative impact on anti-HIV-1 activity. Equation 19 was obtained from GFA-MLR with 100,000 iterations with nine variables. This equation includes six topological descriptors and three indicator variables. In case of G/PLS with 5000 iterations, Equation 20 was obtained. Like PLS, it has low explained variance but good predictive capacity. It consists of two E-state parameters, two kappa shape indices, one indicator, one hydrophobicity parameter and one connectivity index. Table IX shows that the

Table VII. Serial numbers of compounds under different clusters.

Cluster No.	Serial numbers of compounds
1	1;23;25;27;28;29;30;31;32;34;44;45;46;47;48;85;86;87;88;89;96;97;100;110
2	2;11;13;16;17;18;19
3	5;14;21;39;40;41;42;79;80;82;91;92;93;94
4	6;12;20;22;81;83;84;95
5	7;53;66;68;71;72;73;74;75;76;77;78
6	9;10;24;26;33;35;36;37;38;49;50;55;63
7	90;98;101;102;103;104;105;106;107;108;109;111;112;113;114;115;116;117

Table VIII. Different equations obtained from combined set of descriptors using different statistical methods from the training set.

Physicochemical, topological and indicator	Stepwise regression	(16)	$pC = -1.259 + 1.18(\pm 0.322)\,{}^{3}\chi_c + 0.79(\pm 0.478)I_{Z\_NH2} - 0.295(\pm 0.062)S_{\_sOH} - 1.07(\pm 0.334)\pi_p + 1.01(\pm 0.898)S_{\_sssN} + 1.92(\pm 0.630)I_{NH} - 0.223(\pm 0.088)\,S_{\_sNH2} + 2.53(\pm 1.190)\mathcal{J}_X - 1.45(\pm 0.339)I_{Z\_ethoxy} - 0.267(\pm 0.184)Hbondacceptor - 0.57(\pm 0.374)\pi_o + 0.55(\pm 0.472)I_{W\_Cl}$ $n = 72, R^2 = 0.906, R_a^2 = 0.887, Q^2 = 0.853, \mathrm{F} = 47.243, \mathrm{s} = 0.522, PRESS = 0.2647, RMSEP = 0.835$
	PLS	(17)	$pC = 7.463 - 0.302\phi - 0.635\,{}^{3}\chi_c^v + 0.152S_{\_sCH_3} - 1.607S_{\_aaaC} + 1.381S_{\_sssN} - 0.085S_{\_sOH} + 0.926I_{Z\_a\min o} - 0.538I_{Z\_hetami} - 0.810I_{Z\_hethydr} + 0.545B2_{\_m} + 1.215I_{NH}$ $n = 72, R^2 = 0.842, R_a^2 = 0.598, Q^2 = 0.782, \mathrm{F} = 57.650, \mathrm{s} = 0.269, PRESS = 37.070, RMSEP = 0.859$
	FA-MLR	(18)	$pC = 5.478 + 0.927(\pm 0.366)\,{}^{3}\chi_c - 0.479(\pm 0.404)B2_{\_p} - 1.37(\pm 0.612)S_{\_aasN} - 0.243(\pm 0.059)S_{\_sOH} - 1.466(\pm 1.320)\sigma_m$ $n = 72, R^2 = 0.723, R_a^2 = 0.702, Q^2 = 0.677, \mathrm{F} = 34.510, \mathrm{s} = 0.471, PRESS = 54.976, RMSEP = 0.904$
	GFA-MLR	(19)	$pC = 1.433 - 0.180S_{\_sOH} + 2.972\,{}^{2}\chi + 2.091I_{Z\_ethcarb} + 0.264S_{\_aaCH} - 2.857\,{}^{1}\chi + 2.681I_{NH} - 0.0998S_{\_sNH_2} + 0.610I_{W\_Cl} - 1.269\pi_p$ $n = 72, R^2 = 0.883, R_a^2 = 0.866, Q^2 = 0.823,$ $\mathrm{F} = 52.110, \mathrm{s} = 0.566, PRESS = 30.201, RMSEP = 0.799$
	G/PLS	(20)	$pC = 3.824 + 0.193S_{\_sCH_3} - 1.160\,{}^{2}\kappa_{AM} + 0.731\,{}^{2}\chi + 0.680S_{\_aaNH} - 0.161S_{\_sOH} - 0.796B3_{\_p} - 1.176I_{Z\_ethoxy} + 0.06\,{}^{1}\kappa$ $n = 72, R^2 = 0.844, R_a^2 = 0.604, Q^2 = 0.806, \mathrm{F} = 71.610, \mathrm{s} = 0.265, PRESS = 33.010, RMSEP = 0.969$

GFA derived model has the maximum $\mathrm{R}^2_{\mathrm{pred}}$ value. The squared correlation coefficient values between the observed and predicted values of the test set compounds with (r^2) and without (r_0^2) were also noted. All the models have satisfied the requirement of the value of $(\mathrm{r}^2 - \mathrm{r}_0^2)/\mathrm{r}^2$ being less than 0.1 as recommended by Golbraikh and Tropsha [52].

Moreover, R^2_{pred} value is mainly controlled by the value of $(Y_{test} - \bar{Y}_{training})^2$, i.e., mean of training data set. Thus, it may not truly reflect the predictive capability on new dataset. Besides squared regression coefficient (r^2) between observed and predicted values of the test set compounds does not necessarily mean that the predicted values are very near to observed activity (there may be considerable difference between the values though maintaining an overall good intercorrelation). To better indicate external predictive capacity of a model a modified r^2 term ($\mathrm{r}^2_{\mathrm{m}}$) was defined in the following manner [53].

$$r_m^2 = r^2\left(1 - \left|\sqrt{r^2 - r_o^2}\right|\right)$$

In case of good external prediction predicted values will be very close to observed activity values. So, r^2 value will be very near to $\mathrm{r}_{\mathrm{o}}^2$ value. In the best case $\mathrm{r}^2_{\mathrm{m}}$ will be equal to r^2 whereas in the worst case $\mathrm{r}^2_{\mathrm{m}}$ value will be zero. In the present case $\mathrm{r}^2_{\mathrm{m}}$ values of all the models [Equations 16–20] are acceptable.

Artificial neural network. For the development of neural network model we had trained the network with the training set using backpropagation followed by conjugate gradient descent method. The network so developed was used for prediction of anti-HIV-1 activity values of the test set compounds. Using different iterations of backpropagation and conjugate gradient descent, varying numbers of hidden layers and its units per layer, a number of models were developed. Neural networks were optimized using a training subset. A separate subset (the selection subset) was used to halt training to mitigate over-learning, or to select from a number of models trained with different parameters. Then, a third subset (the test subset) was used to perform an unbiased estimation of the network's likely performance. During this study we have first selected certain values for iterations, numbers of hidden layers,

Table IX. Comparison of predictivity parameters of different models obtained from the training set.

Statistical methods	Q^2 (Training set)	R^2 (Training set)	R_a^2 (Training set)	$\mathrm{R}^2_{\mathrm{pred}}$ (test set)	r^2	$\mathrm{r}_{\mathrm{o}}^2$	$(\mathrm{r}^2 - \mathrm{r}_{\mathrm{o}}^2)/\mathrm{r}^2$	$\mathrm{r}^2_{\mathrm{m}}$
FA-MLR	0.677	0.723	0.702	0.693	0.677	0.674	0.004	0.643
Stepwise	**0.853**	**0.906**	0.887	0.738	0.708	0.708	0.000	0.702
PLS	0.782	0.842	0.831	0.722	0.695	0.695	0.000	0.691
GFA	0.823	0.883	0.866	**0.760**	0.736	0.734	0.003	0.704
G/PLS	0.806	0.844	0.604	0.647	0.636	0.635	0.002	0.614

Table X. Comparative study of different networks.

Model No.	No. of hidden layer	No. of unit in different layers	No. of cross validated resampling	No. of epoch in backpropagaion followed by conjugate gradient descent	Absolute error mean	Correlation coefficient (r^2) between Obs. & Pred. values of the test set
1.	3	40 39 38	18	500,200	0.829	0.662
2.	2	39 37	15	800,700	0.686	0.742
3.	1	38	15	500,200	0.721	0.729
4.	**1**	**39**	**15**	800,700	**0.641**	**0.781**
5.	1	39	15	800,600	0.654	0.764

numbers of elements per layer etc. After that we have increased and decreased the numbers of a particular parameter by fixing the other parameters. Here we have presented comparative study of 5 different models in Tables X and XI. The model shown in bold is the best one which we have developed so far. In that model one hidden layer of 39 elements was used. Initialization method selected for network was random uniform. Weight decay was regularized in both phases (decay factor = 0.01, scale factor = 1). Learning rate and momentum of each iteration were adjusted to 0.01 and 0.3 respectively. Number of crossvalidated resampling was set to 15. During 15 resampling, numbers of cases selected for training, selection and test were 36, 19 and 4 respectively. Root mean square error of prediction (RMSEP) of this model was 0.765. In Table XII, r^2 (correlation coefficient between observed and predicted value) of nonlinear method (ANN) was compared with the r^2 value obtained from other linear methods.

Overview

Different statistical methods like stepwise regression, PLS, FA-MLR, GFA-MLR, G/PLS have been applied for modeling anti-HIV-1 activity of indolyl aryl sulfone derivatives using physicochemical and topological descriptors along with indicator variables. Different equations indicate that sulfonyl group flanked between the indole nucleus and phenyl ring, NH group of the indole nucleus, hydrophobicity of the substituents on the phenyl nucleus and chlorine atom at the 5 position of indole moiety are necessary for optimum interaction with reverse transcriptase enzyme. These interactions are also supported by the previously published results of docking studies on this group of compounds [16]. In case of the modeling

Table XI. Comparison of external predictivity characteristics of different ANN models.

Model No.	r^2	r_o^2	$(r^2 - r_o^2)/r^2$	r_m^2
1.	0.662	0.607	0.084	0.506
2.	0.742	0.727	0.020	0.652
3.	0.729	0.716	0.019	0.644
4.	0.782	0.765	0.020	0.682
5.	0.764	0.751	0.017	0.677

Table XII. Comparison of r^2 between observed and predicted values of the test set compounds using different techniques.

Statistical Methods	r^2 value
Stepwise Regression	0.708
PLS	0.695
FA-MLR	0.677
GFA-MLR	0.736
G/PLS	0.636
ANN	**0.781**

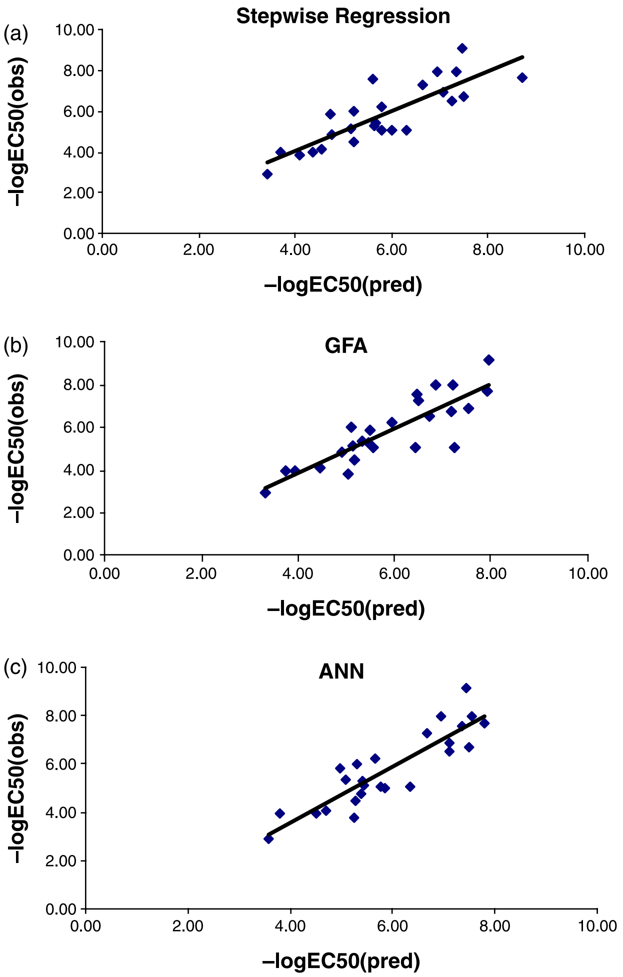


Figure 1. Scatter plots of observed versus predicted values of the test set compounds using (a) stepwise regression model {Equation 16]; (b) GFA-MLR model [Equation 19; (c) artificial neural network model [ANN model 4].

of the whole data set with physicochemical parameters, the best equation based on internal validation characteristics was obtained with GFA-MLR ($Q^2 = 0.821$). When the data set was modeled with topological descriptors, the best model came from stepwise regression ($Q^2 = 0.811$). On using combined set of descriptors, the best model was obtained from stepwise regression ($Q^2 = 0.849$). Again, the whole dataset was divided into training set (72 compounds) and test set (25 compounds). The best model obtained from the training set (stepwise regression) showed good internal predictive power ($Q^2 = 0.853$) which was superior to predictive power of the model ($Q^2 = 0.81$) obtained from 3D-QSAR study published in reference [26]. The external predictive power of the model was also encouraging ($R^2_{pred} = 0.738$). However, the model showing the best external validation parameter was one obtained from GFA-MLR. ($R^2_{pred} = 0.760$). Again, ANN model was developed based on the training set data. The best model obtained from ANN showed a good r^2 value (squared regression coefficient between observed and predicted values) for the test set compounds (0.781) which was superior to the corresponding value (0.736) in case of the best linear model (GFA-MLR). This suggests that nonlinear modeling performs better than the linear technique for this data set. The calculated (or predicted) values of the compounds according to Equations 16 and 19, and ANN model (4) are given in Table I. The scatter plots of observed versus predicted values of the test set compounds according to stepwise regression, GFA and ANN models are shown in Figure 1.

Conclusions

Among the linear models, the best equation based on internal validation was obtained with stepwise regression while the best model based on external validation was obtained from GFA-MLR. Again, ANN models were better than GFA-MLR model based on external validation. Thus, nonlinear modeling performs better than the linear technique for this data set.

Acknowledgements

Financial support under the DST Fast Track Scheme for Young Scientists (DST, Govt. of India, New Delhi) is thankfully acknowledged.

References

[1] Roy K, Leonard J. QSAR modeling of HIV-1 reverse transcriptase inhibitor 2-amino-6-arylsulfonylbenzonitriles and congeners using molecular connectivity and E-state parameters. Bioorg Med Chem 2004;12:745–754.

[2] www.unaids.org

[3] pathmicro.med.sc.edu/lecture/hiv2000.htm

[4] www.rhodes.edu/biology/glindquester/viruses/pagespass/hiv/hiv.html

[5] uhavax.hartfold.edu/bugl/hiv.htm

[6] www.aegis.com/topics/basics/hivandaids.html

[7] Villar S, Santana L, Uriate E. PNN model for in silico evaluation of anti-HIV activity and mechanism of action. J Med Chem 2006;49(3):118–1124.

[8] Prabhakar YS, Rawal RK, Gupta MK, Solomon VR, Katti SB. Topological descriptors in modeling the HIV inhibitory activity of 2-aryl 3-pyridyl-thiazolidin 4-ones. Comb Chem High Throughput Screen 2005;8(5):431–437.

[9] Senese CL, Hopfinger AJ. Receptor-independent 4D-QSAR analysis of a set of norstatine derived inhibitors of HIV-1 protease. J Chem Inf Comput Sci 2003;43:1297–1307.

[10] Makhija MT, Kulkarni VM. 3D-QSAR and molecular modeling of HIV-1 integrase inhibitors. J Comp Aided Mol Des 2002;16(3):181–200.

[11] Buolamwini JK, Assefa H. CoMFA and CoMSIA 3D-QSAR and docking studies on conformationally-restrained cinnamoyl HIV-1 integrase inhibitors: Exploration of a binding mode at the active site. J Med Chem 2002;45(4):841–852.

[12] Niwa T. Prediction of biological targets using PNN and atom-type descriptors. J Med Chem 2004;47(10):2645–2650.

[13] Weekes D, Fogel GB. Evolutionary optimization, backpropagation and data preparation issues in QSAR modeling of HIV inhibition by HEPT derivatives. Biosystems 2003;72: 149–158.

[14] Senese CL, Hopfinger AJ. A simple clustering technique to improve QSAR model selection and predictivity: Application to a receptor independent 4D-QSAR analysis of cyclic urea derived inhibitors of HIV-1 protease. J Chem Inf Comp Sci 2003;43(6):2180–2193.

[15] Pungpo P, Hannongbua S. 3D-QSAR study on HIV-1 RT inhibitors in the class of dipyridodiazepinone derivatives, using CoMFA. J Mol Graph Model 2000;18(6):581–590, & 601.

[16] Ragno R, Artico M, Martino GD, Regina GL, Coluccia A, Pasquali AD, Silvestri R. Docking and 3-D QSAR studies on indolyl aryl sulfones. Binding mode exploration at the HIV-1 reverse transcriptase non-nucleoside binding site and design of highly active N-(2-Hydroxyethyl) carboxamide and N-(2-Hydroxyethyl) carbohydrazide derivatives. J Med Chem 2005; 48:213–223.

[17] Leonard JT, Roy K. Classical QSAR modeling of HIV-1 reverse transcriptase inhibitor 2-amino-6-arylsulfonylbenzonitriles and congener. QSAR Comb Sci 2004;23:23–35.

[18] Leonard JT, Roy K. QSAR modeling of anti-HIV activities of alkenyldiarylmethanes using topological and physicochemical descriptors. Drug Des Discov 2003;18:165–180.

[19] Leonard JT, Roy K. Classical QSAR modeling of CCR5 receptor binding affinity of substituted benzylpyrazoles. QSAR Comb Sci 2004;23:387–398.

[20] Roy K, Leonard JT. Classical QSAR modeling of anti-HIV 2,3-diaryl-1,3-thiazolidin-4-ones. QSAR Comb Sci 2005;24: 579–592.

[21] Roy K, Leonard JT. QSAR by LFER model of cytotoxicity data of anti-HIV 5-phenyl-1-phenylamino-1H-imidazole derivatives using principal component factor analysis and genetic function approximation. Bioorg Med Chem 2005;13:2967–2973.

[22] Roy K, Leonard JT. Topological QSAR modeling of cytotoxicity data of anti-HIV5-phenyl-1-phenylamino-imidazole derivatives using GFA, G/PLS, FA and PCRA techniques. Indian J Chem 2006;45A:126–137.

[23] Roy K, Leonard JT. QSAR analyses of 3-(4-benzylpiperidin-1-yl)-N phenylpropylamine derivatives as potent CCR5 antagonists. J Chem Inf Model 2005;45:1352–1368.

[24] Leonard JT, Roy K. QSAR by LFER model of HIV protease inhibitor mannitol derivatives using FA-MLR, PCRA, and PLS techniques. Bioorg Med Chem 2006;14:1039–1046.

[25] Leonard JT, Roy K, Comparative QSAR. modeling of CCR5 receptor binding affinities of substituted 1-(3,3-diphenylpropyl)-piperidinyl amides and ureas. Bioorg Med Chem Lett 2006;16:4467–4474.

[26] Ragno R, Coluccia A, Regina GL, Martino GD, Piscitelli F, Lavecchia A, Novellino E, Bergamini A, Ciaprini C, Sinistro A, Maga G, Crespan E, Artico M, Silvestri R. Design molecular modeling, synthesis, and anti-HIV-1 activity of new indolyl aryl sulfones. Novel derivatives of the indole-2-carboxamide. J Med Chem 2006;49:3172–3184.

[27] Hansch C, Leo A, Hoekman D. Exploring QSAR Hydrophobic, electronic and steric constants. Washington, DC: American Chemical Society; 1995.

[28] Cerius2 Version 4.10 is a product of Accelrys Inc., San Diego, CA.

[29] Eriksson L, Jaworska J, Worth AP, Cronin MTD, McDowell RM, Gramatica P. Methods for reliability and uncertainty assessment and for applicability evaluations of classification and regression-based QSARs. Environ Health Perspect 2003;111:1361–1375.

[30] Guha R, Jurs PC. Determining the validity of a QSAR Model-A classification approach. J Chem Inf Model 2005;45:65–73.

[31] Leonard JT, Roy K. On selection of training and test sets for the development of predictive QSAR models. QSAR Comb Sci 2006;25(3):235–251.

[32] Everitt BS, Landau S, Leese M. Cluster analysis. Edward Arnold: London; 2001.

[33] Kowalski RB, Wold S. Handbook of statistics. Amsterdam: North Holland Publishing Company; 1982.

[34] Downs GM, Willett P. In: van de Waterbeemd H, editor. Advanced computer assisted techniques in drug discovery. Weinheim (Ger.): VCH; 1995. p 111–130.

[35] Darlington RB. Regression and linear models. New York: McGraw-Hill; 1990.

[36] Wold S. In: van de Waterbeemd H, editor. Chemometric methods in molecular design. Weinheim: VCH; 1995. p 195.

[37] Fan Y, Shi LM, Kohn KW, Pommier Y, Weinstein JN. Quantitative structure–antitumor activity relationships of camptothecin analogues: Cluster analysis and genetic algorithm-based studies. J Med Chem 2001;44:3254.

[38] Franke R. Theoretical drug design methods. Amsterdam: Elsevier; 1984. p 184.

[39] Franke R, Gruska A. In: van de Waterbeemd H, editor. Chemometric methods in molecular design. Weinheim: VCH; 1995. p 113.

[40] Rogers D, Hopfinger AJ. Application of genetic function approximation to quantitative structure–activity relationships and quantitative structure–property relationships. J Chem Inf Comput Sci 1994;34:854–866.

[41] Tang Y, Jiang HL, Chen KX, Ji RY. QSAR study of artemisinin (Qinghaosu) derivatives using neural network method. Indian J Chem 1996;35B:325–332.

[42] Snedecor GW, Cochran WG. In: van de Waterbeemd H, editor. Statistical methods. New Delhi: Oxford and IBH; 1967. p 381.

[43] Wold S, Eriksson L. In: van de Waterbeemd H, editor. Chemometric methods in molecular design. Weinheim: VCH; 1995. p 312.

[44] Debnath AK. In: Ghose AK, Viswanadhan VN, editors. Combinatorial library design and evaluation. New York: Marcel Dekker Inc. 2001. p 73.

[45] MINITAB is a statistical software of Minitab Inc.; USA.

[46] SPSS is a statistical software of SPSS Inc.; USA.

[47] STATISTICA is a statistical software of STATSOFT Inc.; USA.

[48] Livingstone DJ, Salt DW. Judging the significance of multiple linear regression models. J Med Chem 2005;48:661–663.

[49] http://www.port.ac.uk/research/cmd/research/selectionbiasin-multipleregression/

[50] Gramatica P. Principles of QSAR models validation: Internal and external. QSAR Comb Sci 2007;26:694–701.

[51] Roy K. On some aspects of validation of predictive QSAR models. Expert Opin Drug Discov 2007;2:1567–1577.

[52] Golbraikh A, Tropsha A. Beware of q^2!. J Mol Graphics Mod 2002;20:269–276.

[53] Roy P, Roy K. On some aspects of variable selection for partial least squares regression models. QSAR Comb Sci 2007;26: http://dx.doi.org/10.1002/qsar.200710043.