



Pergamon

Bioorganic & Medicinal Chemistry 10 (2002) 4169–4183

BIOORGANIC &
MEDICINAL
CHEMISTRY

QSAR Studies of HIV-1 Integrase Inhibition

Hongbin Yuan and Abby L. Parrill*

Department of Chemistry, University of Memphis, Memphis, TN 38152, USA

Received 16 August 2001; accepted 8 March 2002

Abstract—Compounds from a wide variety of structural classes inhibit HIV-1 integrase. However, a single unified understanding of the relationship between the structures and activities of these compounds still eludes researchers. We report herein the development of QSAR models for integrase inhibition. The genetic function approximation (GFA) was utilized to select descriptors for the development of the QSAR models. The best QSAR model derived for the complete set of 11 structural classes had a correlation coefficient (r^2) of only 0.54 and a cross-validated correlation coefficient (q^2) of only 0.42. This indicated that the compounds studied may differ in the exact relationship between structure and inhibition, perhaps through interactions with different subsets of amino acids in the binding pocket, or through the presence of non-overlapping binding pockets. Descriptor-based cluster analysis indicated that the 11 structural classes of integrase inhibitors studied belonged to two clusters, one consisting of five structural classes, and the other six. QSAR models for these two clusters had r^2 values of 0.79 and 0.82 and q^2 values of 0.71 and 0.74, a significant improvement over models obtained for the complete set of compounds. The two models were applied to predict the activities of compounds from the same structural classes as those used to build the models, giving r^2 values of 0.65 and 0.78. The models were also used to predict the activities of compounds shown in crystallographic or docking studies to interact near the active site metal ion. The model describing the larger cluster of structural classes was better able to reproduce the biological activities of these five structures with an average percent residual error of 7.9 compared with the 19.3% residual error for predictions from the other model. This indicated that the six structural classes comprising the larger cluster may bind near the metal ion in a fashion similar to that observed in one publicly available co-crystal structure of an inhibitor bound to HIV-1 integrase. Flexible alignment of inhibitors in the two clusters found different pharmacophores that are consistent with previously published pharmacophores developed on the basis of individual structural classes that have produced novel inhibitory compounds. Thus we expect that these two QSAR models can be used in the search for novel HIV-1 integrase inhibitors as well as to provide insight into the binding modes of such diverse chemical compounds.

© 2002 Elsevier Science Ltd. All rights reserved.

Introduction

The acquired immunodeficiency syndrome (AIDS), which is the final and most serious stage of human immunodeficiency virus (HIV) infection, renders the body susceptible to a variety of normally manageable infections, cancers, and other diseases.^{1–4} Reverse transcriptase, protease and integrase are three enzymes required in the HIV replication cycle.¹ HIV integrase (IN) is currently recognized as an attractive target against AIDS.^{2,5} It catalyzes the integration of viral DNA into host DNA in two steps: 3'-processing and strand transfer. First, integrase cleaves the last two nucleotides from each 3'-end of the linear viral DNA. The subsequent DNA strand transfer reaction involves the nucleophilic attack of these 3'-ends on host chromosomal DNA.⁶

A number of compounds have been reported recently to inhibit HIV-1 integrase in biochemical assays.^{7–17} The most potent compounds tend to contain multiple aromatic rings and aryl *ortho*-hydroxylation. It has been proposed that these inhibitors could block the reaction through inhibiting the glycerolysis, hydrolysis, and circular nucleotide formation that are involved in the 3'-processing step.^{7,11} Most compounds reported to date are not selective for IN and the practical utility of those catechol-containing inhibitors is severely reduced by cytotoxicity even though they have been found to inhibit HIV-1 integrase *in vitro*.^{5,15,18} Thus predictive models describing the relationship between structure and inhibition applicable to diverse sets of structures could be valuable in the search for novel HIV IN inhibitors.

Divalent cations in the IN active site are important in both catalysis and inhibition.^{5,18–20} However, the mechanism of their effect on inhibition is not very clear. A previous study has implied that salicylhydrazines

*Corresponding author. Tel.: +1-901-678-2638; fax: +1-901-678-3447; e-mail: aparrill@memphis.edu

inhibit HIV-1 integrase by chelating to the metal at the active site as they are active only when Mn^{2+} is used as a cofactor.¹⁵ However, thiazolothiazepines showed equal activities in the presence of Mg^{2+} or Mn^{2+} , thus indicating that they differ from salicylhydrazines and perhaps act at a different site on HIV-1 integrase.¹³ For those inhibitors that may interact with both the IN molecule and Mg^{2+} or Mn^{2+} , several types of metal–inhibitor interactions are possible. The aromatic moiety common to many inhibitors has been proposed to interact with the divalent cation in a ‘cation- π ’ type interaction.⁹ There is also a possibility of a typical charge–charge interaction between the metal ions and ionic or partial charges of the ligands.^{9,15} It has been shown that both types of interactions can co-exist in a binding site.²¹

A recent crystallographic study has shown that the inhibitor 1-(5-chloroindol-3-yl)-3-(tetrazolyl)-1,3-propanedione enol (5CITEP) binds in the middle of the active site of the enzyme, lying between the three catalytic acidic residues, Asp64, Asp152 and Glu152, in the vicinity of the active site metal ion.²² This structure supports the speculation that the interactions between the inhibitors and integrase mimic the normal interactions with viral DNA substrate during the 3'-processing reaction. Additionally, a structure of the avian sarcoma virus integrase core domain in complex with 4-acetyl-amino-5-hydroxynaphthalene-2,7-disulfonic acid (Y-3), an inhibitor found to be active against the structurally homologous ASV IN and HIV-1 IN enzymes, has been studied.²³ Y-3 binds more distantly from the active site metal ion than 5CITEP on the other side of the catalytic loop. In another study, a small-molecule family consisting of a core of arsenic or phosphorus surrounded by four aromatic groups was identified to have a binding site at the dimer interface of the HIV integrase catalytic domain, which is different from the previous two sites.²⁴ These results provide support for the possibility that structurally different inhibitors interact at different sites.

QSAR modeling is a mathematical analysis, first developed by Hansch,²⁵ to elucidate a quantitative correlation between chemical structure and biological activity. The fundamental hypothesis of QSAR is that biological properties are functions of molecular structure. Molecules with similar structures can reasonably be expected to show similar biological activity and their structure–activity relationships can be explored using descriptors, numerical representations that characterize structures. A descriptor can be any quantitative property that depends on the molecular structure such as molecular weight, van der Waals surface area, dipole moment or number of hydrogen atoms.

In QSAR studies of large data sets, variable selection and model building are difficult and time-consuming procedures. Different strategies have been proposed for variable selection. Genetic algorithms (GA) are relatively new techniques for variable selection.^{26,27} They are inspired by Darwin's theory of natural selection, in which the members of a species struggle for survival and individuals having a high fitness survive to pass their

genes to the next generation. The best individuals are reproduced by crossover and random mutations. Genetic function approximation (GFA), a combination of GA and the SPLINES (multivariate adaptive regression splines algorithm) techniques, provides multiple models with high predictive ability.²⁸

In this paper, we assigned different classes of inhibitors into two clusters using cluster analysis after finding that a single predictive model could not be developed for all classes together. Two models were constructed using GFA to predict the activities of the inhibitors from each cluster. Possible pharmacophores were also identified for the two clusters. These results provided additional evidence that there are probably at least two different binding sites or binding modes for different inhibitors to interact with HIV integrase as well as defining exactly which structural classes share a common binding mode. They supplied more knowledge of the inhibitors previously studied and a route to compare the structural diversities of different sets of inhibitors, which result in different interaction between the enzyme and inhibitors and hence possibly various binding sites or modes. We anticipate that these models can be used to predict biological activities to prioritize experimental efforts in the search for novel integrase inhibitors.

Experimental Methods

In this paper, all QSAR studies were performed with the MOE²⁹ and Cerius2 programs.³⁰

Conformational search and descriptor calculation

One hundred and seventy-four inhibitors in 11 structural classes have been studied. All structures were modeled using the ionization states that would be observed in aqueous solution at a pH of 7. Once each structure was constructed, energy minimization was performed to a gradient of 0.01 kcal/mol Å using the MMFF94³¹ force field in the MOE program. These energy-minimized structures were recorded as starting points for molecular modeling studies. Conformational searches were conducted by using the Random Incremental Pulse Search (RIPS) method³² implemented in the MOE program. The lowest energy conformation of each molecule was saved for subsequent study.

In order to build QSAR models it is necessary to construct numerical representations, or descriptors, of molecules. These descriptors can be classified into three groups.³³ First, topological descriptors are derived solely from connectivity and composition of the structure: examples include Kier's molecular shape index and the Wiener index. Second, geometrical descriptors are derived from the 3-D molecular geometry: examples include molecular volume and the solvent-accessible surface area. Finally, electronic descriptors reflect the electronic structure of the molecule and overall characteristics of the partial charge distribution: examples include the dipole moment and the sum of partially positive surface area.

The MOE program was used to calculate over 180 descriptors for each conformation. All of these descriptors were imported into the Cerius2 program for cluster analysis and variable selection by GFA.

Cluster analysis

The purpose of cluster analysis is to partition a data set into classes or categories consisting of similar elements. In QSAR modeling, it can be used in two ways to study structural similarity based on descriptors. First, inhibitors in different classes were joined to form a training set. The same numbers of inhibitors in each class were selected from different clusters obtained by hierarchical cluster analysis (HCA) in the Cerius2 program to maintain the same weight for each class while maximizing structural diversity in the training set for modeling. Second, the inhibitors in different classes are structurally diverse. It was helpful to compare the molecular dissimilarity with Euclidean distances between these classes to align them into the same or different groups. To do that, one structure was selected to represent each class with HCA. Then all of these representative structures were clustered into different subgroups, which were modeled separately.

GFA and model building

GFA,²⁸ a genetic algorithm based method, is a technique developed for model building. It begins with an initial population of QSAR models using randomly selected features. Least-squares regression is used to generate the coefficients. The population is evolved by building new models based on variables of two better-scored models. The worst models in the populations are replaced by new models. The average fitness of the models increases as evolution proceeds.

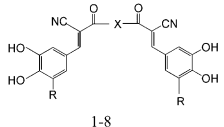
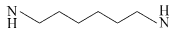
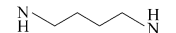
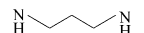
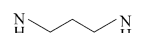
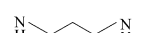

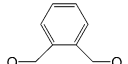
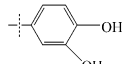
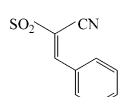
All descriptors of each inhibitor calculated in the MOE program were imported into the Cerius2 program where GFA was employed to perform QSAR modeling. The equation term type was set to linear polynomial and the mutation probability was specified as 50%. The length of the equations was set to five terms and a constant. The population size was established as 100. All equations were sorted by a statistical term, the correlation coefficient (r^2). The best equations were saved for subsequent studies, which include the comparison of predicted results of the test sets and examination of the descriptors explored.

Results and Discussion

Cluster analysis and QSAR modeling

Eleven classes of compounds with previously published experimental activities determined by the same research group have been included in our QSAR study (see Tables 1–11). Their biological activities (IC_{50} for 3'-processing) range from 0.1 μ M to greater than 300 μ M. In the QSAR study, the structural features of each inhibitor were described numerically using

Table 1. Structures and activities of tyrphostins⁷

No.	X	R	3'-Processing IC_{50} (μ M)	
			Experimental ^b	Predicted
1		H	1.9	3.11
2		H	1.35 \pm 0.6	4.61
3		OH	0.66 \pm 0.5	1.59
4 ^a		Br	0.8 \pm 0.4	0.88
5 ^a		NO ₂	3.3	1.16
6		OH	0.4 \pm 0.1	1.42
7 ^a		OH	0.45 \pm 0.1	1.74
8 ^a		H	3	4.01
9 ^a		H	1.0 \pm 0.5	18.4
10 ^a		H	4.7 \pm 1.1	1.29

^aStructures in the training set; others in the test set.

^bUse average value if there were multiple tests.

descriptors in several categories. GFA was employed to optimize the subset of these descriptors used in the model. Cluster analysis was performed in the Cerius2 program to select the same number of representative structures from each class of inhibitors for the training set in the QSAR modeling. The other compounds formed the test sets that were used for the evaluation of the models.


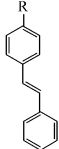
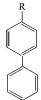
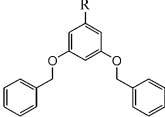
Initially, we constructed a single model to predict all of these compounds. However, neither the correlation coefficient ($r^2 = 0.542$) nor the cross-validated (leave-one-out) correlation coefficient ($q^2 = 0.418$) was satisfactory (data not shown). Previous crystallographic studies on three HIV and ASV integrase inhibitors occupying three distinct binding sites in the two enzymes indicated the possibility of more than one binding site in HIV-1 integrase.^{22–24} QSAR is based on

Table 2. Structures and activities of coumarins⁸

11-16			17-26			27-40		
No.	Structures	3'-Processing IC ₅₀ (μM)		No.	Structures	3'-Processing IC ₅₀ (μM)		
		Experimental ^b	Predicted			Experimental ^b	Predicted	
11 ^a		1.5 ± 0.5	1.30	24		7.0 ± 0.07	1.27	
12		43.4 ± 23.7	39.0	25		121, 127	86.1	
13		80.6 ± 23.0	72.9	26 ^a		35.7 ± 2.5	65.6	
14		46.0, 45.0	77.4	27 ^a		141, 128	31.0	
15		34.0, 40.0	75.2	28 ^a		100, 88	204	
16		3.0, 2.9	1.46	29		66, 48	30.5	
17 ^a		300	337	30		75, 34	28.2	
18		46.3 ± 24	100.2	31		19, 20	30.1	
19		17.2 ± 11.2	24.0	32		10, 11	35.3	
20		0.37 ± 0.10	0.18	33		19, 8.5	31.2	
21		94, 84	54.3	34 ^a		198, 177	83.6	
22 ^a		62, 116	81.8	35		58, 50	27.9	
23		4.2 ± 0.74	1.71	36		100, 110	18.2	

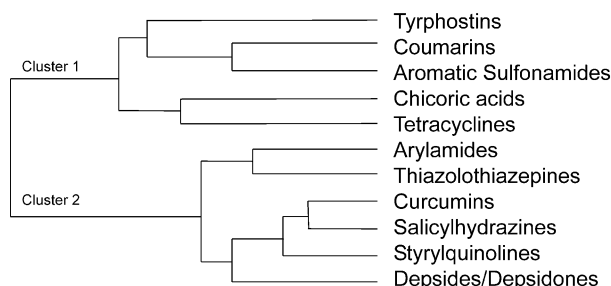
(continued on next page)

Table 2 (continued)

No.	Structures	3'-Processing IC ₅₀ (μM)		No.	Structures	3'-Processing IC ₅₀ (μM)	
		Experimental ^b	Predicted			Experimental ^b	Predicted
37		40, 29	60.1	39		5.5, 9.2	2.52
38		14, 16	22.3	40		8.0, 11	22.2

^aStructures in the training set; others in the test set.^bUse average value if there were multiple tests.

the assumption that similar structures will interact with the target at the same site. If two or more binding sites do exist, activities of these compounds should be more accurately reflected by separate QSAR models. Hence cluster analysis was used to classify inhibitors and align them into different groups according to descriptors. The result showed that these inhibitors could be divided into two clusters. Among 11 classes of inhibitors, tyrphostins (Table 1), coumarins (Table 2), aromatic sulfonamides (Table 3), chicoric acids (Table 4), and tetracyclines (Table 5) are members of the first cluster. The other six classes of inhibitors, arylamides (Table 6), thiazolo-thiazepines (Table 7), curcumins (Table 8), salicylhydrazines (Table 9), styrylquinolines (Table 10), and depsides/depsidones (Table 11) comprise the second cluster (see Fig. 1). QSAR modeling was performed for each cluster separately using GFA. As we mentioned earlier, GFA gives a set of equations. Although the correlation coefficients for the top several models are nearly equal, we only discuss the best model for each cluster in this paper. This is justifiable because the models with similar correlation coefficients (at least three to five models) have only one variable out of five different from each other. Most of these different variables in models belong to the same sub-class of descriptors and therefore reflect similar aspects of the structures. Thus the top several models are minor variants of each other.

**Figure 1.** Cluster analysis of 11 classes of inhibitors shown in Tables 1–11.

Model 1 is the equation built for the training set of the first cluster, which includes six representative structures in each of five classes. The fits of model 1 to the training set and test set of cluster 1 are shown in Figures 2 and 3 and the predicted IC₅₀ values are shown in Tables 1–5. The model relates five descriptors to inhibition. The first descriptor is a-base, the number of basic (positive) atoms, which is a pharmacophore feature descriptor. The second one is SMR-VSA6, the sum of the van der Waals surface area for atoms such that the contribution to molar refractivity is in the range of (0.485, 0.56). The third descriptor is Q-VSA-FPOL, the fractional polar van der Waals surface area. The fourth one is PEOE-VSA + 4, the sum of the van der Waals surface area for atoms having partial charges in the range (0.20, 0.25). The final one, weinerPath, is the sum of bond distances between all heavy atom pairs in the molecule, an adjacency and distance matrix descriptor.

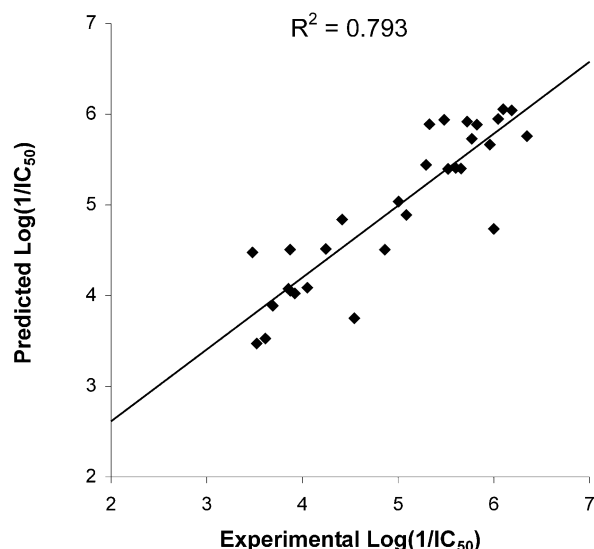
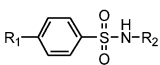
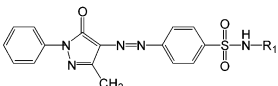
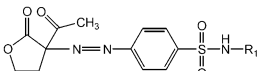
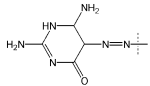

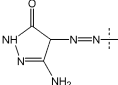


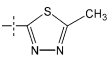
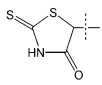

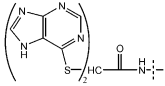
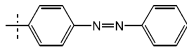


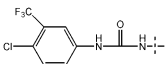

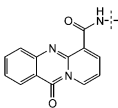

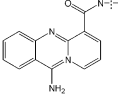


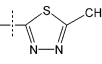



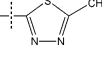
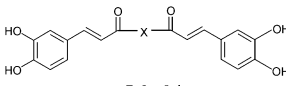
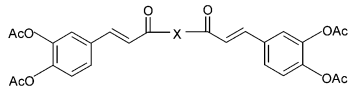
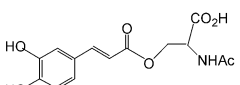
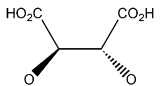
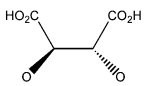
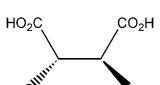
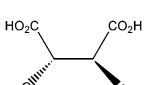
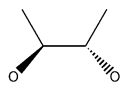
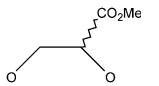
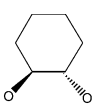
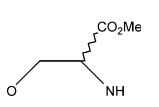
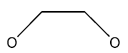
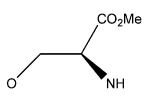

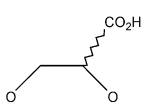
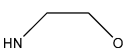
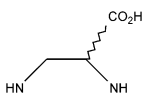
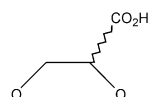
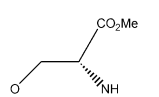
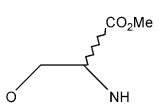
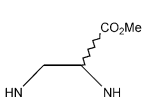
**Figure 2.** Predicted versus experimental log(1/IC₅₀) values of the training set for model 1.

Table 3. Structures and activities of aromatic sulfonamides⁹

No.	R ₁	R ₂	3'-Processing IC ₅₀ (μM)	
			Experimental ^b	Predicted
	 41-49	 50-52	 53-55	
41 ^a			120 ± 32	94.6
42			71 ± 44	131
43			24 ± 8.2	103
44 ^a			28.6 ± 11.6	195
45 ^a			140 ± 10	84.3
46 ^a			8.2 ± 2.4	12.9
47			122 ± 22	372
48			75.5 ± 17.3	364
49			48.3 ± 25.8	764
50			124, 132.7	434
51 ^a			134 ± 36	88.5
52			49.0 ± 9.5	456
53			70 ± 33	652
54			193	778
55 ^a			244	297

^aStructures in the training set; others in the test set.^bUse average value if there were multiple tests.

Table 4. Structures and activities of chicoric acids¹⁰

<div></div> 56-64				<div></div> 65-73				<div></div> 74			
No.	X	3'-Processing IC ₅₀ (μM)		No.	X	3'-Processing IC ₅₀ (μM)					
		Experimental ^b	Predicted			Experimental ^b	Predicted				
56		1.1, 2.3	2.2	65 ^a		9.8, 10	9.2				
57 ^a		1.1	2.2	66		33.3, 10	9.2				
58 ^a		38.4	14.5	67		10.0, 10.0	6.8				
59		24.8	14.5	68 ^a		2.5±0.6	3.9				
60 ^a		0.4, 0.9	0.9	69		2.8±0.8	3.9				
61		27.5, 16	1.2	70		2.1±0.3	5.9				
62		27.5	0.63	71		4.1, 9.4	2.2				
63		4.2±2.5	1.4	72		2.8, 3.0	3.9				
64		3.3±2.6	0.8	73		10.9, 12.3	2.6				
74 ^a		333	33.3								

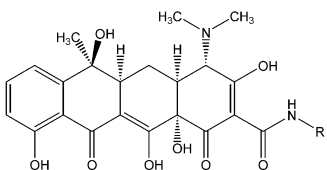
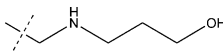
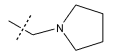
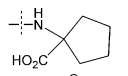
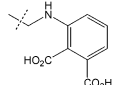
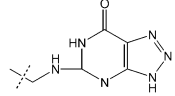
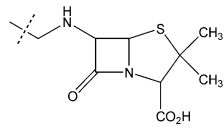
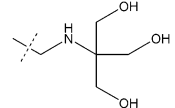
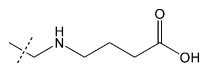
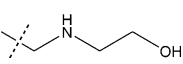
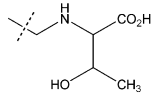
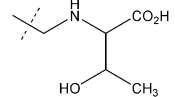
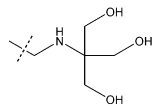
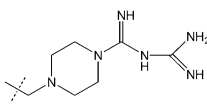
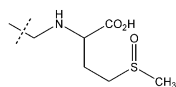
^aStructures in the training set; others in the test set.^bUse average value if there were multiple tests.Model 1: $r^2=0.793$, $q^2=0.710$

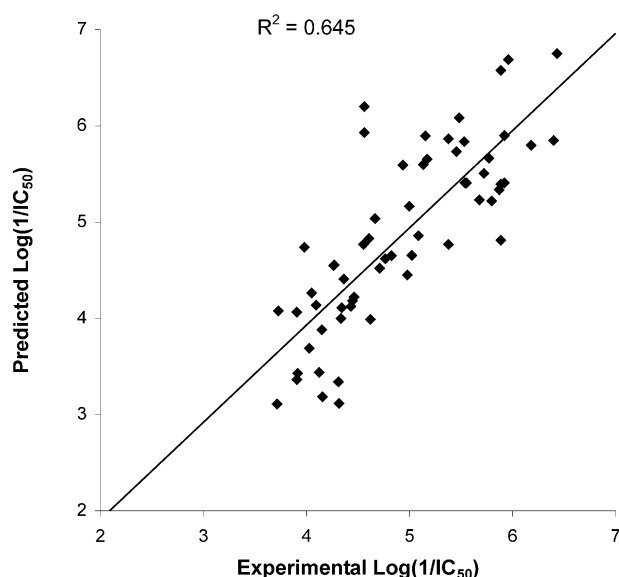
$$\begin{aligned} \text{Log}(1/\text{IC}_{50}) &= 6.0831 \\ &+ 1.3565 \times \text{'a-base'} \\ &+ 0.023037 \times \text{'SMR-VSA6'} \\ &- 6.71661 \times \text{'Q-VSA-FPOL'} \\ &+ 0.026054 \times \text{'PEOE-VSA + 4'} \\ &+ 7.8\text{e-}05 \times \text{'weinerPath'} \end{aligned}$$

Model 2 is the equation built for the training set of the second cluster, which includes five representative

structures in each of six classes. The fits of model 2 to the training set and test set of cluster 2 are shown in Figures 4 and 5 and the predicted IC₅₀ values are shown in Tables 6–11. Five different descriptors were chosen. The first descriptor is PEOE-VSA + 5, the sum of the van der Waals surface area for atoms with partial charges in the range (0.25, 0.30). The second descriptor is PmiY, the y component of the principal moment of inertia (external coordinates). The third descriptor is a-aro, the number of aromatic atoms. The fourth descriptor is a-hyd, the number of hydrophobic atoms. The fifth descriptor is E-oop, the out-of-plane potential energy.

Table 5. Structures and activities of tetracyclines¹¹

							
No.	R	3'-Processing IC ₅₀ (μM)		No.	R	3'-Processing IC ₅₀ (μM)	
		Experimental ^b	Predicted			Experimental ^b	Predicted
75 ^a	H	204.0 ± 37.4	129	83		1.2 ± 0.2	3.9
76		28.0 ± 22.5	17.0	84		1.3 ± 1.0	4.0
77 ^a		1.7 ± 1.2	1.9	85		1.6 ± 1.4	6.0
78 ^a		2.2 ± 0.9	4.0	86		1.3 ± 0.4	0.3
79 ^a		1.9 ± 0.9	1.2	87 ^a		5.1	3.6
80		3.5 ± 1.3	1.9	88		1.3 ± 0.3	15.4
81		1.1 ± 0.6	0.2	89 ^a		0.9 ± 0.6	1.1
82		1.2 ± 0.7	1.3				

^aStructures in the training set; others in the test set.^bUse average value if there were multiple tests.**Figure 3.** Predicted versus experimental $\log(1/IC_{50})$ values of the test set for model 1.Model 2: $r^2 = 0.822$, $q^2 = 0.742$

$$\begin{aligned} \text{Log}(1/IC_{50}) &= 3.35197 \\ &+ 0.034373 \times \text{'PEOE-VSA + 5'} \\ &+ 0.000164 \times \text{'pmiY'} \\ &+ 0.231308 \times \text{'a-aro'} \\ &- 0.148344 \times \text{'a-hyd'} \\ &+ 0.397343 \times \text{'E_oop'} \end{aligned}$$

Both models 1 and 2 can predict activities for inhibitors from the same cluster as shown in Figures 2–5. However, they cannot be applied to predict the activities of compounds from the other clusters. Model 2 poorly predicts the activities of compounds in cluster 1 ($r^2 = 0.02$) and model 1 unsuccessfully predicts the activities of compounds in cluster 2 ($r^2 = 0.03$). Different descriptors have been selected by GFA for models 1 and 2. The differences in the dependence of biological activity on the structural features reflected in these descriptors could arise in two ways. First, the two clusters of

Table 6. Structures and activities of arylamides and naphthalene-based compounds¹²

No.	Structures	3'-Processing IC ₅₀ (μM)	
		Experimental ^b	Predicted
90 ^a		0.98 ± 0.5	0.85
91		0.23 ± 0.05	0.19
92 ^a		172.3	41.5
93 ^a		5.4 ± 0.8	29.0
94		53.3 ± 11.9	25.2
95 ^a		58.4 ± 0.8	11.4
96 ^a		33	54.3

^aStructures in the training set; others in the test set.^bUse average value if there were multiple tests.**Table 7.** Structures and activities of thiazolothiazepines¹³

No.	R ₁	R ₂	-X-Y-	3'-Processing IC ₅₀ (μM)	
				Experimental ^b	Predicted
97 ^a	H	H	-S-CH ₂ -	110 ± 12	330
98	H	Cl	-S-CH ₂ -	151, 105	428
99	H	Br	-S-CH ₂ -	58 ± 15	316
100	H	Me	-S-CH ₂ -	64 ± 47	394
101	H	H	-CH ₂ -S-	208 ± 24	169
102 ^a	H	Cl	-CH ₂ -S-	158, 111	205
103	H	Br	-CH ₂ -S-	87 ± 24	155
104	H	Me	-CH ₂ -S-	52	12.6
105	NO ₂	H	-CH ₂ -S-	90 ± 27	50.9
106 ^a	H	OMe	-CH ₂ -S-	155, 275	137
107	OMe	OMe	-CH ₂ -S-	670, 630	117
108	H	H	-(CH ₂) ₂ -	406, 495	239
109	H	Me	-S(O)-CH ₂ -	590 ± 350	65.2
110 ^a	H	H	-CH ₂ -S(O)-	200, 185	105
111	H	Cl	-CH ₂ -S(O)-	260, 215	143
112	H	OMe	-CH ₂ -S(O)-	84.5	16.3
113			-S-CH ₂ -	40 ± 10	12.3
114 ^a			-CH ₂ -S-	92 ± 30	65.0
115	Ph		-CH ₂ -S-	372, 111	92.0
116			-CH ₂ -S-	590	46.6

^aStructures in the training set; others in the test set.^bUse average value if there were multiple tests.

inhibitors could interact at completely non-overlapping sites of the IN enzyme. Second, their binding sites could overlap, but require interaction with different sets of amino acids at that site.

The descriptors appearing in two models provide some insight into the nature of the binding sites. First of all, the positive correlation with the descriptor (a-base) in model 1 implies that the inhibitors in cluster 1 are not expected to interact with the divalent metal ion as

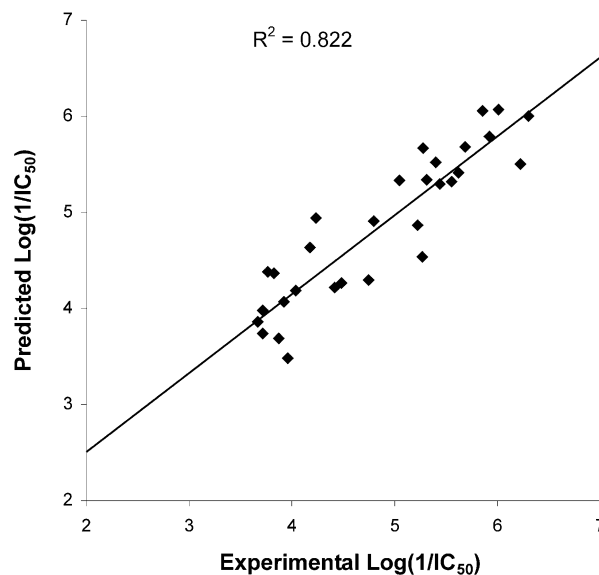
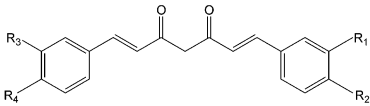
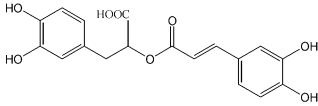
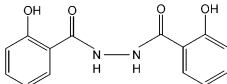
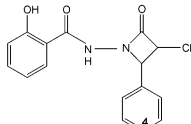
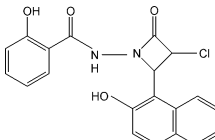
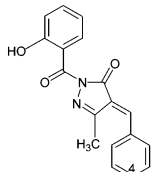
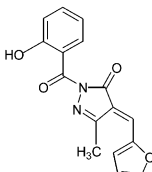
**Figure 4.** Predicted versus experimental log(1/IC₅₀) values of the training set for model 2.

Table 8. Structures and activities of curcumins¹⁴

 117-121					 122	
No.	R ₁	R ₂	R ₃	R ₄	3'-Processing IC ₅₀ (μM)	
					Experimental ^b	Predicted
117 ^a	H	OH	H	OH	120	85.1
118	H	OH	OCH ₃	OH	140	68.2
119 ^a	OCH ₃	OH	OCH ₃	OH	150	43.1
120 ^a	OH	OH	OH	OH	6.0 ± 1.5	13.6
121 ^a	OCH ₃	OH	OH	OH	18.0 ± 9.0	50.5
122 ^a					9 ± 7	4.65

^aStructures in the training set; others in the test set.^bUse average value if there were multiple tests.**Table 9.** Structures and activities of salicylhydrazines¹⁵

	123		124-131		132		133-141		142
No.	R	3'-Processing IC ₅₀ (μM)							
		Experimental ^b		Predicted					
123 ^a		2.07 ± 0.75		2.08					
124	4-OCH ₃	288		1169					
125	4-NO ₂	192, 168		120					
126	2-OH	300		154					
127	2-Cl	241, 248		573					
128	3-Cl	155, 202		209					
129 ^a	3-OH	180, 225		166					
130	3,4-(OCH ₃) ₂	125, 109		199					
131	3,4,5-(OCH ₃) ₃	68.8, 75.7		195					
132		127 ± 42		209					
133	2-OH	0.6		1.8					
134	4-OCH ₃	0.9		6.1					
135	4-NO ₂	0.8		3.2					
136 ^a	3-NO ₂	1.4		0.87					
137 ^a	4-OH	0.6		3.1					
138	3-OH	0.9		4.2					
139	3-OCH ₃ , 4-OH	0.8		1.0					
140 ^a	3,4-(OCH ₃) ₂	0.5		1.0					
141	3,4,5-(OCH ₃) ₃	2.0		0.9					
142		2.7		3.2					

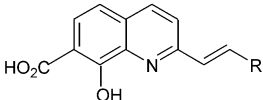
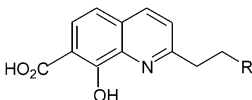

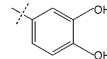

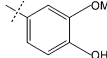

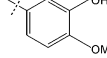

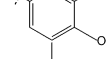

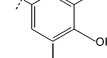

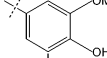

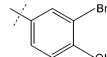

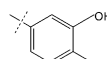

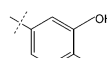

^aStructures in the training set; others in the test set.^bUse average value if there were multiple tests.

increased numbers of basic atoms would produce repulsive interactions with the positive charge. Thus inhibitors in cluster 1 may bind at a site distant from the metal in the enzyme. The negative contribution of the descriptor (a-hyd) in model 2 indicates that excessive numbers of hydrophobic atoms reduce the biological activity. The aromatic descriptor (a-aro) in the same model supplies evidence that the inhibitors might interact with the divalent cation in a 'cation-π' type interaction, indicating proximity to the metal. These

speculations are consistent with the proposition that at least two distinct binding sites exist.

The descriptors to describe the van der Waals surface area played different roles in models 1 and 2. Model 1 includes three surface area descriptors, SMR-VSA6, Q-VSA-FPOL and, PEOE-VSA + 4; model 2 has only one, PEOE-VSA + 5. While Q-VSA-FPOL in model 1 has a negative sign that suggests the polarity of structures in cluster 1 decreases their activities, SMR-VSA6,

Table 10. Structures and activities of styrylquinolines¹⁶

<div style="display: flex; justify-content: space-around; align-items: center;"> <div style="text-align: center;">  <p>143-160</p> </div> <div style="text-align: center;">  <p>161</p> </div> </div>							
No.	R	3'-Processing IC ₅₀ (μM)		No.	R	3'-Processing IC ₅₀ (μM)	
		Experimental	Predicted			Experimental	Predicted
143 ^a		5.3	2.1	153 ^a		2.4	3.9
144		1.9	4.6	154 ^a		2.8	4.8
145		3.4	5.8	155		0.9	0.8
146		4.1	2.8	156		0.3	1.1
147 ^a		1.2	1.6	157		0.7	3.7
148		3.5	4.3	158		4.9	4.0
149		1.4	5.6	159		1.3	3.1
150		1.6	4.6	160 ^a		4.0	3.0
151		3.2	2.4	161		2.3	3.8
152		3.7	3.3				

^aStructures in the training set; others in the test set.

PEOE-VSA +4 and weinerPath, representing molar refractivity, partial charges, and geometry of the molecules, respectively, have positive effects on the biological activities. In model 2, except descriptors a-aro and a-hyd discussed above, the other three descriptors, which describe the van der Waals surface area for atoms with specific partial charges, the principal moment of inertia and potential energy, give positive contribution. These structural features can provide various capabilities for inhibitors to interact with different residues in the enzyme through van der Waals and hydrogen binding interactions.

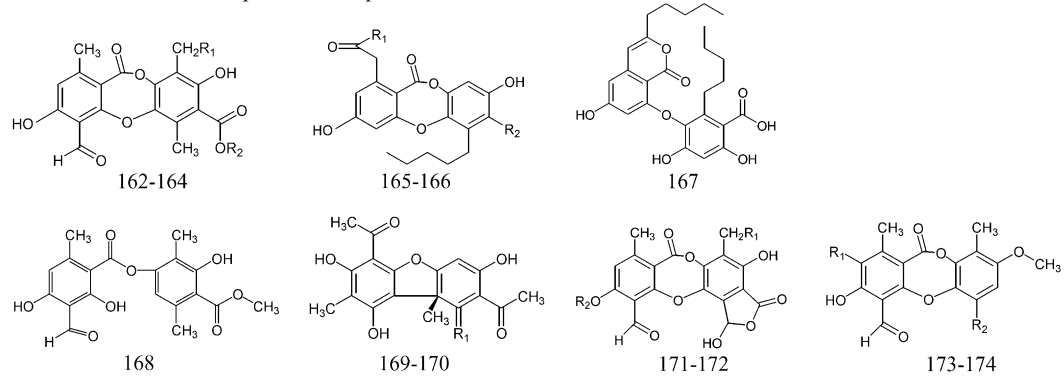
In a recently published paper, docking studies were performed for five different structures.³⁴ The results showed a favorable binding site for all structures close to the active site metal ion. Among these five structures,

L-chicoric acid (LCH) was included in the first cluster in our QSAR modeling. None of the other structures were used in developing either model. Model 1 and model 2 have been applied to predict the activities for all five structures (see Tables 12 and 13).

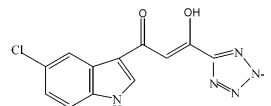
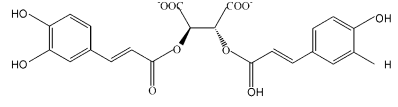
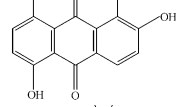
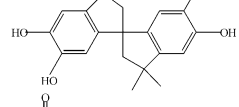
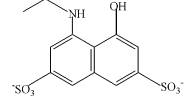
Interestingly, both models 1 and 2 can predict the 3'-IC₅₀ for 5CITEP and LCH reasonably well. For the other three structures, model 2 gave more accurate predictions. Including TMS, which has the worst predicted result, the average percent residuals using models 1 and model 2 are 19.3 and 7.9%, respectively.

Another study explored the binding mode of styrylquinolines,¹⁶ members of the second cluster in our study. The results of that study indicated the styrylquinolines possibly coordinate to the metal cation,

Table 11. Structures and activities of depsides and depsidones¹⁷

				
No.	R ₁	R ₂	3'-Processing IC ₅₀ (μM) ^b	
			Experimental	Predicted
162	H	H	4.6 ± 1.6	9.7
163	H	CH ₃	5.4 ± 1.9	3.8
164 ^a	OCOCH=CHCOOH	H	4.9 ± 2.7	4.6
165 ^a	(CH ₂) ₄ CH ₃	COOH	38.5 ± 25.7	60.5
166	(CH ₂) ₃ CH ₃	H	59.9, 42.2	96.4
167			52.2 ± 11.4	70.0
168			61.0, 62.4	7.4
169	O		126.4 ± 23.5	179
170 ^a	NNHCOC ₅ H ₄ N		73.0, 60.8	23.2
171	OH	H	19.0, 16.0	3.6
172 ^a	H	CH ₃	4.4, 2.9	5.0
173 ^a	H	COOH	17.0, 15.1	12.3
174	Cl	CH ₃	2.2, 2.7	24.6

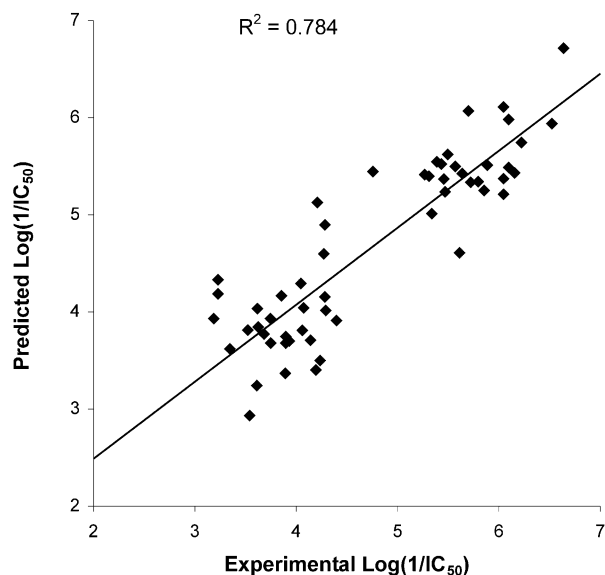
^aStructures in the training set; others in the test set.^bUse average value if there were multiple tests.**Table 12.** Comparison of experimental and predicted log (1/IC₅₀) of five test structures

Structures	Experimental value		Predicted log (1/IC ₅₀)			
	3'-Processing IC ₅₀ (μM)	Log (1/IC ₅₀)	Model 1	% Residuals	Model 2	% Residuals
5CITEP ²² 	2.20	5.66	5.84	3.2	5.97	5.5
LCH ¹⁰ 	1.10	5.96	5.66	5.0	5.97	0.2
QLZ ³⁵ 	4.00	5.40	4.19	22.4	5.05	6.5
TMS ³⁶ 	17.00	4.77	2.29	52.0	3.71	22.2
Y-3 ²³ 	16.20	4.79	4.12	14.0	5.03	5.0
Average percent residuals			19.3%		7.9%	

5CITEP, 1-(5-chloroindol-3-yl)-3-(tetrazolyl)-1,3-propanedione enol; LCH, L-chicoric acid; QLZ, quinalizarin; TMS, 3,3,3',3'-tetramethyl-1,1'-spirobis(indan)-5,5',6,6'-tetrol; Y-3, 4-acetylamino-5-hydroxynaphthalene-2,7-disulfonic acid % Residuals = 100 × Abs(predicted log(1/IC₅₀) - experimental log(1/IC₅₀)) / experimental log(1/IC₅₀).

Table 13. Model parameters and compound numbers in training sets and test sets

Model	r^2	q^2	Training set	Test set
1	0.793	0.710	30 compounds (Tables 1–5)	59 compounds (Tables 1–5)
2	0.822	0.742	30 compounds (Tables 6–11)	55 compounds (Tables 6–11)

**Figure 5.** Predicted versus experimental $\log(1/IC_{50})$ values of the test set for model 2.

which is required by HIV IN as a cofactor for its enzymatic activity. A docking study of styrylquinolines in Rous sarcoma virus (RSV) integrase also showed that the inhibitors bind closed to the crystallographic catalytic divalent cation.³⁷ These results support our conclusion that the other members in the second cluster also interact with the HIV IN at the active site near the metal cation. Model 2 reasonably predicted the five compounds that have similar presumed binding sites even though they were not included in development of model 2. LCH, a member of the first cluster, can be predicted even better using model 2. We assume this compound may bind with HIV IN in two different sites or with two different binding modes.

Flexible alignment

The two inhibitor clusters have been studied for possible pharmacophores. The most active structure in each class has been selected as a representative structure and superimposed on other members of the same clusters using MOE's flexible alignment based on several similarity

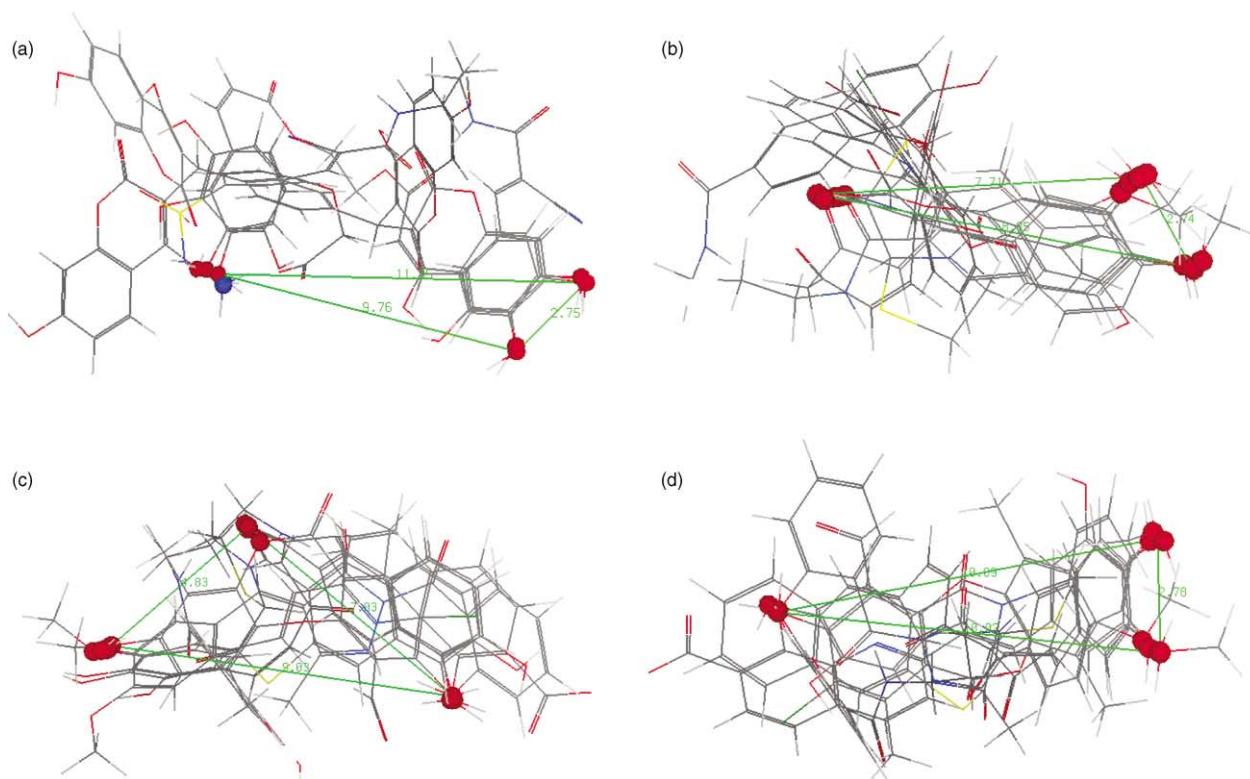


Figure 6. Possible pharmacophores for two clusters. (a) A possible pharmacophore of the first cluster: three hydrogen bond acceptors/donors positioned 11.35, 9.76, and 2.75 Å apart. (b) The first possible pharmacophore of the second cluster: three hydrogen bond acceptors/donors positioned 8.85, 7.71, and 2.74 Å apart. (c) The second possible pharmacophore of the second cluster: three hydrogen bond acceptors/donors positioned 9.03, 7.03, and 4.83 Å apart. (d) The third possible pharmacophore of the second cluster: three hydrogen bond acceptors/donors positioned 10.09, 10.02, and 2.78 Å apart.

terms including H-bond donor and acceptor, aromaticity, and partial charge. Flexible alignment uses a stochastic search procedure to superimpose similar functionality in these structures as defined by the similarity terms while allowing each structure full conformational flexibility. The RMSD tolerance was set to 0.5.

We found a possible pharmacophore for the first cluster with a distance pattern of 11.35, 9.76, and 2.75 Å separating hydrogen bond acceptors or donors (Fig. 6a). This is essentially identical to a pharmacophore (11.5 Å, 9.5 Å, 2.6 Å) constructed from the structures of dicaffeoylquinic acid and chicoric acid.¹¹ The latter compound is a member of the first cluster but could be predicted equally well by model 2. Three possible pharmacophores for the second cluster were identified (Fig. 6b–d). The first and second three-point pharmacophores with distance patterns of 8.85, 7.71, 2.74 Å and 9.03, 7.03, 4.83 Å separating oxygen atoms are close to the pharmacophores (8.73 ± 0.7 Å, 8.01 ± 0.7 Å, 2.71 ± 0.7 Å) and (8.73 ± 0.7 Å, 7.41 ± 0.7 Å, 3.96 ± 0.7 Å),¹⁷ which were derived from the structures of depsides and depsidones, members of the second cluster. The third pharmacophore derived from cluster 2 with distances of 10.09, 10.02, 2.78 Å separating hydrogen bond acceptors or donors has not previously been proposed. Thus the cluster analysis described here provides added support for the use of more than one pharmacophore in the search for novel integrase inhibitors.

Conclusion

In our QSAR modeling of HIV-1 integrase inhibition, a single model developed for all classes of inhibitors cannot adequately describe the relationship between their structures and activities. Accordingly, two clusters of inhibitors have been identified and predictive QSAR models have been developed for each cluster. This finding can be rationalized in two ways. Either the two clusters of inhibitors interact at two different sites of HIV-1 integrase or an overlapping site with different sets of amino acids. Inspection of the physical meaning of the descriptors in the context of the crystallographic structure of the enzyme catalytic core lead us to conclude that two distinct binding sites are likely. The results of cluster analysis of these inhibitors using the Cerius2 program are consistent with our previous QSAR study of inhibitors in the tyrphostin and salicylhydrazine classes,³⁸ which are members of two different clusters. These previous results showed that their activities cannot be predicted by the same model. Other researchers proposed that styrylquinolines, members of the second cluster, have a binding site near the metal.¹⁶ This is consistent with observation that model 2 more accurately predicted the activities of five compounds docked near the metal by Sottriffer et al.³⁴

QSAR models have been constructed using different descriptors for each cluster of inhibitors. These descriptors demonstrated various structural characteristics that are important to the activities for two different clusters of inhibitors. Molecular flexible alignment found different

possible pharmacophores for two clusters. They are close to those pharmacophores derived previously from the individual classes of inhibitors.^{11,17} Thus this work provides a clear reason why different pharmacophores have been successful in finding new inhibitors of integrase.

The results of our study supplied a clearer understanding of the HIV-1 integrase inhibitors and an approach to assign and compare different sets of inhibitors. QSAR models and cluster analysis can also predict the biological activities for prospective inhibitors and anticipate their possible binding sites. These models can thus be used in the continuing search for novel integrase inhibitors.

Acknowledgements

Support from NIH/NIAID (Grant R15 AI 45984-01) and NSF (STI-9602656, CHE-9708517) are gratefully acknowledged. We thank the Chemical Computing Group for their donation of the MOE program. This work was also supported by the funds from the University of Memphis.

References and Notes

- Gait, M. J.; Karn, J. *Trends. Biotechnol.* **1995**, *13*, 430.
- Robinson, W. E. J. *Infect. Med.* **1998**, *15*, 129.
- Andrake, M. D.; Skalka, A. M. *J. Biol. Chem.* **1996**, *271*, 19633.
- Nanni, R. G.; Ding, J.; Jacobo-Molina, A.; Hughes, S. H.; Arnold, E. *Perspect. Drug Disc. Des.* **1993**, *1*, 129.
- Pommier, Y.; Pilon, A.; Bajaj, K.; Mazumder, A.; Neamati, N. *Chemotherapy* **1997**, *8*, 483.
- De Clercq, E. *J. Med. Chem.* **1995**, *38*, 2491.
- Mazumder, A.; Gazit, A.; Levitzki, A.; Nicklaus, M.; Yung, J.; Kohlhagen, G.; Pommier, Y. *Biochemistry* **1995**, *35*, 15111.
- Zhao, H.; Neamati, N.; Hong, H.; Mazumder, A.; Wang, S.; Sunder, S.; Milne, G. W.; Pommier, Y.; Burke, T. R., Jr. *J. Med. Chem.* **1997**, *40*, 242.
- Nicklaus, M. C.; Neamati, N.; Hong, H.; Mazumder, A.; Sunder, S.; Chen, J.; Milne, G. W.; Pommier, Y. *J. Med. Chem.* **1997**, *40*, 920.
- Lin, Z.; Neamati, N.; Zhao, H.; Kiryu, Y.; Turpin, J. A.; Aberham, C.; Strebel, K.; Kohn, K.; Witvrouw, M.; Pannecoque, C.; Debyser, Z.; De Clercq, E.; Rice, W. G.; Pommier, Y.; Burke, T. R., Jr. *J. Med. Chem.* **1999**, *42*, 1401.
- Neamati, N.; Hong, H.; Sunder, S.; Milne, G. W. A.; Pommier, Y. *Mol. Pharmacol.* **1997**, *52*, 1041.
- Zhao, H.; Neamati, N.; Mazumder, A.; Sunder, S.; Pommier, Y.; Burke, T. R., Jr. *J. Med. Chem.* **1997**, *40*, 1186.
- 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.
- Mazumder, A.; Neamati, N.; Sunder, S.; Schulz, J.; Pertz, H.; Eich, E.; Pommier, Y. *J. Med. Chem.* **1997**, *40*, 3057.
- Neamati, N.; Hong, H.; Owen, J. M.; Sunder, S.; Winslow, H. E.; Christensen, J. L.; Zhao, H.; Burke, T. R., Jr.; Milne, G. W.; Pommier, Y. *J. Med. Chem.* **1998**, *41*, 3202.
- Zouhiri, F.; Mouscadet, J. F.; Mekouar, K.; Desmaele, D.; Savoure, D.; Leh, H.; Subra, F.; Le Bret, M.; Auclair, C.; d'Angelo, J. *J. Med. Chem.* **2000**, *43*, 1533.

17. Neamati, N.; Hong, H.; Mazumder, A.; Wang, S.; Sunder, S.; Nicklaus, M. C.; Milne, G. W.; Proksa, B.; Pommier, Y. *J. Med. Chem.* **1997**, *40*, 942.
18. Neamati, N.; Sunder, S.; Pommier, Y. *Drug Discov. Today* **1997**, *2*, 487.
19. Katz, R. A.; Skalka, A. M. *Annu. Rev. Biochem.* **1996**, *63*, 133.
20. Rice, P.; Craigie, R.; Davies, D. R. *Curr. Opin. Struct. Biol.* **1996**, *6*, 76.
21. Dougherty, D. A. *Science* **1996**, *271*, 163.
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. Lubkowski, J.; Yang, F.; Alexandratos, J.; Wlodawer, A.; Zhao, H.; Burke, T. R., Jr.; Neamati, N.; Pommier, Y.; Merkel, G.; Skalka, A. M. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 4831.
24. Molteni, V.; Greenwald, J.; Rhodes, D.; Hwang, Y.; Kwiatkowski, W.; Bushman, F. D.; Siegel, J. S.; Choe, S. *Acta Crystallogr. D Biol. Crystallogr.* **2001**, *57*, 536.
25. Hansch, C. *Acc. Chem. Res.* **1969**, *2*, 232.
26. Hasegawa, K.; Miyashita, Y.; Funatsu, K. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 306.
27. Leardi, R.; Boffia, R.; Terrile, M. *J. Chemometrics* **1992**, *6*, 267.
28. Rogers, D.; Hopfinger, A. J. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854.
29. Chemical Computing Group Inc. *MOE Program*, 2001.
30. Molecular Simulations Inc., Cerius2 program, 1999.
31. Halgren, T. A. *J. Comp. Chem.* **1996**, *17*, 490.
32. Ferguson, D. M.; Raber, D. J. *J. Am. Chem. Soc.* **1989**, *111*, 4371.
33. Katritzky, A. R.; Gordeeva, E. V. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 835.
34. Sottriffer, C. A.; Ni, H.; McCammon, J. A. *J. Med. Chem.* **2000**, *43*, 4109.
35. Farnet, C. M.; Wang, B.; Lipford, J. R.; Bushman, F. D. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 9742.
36. Molteni, V.; Rhodes, D.; Rubins, K.; Hansen, M.; Bushman, F. D.; Siegel, J. S. *J. Med. Chem.* **2000**, *43*, 2031.
37. Ouali, M.; Laboulais, C.; Leh, H.; Gill, D.; Desmaele, D.; Mekouar, K.; Zouhiri, F.; d'Angelo, J.; Auclair, C.; Mouscadet, J. F.; Le Bret, M. *J. Med. Chem.* **2000**, *43*, 1949.
38. Yuan, H.; Parrill, A. L. *J. Mol. Struct. (THEOCHEM)* **2000**, *529*, 273.