



QSAR of HIV-1 Integrase Inhibitors by Genetic Function Approximation Method

Mahindra T. Makhija and Vithal M. Kulkarni*

Pharmaceutical Division, Department of Chemical Technology, University of Mumbai, Matunga, Mumbai 400 019, India

Received 10 September 2001; accepted 22 November 2001

Abstract—Quantitative structure–activity relationship (QSAR) paradigm, using genetic function approximation (GFA) technique was used to examine the correlations between the calculated physicochemical descriptors and the in vitro activities (3'-processing and 3'-strand transfer inhibition) of a series of human immunodeficiency virus type 1 (HIV-1) integrase inhibitors. Depending on the chemical structure, all molecules were divided into two classes—catechols and noncatechols. Eighty-one molecules were used in the present study and they were divided into training set and test set. The training set in each class consisted of 35 molecules and QSAR models were generated separately for both catechols and noncatechols. Equations were evaluated using internal as well as external test set predictions. Models generated for catechols show that electronic, shape related, and thermodynamic parameters are important whereas for noncatechols, spatial, structural, and thermodynamic properties play an important role for the activity. © 2002 Elsevier Science Ltd. All rights reserved.

Introduction

Acquired immunodeficiency syndrome (AIDS) is the most devastating pandemic with no holds barred. Human immunodeficiency virus (HIV) is an etiologic agent of AIDS, which expresses its effects through the genetic direction of viral polypeptides prepared by the host. Several key enzymes in the replication cycle of the HIV can be targeted for chemotherapeutic intervention, most notably, reverse transcriptase and protease.¹ HIV-1 integrase (HIV-1 IN) is another such enzyme whose inhibition may be efficacious in the treatment of AIDS, since this enzyme is required for viral replication, yet it is not indigenous to the human host.^{2,3}

HIV-1 integrase is an enzyme that mediates the integration of HIV-1 DNA into a host chromosome^{4–6} and is essential to replication of the virus.^{7,8} HIV-1 IN functions in a two-step manner by initially removing a dinucleotide unit from the 3'-ends of the viral DNA (termed 3'-processing). The 3'-processed strands are then transferred from the cytoplasm to the nucleus where they are introduced into the host DNA following 5 base-pair offset cleavages of opposing host strands (termed 3'-strand transfer or end joining). There is

obviously a requirement for a functional integrase in HIV-1 replication. This enzyme is therefore thought to be a suitable target for chemotherapeutic intervention and has become a focus of anti-AIDS drug design efforts.

Broadly, all inhibitors of HIV-1 integrase can be classified as those containing catechol substructure and those lacking it. Recently, we used 3D QSAR methods like EVA⁹ and CoMSIA with a novel type of alignment technique based on molecular electrostatic potentials.¹⁰ In order to gain insight into the physicochemical requirements of these catecholic and noncatecholic inhibitors, genetic function approximation (GFA) technique has been used which generates different QSAR models from various descriptors calculated using Cerius2 molecular modeling software. GFA was used since it generates a population of equations rather than one single equation for correlation between biological activity and physicochemical properties. GFA developed by Rogers involves combination of Friedman's multivariate adaptive regression splines (MARS) algorithm with Holland's genetic algorithm to evolve a population of equations that best fit the training set data.^{11–17} This is done as follows:

- i. An initial population of equations is generated by random choice of descriptors. The fitness of each equation is scored by lack-of-fit (LOF) measure

*Corresponding author. Tel.: +91-22-414-5616; fax: +91-22-4145614; e-mail: vithal@biogate.com

$$\text{LOF} = \text{LSE} / \{1 - (c + d^*p)/m\}^2$$

where LSE is least square error, c is the number of basis functions in the models, d is the smoothing parameter which controls the number of terms in the equation, p is the number of features contained in all terms of the models, and m is the number of compounds in the training set.

- ii. Pairs from the population of equation are chosen at random and 'crossovers' are performed and progeny equations are generated.
- iii. The fitness of each progeny equation is assessed by LOF measure.
- iv. If the fitness of the new progeny equation is better, then it is preserved.

A distinctive feature of GFA is that instead of generating a single model, as do most other statistical methods, it produces a population of models (e.g., 100). The range of variation in this population gives added information on the quality of fit and importance of descriptors. By examining these models, additional information can be discerned. For example, the frequency of use of a particular descriptor in the population of equations may indicate how relevant the descriptor is to the prediction of activity.

Results

Eighty-one molecules, depending upon their structure were divided in two classes as catechols and non-catechols. Both 3'-processing and 3'-strand transfer inhibitory activity values were considered. Molecules in each class were divided into training set and test set. The rotatable bonds in the molecules were searched using random sampling technique in order to obtain sterically accessible conformations within optimum computational time. Conformational search using random sampling was performed during molecular shape analysis (MSA) technique and the lowest energy conformers were aligned using MSA technique. Molecules in a particular class were superimposed on the lowest energy conformer of the molecule with highest biological activity in that class.

GFA technique was used for generating QSAR models for both classes with 500,000 crossovers and the smoothness value (d) of 1.0 was used during the equation generation.

The list of descriptors used in the present study is given in Table 1. The structures and activities of non catechol molecules used in the training set are given in Tables 2–7 and those for the test set are found in Table 8. For catechols, the structures and activities of training set molecules are given in Table 9 and test set molecules in Table 10. All statistically significant equations for both classes are given in Table 11. The term BA in these equations represents biological activity expressed as pIC_{50} values.

Noncatechols

This class consisted of 35 training set and six test set molecules. Various types of descriptors were calculated and QSAR equations were generated for both the activities.

3'-Processing inhibition. The sets of QSAR equations were evaluated for their predictive ability. Observation of the variable usage graph indicated that the terms PMI_Y, NCOSV, Radius of gyration and MolRef contribute more significantly than all other descriptors. The best equation from the set of equations was selected on the basis of predictivity, variables and LOF value. All three equations show more or less similar internal predictivity, but eq (3) shows better external predictivity for test set molecules. The purpose of any QSAR study is to develop a predictive model. Eq (3) is certainly better than eqs (1) and (2), because other statistical parameters do not differ much for the three equations, but r^2_{pred} increases from 0.504 to 0.628. Other statistical parameters for example conventional r^2 is around 0.7 for the three equations, r^2_{cv} is around 0.6, BS r^2 is approximately 0.7, and difference in LOF is also not much and varies from 0.289 to 0.339. All the three equations pass F-test of significance and hence all are equally significant. Thus, eq (3) on the basis of its high r^2_{pred} value (0.628) can be regarded as a model to describe the activity for this class. MolRef, Foct, NCOSV, and PMI_Y contribute to this equation and explain about 70% variance in the activity. The magnitude of coefficients in these equations describes the relative importance of a particular descriptor for explaining the biological activity for that class of compounds. These variables show low correlation among themselves indicating low probability of chance correlation. Eq (3) with better r^2_{pred} (0.628) describes the QSAR model for inhibition of 3'-processing activity of HIV-1 integrase by noncatechols. Observed versus predicted activity values are given in Tables 12 and 13.

3'-Strand transfer inhibition. The training set and the test set were same as used for 3'-processing activity. QSAR models were generated and equations were analyzed on the basis of r^2_{cv} , LOF, r^2_{pred} and variable usage graph. PMI_Y, MolRef, Apol, and H bond donor contribute significantly to these equations. All three equations show good r^2_{cv} , but only eq (4) shows good predictivity for the test set molecules. MolRef, PMI_Y and H bond donor contribute to this equation. The three variables are not intercorrelated with each other and they explain 67% variance in the activity. Eq (4) with better r^2_{pred} (0.787) is proposed as the best equation describing inhibition of 3'-strand transfer activity of HIV-1 integrase by noncatechols. Observed versus predicted activity values are given in Tables 12 and 13.

Catechols

This class comprised of 35 molecules in the training set and five molecules in the test set. Using GFA, various models were generated for both the activities and the most significant of them are given in Table 11.

Table 1. Table of all descriptors used in this study

Sr. no.	Descriptor	Type	Description
1	DIFFV	MSA	Difference volume
2	COSV	MSA	Common overlap steric volume
3	Fo	MSA	Common overlap volume ratio
4	NCOSV	MSA	Non-common overlap steric volume
5	ShapeRMS	MSA	RMS to shape reference
6	SRVol	MSA	Volume of shape reference compound
7	Vm	Spatial	Molecular volume
8	Area	Spatial	Molecular surface area
9	Density	Spatial	Molecular density
10	RadOfGyr	Spatial	Radius of gyration
11	PMI-mag	Spatial	Principal moment of inertia
12	PMI_X	Spatial	Principal moment of inertia X-component
13	PMI_Y	Spatial	Principal moment of inertia Y-component
14	PMI_Z	Spatial	Principal moment of inertia Z-component
15	Charge	Electronic	Sum of partial charges
16	Apol	Electronic	Sum of atomic polarisabilities
17	Dipole-mag	Electronic	Dipole moment
18	Dipole-X	Electronic	Dipole moment X-component
19	Dipole-Y	Electronic	Dipole moment Y-component
20	Dipole-Z	Electronic	Dipole moment Z-component
21	HOMO	Electronic	Highest occupied molecular orbital energy
22	LUMO	Electronic	Lowest unoccupied molecular orbital energy
23	Sr	Electronic	Superdelocalisability
24	MW	Structural	Molecular weight
25	RotlBonds	Structural	Number of rotatable bonds
26	HbondAcc	Structural	Number of hydrogen bond acceptors
27	HbondDon	Structural	Number of hydrogen bond donors
28	AlogP	Thermodynamic	Logarithm of partition coefficient
29	Fh ₂ o	Thermodynamic	Desolvation free energy for water
30	Foct	Thermodynamic	Desolvation free energy for octanol
31	Hf	Thermodynamic	Heat of formation
32	MolRef	Thermodynamic	Molar refractivity

Table 2. Structures and activities of coumarins used in the training set for noncatechols

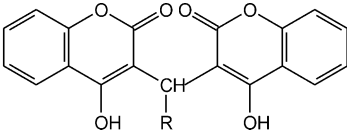
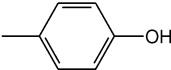
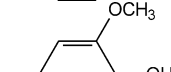
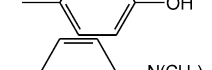
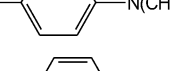
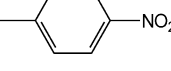
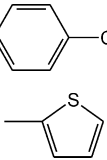
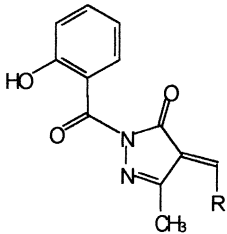
			
Sr. no.	R	3'-Processing IC ₅₀ (mM)	3'-Strand transfer IC ₅₀ (mM)
1		0.134	0.074
2		0.187	0.092
3		0.094	0.069
4		0.054	0.018
5		0.057	0.051
6		0.105	0.148

Table 3. Structures and activities of salicylhydrazines used in the training set for noncatechols


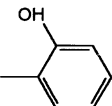
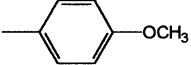
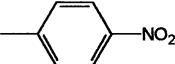
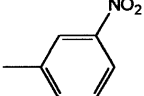
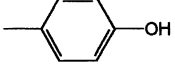
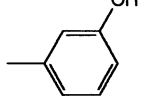
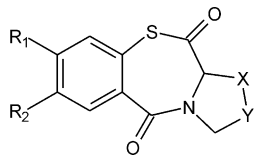

Sr. no.	R	3'-Processing IC ₅₀ (mM)	3'-Strand transfer IC ₅₀ (mM)
7		0.0006	0.0028
8		0.0009	0.0006
9		0.0008	0.0006
10		0.0014	0.0026
11		0.0006	0.0009
12		0.0009	0.0074

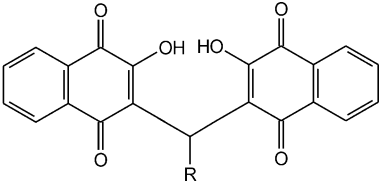
Table 4. Structures and activities of thiazolothiazepines used in the training set for noncatechols


Sr. no.	R ₁	R ₂	-X-Y-	3'-Processing IC ₅₀ (mM)	3'-Strand transfer IC ₅₀ (mM)
13	H	Cl	-S-CH ₂ -	0.128	0.090
14	H	Br	-S-CH ₂ -	0.058	0.048
15	H	CH ₃	-S-CH ₂ -	0.064	0.055
16	H	OCH ₃	-CH ₂ -S-	0.215	0.200
17	OCH ₃	OCH ₃	-CH ₂ -S-	0.650	0.331
18 ^a			-S-CH ₂ -	0.040	0.047

^aFused ring system (naphthalene derivative).

3'-Processing inhibition. Various equations generated by GFA were analyzed on the basis of LOF, r_{cv}^2 , r_{pred}^2 , and variables that were used more frequently by GFA for generating models using variable usage graph. Eq (7) shows better internal predictive ability ($r_{cv}^2=0.728$). Fh₂o, Apol, H bond donor, and Density were found to be the most frequently used descriptors for this class. Out of three, eq (7) shows better external predictivity for the test set molecules ($r_{pred}^2=0.688$). The terms Vm,

Fh₂o, and Apol contribute to this equation and explain about 78% variance in biological activity. Although eqs (8) and (9) show good r_{cv}^2 , but their ability to predict the activities of the test set is not very impressive. Eq (7) with high r_{cv}^2 and r_{pred}^2 value coupled with low LOF value is proposed as a model that best describes the inhibition of 3'-processing activity of integrase by catechol containing inhibitors. Observed versus predicted activity values are given in Tables 14 and 15.

Table 5. Structures and activities of quinones used in the training set for noncatechols


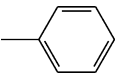
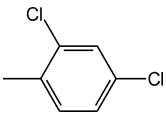
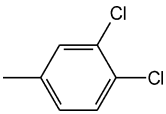
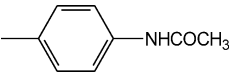
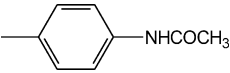
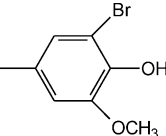
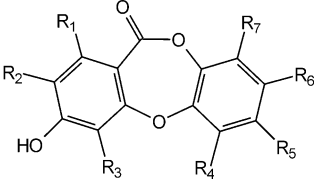
Sr. no.	R	3'-Processing IC ₅₀ (mM)	3'-Strand transfer IC ₅₀ (mM)
19		0.068	0.048
20		0.037	0.040
21		0.090	0.052
22		0.086	0.078
23		0.092	0.060
24		0.032	0.020

Table 6. Structures and activities of lichen acids used in the training set for noncatechols


No.	R ₁	R ₂	R ₃	R ₄	R ₅	R ₆	R ₇	^a IC ₅₀	^b IC ₅₀
25	Me	H	CHO	Me	CO ₂ H	OH	Me	0.0046	0.0065
26	Me	H	CHO	Me	CO ₂ Me	OH	Me	0.0054	0.0044
27	Me	H	CHO	Me	CO ₂ H	OH	CH ₂ OCOCH/CHCO ₂ H	0.0049	0.0046
28	CH ₂ COC ₅ H ₁₁	H	H	<i>n</i> -C ₅ H ₁₁	CO ₂ H	OH	H	0.0385	0.0309
29	COC ₄ H ₉	H	H	<i>n</i> -C ₅ H ₁₁	H	OH	H	0.0510	0.0336
30	Me	H	CHO	CO ₂ H	H	OMe	Me	0.0160	0.0088

^aIC₅₀ (mM) values for 3'-processing activity.^bIC₅₀ (mM) values for 3'-strand transfer activity.

Table 7. Structures and activities of hydrazides used in the training set for noncatechols

Sr. no.	R ₁	R ₂	R ₃	R ₄	3'-Processing IC ₅₀ (mM)	3'-Strand transfer IC ₅₀ (mM)
31		OH	H	H	0.00207	0.00073
32		OH	H	H	0.006	0.0052
33	-NH ₂	OH	H	H	0.080	0.038
34 ^a		OH		H	0.0023	0.0011
35		OH	H	H	0.0091	0.0058

^aFused ring system (naphthalene derivative).

3'-Strand transfer inhibition. The models that describe the activity for this class are represented by the three equations [eqs (10)–(12)] in Table 11. The frequent occurrence of Fh₂O, H bond donor, Apol and Area clearly underlines the importance of thermodynamic and electronic parameters for inhibition of 3'-strand transfer activity. Eqs (10) and (11) show high r_{cv}^2 and r_{pred}^2 value, whereas the predictive ability of eq (12) for the test set is low. Eq (10) with low LOF value and high internal and external predictive abilities describes the QSAR model for this class and the terms DIFFV, Fh₂O, and Apol explain about 80% variation in biological activity. Observed versus predicted activity values are given in Tables 14 and 15.

Discussion

Noncatechols

3'-Processing inhibition. Eq (3) describes the biological activity for this class. This equation has thermodynamic parameters—MolRef and Foct, shape parameters—NCOSV, and spatial descriptor—PMI_Y, which con-

tribute significantly to biological activity. The role of MolRef in drug–receptor interactions may be viewed as an ambivalent one. It may be that MolRef represents dispersion forces aiding the binding of an inhibitor to the enzyme. In such a case, one would expect a positive coefficient for this term. Alternatively, since MolRef is to a large degree a measure of volume, it may measure the ability of an inhibitor to induce the conformational change in an enzyme and occupy the active site in such a way as to preclude union with a natural substrate (i.e., viral and host DNA). Since the conformational change thus brought about is detrimental, a negative coefficient for MolRef should result in this case. MolRef is an approximation of molecular volume and its inverse relationship with activity indicates that the size of the active site is limited and smaller molecules would be more active than the larger ones. Foct is the 1-octanol desolvation free energy and negative correlation of this term with activity indicates that increase in the Foct value would result in decrease in activity. NCOSV is the volume of the individual molecule minus the common overlap steric volume. Negative correlation of this descriptor with biological activity again stresses the importance of size of the molecules for anti-integrase

Table 8. Structures and activities of molecules used in the test set for noncatechols

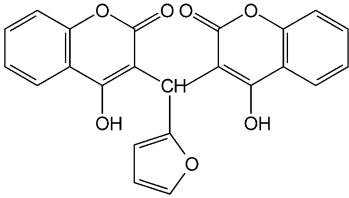
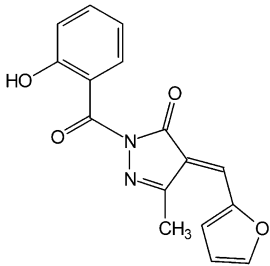
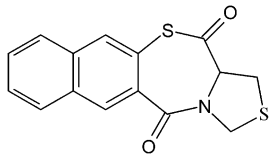
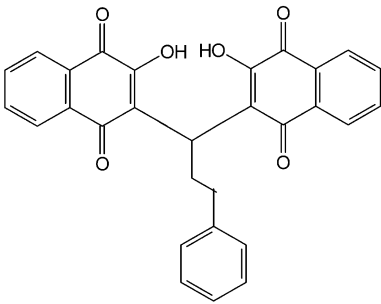
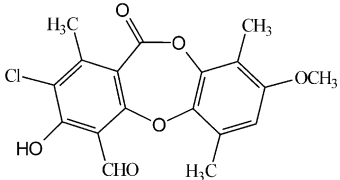
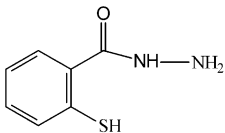
Sr. no.	Structure	3'-Processing IC ₅₀ (mM)	3'-Strand transfer IC ₅₀ (mM)
36		0.0545	0.1225
37		0.0027	0.0020
38		0.0920	0.1000
39		0.0830	0.0900
40		0.0024	0.0021
41		0.0748	0.0537

Table 9. Structures and activities of catechols used in the training set

Sr. no.	Structure	^a IC ₅₀ (mM)	^b IC ₅₀ (mM)
1		0.0011	0.0008
2		0.0384	0.0078
3		0.0248	0.0076
4		0.0004	0.0004
5		0.0275	0.0075
6		0.0275	0.0065
7		0.0098	0.0062
8		0.333	0.333
9		0.333	0.333
10		0.100	0.100
11		0.0041	0.0041
12		0.0042	0.0017
13		0.0033	0.0017
14		0.0109	0.0100

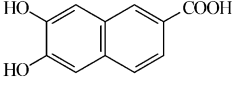
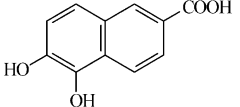
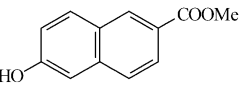
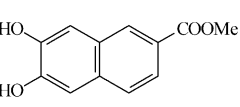
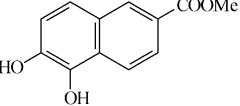
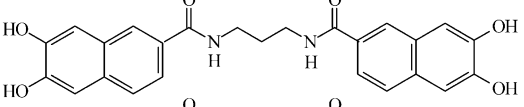
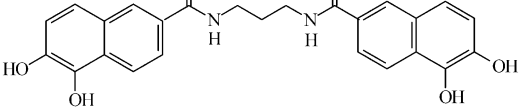
(continued on next page)

Table 9 (continued)

Sr. no.	Structure	^a IC ₅₀ (mM)	^b IC ₅₀ (mM)
15		1.000	1.000
16		1.000	1.000
17		0.333	0.333
18		0.333	0.111
19		0.300	0.300
20		0.120	0.080
21		0.300	0.300
22		0.140	0.120
23		0.150	0.140
24		0.006	0.0031
25		0.0180	0.0090
26		0.300	0.300
27		0.009	0.004
28		0.1723	0.1723

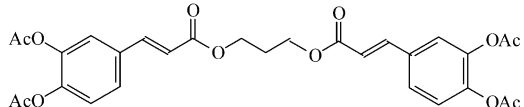
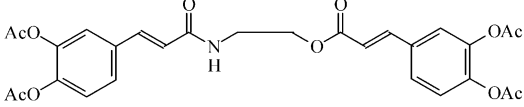
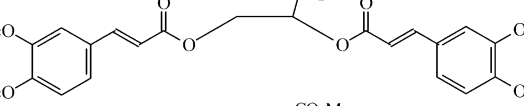
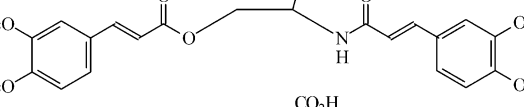
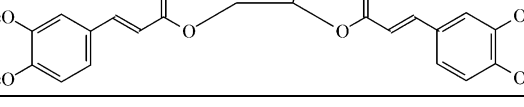
(continued on next page)

Table 9 (continued)

Sr. no.	Structure	^a IC ₅₀ (mM)	^b IC ₅₀ (mM)
29		0.0054	0.0047
30		0.0533	0.0624
31		0.200	0.200
32		0.200	0.200
33		0.0584	0.0527
34		0.00098	0.00081
35		0.00023	0.00011

^aIC₅₀ (mM) values for 3'-processing activity.^bIC₅₀ (mM) values for 3'-strand transfer activity.

Table 10. Structures and activities of molecules used in the test set for catechols

Sr. no.	Structure	^a IC ₅₀ (mM)	^b IC ₅₀ (mM)
36		0.333	0.333
37		0.333	0.100
38		0.010	0.010
39		0.0025	0.0032
40		0.0021	0.0028

^aIC₅₀ (mM) values for 3'-processing activity.^bIC₅₀ (mM) values for 3'-strand transfer activity.

Table 11. QSAR equations generated using genetic function approximation for the training set of the molecules in each class

No.	Equation	LOF	r^2	$^a r_{cv}^2$	$^b BSr^2$	F-test	r_{pred}^2
Noncatechols							
A. 3'-Processing inhibition							
1	BA = 1.22486 + 0.001997 (PMI_Y) + 0.216583 (H bond Don) – 0.007969 (NCOSV)	0.289	0.719	0.634	0.720	26.455	0.507
2	BA = 2.44492 + 0.003121 (PMI_Y) + 0.288741 (H bond Don) + 0.291626 (AlogP) – 0.011581 (Area)	0.315	0.734	0.620	0.735	20.733	0.504
3	BA = 2.14485 – 0.020682 (MolRef) – 0.022272 (Foct) – 0.005841 (NCOSV) + 0.002704 (PMI_Y)	0.339	0.714	0.616	0.715	18.748	0.628
B. 3'-Strand transfer Inhibition							
4	BA = 2.44583 – 0.034371 (MolRef) + 0.003125 (PMI_Y) + 0.214551 (H bond Don)	0.272	0.673	0.544	0.674	21.241	0.787
5	BA = 2.49298 + 0.003072 (PMI_Y) – 0.008991 (MW) + 0.202332 (H bond Don)	0.269	0.677	0.562	0.678	21.621	0.670
6	BA = 0.615708 + 0.261683 (H bond Don) – 0.003914 (PMI_X) + 0.001556 (PMI-mag)	0.265	0.681	0.592	0.682	22.071	0.440
Catechols							
A. 3'-Processing inhibition							
7	BA = –0.931143 – 0.045436 (Vm) – 0.040505 (Fh ₂ o) + 0.001071 (Apol)	0.292	0.780	0.728	0.781	36.669	0.688
8	BA = –1.32243 + 0.271498 (H bond Don) + 0.000999 (Apol) – 0.040244 (Vm) + 0.002789 (COSV)	0.358	0.766	0.678	0.767	24.566	0.554
9	BA = –15.8228 – 0.041533 (DIFFV) + 0.282594 (H bond Don) + 0.001033 (Apol)	0.327	0.754	0.699	0.754	31.633	0.534
B. 3'-Strand transfer inhibition							
10	BA = –16.4248 – 0.043838 (DIFFV) – 0.051796 (Fh ₂ o) + 0.00102 (Apol)	0.255	0.837	0.796	0.837	52.962	0.736
11	BA = –0.74026 – 0.046912 (Fh ₂ o) – 0.025759 (Area) + 0.000765 (Apol)	0.270	0.827	0.781	0.827	49.319	0.702
12	BA = –1.19031 + 0.362177 (H bond Don) + 0.00095 (Apol) – 0.038007 (Vm) + 0.001169 (COSV)	0.346	0.808	0.749	0.808	31.511	0.533

BA = biological activity expressed as pIC₅₀ (i.e., –log IC₅₀).^aLeave one out (LOO) cross-validated r^2 .^bBootstrapped r^2 .

activity. PMI_Y, a spatial descriptor shows a positive correlation with the activity. Principal moment of inertia describes the orientation of molecules reflecting that the conformation of molecules is important for activity. Thus, for all the molecules, similar orientation of the important pharmacophoric groups is essential for activity.

3'-Strand transfer inhibition. One important observation is the occurrence of H bond Don as common descriptor in all the significant equations [eqs (4)–(6)] describing activity for this class. Most of the HIV-1 IN inhibitors contain groups like OH, COOH, SO₃H and are known to act by forming H bonds with the acceptor residues in the enzymes active site. Positive correlation of this term with activity indicates more the number of H bond donor groups in the molecule, more active it would be. Since PMI_Y shows positive correlation with activity, it indicates that the proper spatial orientation of H bond donor groups is important for them to interact with the acceptor groups in the active site of an enzyme. MolRef being a measure of molecular volume dictates the size requirements for the molecules due to the limited size of active site and increase in the size of molecules would be detrimental to the activity. Also, in eq (5) there is a negative correlation of activity with the MW, which again emphasizes the importance of molecular size. Eq

(4) on the basis of its high r_{pred}^2 value can be considered as model to describe the activity for this class.

Catechols

3'-Processing inhibition. Three descriptors namely Vm, Fh₂o, and Apol, significantly explain the variance in the biological activity for this class. Apol is an electronic descriptor and is present in all the equations for this class. It is a sum of atomic polarisabilities and is proportional to the number of valence electrons in a molecule as well as on how tightly these valence electrons are bound to their nuclei. Most of the active molecules in this class have two or more free hydroxyl groups attached on adjacent carbon atoms. These molecules act by forming a complex with Mg²⁺ in the active site of HIV-1 IN.¹⁸ Since complexation is a phenomenon that involves valence electrons, there is a positive correlation between Apol and HIV-1 IN inhibitory activity. As mentioned earlier, the active site of IN is small and hence there is a negative correlation with Vm, which is a spatial descriptor. Fh₂o is a thermodynamic parameter, which represents aqueous desolvation free energy of the molecule. Since most of the integrase inhibitors are polar molecules containing groups like OH, COOH, SO₃H, CONH and so on, and act by forming H bonds

Table 12. Observed versus predicted pIC₅₀ values for noncatechols training set

Sr. no.	3'-Processing		3'-Strand transfer	
	Observed pIC ₅₀	Predicted pIC ₅₀ ^a	Observed pIC ₅₀	Predicted pIC ₅₀ ^b
1	0.871	1.791	1.127	1.646
2	0.726	1.662	1.036	1.517
3	1.026	1.194	1.161	1.042
4	1.267	1.185	1.744	1.308
5	1.244	1.567	1.288	1.547
6	0.978	1.700	0.828	1.545
7	3.221	2.589	2.552	2.177
8	3.045	3.232	3.221	2.702
9	3.096	3.270	3.221	3.103
10	2.853	2.909	2.585	2.772
11	3.221	3.179	3.045	2.583
12	3.045	2.728	2.130	2.354
13	0.892	1.068	1.045	1.202
14	1.236	1.439	1.318	1.617
15	1.193	0.798	1.259	0.929
16	0.667	0.937	0.698	1.052
17	0.187	1.088	0.479	1.222
18	1.397	1.174	1.327	1.086
19	1.167	1.007	1.318	1.227
20	1.431	1.059	1.397	1.065
21	1.045	0.805	1.283	1.025
22	1.065	0.744	1.107	1.226
23	1.036	1.354	1.221	1.583
24	1.494	1.195	1.698	1.499
25	2.337	1.717	2.187	1.927
26	2.267	1.591	2.356	1.822
27	2.309	2.427	2.337	3.170
28	1.414	1.334	1.510	1.769
29	1.292	1.434	1.473	1.455
30	1.794	1.438	2.055	1.494
31	2.684	2.047	3.136	2.174
32	2.173	1.934	2.283	2.329
33	1.096	1.913	1.420	2.238
34	2.638	2.188	2.958	2.597
35	2.040	1.753	2.236	2.031

^aResults from eq (3) for 3'-processing activity.^bResults from eq (4) for 3'-strand transfer activity.**Table 13.** Observed versus predicted pIC₅₀ values for noncatechols test set

Sr. no.	3'-Processing		3'-Strand transfer	
	Observed pIC ₅₀	Predicted pIC ₅₀ ^a	Observed pIC ₅₀	Predicted pIC ₅₀ ^b
36	1.263	1.945	0.911	0.732
37	2.568	1.961	2.698	1.948
38	1.036	1.156	1.000	1.087
39	1.080	0.681	1.045	1.415
40	2.610	2.326	2.667	2.586
41	1.126	1.120	1.269	1.024

^aResults from eq (3) for 3'-processing.^bResults from eq (4) for 3'-strand transfer.

and steric and electrostatic interactions, there is hardly any contribution from desolvation entropy during binding. Hence, there is a negative correlation with this term.

3'-Strand transfer inhibition. The requirements of this class for 3'-strand transfer inhibition are almost similar as that for 3'-processing. This is due to the fact that catechols are nonspecific inhibitors that act by chelation of divalent metal ion (Mg²⁺). Again here as chelation is the main mechanism of action, which is an electronic

phenomenon, Apol is found to have positive correlation with activity. Similarly, due to the polar nature of these compounds, there is negative correlation with aqueous desolvation energy. DIFFV is an MSA parameter, which is defined as the difference between the volume of the individual molecule and the volume of the shape reference compound. Negative correlation of this shape descriptor indicates that there is a strict requirement in terms of molecular similarity with the reference molecule for volume to show HIV-1 IN inhibitory activity.

Table 14. Observed versus predicted pIC₅₀ values for catechols training set

Sr. no.	3'-Processing		3'-Strand transfer	
	Observed pIC ₅₀	Predicted pIC ₅₀ ^a	Observed pIC ₅₀	Predicted pIC ₅₀ ^b
1	2.958	3.068	3.096	3.494
2	1.415	1.641	2.107	1.929
3	1.605	1.546	2.119	1.886
4	3.397	2.130	3.397	2.393
5	1.560	1.874	2.124	2.153
6	1.560	2.151	2.187	2.490
7	2.008	2.053	2.207	2.187
8	0.477	0.626	0.477	0.672
9	0.477	0.574	0.477	0.585
10	1.000	1.112	1.000	1.156
11	2.387	1.638	2.387	1.768
12	2.376	2.588	2.769	2.918
13	2.481	2.653	2.769	3.016
14	1.962	1.265	2.000	1.321
15	0.000	0.433	0.000	0.569
16	0.000	0.827	0.000	0.951
17	0.477	-0.09	0.477	-0.01
18	0.477	0.816	0.954	1.037
19	0.522	0.620	0.522	0.442
20	0.920	1.431	1.096	1.454
21	0.522	0.492	0.522	0.392
22	0.853	1.354	0.920	1.399
23	0.823	1.293	0.853	1.384
24	2.221	2.176	2.508	2.415
25	1.744	1.739	2.045	1.881
26	0.522	0.018	0.522	0.136
27	2.045	2.182	2.397	2.505
28	0.763	0.899	0.763	0.960
29	2.267	1.285	2.327	1.435
30	1.273	1.284	1.204	1.437
31	0.698	0.592	0.698	0.638
32	0.698	0.982	0.698	1.119
33	1.233	0.971	1.278	1.123
34	3.008	3.148	3.091	3.409
35	3.638	3.154	3.958	3.462

^aResults from eq (7) for 3'-processing activity.^bResults from eq (10) for 3'-strand transfer activity.**Table 15.** Observed versus predicted pIC₅₀ values for catechols test set

Sr. no.	3'-Processing		3'-Strand transfer	
	Observed pIC ₅₀	Predicted pIC ₅₀ ^a	Observed pIC ₅₀	Predicted pIC ₅₀ ^b
36	0.477	0.920	0.477	0.933
37	0.477	1.166	1.000	1.209
38	1.995	1.162	1.995	1.209
39	2.602	2.215	2.494	2.281
40	2.677	2.545	2.552	2.627

^aResults from eq (7) for 3'-processing activity.^bResults from eq (10) for 3'-strand transfer activity.

Recently, we performed docking studies of these inhibitors using crystal structure of HIV-1 integrase (1QS4.pdb). We found the importance of hydrogen bonding, hydrophobic interactions, and electrostatic interactions in the inhibition of HIV-1 integrase.¹⁹ Also, docking has been performed by two other groups, that is Neamati et al.²⁰ and McCammon et al.²¹ They also found the importance of hydrogen bonding, hydrophobic interactions, and electrostatic interactions for HIV-1 integrase inhibition. The importance of hydrogen bonding and van der Waals interactions has also been described by Goldgur et al.²² who have determined the crystal structure of HIV-1 integrase in complex with the

inhibitor. These interactions are clearly described by the descriptors that we have obtained in our QSAR study as shown in Table 11. Thus, our QSAR results are consistent with the crystal structure of HIV-1 integrase.

Conclusion

QSAR analysis of 81 molecules for their activity against HIV-1 integrase was performed using various physico-chemical descriptors and GFA technique. The molecules were divided into two classes—catechols and noncatechols depending on their chemical structure.

Separate analysis was performed for both the classes and QSAR equations were generated. These equations were then analyzed for their statistical significance and test set predictions. The results indicate that for noncatechols spatial, thermodynamic and shape descriptors appear to be contributing significantly to 3'-processing inhibition. For 3'-strand transfer inhibition, apart from spatial and thermodynamic descriptors, structural parameters such as number of H bond donor groups are also important. Molecules containing catechol substructure act mainly through chelation of divalent metal ions. Hence, apart from electronic descriptor (Apol), shape related physicochemical properties based on MSA and thermodynamic descriptors describing desolvation phenomenon during binding contribute significantly toward inhibition of both 3'-processing and 3'-strand transfer activities. On the basis of the physicochemical descriptors thus obtained from the QSAR analyses, novel molecules can be designed that are predicted to possess improved HIV-1 IN inhibitory activity.

Experimental

Molecules

Eighty-one molecules belonging to different chemical classes were used in the study. These were divided as catechols and noncatechols. Catechols comprise mainly chicoric acid derivatives^{23–25} (total 40) and noncatechol containing compounds used were salicylhydrazines,²⁶ lichen acids,²⁷ coumarins,²⁸ quinones,²⁹ hydrazides,³⁰ thiazolothiazepines¹⁸ (total 41) that inhibit integrase function at low micromolar concentrations. These compounds have been reported and tested by the same group (Neamati and Pommier, NCI, NIH, USA). The method of activity testing for all these compounds and the conditions for testing are all exactly similar. Hence, they can be combined in the same study. The molecules in both classes were then divided into training set and test set. Both activities, that is 3'-processing and 3'-strand transfer were used in this study. The structures of different molecules belonging to various chemical classes along with their biological activities are given in Tables 2–10.

Biological activity

All biological activities used in the present study were expressed as:

$$pIC_{50} = -\log_{10} IC_{50}$$

where IC_{50} is the millimolar concentration of the inhibitor producing 50% inhibition.

Molecular modeling

All molecular modeling studies were carried out using Cerius2 (version 3.5) running on Silicon Graphics O2 R5000 workstation.³¹ All the molecules were constructed and partial charges were assigned using the charge equilibration method within Cerius2.³² The

molecules were subsequently minimized until root mean square deviation 0.01 kcal/mol Å was achieved and used in the study.

Calculation of descriptors. Different types of descriptors were calculated for each molecule in the study table using default settings within Cerius2. These descriptors included electronic, spatial, structural, thermodynamic, and molecular shape analysis (MSA) descriptors. A complete list of descriptors used in the study is given in Table 1.

MSA descriptors³³

MSA descriptors were calculated using MSA module within Cerius2. Conformational analysis on all the molecules was performed using random sampling search with maximum number of conformers set equal to 100. Lowest energy conformer of the molecule with highest biological activity was used as reference for calculation of MSA descriptors. Compounds **11** and **9** (Table 3) were used as reference molecules for 3'-processing and 3'-strand transfer for noncatechols and compound **35** (Table 9) was used as reference molecule for catechol series.

Generation of QSAR models

QSAR analysis is an area of computational research, which builds models of biological activity using physicochemical properties of a series of compounds. The underlying assumption is that the variations of biological activity within a series can be correlated with changes in measured or computed molecular features of the molecules. In the present study, QSAR model generation was performed by GFA technique. Application of the GFA algorithm allows the construction of higher-quality predictive models and makes available additional information not provided by standard regression techniques, even for data sets with many features. GFA was performed using 500,000 crossovers, smoothness value of 1.00 and other default settings for each chemical class. GFA was asked to consider predetermined number of terms in the equation depending upon the number of molecules in the training set. The set of equations generated for each class was evaluated on the following basis:

- LOF measure
- Variable terms in the equation
- Internal as well as external predictivity of the equation.

Predictive r^2 value. The predictive r^2 was based only on molecules not included in the training set and is defined as:

$$r^2_{pred} = (SD - PRESS)/SD$$

where SD is the sum of the squared deviations between the biological activity of molecules in the test set and the mean biological activity of the training set molecules and PRESS is the sum of the squared deviations

between predicted and actual activity values for every molecule in the test set. Like r_{cv}^2 , the predictive r^2 can assume a negative value reflecting a complete lack of predictive ability of the training set for the molecules included in the test set.^{34,35}

Acknowledgements

The authors gratefully acknowledge support for this research from the University Grants Commission (UGC), New Delhi, under its DSA and COSIST programmes. Authors thank IPCA Laboratories for partial financial support. M.M. thanks UGC for the award of a senior research fellowship.

References and Notes

1. Hariprasad, V.; Talele, T. T.; Kulkarni, V. M. *Pharm. Pharmacol. Commun.* **1998**, *4*, 365.
2. Varmus, H. and Brown, P. Retroviruses. In *Mobile DNA*; Berg, D., Howe, M., Eds.; American Society of Microbiology: Washington, DC, 1989; p 53.
3. Katz, R. A.; Skalka, A. M. *Annu. Rev. Biochem.* **1994**, *63*, 137.
4. Goff, S. P. *Annu. Rev. Genet.* **1992**, *26*, 527.
5. Vink, C.; Plasterk, R. H. A. *Trends Genet.* **1993**, *9*, 433.
6. Craigie, R. *Trends Genet.* **1992**, *8*, 187.
7. LaFemina, R. L.; Schneider, C. L.; Robbins, H. L.; Callahan, P. L.; LeGrow, K.; Roth, E.; Schleif, W. A.; Emini, E. A. *J. Virol.* **1992**, *66*, 7414.
8. Sakai, H.; Kawamura, M.; Sakuragi, J.; Shibata, R.; Ishimoto, A.; Ono, N.; Veda, S.; Adachi, A. *J. Virol.* **1993**, *67*, 1169.
9. Makhija, M. T. and Kulkarni, V. M. *J. Chem. Inf. Comput. Sci.* **2001** (published online in 5th October issue).
10. Makhija, M. T. and Kulkarni, V. M. *J. Comput.-Aided Mol. Des.* **2002**. In press.
11. Rogers, D.; Hopfinger, A. J. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854.
12. Shi, L. M.; Fan, Y.; Myers, T. G.; O'Connor, P. M.; Paull, K. D.; Friend, S. H.; Weinstein, J. N. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 189.
13. Venkatarangan, P.; Hopfinger, A. J. *J. Med. Chem.* **1999**, *42*, 2169.
14. Tokarski, J. S.; Hopfinger, A. J. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 779.
15. Tokarski, J. S.; Hopfinger, A. J. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 792.
16. Hahn, M.; Rogers, D. *J. Med. Chem.* **1995**, *38*, 2091.
17. Gokhale, V. M.; Kulkarni, V. M. *Bioorg. Med. Chem.* **2000**, *8*, 2487.
18. Neamati, N.; Turpin, J. A.; Winslow, H. E.; Christensen, J. L.; Williamson, K.; Orr, A.; Rice, W. G.; Pommier, Y.; Garofalo, A.; Brizzi, A.; Campiani, G.; Fiorini, I.; Nacci, V. *J. Med. Chem.* **1999**, *42*, 3334.
19. Makhija, M. T. and Kulkarni, V. M. *J. Comput.-Aided Mol. Des.* **2001**.
20. Chen, J. I.; Neamati, N.; Nicklaus, M.; Orr, A.; Anderson, L.; Barchi, J. J.; Kelley, J. Y.; Pommier, Y. and MacKerell, A. D. *Bioorg. Med. Chem.* **2000**, *8*, 2385.
21. Sottriffer, C. A.; McCammon, J. A. *J. Med. Chem.* **2000**, *43*, 4109.
22. Goldgur, Y.; Craigie, R.; Cohen, G. H.; Fujiwara, T.; Yoshinaga, T.; Fujishita, T.; Sugimoto, H.; Endo, T.; Murai, H.; Davies, D. R. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 13040.
23. Lin, Z.; Neamati, N.; Zhao, H.; Kiryu, Y.; Turpin, J. A.; Aberham, C.; Strebel, K.; Kohn, K.; Witvrouw, M.; Pannecouque, C.; Debyser, Z.; Clercq, D. E.; Rice, W. G.; Pommier, Y.; Burke, T. R., Jr. *J. Med. Chem.* **1999**, *42*, 1401.
24. Mazumder, A.; Neamati, N.; Sunder, S.; Schulz, J.; Pertz, H.; Eich, E.; Pommier, Y. *J. Med. Chem.* **1997**, *40*, 3057.
25. Zhao, H.; Neamati, N.; Mazumder, A.; Sunder, S.; Pommier, Y.; Burke, T. R., Jr. *J. Med. Chem.* **1997**, *40*, 1186.
26. Neamati, N.; Hong, H.; Owen, J. M.; Sunder, S.; Winslow, H. E.; Christensen, J. L.; Zhao, H.; Burke, T. R., Jr.; Milne, G. W. A.; Pommier, Y. *J. Med. Chem.* **1998**, *41*, 3202.
27. Neamati, N.; Hong, H.; Mazumder, A.; Wang, S.; Sunder, S.; Nicklaus, M. C.; Milne, G. W. A.; Proksa, B.; Pommier, Y. *J. Med. Chem.* **1997**, *40*, 942.
28. Zhao, H.; Neamati, N.; Hong, H.; Mazumder, A.; Wang, S.; Sunder, S.; Milne, G. W. A.; Pommier, Y.; Burke, T. R., Jr. *J. Med. Chem.* **1997**, *40*, 242.
29. Mazumder, A.; Wang, S.; Neamati, N.; Nicklaus, M.; Sunder, S.; Chen, J.; Milne, G. W. A.; Rice, W. G.; Burke, T. R., Jr.; Pommier, Y. *J. Med. Chem.* **1996**, *39*, 2472.
30. Zhao, H.; Neamati, N.; Sunder, S.; Hong, H.; Wang, S.; Milne, G. W. A.; Pommier, Y.; Burke, T. R., Jr. *J. Med. Chem.* **1997**, *40*, 937.
31. Cerius2 version 3.5 is available from Molecular Simulations Inc., 9685, Scranton Road, San Diego, CA 92121, USA.
32. Rappe, A. K.; Goddard, W. A. *J. Phys. Chem.* **1991**, *95*, 3358.
33. Hopfinger, A. J. *J. Am. Chem. Soc.* **1980**, *102*, 7196.
34. Waller, C. L.; Oprea, T. L.; Giolliti, A.; Marshall, G. R. *J. Med. Chem.* **1993**, *36*, 4152.
35. Cramer, R. D. III; Bunce, J. D.; Patterson, D. E. *Quant. Struct. Act. Relat.* **1988**, *7*, 18.