QSAR Studies on the use of 5,6-dihydro-2-pyrones as HIV-1 protease inhibitors[#]

Vijay K Agrawal[a], Jyoti Singh[a], Krishna C Mishra[a] and Padmakar V. Khadikar,[b]* and Yusuf Ali Jaliwala[c]

[a]QSAR and Computer Chemical Laboratories, A.P.S. University, Rewa-486 003, India
[b,]Research Division, Laxmi Fumigation and Pest Control Pvt. Ltd.*
3 Khatipura, Indore-452 007, India
[c]Risiraj college of Pharmacy Sawer Road Indore-452010 India
E-mail: pvkhadikar@rediffmail.com, vijay-agrawal@lycos.com, jyoti_singh07@rediffmail.com

Abstract

The paper describes QSAR studies on HIV-1 protease inhibitors using distance-based topological indices. A series of 5,6-dihydro-2-pyrones were used as HIV-1 protease inhibitors. The regression analysis of the data has shown that a tetra-parametric model containing topological indices and a pair of indicator parameters gives excellent results. The inhibitory activity is expressed as $\log IC_{50}$.

Keywords: QSAR, HIV-1 inhibitors, dihydropyrones, topological indices, molecular descriptors

Introduction

The World Health Organization has warned about the danger of AIDS, which has killed more than 2.5 million people world-wide.[1] Hagen and coworkers[1] have reported on advancements in the treatment of HIV infection by the use of HIV protease inhibitors. They have synthesized a series of dihydropyrones substituted with one or more groups and assayed them for antiviral activity. They observed that 5,6-dihydro-4-hydroxy-1-pyrones are very effective as HIV-1 protease inhibitors and have also discussed the profound effect of polarity on antiviral activity and preliminary studies indicated that these compounds are effective against protease resistant strains of HIV, however, to date, no attempt has been made to investigate the structure-activity relationship (SAR) on the biological activity of this series of compounds using distance-based topological indices. Our earlier research has shown that topological indices are useful tools for modeling biological activities of organic compounds acting as drugs[2–12]. Topological indices can be used successfully for QSAR studies including modeling of anti-HIV activity[13,14].

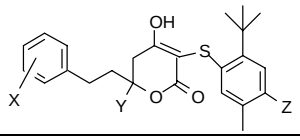
A plethora of such topological indices are available in the literature, which are widely used in drug modeling and also in establishing quantitative structure-property-activity –toxicity relationships (QSAR/QSPR/QSTR) [13–23]. Therefore, we thought it worthy to use distance-based topological indices for modeling 5,6-dihydro-4-hydroxy-1-pyrones as HIV-1 inhibitors, i.e. to estimate their inhibitory activity ($\log IC_{50}$). For this purpose, we have adopted the activities as $\log IC_{50}$, which were reported by Hagen and coworkers.[1] Since we have used a cross-validation method there is no need to work on training and test sets separately.

Results and Discussion

The success of QSAR studies mainly depends whether or not the molecular descriptors chosen are appropriate to explain the biological activity. Based on our earlier studies we have used Wiener [18] (W)-, Szeged [19,20](Sz)-, Balaban [21] (J)-, first-order connectivity [22] ($^1\chi$)-, branching [23] (B)- and logRB[12] indices (Table 2) in the present study. In addition, we have also used three indicator parameters (Ip_1, Ip_2, Ip_3) related to substitution at X, Y and Z respectively (Table 1). These indicator parameters account for those structural features which are not covered in the topological indices used. The details of these indicator parameters are given in the experimental section. It is worth mentioning that, though the indictor parameters are dummy parameters, they are commonly used in QSAR studies.

The inter-correlatedness among the topological indices, indicator parameters and their correlation with the activity ($\log IC_{50}$) is demonstrated in Table 3. A close look at Table 3 shows that W, $^1\chi$, B and logRB indices are highly linearly correlated. This means that a model in which any combination of these indices occurs may suffer from the defect due to collinearity[24]. However, the use of highly correlated parameters in a model has been critically examined by Randic[25] and we will use his recommendations to explain the models where highly linearly correlated topological indices are combined together. Randic[25] argued that "selection of the descriptors to be used in structure-property-activity relationships should not be delegated solely to the computers although the statistical criteria will continue to be useful for preliminary screening of descriptors taken from a large pool.

Table 1. Structural detail, biological activity and indictor parameters for the compounds used in the present study



Compd No	X	Y	Z	logIC_{50}	Ip_1	Ip_2	Ip_3
1.	H	Ph	H	1.5440	0	1	0
2.	**(*)** H	Ph	OH	1.5185	0	1	1
3.	H	Ph	$O(CH_2)_2OH$	0.8325	0	1	1
4.	H	Ph	CH_2OH	0.8195	0	1	1
5.	H	Ph	OCH_3	1.1760	0	1	1
6.	4-OH	Ph	H	1.0413	1	1	0
7.	4-NH_2	Ph	H	1.3802	1	1	0
8.	H	Ph-OH	H	1.6020	0	1	0
9.	H	Ph-NH_2	H	1.5051	0	1	0
10.	H	$PhO(CH_2)_2OH$	H	1.0792	0	1	0
11.	4-OH	Ph	CH_2OH	0.2304	1	1	1
12.	3-OH	Ph	CH_2OH	0.3979	1	1	1
13.	4-NH_2	Ph	CH_2OH	0.4913	1	1	1
14.	3-NH_2	Ph	CH_2OH	0.6020	1	1	1
15.	**(*)** H	$PhO(CH_2)_2OH$	CH_2OH	0.1461	0	1	1
16.	H	$PhO(CH_2)_2OH$	$O(CH_2)_2OH$	0.8061	0	1	1
17.	H	$PhO(CH_2)_2OH$	OH	0.5682	0	1	1
18.	H	$PhO(CH_2)_2OH$	CH_2OCH_3	0.6532	0	1	1
19.	4-OH	PhOH	CH_2OH	2.0791	1	1	1
20.	4-OH	Cyclohexyl	CH_2OH	0.6127	1	0	1
21.	4-OH	Isopropyl	CH_2OH	0.5563	1	0	1
22.	4-OH	Methyl	CH_2OH	0.6334	1	0	1
23.	4-NH_2	Cyclohexyl	CH_2OH	0.5051	1	0	1
24.	4-NH_2	Isopropyl	CH_2OH	0.4313	1	0	1

The indictor parameter Ip_1 =1 when substituents at X are other than hydrogen,
Ip_2 =1 when substituent at Y is phenyl and Ip_3 =1 when substituent at Z is other than hydrogen.
(*) = Show the deleted compound

Often in an automated selection of descriptors a descriptor will be discarded because it is highly correlated with another descriptor already selected. But what is important is not whether two descriptors parallel one another, i.e. duplicate much of the same structural information but whether they in those parts that are important for structure-property-activity correlations. If they

differ in the domain, which is important for the property/ activity considered, both descriptors should be retained. If they differ in parts that are not relevant for the correlation of considered property/ activity then one of them can be discarded. Hence, the residual of the correlation between two descriptors should be examined and kept or discarded depending on how well it can improve the correlation based on already selected descriptors. Alternately, one should replace the set of descriptors used by descriptors that can be extracted from them through the orthogonalization procedure that has been introduced in regression analysis."

Table 2. Calculated values of topological indices for the set of compounds used in the present study

Compd. No	W	$^{1}\chi$ (=B)	J	Sz	log RB
1.	3420	16.1772	1.4640	5272	906.8839
2.	3694	16.5879	1.4714	5704	971.9099
3.	4730	18.1259	1.4573	7214	1199.4270
4.	4003	17.1259	1.4720	6171	1042.5930
5.	4003	17.1259	1.4720	6171	1042.5930
6.	3740	16.5711	1.4561	5758	977.6093
7.	3740	16.5711	1.4561	5758	977.6093
8.	3702	16.5711	1.4716	5720	973.9668
9.	3702	18.1091	1.4716	5720	973.9668
10.	4762	17.5198	1.4563	7278	1205.8550
11.	4356	17.5198	1.4650	6708	1118.9240
12.	4326	17.5198	1.4737	6648	1115.8360
13.	4356	17.5198	1.4650	6708	1118.9240
14.	4326	17.5198	1.4737	6648	1115.8360
15.	5473	19.0578	1.4666	8777	1363.7220
16.	6344	20.0578	1.4572	9636	1543.6610
17.	5098	18.5198	1.4644	7808	1281.8340
18.	5888	19.5578	1.4640	8986	1451.0900
19.	4679	17.9136	1.4741	7221	11193.8390
20.	4356	17.5198	1.4650	6708	1118.9240
21.	3564	15.8580	1.6709	5346	921.4896
22.	3130	14.9146	1.5936	4736	806.3530
23.	4356	17.5198	1.4650	6708	1118.9240
24.	3564	15.8580	1.6709	5346	921.4896

A perusal of Table 3 shows that the Balaban index[21] (J) does not correlate with any of the other topological indices used. This is also the case for all the three indicator parameters used.

Furthermore, none of the topological indices (and indicator parameters too) correlate with $logIC_{50}$(nM) singly to yield one variable model, i.e. no statistically significant mono-parametric model is possible for modeling the antiviral activity ($logIC_{50}$).

Table 3. Correlation matrix for the inter-correlation of structural descriptors and their correlation with the activity

	$logIC_{50}$	W	$^1\chi$ (=B)	J	Sz	log RB	Ip_1	Ip_2	Ip_3
$logIC_{50}$	1.0000								
W	-0.2823	1.0000							
$^1\chi$(=B)	-0.1634	0.9357	1.0000						
J	-0.2540	-0.4245	-0.5665	1.0000					
Sz	-0.2805	0.9972	0.9420	-0.4527	1.0000				
logRB	0.4793	0.1853	0.1742	-0.0874	0.1892	1.0000			
Ip_1	-0.2797	-0.3405	-0.4206	0.3695	-0.3512	0.1759	1.0000		
Ip_2	0.3518	0.3384	0.4658	-0.7140	0.3598	0.1373	-0.5130	1.0000	
Ip_3	-0.5595	0.3433	0.2332	0.2426	0.3337	0.1473	0.1924	-0.2962	1.0000

Consequently, we have carried out multi-parametric regression analysis using the maximum-R^2 method [24]. The statistically significant models obtained are summarized in Table 4. In obtaining these models we found that compounds **2** and **15** are outliers, therefore, they were deleted from the further regression analysis. The results recorded in Table 4 show that there are six bi-parametric regression models out of which the one containing logRB and Ip_3 gave the best results. This model is found as:

$$logIC_{50} = 1.4490\times10^{-4} (\pm 2.3552\times10^{-5})logRB - 0.7551(\pm 0.1116)Ip_3 + 1.2142$$

n=22, Se=0.2302, R=0.8893, F=35.917, Q=3.8632 (1)

Here and there after n is the number of compounds used, Se is the standard error of estimation, R is the multiple correlation coefficients, F is the F-ratio and Q is the quality factor[26,27]. The quality factor Q is used to decide the predictive potential of the proposed models. The quality factor Q is defined as the ratio of correlation coefficient to the standard error of estimation. Though its use is criticized[28] we found it to be a good parameter to explain the predictive potential of the models proposed by us. The higher the value of Q the better is the predictive potential of the models.

Table 4. Regression parameters and quality of the proposed models

Model No	Parameters used	A_i $_i$=1,2,3,4,5,6	B (Inter-cept)	Se	R Corr. coeff	F-Ratio	Q=R/Se
1.	logRB	1.1874×10^{-4} ($\pm4.1839\times10^{-5}$)	0.7043	0.4143	0.5358	8.054	1.2933
2.	W	-1.0431×10^{-4} ($\pm1.3608\times10^{-4}$)	1.3349	0.4836	0.1689	0.588	0.3851
3.	Ip_2	0.4409(±0.2293)	0.5478	0.4507	0.3950	3.698	0.8764
4.	Ip_3	-0.6464(±0.1851)	1.3586	0.3867	0.6154	12.194	1.5915
5.	logRB	1.3493×10^{-4}($\pm3.7533\times10^{-5}$)	0.9018	0.3663	0.6859	8.438	1.8725
	Ip_1	-0.4081(±0.1591)					
6.	log RB	1.0818×10^{-4} ($\pm4.0173\times10^{-5}$)	0.4420	0.3934	0.6238	6.053	1.5856
	Ip_2	0.3606(±0.2024)					
7.	W	-1.7518×10^{-4} ($\pm1.1511\times10^{-4}$)	1.4350	0.4013	0.6038	5.450	1.5046
	logRB	1.3095×10^{-4} ($\pm4.1314\times10^{-5}$)					
8.	$^1\chi$	0.0585(±0.0792)	1.7111	0.4190	0.5541	4.209	1.3224
	logRB	1.2452×10^{-4} ($\pm4.3039\times10^{-4}$)					
9.	Sz	-1.1036×10^{-4} ($\pm7.6084\times10^{-5}$)	1.4088	0.4033	0.5985	5.301	1.4840
	logRB	1.3107×10^{-4} (4.1609×10^{-5})					
10.	logRB	1.4490×10^{-4} ($\pm2.3552\times10^{-5}$)	1.2142	0.2302	0.8893	35.917	3.8632
	Ip_3	-0.7551(±0.1116)					
11.	logRB	1.5207×10^{-4} ($\pm1.9889\times10^{-5}$)	1.2930	0.1927	0.9280	37.217	4.8156
	Ip_1	-0.2604(±0.0862)					
	Ip_3	-0.6852(±0.0963)					
12.	W	-6.4633×10^{-4} ($\pm2.8782\times10^{-4}$)	-2.3688	0.3805	0.6773	5.084	1.7800
	$^1\chi$	0.3355(±0.1896)					
	logRB	$1.3065X10^{-4}$(±3.9177)					
13.	$^1\chi$	-0.1759(±0.0881)	9.1105	0.3790	0.6804	5.173	1.7953
	J	-3.014(±1.5751)					
	logRB	1.2609×10^{-4} ($\pm3.8931\times10^{-5}$)					

14.	W	-3.3058×10^{-4} ($\pm8.7616\times10^{-5}$)	2.3751	0.2812	0.8392	14.290	2.9844
	logRB	1.6573×10^{-4} ($\pm2.9948\times10^{-5}$)					
	Ip_1	-0.6033(±0.1327)					
15.	Sz	-2.1419×10^{-4} ($\pm5.8580\times10^{-5}$)	2.3637	0.2851	0.8343	13.744	2.9263
	logRB	$1.6664X10^{-4}$ ($\pm3.0470\times10^{-5}$)					
	Ip_1	-0.6034(±0.1349)					
16.	W	-1.6464×10^{-4} ($\pm5.7758\times10^{-5}$)	1.9551	0.1631	0.9519	40.994	5.8363
	logRB	1.6427×10^{-4} ($\pm1.7370\times10^{-5}$)					
	Ip_1	-0.3847(±0.0850)					
	Ip_3	-0.5597(±0.0926)					
17.	Sz	-1.0647×10^{-4} ($\pm3.8038\times10^{-5}$)	1.9512	0.1641	0.9513	40.470	5.7971
	logRB	1.6484×10^{-4} ($\pm1.7536\times10^{-5}$)					
	Ip_1	-0.3834(±0.0856)					
	Ip_3	-0.5653(±0.0925)					

Ai is the correlation coefficient of the I^{th} parameter.

Table 5. Various correlation models and their qualities of correlations

Model No.	Regression expression
1.	$\log IC_{50}=1.1874\times10^{-4}(\pm4.1839\times10^{-5})\log RB+0.7043$
2.	$\log IC_{50}=-1.0431\times10^{-4}(\pm1.3608\times10^{-4})W+1.3349$
3.	$\log IC_{50}=0.4409(\pm0.2293)Ip_2+0.5478$
4.	$\log IC_{50}=-0.6464(\pm0.185)Ip_3+1.3586$
5.	$\log IC_{50}=1.3493\times10^{-4}(\pm3.7533\times10^{-5})\log RB -0.4081(\pm0.1591)Ip_1+0.9018$
6.	$\log IC_{50=}1.0818\times10^{-4}(\pm4.0173\times10^{-5})\log RB+0.3606(\pm0.2024)Ip_2+0.4420$
7.	$\log IC_{50=1.}7518\times10^{-4}(\pm1.1511\times10^{-4})W+1.3095\times10^{-4}(\pm4.1314\times10^{-5})\log RB+1.4350$
8.	$\log IC_{50=}0.0585(\pm0.0792)\ ^1\chi+1.2452\times10^{-4}(\pm4.3039\times10^{-4})\log RB+1.7111$
9.	$\log IC_{50=}1.1036\times10^{-4}(\pm7.6084\times10^{-5})Sz+1.3107\times10^{-4}(4.1609\times10^{-5})\log RB+1.4088$
10.	$\log IC_{50=}1.4490\times10^{-4}\ (\pm2.3552\times10^{-5})\log RB--07551(\pm0.1116)Ip_3+1.2142$
11.	$\log IC_{50}=-1.5207\times10^{-4}(\pm1.9889\times10^{-5})\log RB-0.2604(\pm0.0862)Ip_1-0.6852(\pm0.0963)Ip_3+1.9512$
12.	$\log IC_{50}=-6.4633\times10^{-4}(\pm2.8782\times10^{-4})W+0.3355(\pm0.1896)\ ^1\chi+1.3065\times10^{-4}(\pm3.9177)\log RB-2.3688$
13.	$\log IC_{50}=-0.1759(\pm0.0881)\ ^1\chi-3.6014(\pm1.5751)J+1.2609\times10^{-4}(\pm3.8931\times10^{-5})\log RB+9.1105$
14.	$\log IC_{50}=-3.3058\times10^{-4}(\pm8.7616\times10^{-5})W+1.6573\times10^{-4}(\pm2.9948\times10^{-5})\log RB-0.6033(\pm0.1327)Ip_1+2.3751$
15.	$\log IC_{50}=-2.1419\times10^{-4}(\pm5.8580\times10^{-5})Sz+1.6664\times10^{-4}(\pm3.0470\times10^{-5})\log RB-0.6034(\pm0.1349)Ip_1+2.3637$
16.	$\log IC_{50}=-1.6464\times10^{-4}(\pm5.7758\times10^{-5})W+1.6427\times10^{-4}(\pm1.7370\times10^{-5})\ \log RB-0.3847(\pm0.0850)Ip_1-0.5597(\pm0.0926)Ip_3+1.9551$
17.	$\log IC_{50}=-1.0647\times10^{-4}(\pm3.8038\times10^{-5})Sz+1.6484\times10^{-4}(\pm1.7536\times10^{-5})\log RB-0.3834(\pm0.0856)Ip_1-0.5653(\pm0.0925)Ip_3+1.9512$

Table 6. Cross- validation parameters for the proposed models

S.N.	Parameters Used	PRESS/SSY	R^2cv	S_{PRESS}	PSE
1. (10)	log(RB), Ip_3	0.2646	0.7355	0.2238	0.2140
2. (11)	log(RB), Ip_1, Ip_3	0.1612	0.8388	0.1825	0.1744
3. (16)	W, log(RB), Ip_1, Ip_3	0.1037	0.8964	0.1501	0.1434
4. (17)	Sz, log(RB), Ip_1, Ip_3	0.1050	0.8950	0.1510	0.1443

PRESS- Predicted residual sum of squares; SSY-Sum of the squares of the response value; R^2cv- Cross validation correlation coefficient; S_{PRESS}- Uncertainty of prediction; PSE- Predictive square error.

Table 7. Observed and estimated logIC_{50} values using model-16 and -17

Compd No	Obs. log IC_{50}	Estimated log IC_{50}	
		Model-16	Model-17
1.	1.544	1.541	1.539
2.	0.833	0.814	0.816
3.	0.820	0.908	0.901
4.	1.176	0.908	0.901
5.	1.041	1.115	1.116
6.	1.380	1.115	1.116
7.	1.602	1.506	1.503
8.	1.505	1.506	1.503
9.	1.079	1.369	1.375
10.	0.230	0.477	0.473
11.	0.398	0.482	0.479
12.	0.491	0.477	0.473
13.	0.602	0.482	0.479
14.	0.806	0.605	0.614
15.	0.568	0.767	0.766
16.	0.653	0.664	0.668
17.	2.079	2.079	2.079
18.	0.613	0.477	0.473
19.	0.556	0.575	0.585
20.	0.633	0.628	0.631
21.	0.505	0.477	0.473
22.	0.431	0.575	0.585

The aforementioned model [eq.(1)] shows that the extent of branching is favorable for the exhibition of the activity, while Ip_3 (i.e. presence of subsistent at Z other than hydrogen) has the

retarding effect for the same. For the set of compounds used, the branching is possible only in the substituents, clearly meaning that the branching in the substitution affects the exhibition of the activity. The negative sign associated with Ip_3 indicates that an increase in the branching in the substitution is not favorable for the exhibition of the activity .

Successive regression analysis resulted in six tri-parametric models out of which two models gave slightly better statistics than the bi-parametric model discussed above eq.(1). One of the two better quality tri-parametric models the model containing logRB, Ip_1 and Ip_3 gave better results. This model is found as:

$$\log IC_{50} = -1.5207\times10^{-4}(\pm1.9889\times10^{-5})\log RB - 0.2604(\pm0.0862)Ip_1 - 0.6852(\pm0.0963)\,Ip_3 + 1.9512$$

n=22, Se=0.1927, R=0.9280, F=37.217, Q=4.8156 (2)

This model [eq.(2)] further supports that the extent of branching is responsible for the exhibition of activity. The physical significance of log RB and the indicator parameter Ip_3 is the same as in model 1 [eq.1]. The indicator parameter Ip_1 stands for the substitution at X, and is other than hydrogen. The negative sign of this parameter indicated that such a substitution is not favorable for the exhibition of activity.

The step-wise regression finally resulted in two tetra-parametric models having better statistics than both the models discussed above. These two models differ only due to the occurrence of W and Sz terms {while the remaining parameters are common for both of them}. We can use these two models for investigating the relative correlation potential of W and Sz indices in modeling the activity. It is worthy of mention that W and Sz indices are highly correlated parameters (r=0.9972). In such a situation one can replace the other and they have to be equally good or equally bad in modeling the activity under investigation. Furthermore, if the model contains both of these indices then the model may suffer from the defect due to collinearity and need to be dealt with using the recommendations of Randic[25]. Regression qualities of the models under study show that W and Sz indices are more or less equally good for this purpose, W index being slightly better than Sz index. Therefore, the tetra-parametric model containing logRB, W, Ip_1 and Ip_3 as is the correlating parameters is good for modeling the inhibitory activity. This model is found as:

$$\log IC_{50} = -1.6464\times10^{-4}(\pm5.7758\times10^{-5})W + 1.6427\times10^{-4}(\pm1.7370\times10^{-5})\log RB - 0.3847(\pm0.0850)Ip_1 - 0.5597(\pm0.0926)Ip_3 + 1.9551$$

n=22, Se=0.1631, R=0.9519, F=40.994, Q=5.8363 (3)

The other tetra-parametric model having almost the same correlation potential includes Sz in place of W. The model is as below:

$$\log IC_{50} = -1.0647\times10^{-4}(\pm3.8038\times10^{-5})Sz + 1.6484\times10^{-4}(\pm1.7536\times10^{-5})\log RB - 0.3834(\pm0.0856)Ip_1 - 0.5653(\pm0.0925)Ip_3 + 1.9512$$

n=22, Se=0.1641, R=0.9513, F=40.470, Q=5.7971 (4)

It is interesting to record that in all the models discussed above, as well as those recorded in Table 4, we observed that the coefficient of the indicator parameter terms are negative. This means that they have a negative role in the exhibition of the activity.

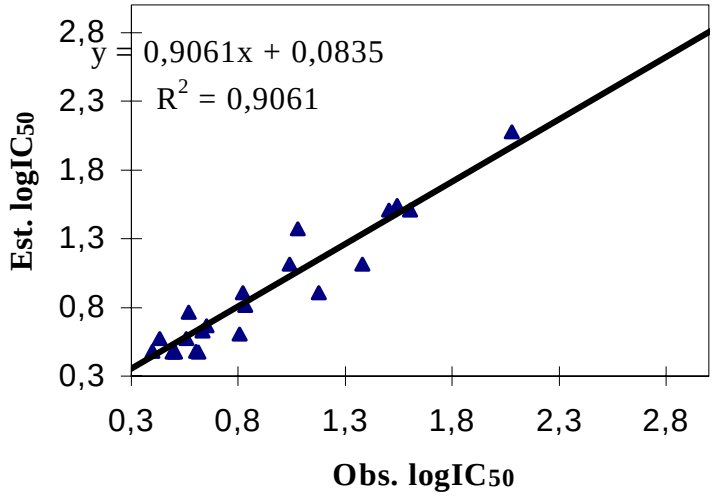


Figure 1. Correlation between observed and estimated $logIC_{50}$-using model 16.

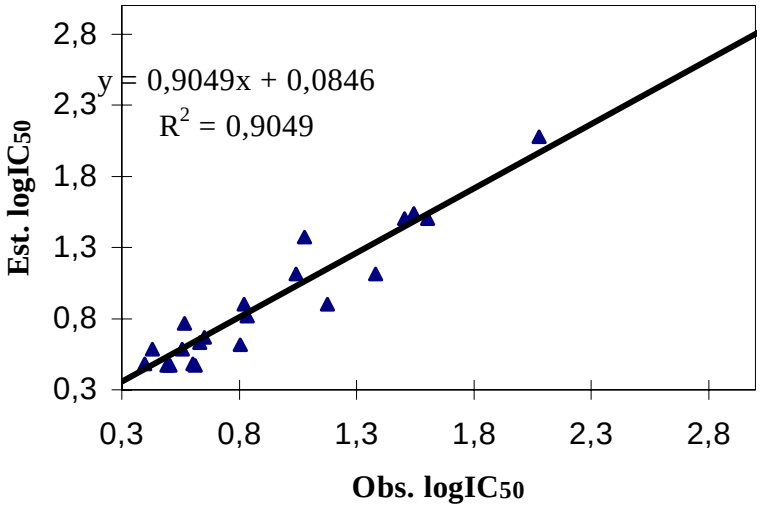


Figure 2. Correlation between observed and estimated $logIC_{50}$-using model 17.

In order to confirm our findings, we have calculated the activity ($logIC_{50}$) from models expressed by eq.(3) and eq.(4) which are discussed above. The calculated activities are then compared with their observed values. Such a comparison is shown in Table 6. The difference between observed and calculated antiviral activity (residue} is the least for the model expressed by eq. (3), showing it to be the most appropriate model for modeling the activity ($logIC_{50}$) of the present set of compounds.

In order to examine the relative potential of the proposed models we have further estimated their predictive correlation coefficients ($R^2_{pred.}$) by plotting graphs between observed and

estimated logIC_{50} values using equations (3) and (4). Such correlations are shown in Figs. 1 and 2 respectively. From Figures 1 and 2 the $R^2_{pred.}$ values are found as 0.9061 and 0.9049, respectively, for the models expressed by eqs. (3) and (4). This finally confirms that the model expressed by eq.(3) has the best predictive potential also.

The aforementioned results show that out of the set of topological indices used by us, the indices logRB and W (or Sz), and the to indicator parameters, are the better parameters for modeling activity. In order to justify the occurrence of highly correlated parameters in the proposed models we have used the recommendations made by Randic[25] wherein, highly inter-correlated descriptors can be used in multi-parametric correlations. The simple reason is that molecular descriptors carry different structural information. By discarding one of the descriptors, which commonly duplicates another, we may be discarding a descriptor that nevertheless may carry useful structural information in the parts in which it does not parallel with the other descriptors. Thus, as suggested by Randic[25] we may safely say that logRB and any other descriptor in combination with this is allowed statistically.

It is interesting to mention that correlation of observed and estimated activity (logIC_{50}) (Fig.1) gave much higher value for predictive correlation coefficient (R^2_{pred}=0.9061). The predictive power, as determined by the Pogliani Q parameter[26,27] for the model expressed by eq.3 (Q=5.8363), confirms that this model has excellent statistics as well as excellent predictive power too. Final support in favor of our findings is obtained by using the cross-validation method[24]. The calculated cross-validated parameters for each of the models are discussed below. The meanings of cross-validated parameters used are given as a footnote to Table 6.

The data presented in Table 6 show that except for the bi-parametric model containing log(RB) and Ip_3, all other models have PRESS/SSY either nearer to 0.4 or smaller than 0.4 and that this ratio for the models expressed by eq. (1) is the smallest. Therefore, we conclude that this model is the best among all the models discussed above. The highest value of R^2_{cv} and the lower value of PSE gave further support to our finding. However, the cross-validated parameter, S_{PRESS} is of no value in this respect as it coincides with the value in the standard error of estimation (Se).

Conclusions

The results and discussion made above lead to the conclusion that the activity (logIC50) of the present set of compounds can be successfully modeled using distance-based topological indices. It was also observed that out of the topological indices used logRB and W are most useful for this purpose. Among the indicator parameters Ip_3 is the best to be used in multi-parametric regression analysis. We also conclude that branching in the substituent has a significant role in the exhibition of the activity.

Experimental Section

Inhibition of HIV-1 protease (IC_{50})

The inhibition of HIV-1 protease values (IC_{50}) (Table 1), which refer to the nanomolar (μM) concentration of the compounds, were adopted from the work of Hagen et al[1]. We have converted IC_{50} values into their log units to get a linear relationship in the equations.

Molecular graphs

The hydrogen suppressed molecular graphs[13-15] were used for the calculation of topological indices W, Sz, $^1\chi$, B, J, logRB (Table 2) employed in the present study.

Topological Indices

Wiener index(W) -The Wiener index (*W*) is a widely used topological index [18] which is based on the vertex-distances of the respective molecular graph. The molecular graph can be denoted by *G* and having v_1, v_2, v_3,...,v_n as its vertices. Let d(v_i,v_j|*G*) stand for the shortest distance between the vertices v_i and v_j. Then the Wiener index is defined as:

$$W = W(G) = \frac{1}{2}\sum_{i=1}^{n\rightarrow}\ \sum_{j=1}^{n} d(v_i, v_j|G) \qquad (5)$$

Szeged index (Sz) - Let *e* be an edge of the molecular graph *G*. Let $n_1(e|G)$ be the number of vertices of *G* lying closer to one end of *e* ; let $n_2(e|G)$ be the number of vertices of *G* lying closer to the other end of *e*. Then the Szeged index [19,20] (*Sz*) is defined as:

$$Sz(G) = Sz = \sum_{e} n_1(e|G)n_2(e|G) \qquad (6)$$

with the summation going over all the edges of *G*.

In cyclic graphs, there are edges equidistant from both the ends of edge *e*; by definition of *Sz* such edges are not taken into account.

First-order connectivity index ($^1\chi$) - The connectivity index $\chi = \chi(G)$ of a graph G is defined by Randic [22] as under :

$$\chi = \chi(G) = \sum_{ij}\left[\delta_i\delta_j\right]^{-0.5} \qquad (7)$$

where δ_i and δ_j are the valence of a vertex i and j, equal to the number of bonds connected to the atoms i and j , in G.

In the case of hetero-systems the connectivity is given in terms of valence delta values δ_i^v and δ_j^v of atoms i and j and is denoted by χ^v. This version of the connectivity index is called the valence connectivity index and is defined [29,30] as under :

$$\chi^v = \chi^v(G) = \sum_{ij}\left[\delta^v{}_i\delta_j{}^v\right]^{-0.5} \qquad (8)$$

where the sum is taken over all bonds i-j of the molecule. Valence delta values are given by the following expression :

$$\delta_i^{\ v} = \frac{Z_i^{\ v} - H_i}{Z_i - Z_j - 1} \qquad (9)$$

where Z_i is the atomic number of atom i, Z_i^v is the number of valence electron of the atom i and H_i is the number of hydrogen atoms attached to atom i.

Now the connectivity and the valence connectivity indices expressed by equations (10) and (11) are termed as first-order connectivity and first-order valence connectivity indices respectively.

Branching index (B) and logRB - The branching index B and logRB have been calculated by the method described in the literature[13].

Balaban index (J) - The Balaban index, J (the average distance sum connectivity index) is defined [21] by :

$$J = J(G)\frac{M}{\mu+1}\Sigma_{bonds}(d_i d_j)^{-1/2} \qquad (10)$$

where M is the number of bonds in a graph G, μ is the cyclomatic number of G and d_i's (i=1,2,3,...,N) are the distance sums (distance degrees) of atoms in G such that

$$d_i = \sum_{j=1}^{N} (D)_{ij} \qquad (11)$$

The cyclomatic number μ of G indicates the number of independent cycles in G and is equal to the minimum number of cuts (removal of bonds) necessary to convert a polycylic structure into an acyclic structure:

$$\mu = M - N + 1 \qquad (12)$$

One way to compute the Balaban index (J) for a hetero-system is to modify the elements of the distance matrix for a hetero-system as follows:

(i) The diagonal elements :

$$(D)_{ij} = 1 - (Z_c / Z_i) \qquad (13)$$

where $Z_c = 6$ and Z_i = atomic number of the given element.

(ii) The off-diagonal elements :

$$d_i = \sum_r k_r \qquad (14)$$

where the summation is over all bonds. The bond parameter k_r is given by :

$$k_r = 1 / b_r (Z_c^2 / Z_i Z_j)$$

where b_r is the bond weight with values : 1 for single bond, 2 for double bond, 1.5 for aromatic bond and 3 for triple bond.

Cross-validation- As opposed to traditional regression methods, the cross-validation evaluates the validity of a model by how well it predicts data rather than how well it fits data. The analysis uses a "leave-one-out" scheme, a model is built with N-1 compounds and the nth compound is predicted.[24] Each compound is left out of the model derivation and predicted in turn. An indication of the performance of the model is obtained from the cross-validated (or predictive r^2_{CV}) method which is defined as:

$$r^2_{cv} = \frac{SD - PRESS}{SD} \qquad (15)$$

where SD is the sum-of-squares deviation for each activity from the mean. PRESS (or predictive sum-of-squares) is the sum of squared difference between the actual and that of predicted values when the compound is omitted from the fitting process. Once a model is developed which has highest cross-validated r^2_{cv}, this method is used to derive the conventional QSAR equation and conventional r^2 and s values. The results of the final model are often visualized as contour maps of the coefficients.

In addition to PRESS, SD, r^2_{cv}, S_{PRESS}, one also needs to evaluate predictive-square-error (PSE) in an attempt to decide the predictive potential of the proposed models. The data of calculation of cross-validated parameters are given in our publications.

Regression Analysis – The maximum-R^2 improvement method was used to propose statistically significant models and to identify prediction models[24]. This method finds the "best" one variable model, the "best" two variable model and so forth for the prediction of property /activity. Several models (combinations of variables) were examined to identify combinations of variables with good prediction capabilities. A variety of statistics associated with residues, i.e. the Wilks-Shapiro test for normality and Cooks D-statistics for outliers, are used to obtain the most reliable results.

Regress-1 software supplied Lukovits, Hungarian Academy of Sciences, Budapest, Hungary.

Computations - All the computations were carried out on a Power Macintosh 9600/233.

Acknowledgments

The authors are thankful to Professor Istvan Lukovits, Hungarian Academy of Sciences, Budapest, Hungary for providing software to carry out regression analysis and to Prof. Ivan Gutman, Faculty of Science, University of Kragujevac, Yugoslavia for introducing one of the authors (Prof. P.V. Khadikar) to this fascinating field of chemical graph theory and topology. The authors are thankful to Prof. A.D.N. Bajpai, Vice-Chancellor of APS University, Rewa, India for providing research facilities and encouragement. The authors are also thankful to University Grants Commission, New Delhi, India for sanctioning a research scheme.

References

[#] On the eve of the 70th birthday of both Prof. Padmakar V. Khadikar and his wife Mrs. Kusum Khadikar.

1. Hagen,S. E; Prasad, J.V.N.V.; Boyer, F. E.; Domagala, J. M.; Ellsworgh, E. L.; Gajda, C.; Hamilton, H. W.; Markoski, L. J.; Steinbaugh, B. A.; Tait, B. D.; Lunney, E. A.; Tummino, P. J.; Ferguson, D.; Hupe, D.; Nouhan, C.; Grachek, S. J.; Saunders, J. M. and Vander Rost S. *J. Med. Chem.* **1997,** *40,* 3707 (and the references therein).
2. Agrawal, V. K.; Khadikar, P. V. *Bioorg. Med. Chem. Letters* **2003,** *13,* 447.

3. Agrawal, V. K.; Sharma, R.; Khadikar, P. V. *Bioorg. Med. Chem.* **2002,** *10,* 3571.
4. Agrawal, V. K.; Sharma, R.; Khadikar, P. V. *Bioorg. Med. Chem.* **2002,** *10,* 2993.
5. Agrawal, V. K.; Singh, J.; Khadikar, P. V. *Bioorg. Med. Chem.* **2002,** *10,* 3981.
6. Khadikar, P. V.; Agrawal, V .K.; Karmarkar, S. *Bioorg. Med. Chem.* **2002,** *10,* 3499.
7. Khadikar, P. V.; Singh, S.; Shrivastava, A. *Bioorg. Med. Chem.* **2002,** *12,* 1125.
8. Khadikar, P. V.; Phadnis, A.; Shrivastava, A. *Bioorg. Med. Chem.* **2002,** *10,* 1181.
9. Khadikar, P. V.; Karmarkar, S.; Agrawal, V. K. *J. Chem. Inf. Comput. Sci.* **2000,** *41,* 934.
10. Agrawal, V. K.; Khadikar, P.V. *Bioorg. Med. Chem.* **2002,** *10,* 3517.
11. Agrawal, V. K. Khadikar, P.V. *Bulg. Chem. Ind.* **2002,** 73, 11.
12. Agrawal, V. K.; Khadikar, P. V. *Modeling of Anti-HIV-1 activity: Acyclouridine Derivatives Oxidation Commun* In Press
13. Todeschini, R.; Cosonni, V. *Handbook of Molecular Descriptors* Wiley-VCH: Weinheim 2000.
14. Trinajstic, N. *Chemical Graph Theory,* CRC Press: Boca Raton, Florida, 1992.
15. Karelson, M. *Molecular Descriptors in QSAR/QSPR,* Wiley: New York, 2000.
16. Basak, S .C.; Mills, D. *Commun. Math. Chem. (MATCH)* **2001,** *44,* 15.
17. Devillers, J.; Balaban, A. T. *Topological Indices and Related Descriptors in QSAR and QSPR,* Gordon & Breach: Williston, VT, 2000.
18. Wiener, H . *J. Am. Chem. Soc.* **1947,** *69,* 17.
19. Gutman, I. *Graph Theory Notes,* New York , **1994,** *27,* 9.
20. Khadikar, P. V.; Deshpande, N.V.; Kale, P. P.; Dobrynin, A.; Gutman, I.; Domotor, G .*J. Chem. Inf. Comput. Sci.* **1995,** *35,* 547.
21. Balaban ,A .T. *Chem. Phys. Lett.* **1982,** *89,* 399.
22. Randic, M . *J. Am. Chem. Soc.* **1975,** *97,* 6609.
23. Diudea, M. V. (Ed) *QSPR/QSAR Studies by Molecular Descriptors,* Babes-Bolyai University, Cluj, Romania, 2000.
24. Chaterjee, S.; Hadi, A. S.; Price, B. *Regression Analysis by Examples,* 3rd Ed., Wiley: New York, 2000.
25. Randic, M. *Croat. Chem. Acta* **1993,** *66,* 289.; *Acta.Chem.Slov*.**1998**,*45,*239.; Basak,S.C.; Balaban,A.T.; Grunwald,G.D.; Gute,B.D. *J. Chem. Inf. Comput. Sci.* **2000,** *40,* 898.
26. Pogliani, L. *Amino Acids* **1994,** 6, 141.
27. Pogliani, L. *J. Phys. Chem.* **1996,** *100,* 18065.
28. Todeschini, R. *Chemometrics Web News,* Milano Chemometric & QSAR Research Group, Feb. 2001.
29. Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure-Activity Relationship,* Wiley: New York, 1986.
30. Kier, L. B. Hall, L .H. *Molecular Structure Description,* Academic Press: New York, 1999.