Treating Chemical Diversity in QSAR Analysis: Modeling Diverse HIV-1 Integrase Inhibitors Using 4D Fingerprints

Manisha Iyer[†,‡] and A. J. Hopfinger*[,§,||]

Laboratory of Molecular Modeling and Design (MC 781), College of Pharmacy, University of Illinois at Chicago, 833 South Wood Street, Chicago, Illinois 60612-7231, Division of Clinical Chemistry, Department of Pathology, Children's Hospital of Pittsburgh, 5834 Main Tower, 200 Lothrop Street, Pittsburgh, Pennsylvania 15213, The Chem21 Group, Inc., 1780 Wilson Drive, Lake Forest, Illinois 60045, and College of Pharmacy, MSC09 5360, 1 University of New Mexico, Albuquerque, New Mexico 87131-0001

Received May 3, 2007

A set of 213 compounds across 12 structurally diverse classes of HIV-1 integrase inhibitors was used to develop and evaluate a combined clustering and QSAR modeling methodology to construct significant, reliable, and robust models for structurally diverse data sets. The trial-descriptor pool for both clustering- and QSAR-model building consisted of 4D fingerprints and classic QSAR descriptors. Clustering was carried out using a combination of the partitioning around medoids method and divisive hierarchical clustering. QSAR models were constructed for members of each cluster by linear-regression fitting and model optimization using the genetic function approximation. The 12 structurally diverse classes of integrase inhbitors were partitioned into five clusters from which corresponding QSAR models, overwhelmingly composed of 4D fingerprint descriptors, were constructed. Analysis of the five QSAR models suggests that three models correspond to structurally diverse inhibitors that likely bind at a common site on integrase characterized by a common inhibitor hydrogen-bond donor, but involving somewhat different alignments and/or poses for the inhibitors of each of the three clusters. The particular alignments for the inhibitors of each of the three QSAR models involve specific distributions of nonpolar groups over the inhibitors. The two other clusters, one for coumarins and the other for depsides and depsidones, lead to QSAR models with less-defined pharmacophores, likely representing an inhibitor binding to a site(s) different from that of the other nine classes of inhibitors. Overall, the clustering and QSAR methodology employed in this study suggests that it can meaningfully partition structurally diverse compounds expressing a common endpoint in such a manner that leads to statistically significant and pharmacologically insightful composite QSAR models.

INTRODUCTION

There is an increasing interest and effort to construct QSAR models from structurally diverse data sets. Much of this interest and effort is driven by the goal of being able to compute adsorption, distribution, metabolism, excretion, and toxicity (ADMET) properties of organic molecules.[1−3] New computational ADMET approaches generally focus on modeling structurally diverse chemical data sets by dealing only with the intramolecular properties of the chemicals. This approach is taken mainly because the geometries of ADMET "receptors" and/or mechanisms of action of the ADMET endpoint are both normally unknown. Still, the QSAR paradigm, regardless of the form in which it is practiced, is most applicable, and works best, for data sets composed of chemical *analogs*, where a common mechanism of action is likely present.[4]

Two major approaches have emerged for performing QSAR analyses on structurally diverse data sets. In one approach, the number of intramolecular properties computed is made as large as possible, and then a data-reduction method is employed as part of the data-fitting process in constructing the QSAR model.[2,5] The idea is that, if enough molecular features are included, the key intramolecular molecular properties for describing the likely multiple mechanisms of action responsible for the ADMET endpoint will be captured and built into the QSAR model without having to perform data overfitting. Unfortunately, once data reduction is performed, it becomes difficult to interpret the resulting QSAR model and, correspondingly, gain insight into possible mechanisms of action. Of course, it also remains an open question whether all necessary intramolecular features for all mechanisms of action are present, and independently expressed for each mode of action, in the QSAR model.

The second approach to QSAR analysis on a structurally diverse data set having a common endpoint measure involves first partitioning the compounds of the data set and then developing individual QSARs for the compounds in each partition.[5] The criteria for the partitioning are usually based upon molecular similarity (MS). Thus, this approach amounts to trying to create semi-analog sub-data sets from the original diverse data set and then building a QSAR model for each sub-data set.

* Corresponding author phone: (505) 272-8474; fax: (505) 272-0704; e-mail: hopfingr@unm.edu.

[†] University of Illinois at Chicago.

[‡] Children's Hospital of Pittsburgh.

[§] The Chem21 Group, Inc.

[||] University of New Mexico.

10.1021/ci700153g CCC: $37.00 © 2007 American Chemical Society
Published on Web 07/28/2007

This second approach to treating diverse data sets falls more in line with the overall QSAR paradigm; it tries to manipulate the diverse data set into collections of pseudo analogs before attempting to construct QSARs. Still, there are two potential drawbacks to this approach: (1) the uncertainty as to which aspects/features of molecular similarity are best suited for the partitioning/clustering and (2) the possible creation of clustered data sets too small for performing meaningful QSAR analyses.

But perhaps most important at this point in the development of QSAR methodologies to treat structurally diverse data sets is the need to first identify data sets that actually permit a meaningful evaluation of a proposed methodology. Most ADMET data sets are not only globally diverse but also locally diverse. That is, there are no substantial groups of analogs, or near analogs, within the complete data set. Thus, it can be difficult to interpret and validate any type of partitioning/clustering process or method. In addition, the quality and reliability of many ADMET endpoint measures are marginal as compared to pharmacological potency endpoint measures like, for example, enzyme inhibition IC_{50} values.

The ideal data set to evaluate methods of QSAR analysis for diverse data sets should include (a) multiple “islands of compounds” where each “island” is made up of semi-analogs, and these sets of semi-analogs are each structurally diverse from one another; (b) biological endpoint measures of the quality of traditional IC_{50} measures; and (c) crystal structure ligand−receptor-binding information to support, if not validate, the resultant QSAR models derived in the study. One such data set close to meeting these three requirements is a composite set of chemically diverse sets of human immunodeficiency virus type 1 (HIV-1) integrase inhibitors.

This paper reports a series of clustering and QSAR studies of a structurally diverse data set of HIV-1 integrase inhibitor compounds. The findings from this study allow an assessment of the performance of a novel clustering−4D-fingerprint QSAR methodology, as well as permitting significant inferences to be made regarding both the common and the distinct mechanistic binding features of HIV-1 integrase inhibition among the structurally diverse sets of inhibitors investigated.

BACKGROUND OF HIV-1 INTEGRASE AND ITS INHIBITION

Understanding the lifecycle of HIV-1 is critical for gaining insight into all potential targets for chemoprevention of AIDS. The HIV-1 virus is a retrovirus that has a unique replication cycle utilizing three key enzymes following infection and entry into the CD4-positive lymphocytes. These enzymes are (i) reverse transcriptase, an RNA-dependent DNA polymerase used by the virus to transcribe the viral genomic DNA into proviral DNA for incorporation into the host cell DNA; (ii) integrase, which functions by inserting the proviral DNA into the host cell genome; and (iii) protease, which is responsible for processing and packaging new virulent viral particles.[6] The HIV-1 integrase enzyme catalyzes two reactions: 3′-processing and strand transfer. The last two nucleotides of the viral DNA 3′ end are cleaved, and the remaining viral fragment is inserted into the host DNA.[7,8]

Antiviral therapy using a combination of protease and reverse transcriptase inhibitors has been shown to improve the efficacy of AIDS therapy. By comparison, the HIV-1 integrase enzyme did not receive as much attention until the past decade. It is now recognized as a safe and promising target for novel drug discovery due to the fact that it has no mammalian counterpart, and that integration of viral DNA into the host chromosome is an important intermediate step in the replication cycle.[7]

There is limited information available about the structure and mechanism of the integrase enzyme. The published crystal structure of the HIV-1 integrase enzyme[9] identifies key residues involved in binding but also presents an ambiguous picture of the binding site. It is believed that the enzyme forms a complex with DNA, but there is no published information about this integrase−DNA complex structure.[10] In addition, and most germane to the study reported in this paper, known HIV-1 integrase inhibitor compounds belong to very diverse structural classes, and current evidence indicates the presence of more than one binding site.[11]

METHOD

A. HIV-1 Integrase Inhibitors. The HIV-1 integrase inhibitors in this study were gathered from various literature sources.[12−24] The biological activity used as the dependent variable in QSAR model building is the IC_{50} (μM) data for 3′ processing. It has been assumed that the IC_{50} measures are independent of the source of measurement, which is supported by near-identical IC_{50} measurements for a limited number of common inhibitors studied in two or more laboratories. Tables 1−12 list the chemical structures of the 12 classes of inhibitors, each divided into a training set and a test set. The test set compounds are selected so as to span the entire activity range of their corresponding training set compounds. The experimental biological data are represented as $-\log(IC_{50})$, or pIC_{50} values. All molecules in the training and test sets are built using the HyperChem 6.0 software package.[25]

The overall data set includes a total of 213 compounds across the 12 structurally diverse inhibitors. The training sets are composed of 148 inhibitors, and the test sets contain a composite total of 65 inhibitors. The data set has an overall limited range of 3 orders of magnitude in inhibition potency, and clusters of compounds created from this data set will, on average, have even smaller ranges in activity. However, it is important to note that the data set represents a relatively large number of compounds, and that these compounds are reasonably evenly distributed over the range of inhibition potency. Thus, the limited range of data is reasonably robust for statistical analysis.

B. Calculation of 4D Molecular Fingerprint and Other QSAR Descriptors. The theory and formalism underlying the 4D molecular fingerprints are summarized below. Further details on the 4D fingerprints are given in refs 26 and 27.

Molecular dynamic simulations, using the Molsim program,[28] are carried out on each molecule to perform a conformational ensemble sampling of the set of compounds. A thousand conformations of each molecule have been sampled in this study. *Sampling* is the pseudo “fourth dimension” of this method.

Table 1. Set 1: Structures and Inhibition Potencies of Cinnamoyl Derivatives

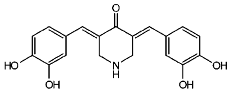
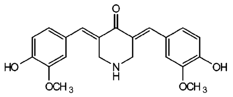
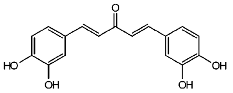
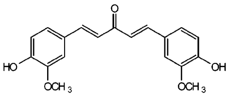
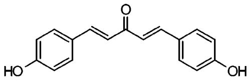
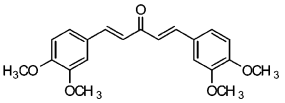
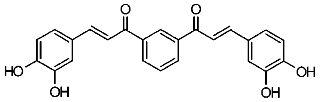
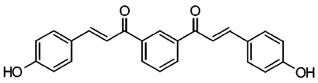
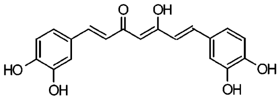
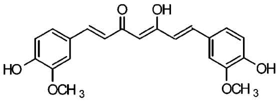
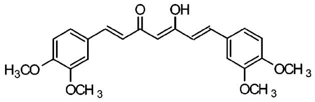
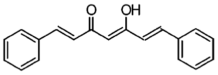
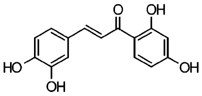
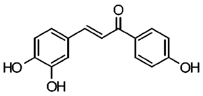
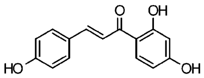
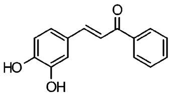
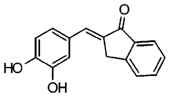
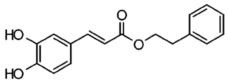
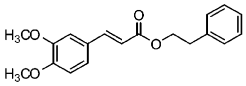
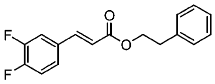
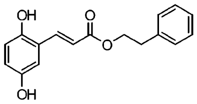
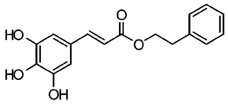
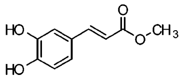
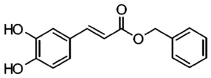
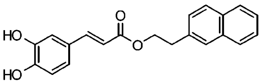
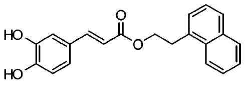
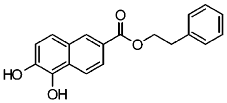
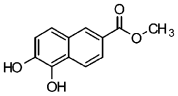
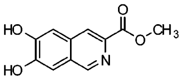
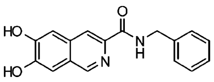
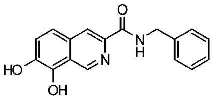
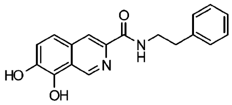
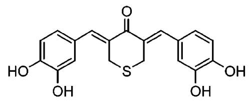
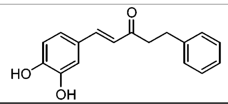
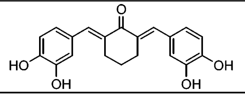
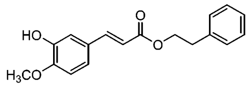
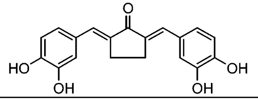
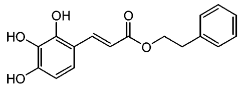
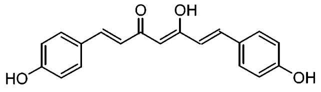
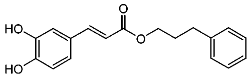
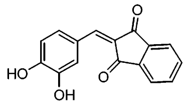
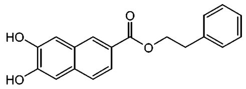
#	Structure	pIC_{50}
	Training Set	
1		**6.70**
2		**3.50**
3		**6.22**
4		**4.43**
5		**3.50**
6		**3.50**
7		**6.10**
8		**4.52**
9		**6.15**
10		**4.52**
11		**3.50**
12		**3.50**
13		**4.70**
14		**5.89**
15		**3.50**
16		**5.64**
	Training Set	
17		**4.07**
18		**5.15**
19		**3.50**
20		**3.50**
21		**3.50**
22		**5.70**
23		**3.50**
24		**5.05**
25		**4.36**
26		**4.82**
27		**4.74**
28		**5.10**
29		**3.50**
30		**3.50**
31		**4.12**
32		**4.40**

Table 1 (Continued)

#	Structure	pIC_{50}	#	Structure	pIC_{50}
	Test Set			*Test Set*	
T1		**6.70**	T6		**3.50**
T2		**6.05**	T7		**4.22**
T3		**6.00**	T8		**5.70**
T4		**3.50**	T9		**5.10**
T5		**5.52**	T10		**4.17**

The molecules are divided into their "functional pieces", called *interaction pharmacophore elements* (IPEs), as defined in Table 13. The IPE 4D-fingerprint descriptors are eigenvalues from the eigenvectors determined for a molecule from its absolute molecular similarity main distance-dependent matrix (MDDM).[27] This matrix captures the intrinsic size, shape, and conformational flexibility of the molecules, and it is constructed for each IPE pair. The elements of the MDDM are defined as

$$E_{(\lambda,dij)} = \mathrm{e}^{(-\lambda<dij>)} \qquad (1)$$

λ is a constant determined by maximizing the difference in the sum of eigenvalues for any two arbitrary molecules with the same number, N, of a maximized IPE type.[27] The term $<dij>$ refers to the average distance between the atom pair i,j of IPE types u and v, such that

$$<dij> = \sum_k dij(k)\, p(k) \qquad (2)$$

where $p(k)$ is the thermodynamic probability of the kth conformer state sampled in the assessment of conformational flexibility, and $dij(k)$ is the corresponding distance between atom pair i,j of IPE type u and v for the kth conformer state.

Eigenvalues are derived by the diagonalization of the MDDM. For the same-term IPE, that is, $u = v$, the MDDMs can be directly diagonalized. To calculate the n normalized eigenvalues for IPE type m of compound α, $\epsilon_{mn}(\alpha)$, the nonscaled eigenvalues $\epsilon'_{mn}(\alpha)$ are scaled relative to the rank of the MDDM

$$\epsilon_{mn}(\alpha) = \frac{\epsilon'_{mn}(\alpha)}{\mathrm{rank}(\alpha)_m} \qquad (3)$$

Thus, $\epsilon_{1,2}(5)$ corresponds to the second eigenvalue of the MDDM for IPE type 1 (*nonpolar* atom) of compound 5.

When the cross-term for IPE terms is different ($u \neq v$), the following square MDDMs are constructed:

$$\mathrm{MDDM}(u,u) = \mathrm{MDDM}(n_u,n_v) \times \mathrm{MDDM}(n_u,n_v)^T \qquad (4)$$

and for $[n_v \times n_u]$

$$\mathrm{MDDM}(v,v) = \mathrm{MDDM}(n_v,n_u) \times \mathrm{MDDM}(n_v,n_u)^T \qquad (5)$$

Since the same rank and trace are present in eqs 4 and 5, both MDDM(u,u) and MDDM(v,v) have the same set of eigenvalues. Consequently, for each pair of IPEs, where $u \neq v$

$$\epsilon(\alpha)_{u,v} = \sqrt{[\epsilon(\alpha)]_{\mathrm{MDDM}(u,u)}} \qquad (6)$$

The estimation of molecular similarity for molecules α and β is a value between 0 and 1. It is calculated as

$$S_{\alpha\beta} = (1 - D_{\alpha\beta})(1 - \varphi) \qquad (7)$$

where $D_{\alpha\beta}$ is the molecular dissimilarity given by

$$D_{\alpha\beta} = \sum_i |\epsilon(\alpha)_i - \epsilon(\beta)_i| \qquad (8)$$

and

$$\varphi = \frac{|\mathrm{rank}(\alpha) - \mathrm{rank}(\beta)|}{|\mathrm{rank}(\alpha) + \mathrm{rank}(\beta)|} \qquad (9)$$

Since the rank of the matrices is essentially the number of atoms of a specific IPE type present, the φ term in eq 9 serves to reincorporate molecular-size information.

The 4D-fingerprint descriptor set for each compound in the training and test sets comprises all the eigenvalues of all IPE eigenvector pairs for the compound. There are 36 possible combinations of the eight IPE types, as listed in Table 13. A threshold cutoff value of 0.001 is applied in this study, and those normalized eigenvalues below the threshold value are disregarded.

Additional "traditional" QSAR descriptors are also calculated and added to the descriptor pool. ClogP is calculated using the Daylight software,[29] and all non-4D-fingerprint-QSAR descriptors listed in Table 14 are calculated using a module of the MI-QSAR program.[30]

Table 2. Set 2: Structures and Inhibition Potencies of Tyrphostins

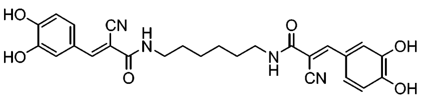
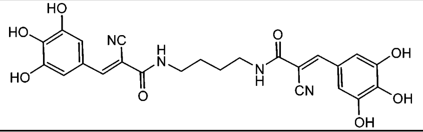
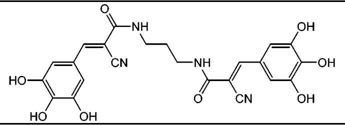
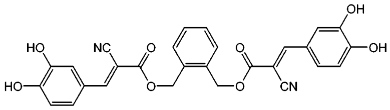
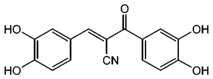
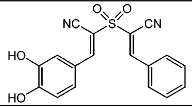
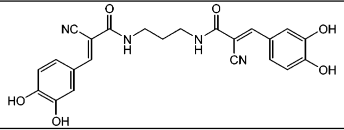
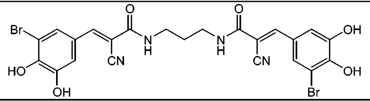
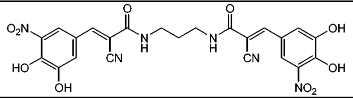
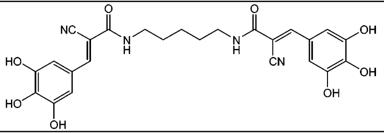
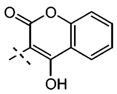
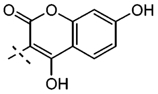
#	Structure	pIC_{50}
	Training Set	
33		5.87
34		6.18
35		6.40
36		5.52
37		6.00
38		5.33
	Test Set	
T11		5.72
T12		6.10
T13		5.48
T14		6.35

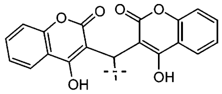
Table 3. Set 3: Structures and Inhibition Potencies of Coumarins



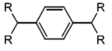
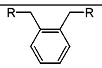
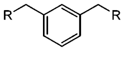
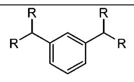
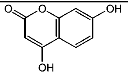
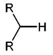
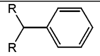
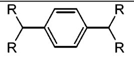
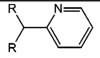
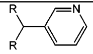
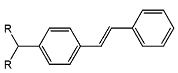
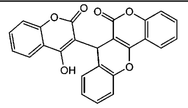
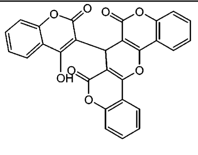
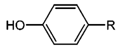
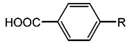
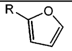
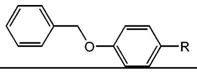
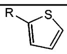
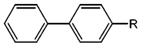
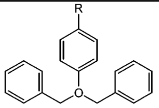
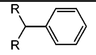
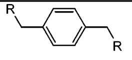
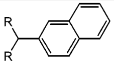
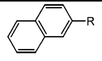
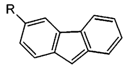
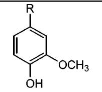
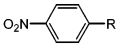
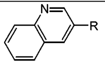
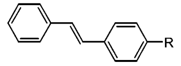
39-42, T15, T16



43-51, T17



52-58, T18-T23

#	Structure	pIC_{50}	#	Structure	pIC_{50}
		Training Set			
39		5.82	40		4.34
41		4.43	42		5.53
43		3.52	44		4.33
45		4.76	46		6.43
47		4.05	48		4.05
49		5.15	50		3.91
51		4.45	52		3.87
53		4.24	54		4.26
55		4.86	56		3.98
57		4.82	58		5.02
		Test Set			
T15		4.36	T16		4.09
T17		5.38	T18		4.71
T19		4.98	T20		3.73
T21		4.27	T22		4.46
T23		5.13			

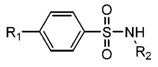
C. Cluster Analysis. Clustering algorithms can be divided into partitioning and hierarchical methods. One partitioning and one hierarchical method of clustering were employed in this study.

Partitioning Around Medoids (PAM).[31] PAM is used when a data set is classified into k clusters, where k is fixed. In order to obtain the k clusters, the method selects k objects (called the representative objects/compounds, or medoids) of the data set. Assigning each other compound to the nearest representative medoid identifies the corresponding groups. The selection of medoids and the assignment of groups are done in such a way that the total dissimilarity of all objects to their nearest medoid is minimal. The software performing the algorithm can generate a novel graphical display of this clustering, called the silhouette plot.[32] The silhouette value determined from the silhouette plot permits the rational selection of the optimum number of clusters to describe the data set.[33] In other words, the silhouette value, $s(i)$, provides a measure of the certainty with which a compound is assigned to the cluster. Given cluster A formed using PAM, and an object i assigned to it, the average dissimilarity of i to all objects j in A is given by

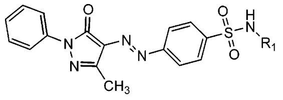
$$a(i) = \sum\nolimits_j d(i,A_j)/\text{number of objects in } A \qquad (10)$$

where $d(i,A_j)$ is the distance of each object j in A from i. The lowest-corresponding dissimilarity of i to any other

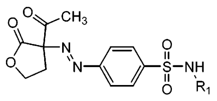
Table 4. Set 4: Structures and Inhibition Potencies of Aromatic Sulfonamides



59-65, T24, T25



66, T26, T27



67,68, T28

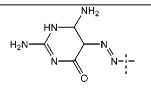
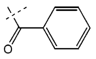
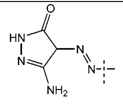
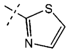
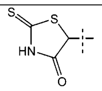
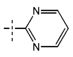
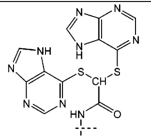
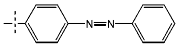
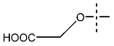
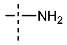
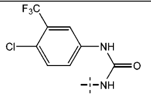
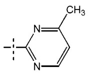
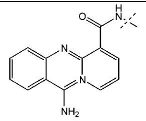
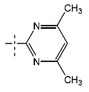
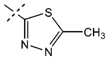
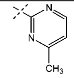
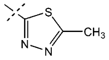
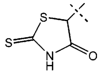
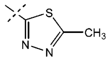
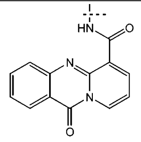
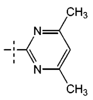
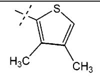
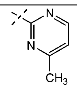
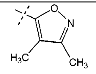
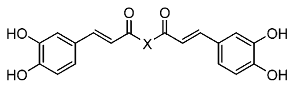
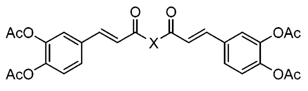
#	Structure: R1	R2	pIC_{50}
Training Set			
59			3.92
60			4.15
61			4.54
62			3.85
63			5.09
64			3.91
65			4.32
66			3.87
67			4.15
68			3.61
Test Set			
T24			4.62
T25			4.12
T26			3.89
T27			4.31
T28			3.71

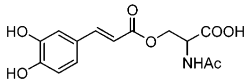
Table 5. Set 5: Structures and Inhibition Potencies of Chicoric Acids



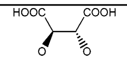
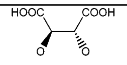
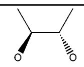
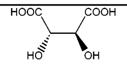
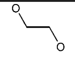
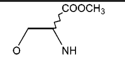
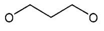
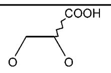
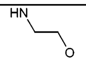
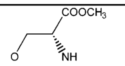
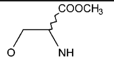
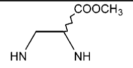
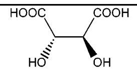
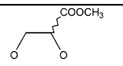
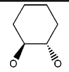
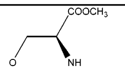
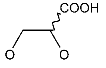
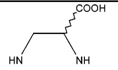
69-74, T29-T31



75-80, T32-T34



81

#	Structure: X	pIC_{50}	#	Structure: X	pIC_{50}
Training Set					
69		5.77	75		5.00
70		4.42	76		4.66
71		6.19	77		5.60
72		4.66	78		5.68
73		4.56	79		5.54
74		5.48	80		4.94
			81		3.48
Test Set					
T29		5.96	T32		5.00
T30		4.61	T33		5.55
T31		5.38	T34		5.17

cluster, $b(i)$, is also calculated. If i is more similar to objects in a cluster B than in cluster A, then

$$b(i) = \Sigma_j d(i, B_j)/\text{number of objects in } B \qquad (11)$$

Hence, the silhouette value for object i is defined as

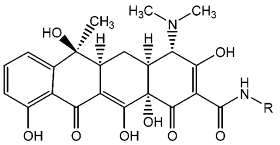
$$s(i) = [b(i) - a(i)]/\max\{a(i),b(i)\} \qquad (12)$$

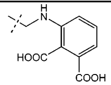
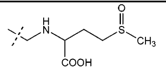
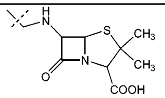
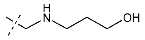
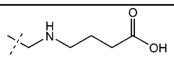
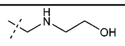
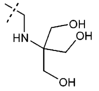
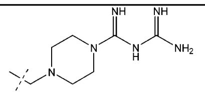
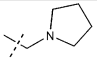
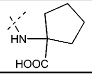
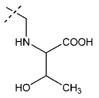
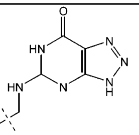
where $\max\{a(i),b(i)\}$ refers to the maximum, or larger, value between $a(i)$ and $b(i)$.

The silhouette values lie between 0 and 1, with values closer to 1 indicating a higher probability of accurate cluster assignment. The average silhouette value of all clustered objects is a measure of the overall "goodness" of clustering across the data set. Thus, it is possible to achieve appropriate clustering by comparing and selecting the clustering solution with the highest average silhouette value.

Divisive Hierarchical Clustering.[31] Hierarchical clustering algorithms proceed by combining or dividing existing groups of data, then producing a hierarchical structure displaying the order in which groups are merged or divided. Hierarchical methods can be either agglomerative or divisive. Agglomerative methods start with each observation in a separate group and proceed until all observations are in a single group. Divisive methods work almost the same way, but in the reverse direction. The divisive approach has an advantage

Table 6. Set 6: Structures and Inhibition Potencies of Tetracyclines



#	Structure: R	pIC50	#	Structure: R	pIC50
Training Set					
82	H	**3.69**			
83		**5.77**	87		**5.92**
84		**5.66**	88		**5.92**
85		**5.72**	89		**5.29**
86		**5.96**	90		**6.05**
Test Set					
T35		**4.55**	T37		**5.89**
T36		**5.46**	T38		**5.80**

over other hierarchical methods because it provides useful information about the main structure of the data split into a few large clusters, as opposed to a detailed description about the individual objects.

The hierarchy obtained from the divisive cluster analysis can be graphically represented in the form of a "cluster tree" structure called a dendrogram. Dendrograms contain information about the objects grouped together at various levels of dissimilarity. At the bottom of the dendrogram, each object is in its own cluster. Vertical lines extend upward from each individual cluster and, at various levels, connect to other vertical lines through horizontal ones. This grouping continues to the top of the dendrogram, where all objects are part of one parent group.

One of the most useful features of hierarchical clustering is that it allows the derived cluster to be divided at any level to increase or decrease the Euclidean distance between clusters. Subsequently, this can alter the number of clusters present in the data. The "correct" number of clusters is determined to be that which best fits the application of the data.[31]

Cluster Analysis and HIV-1 Integrase Inhibition. It is evident from an earlier study that not all chemical classes of HIV-1 integrase inhibitors have a common mechanism of action.[34] One objective of performing cluster analysis is to explore whether two or more of the 12 structurally diverse inhibitor classes inhibit the integrase enzyme in a similar manner. In other words, the task is to determine if some members of the set of structurally diverse inhibitors share key structural features that may not be evident from the QSAR analysis of each of these individual classes of inhibitors.

Only molecular similarity eigenvectors for atom pairs with the same IPE types are used in the cluster analysis. That is, eight sets of 4D fingerprints, one corresponding to each IPE type, are considered as the cheminformatics representation of each compound. Since one objective of this study is to try to capture those 4D fingerprints essential for biological activity, the compound exhibiting the highest activity is selected from each of the 12 structural classes. The eigenvalues of these 12 compounds are clustered using both PAM and a divisive hierarchical clustering. The procedure is repeated using the top two most potent compounds from each structural class to confirm the results of the original cluster analysis. In this study, cluster analyses are performed using the S-plus statistical package.[35] PAM functions by first determining the "medoids," which are actual objects in the data set representative of each of their respective clusters.

D. QSAR Model Construction and Evaluation. The dependent variable used in QSAR model construction is pIC_{50} ($-\log IC_{50}$) for integrase enzyme inhibition. QSAR models for both the complete and partitioned HIV-1 integrase inhibitor data sets are constructed using multidimensional linear regression fitting. Model optimization is achieved using the gentic function approximation (GFA) algorithm.[36,37] Only linear representations of each of the 4D fingerprints and classical descriptors are included in model construction. Both the correlation coefficient of fit, R^2, and the cross-validated correlation coefficient, Q^2, are considered in evaluating the quality of a QSAR model.

E. Test Set Predictions. The optimized QSAR model of a particular cluster is used to predict the activities of the corresponding test-set compounds belonging to the cluster, as well as test set compounds from other clusters. If the specified clusters are distinctive and significant, the test-set compounds belonging to a particular cluster will be best predicted by the corresponding QSAR model. This strategy provides another way to validate the clustering process, in addition to the silhouette values for PAM and the divisive coefficient for hierarchical clustering. To further evaluate the predictive power of the 4D-fingerprint QSAR models, the R^2 for test-set prediction, termed $R_{\text{prediction}}{}^2$, is calculated as

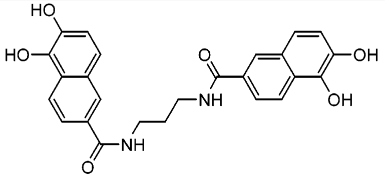
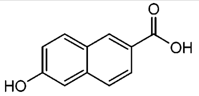
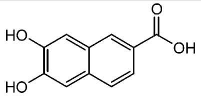
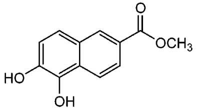
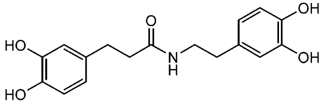
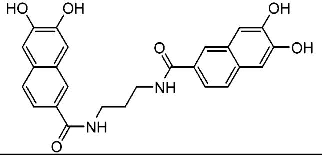
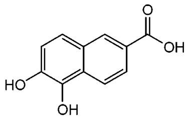
$$R_{\text{prediction}}{}^2 = 1 - \sum [y_{i(\text{pred})} - y_{i(\text{obs})}]^2 / \sum (y_{i(\text{obs})} - K)^2 \tag{13}$$

where K is the average of the observed (experimental) test-set values and $y_{i(\text{pred})}$ and $y_{i(\text{obs})}$ are the predicted and observed activities for the ith test-set compound, respectively. From eq 13, it can be inferred that the $R_{\text{prediction}}{}^2$ metric is greater than zero for a model in which the predicted activities are more accurate than using the average activity, K, of the training set as each predicted activity value in eq 13.

RESULTS

A. PAM and Divisive Hierarchical Clustering. The 148 training set compounds from all 12 diverse structural classes of inhibitors are first combined into one large parent training

Table 7. Set 7: Structures and Inhibition Potencies of Arylamides and Naphthalene-Based Compounds

#	Structure: R	pIC_{50}	#	Structure: R	pIC_{50}
		Training Set			
91					6.64
92		3.76	93		5.27
94		4.23	95		4.48
		Test Set			
T39		6.01	T40		4.27

set for building integrase inhibition QSAR models. The most significant optimized QSAR model is

$$pIC_{50} = 4.483 - 194.81[\epsilon_{11}(\text{hs,hs})] + 43.43[\epsilon_8(\text{p+,hba})] + 38.54[\epsilon_{10}(\text{aro,aro})] - 38.83[\epsilon_{11}(\text{hs,np})] + 10.63[\epsilon_2(\text{np,hba})] + 30.30[\epsilon_5(\text{hs,np})] + 9.05[\epsilon_4(\text{hs,hba})]$$

$$N = 148,\ R^2 = 0.51,\ Q^2 = 0.47 \qquad (14)$$

The lack of significance of a QSAR model for the parent-training set, as concluded from the low R^2 and Q^2 values, shows that there exists no combination of structural features common to the 12 data sets (at least as represented by the trial descriptor pool of 4D-QSAR fingerprints and classic QSAR descriptors) that can effectively describe the binding and/or inhibition mechanism for the HIV-1 integrase enzyme. The underlying assumption is that, if such common features did exist, they would be captured by a subset of the trial pool of QSAR descriptors. It is also worth noting that none of the non-4D-fingerprint descriptors in Table 14 are found in eq 14. This indicates that the maximum extraction of SAR information in optimizing the QSAR given by eq 14, however limited as indicated by its low R^2 value, involves only the use of inhibitor 4D fingerprints.

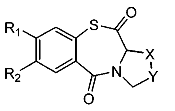
To investigate the best clustering method to be applied in this study, both PAM and divisive hierarchical clusterings are performed on the two most potent compounds from each of the 12 structural classes of inhibitors using their 4D fingerprints. The average PAM silhouette values of partitioning the compounds into the 2, 3, 4, 5, and 6 subsets, with respect to their 4D fingerprints of the eight IPE types, are listed in Table 15. The largest silhouette value, based on 4D fingerprints of a particular IPE type for a particular number of clusters, identifies the optimum method for clustering the data set. The silhouette values listed in Table 15 indicate that the best clustering using PAM is achieved by partitioning the data set into three clusters for IPE type 6 (aromatic atoms). Silhouette values close to 1.00 are considered significant.[31] Thus, as evident from Table 15, no other IPE type performs as well as the aromatic IPE for clustering discrimination. However, the silhouette values are only a statistical measure of clustering, and the real quality and significance of the clusters must be assessed from the quality of the QSAR models constructed from the compounds within each of the clusters and the corresponding predictive power of each of these QSAR models.

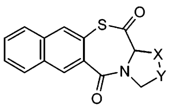
The individual 12 structural classes of inhibitors in this study involve analog series of compounds presumably having a common binding site. Hence, all training- and test-set analogs of a given structural class are assigned to the same cluster as the two corresponding, most-potent inhibitor analogs used in developing the clusters. Table 16 shows the 12 structural classes of HIV-1 integrase inhibitors as clustered into three groups by the PAM analysis.

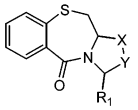
The PAM analysis, which is a nonhierarchical partitioning method, is compared to a hierarchical type of clustering by using the same 24 representative compounds in a divisive hierarchical clustering analysis. Table 17 presents the results of this hierarchical clustering analysis in the form of the divisive coefficients obtained by clustering according to 4D fingerprints of each IPE type.

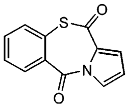
The results of hierarchical clustering are judged not only by their divisive, or agglomerative, coefficients but also by evaluating the corresponding dendrogram. This method of clustering splits the parent data set until there is only one object in each cluster. Hence, clusters are determined on the basis of the distance, or distinction required, between any two clusters. Figure 1 shows the dendrogram obtained by divisive hierarchical clustering for IPE type 6 (aromatic) 4D fingerprints. Comparison of the dendrograms obtained from clustering, based on 4D fingerprints of each IPE type, indicates that the aromatic IPE (type 6) dendrogram is the most significant. But again, the actual significance of the

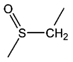
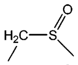
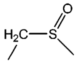
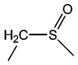
Table 8. Set 8: Structures and Inhibition Potencies of Thiazolothiazepines


96-106, T41-T45


107, T46


108


109

#	R1	R2	-X-Y	pIC_{50}
		Training Set		
96	H	H	$-S-CH_2-$	**3.96**
97	H	Cl	$-S-CH_2-$	**3.89**
98	H	Br	$-S-CH_2-$	**4.24**
99	H	CH_3	$-S-CH_2-$	**4.19**
100	H	H	$-CH_2-S-$	**3.68**
101	H	Br	$-CH_2-S-$	**4.06**
102	NO_2	H	$-CH_2-S-$	**4.05**
103	OCH_3	OCH_3	$-CH_2-S-$	**3.19**
104	H	CH_3		**3.23**
105	H	H		**3.72**
106	H	Cl		**3.62**
107			$-S-CH_2-$	**4.40**
108	Ph		$-CH_2-S-$	**3.62**
109				**3.23**
		Test Set		
T41	H	Cl	$-CH_2-S-$	**3.87**
T42	H	CH_3	$-CH_2-S-$	**4.28**
T43	H	OCH_3	$-CH_2-S-$	**3.67**
T44	H	H	$-H_2C-CH_2-$	**3.35**
T45	H	OCH_3		**4.07**
T46			$-S-CH_2-$	**4.04**

clustered groups obtained by any type of clustering/categorical method is ultimately determined by evaluating the QSAR models derived from each clustered group.

Comparing the dendrogram in Figure 1 to the results of the PAM analysis in Table 16 makes it clear that the same three main clusters have been identified by both clustering methods. Each method identifies a singlet, which is a remote cluster having only one object: the coumarin structural class. The chemical structures of the coumarins of this study have a large core structure, made up of five or more aromatic rings, which significantly differs from all other classes of HIV-1 integrase inhibitors considered in this study.

B. 4D-Fingerprint QSAR Models. The main cluster, number 1 of Table 16, in the PAM analysis has been divided by hierarchical clustering into three subsets of three, two, and one structural class, respectively. These clusters have been termed 1A, 1B, and 1C. QSAR models have been built using a combination of the 4D-fingerprint and classical intramolecular descriptors for each cluster. Table 18 shows the optimized QSAR models, along with their R^2 and Q^2 values. A detailed look at these models shows each has a statistically higher quality than the QSAR model for the parent training set, eq 14. That is, the QSAR modeling of cluster 1 identified by PAM improves considerably upon splitting cluster 1 into the three subsets identified by hierarchical clustering. Cluster 1A consists of the structural classes of cinnamoyl derivatives, chicoric acids, and curcumins. Cluster 1B contains the tyrphostins and salicylhydrazines. And cluster 1C effectively consists of only one structural class, the depsides and depsidones. The number of compounds in each training-set cluster is given in Table 18.

The descriptors appearing in the QSAR models of Table 18 are overwhelmingly 4D fingerprints, with Ecoh, the cohesive packing energy of the inhibitor molecules and Chi-8, a Kier and Hall connectivity index,[38] being the only non-4D-fingerprint descriptors. The 4D-fingerprint descriptors in the QSAR models are named in terms of their corresponding eigenvalues. For example, ϵ_5(np,aro) represents the fifth-largest eigenvalue from the MDDM of u = (nonpolar atom) and v = (aromatic atom). Hence, the eigenvalues of (np,aro) are the “fingers” from which the relative molecular similarity of any pair of inhibitors, with respect to their nonpolar and aromatic atoms, can be estimated.[27]

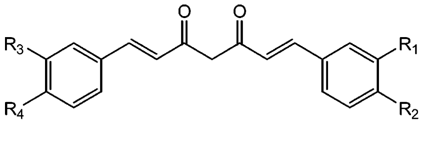
The eigenvalues (4D fingerprints) in the QSAR models of Table 18 that survive as descriptors in the model-optimization process, and thus are significant features in describing HIV-1 integrase inhibition, are found in varying positions near the beginning, middle, or end of their respective eigenvectors. For example, the model for cluster 3 contains the second eigenvalue from the (p+,hbd) eigenvector and the 39th eigenvalue from the (np,np) eigenvector.

An analysis of the correlations between all pairs of 4D fingerprints that appear as descriptors in a QSAR model is helpful in establishing the significance of these descriptors for the QSAR model. The descriptor correlation matrices for the QSAR models built from the two largest training sets, cluster 1A and cluster 3, are given in Table 19a and b. No significant correlation is observed among any pair of descriptors in the QSAR model for cluster 3. A moderate correlation (R^2 = 0.65) is found between one pair of 4D fingerprints of the QSAR model for cluster 1A. However, eliminating either of these two descriptors results in statistically inferior models compared to when both descriptors are present. The R^2 drops from 0.78 to 0.63 when eliminating ϵ_{19}(np,any), and to 0.55 upon eliminating ϵ_{19}(hs,any). This analysis establishes the need for retaining both 4D-fingerprint descriptors in order to generate a significant QSAR model for cluster 1A. Perhaps even more significant is how, overall, nearly orthogonal (small cross-correlation values between pairs of 4D fingerprints) the 4D-fingerprint descriptors of the QSAR models are to one another.

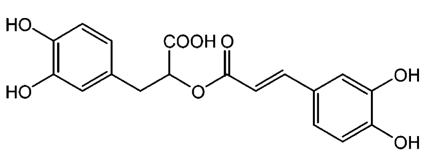
The absence of any significant QSAR models from (200 different) random scrambling samplings of the dependent variables of each of the training sets strongly suggests only a remote possibility of any chance correlation QSAR models. Thus, these findings indicate that the regression-fit and GFA-optimized QSAR models reported here are robust models.

Another diagnostic to estimate the significance of the various 4D fingerprints (descriptors) in a QSAR model is to generate and compare the corresponding set of variance vectors.[39] Variance vectors are calculated by multiplying the regression coefficient (from the particular QSAR model) by

Table 9. Set 9: Structures and Inhibition Potencies of Curcumins



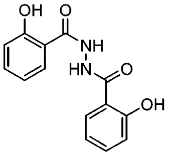
110-113, T47



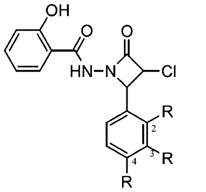
T48

#	R1	R2	R3	R4	pIC$_{50}$
			Training Set		
110	H	OH	H	OH	3.92
111	OCH$_3$	OH	OCH$_3$	OH	3.82
112	OH	OH	OH	OH	4.22
113	OCH$_3$	OH	OH	OH	4.74
			Test Set		
T47	H	OH	OCH$_3$	OH	3.85
T48					5.05

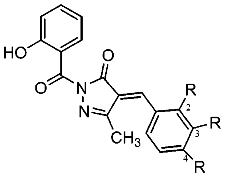
Table 10. Set 10: Structures and Inhibition Potencies of Salicylhydrazines



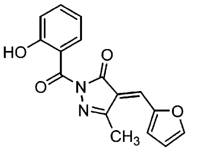
114



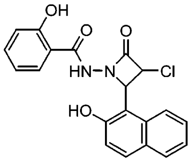
115-120, T49, T50



T51



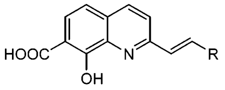
121-124, T52-T55



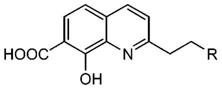
126

#	structure: R	pIC$_{50}$	#	structure: R	pIC$_{50}$
		Training Set			
114		**5.68**	121	4-OCH$_3$	**6.05**
115	4-OCH$_3$	**3.54**	122	4-OH	**6.22**
116	2-OH	**3.52**	123	3-OH	**6.05**
117	3-Cl	**3.75**	124	3-OCH$_3$, 4-OH	**6.30**
118	3-OH	**3.69**	125	3,4,5-(OCH$_3$)$_3$	**5.70**
119	3,4-(OCH$_3$)$_2$	**3.93**	126		**5.57**
120	3,4,5-(OCH$_3$)$_3$	**4.14**			
		Test Set			
T49	4-NO$_2$	**3.74**	T53	4-NO$_2$	**6.10**
T50	2-Cl	**3.61**	T54	3-NO$_2$	**5.85**
T51		**3.90**	T55	3-OCH$_3$, 4-OH	**6.10**
T52	2-OH	**6.22**			

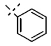
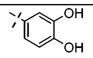
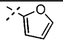
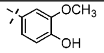
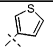
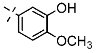
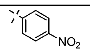
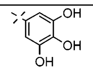
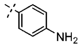
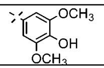
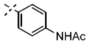
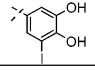
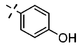
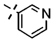
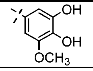
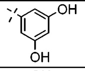
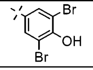
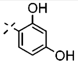
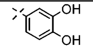
Table 11. Set 11: Structures and Inhibition Potencies of Styrylquinolines



127-139, T56-T60



T61

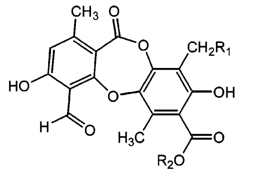
#	Structure: R	pIC$_{50}$	#	Structure: R	pIC$_{50}$
		Training Set			
127		5.28	134		5.62
128		5.72	135		5.55
129		5.47	136		6.05
130		5.92	137		6.52
131		5.46	138		5.31
132		5.85	139		5.40
133		5.80			
		Test Set			
T56		5.39	T59		6.15
T57		5.49	T60		5.89
T58		5.43	T61		5.64

the corresponding descriptor value for each compound in the training sets. The range in descriptor variance reflects its net contribution to the dependent variable (pIC$_{50}$ of HIV-1 integrase inhibition, in this case), as calculated by the respective QSAR model. Descriptors exhibiting similar ranges in their variance vectors are considered similar in significance.

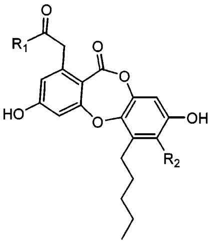
Descriptor significance by variance vectors has been determined for the QSAR models of the two largest clusters (1A and 3). For cluster 1A, two of the 4D fingerprints, ϵ_{19}(np,any) and ϵ_{19}(hs,any), have a higher significance than other 4D-fingerprint descriptors in the model. 4D fingerprints for the model of cluster 3 are all in the same range and, hence, about equally significant in this QSAR model.

C. Test Set Predictions. The test set compounds (see Tables 1−12) of each structural class have been assigned to the same cluster as the training set compounds of the same structural class. The observed and predicted activities of each of the 65 test set compounds are given in Table 20, along with the structural class and assigned cluster. In addition, $R_{\text{prediction}}^2$ values have been calculated using eq 13 as a criterion to evaluate the predictive power of the

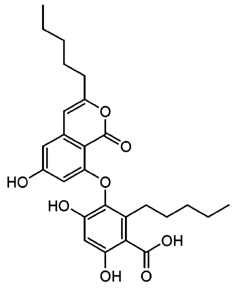
Table 12. Set 12: Structures and Inhibition Potencies of Depsides and Depsidones



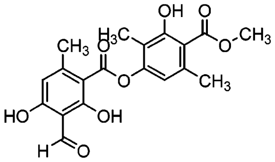
140-141, T62



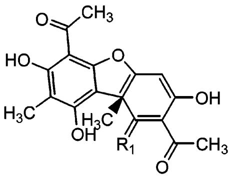
142, T63



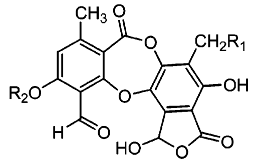
143



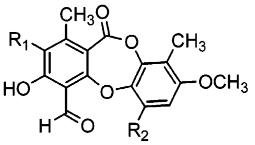
144



145, T64



146, T65



147-148

#	R1	R2	pIC$_{50}$	#	R1	R2	pIC$_{50}$
			Training Set				
140	H	H	**5.34**	145	O		**3.90**
141	OCOCH=CHCOOH	H	**5.31**	146	OH	H	**4.76**
142	$(CH_2)_4CH_3$	COOH	**4.41**	147	H	COOH	**4.79**
143			**4.28**	148	Cl	CH_3	**5.61**
144			**4.21**				
			Test Set				
T62	H	CH_3	**5.27**	T64	$NNHCOC_5H_4N$		**4.17**
T63	$(CH_2)_4CH_3$	H	**4.29**	T65	H	CH_3	**5.44**

Table 13. Interaction Pharmacophore Elements Used in 4D-QSAR Analysis

IPE description/abbreviation	IPE Code
all atoms of the molecule (any)	0
nonpolar atoms (np)	1
polar positive atoms (p+)	2
polar negative atoms (p−)	3
hydrogen-bond acceptor atoms (hba)	4
hydrogen-bond donor atoms (hbd)	5
aromatic atoms (a)	6
non-hydrogen atoms (hs)	7

QSAR models, and to distinguish between the clusters formed.

Table 21 lists the composite average predictivity, as given by the $R_{\text{prediction}}^2$ values, for the QSAR models for each of the five clusters identified by divisive clustering analysis. The QSAR model for cluster 1A shows an $R_{\text{prediction}}^2$ value of 0.60, which increases to 0.67 with the elimination of two outliers. Clusters 1B and 1C have $R_{\text{prediction}}^2$ values of 0.80 and 0.62, respectively. Cluster 1C has a training set of only nine compounds, which is too small a set from which to generate a reliable QSAR model. Clusters 2 and 3 have $R_{\text{prediction}}^2$ values of 0.65 and 0.75, respectively. Thus, all of the QSAR models, composed largely of 4D-fingerprint descriptors, have reasonable to good predictive power and, therefore, can be considered significant and robust models. It can also be seen in Table 21 that each QSAR model for each cluster performs very poorly in predicting the test set compounds assigned to clusters other than its own. This is an important endorsement of the accuracy, specificity, and significance of the cluster analyses based on aromatic atom IPE types.

DISCUSSION

An analysis of the clustered chemical classes of inhibitors, and the corresponding QSAR models for each of the clusters, permits inferences regarding (a) the number of different inhibitor binding sites on integrase, (b) possible common binding features among the chemical classes of inhibitors, and (c) the relative alignments and poses of different classes of integrase inhibitors bound at a common receptor binding site. Table 22 reports each cluster and its member chemical classes of inhibitors. This table also shows the specific 4D-fingerprint descriptors, discussed in terms of their eigenvalue properties, composing the corresponding QSAR models. Only eigenvalues based completely on a *specific* type(s) of IPEs [pharmacophore atom type(s)] are listed in Table 22. QSAR eigenvalues involving the nonspecific IPE types “any” and “hs” are *not* included in Table 22. The signs in brackets next to the descriptions of the specific IPEs in Table 22

Table 14. General Intramolecular Solute Descriptors Used as Part of the Trial MI-QSAR Descriptor Pool and Dissolution and Solvation Descriptors

Part A: General Intramolecular Solute Descriptors

Symbol	Description
HOMO	highest occupied molecular orbital energy
LUMO	lowest unoccupied molecular orbital energy
Dp	dipole moment
Vm	molecular volume
SA	molecular surface area
Ds	density
MW	molecular weight
MR	molecular refractivity
N(hba)	number of hydrogen bond acceptors
N(hbd)	number of hydrogen bond donors
N(B)	number of rotatable bonds
JSSA(X)	Jurs-Stanton surface area descriptors
PSA	polar surface area
Chi-N, Kappa-M	Kier and Hall topological descriptors
Rg	radius of gyration
Pm	principle moment of inertia
Se	conformational entropy
Q(I)	partial atomic charge densities

Part B: Dissolution and Solvation Descriptors

dissolution and solvation solute descriptors: symbols	description of the dissolution/solvation solute descriptors
F(H2O)	the aqueous solvation free energy
F(OCT)	the 1-octanol solvation free energy
ClogP	the 1-octanol/water partition coefficient
E(coh)	the cohesive packing energy of the solute molecules
T_M	the hypothetical crystal-melt transition temperature of the solute
T_G	the hypothetical glass transition temperature of the solute

Table 15. Silhouette Values of the PAM Analysis Based on Each IPE Type

	average silhouette value							
no. of clusters	IPE 0	IPE 1	IPE 2	IPE 3	IPE 4	IPE 5	IPE 6	IPE 7
2	0.42	0.57	0.40	0.31	0.31	0.51	0.50	0.59
3	0.23	0.33	0.18	0.21	0.11	0.53	**0.84**	0.27
4	0.18	0.35	0.23	0.23	0.10	0.32	0.72	0.17
5	0.18	0.23	0.22	0.19	0.1	0.39	0.62	0.16
6	0.15	0.19	0.20	0.19	0.09	0.44	0.53	0.18

Table 16. Structural Classes of HIV-1 Integrase Inhibitors as Grouped into Three Clusters by PAM Analysis

cluster 1	cluster 2 (singlet)	cluster 3
cinnamoyl derivatives	coumarins	aromatic sulfonamides
chicoric acids		tetracyclines
curcumins		arylamides and naphthalenes
tyrphostins		thiazolothiazepines
salicylhydrazines		styrylquinolines
depsides and depsidones		

Table 17. Divisive Coefficients from Hierarchical Clustering Based on Each Type of 4D Fingerprint IPE Type

divisive coefficient							
IPE 0	IPE 1	IPE 2	IPE 3	IPE 4	IPE 5	IPE 6	IPE 7
0.64	0.71	0.63	0.68	0.63	0.72	**0.91**	0.69

indicate if the presence of the type of IPE pair increases [+] or decreases [−] inhibition potency on the basis of the QSAR model.

The eigenvalues (4D-fingerprint descriptors) in Table 22 are also partitioned with respect to *short-range*, *medium-range*, and *long-range* spatial distributions of the IPE pairs over the inhibitors. From the manner in which the eigenvalues are derived in the 4D-molecular-similarity formalism,[27] eigenvalues with small eigenvector numbers correspond to IPE pairs of an inhibitor close to one another in space, while eigenvalues with large eigenvector numbers capture information about inhibitor IPE pairs distant from each other in the molecule. This qualitative assessment of the spatial distributions of IPE pairs of eigenvalues is being investigated in our laboratory in a more quantitative manner. At this time, we can roughly group, for most organic compounds in the 300−700 amu range of molecular weight, eigenvalues $\epsilon_j(x,y)$ as capturing IPE x,y pair interaction distributions which are short-range if $j = 1-5$, medium-range for $j = 6-14$, and long-range $j > 14$. Further, short-range corresponds to IPE pairs distributed in the 3−5 Å range of separation; medium-range IPE pairs are distributed over the 7−9 Å range of separation, and long-range IPE pairs are distributed at distances of more than 14 Å from one another. The gaps in the distance ranges given above for short-, medium-, and long-range interactions reflect our current uncertainty in the relationships of eignevalue number and corresponding separation distance.

Short-range eigenvalues for clusters 1A, 1B, and 3 all involve hydrogen-bond donors. This eigenvalue IPE type may be indicative of a common inhibitor hydrogen-bonding site influencing inhibition potency on all inhibitor classes captured within these three clusters. The presence of a polar positive IPE with the hydrogen-bond donor IPE among the short-range eigenvalues also increases inhibition potency. However, the presence of a second hydrogen-bond donor

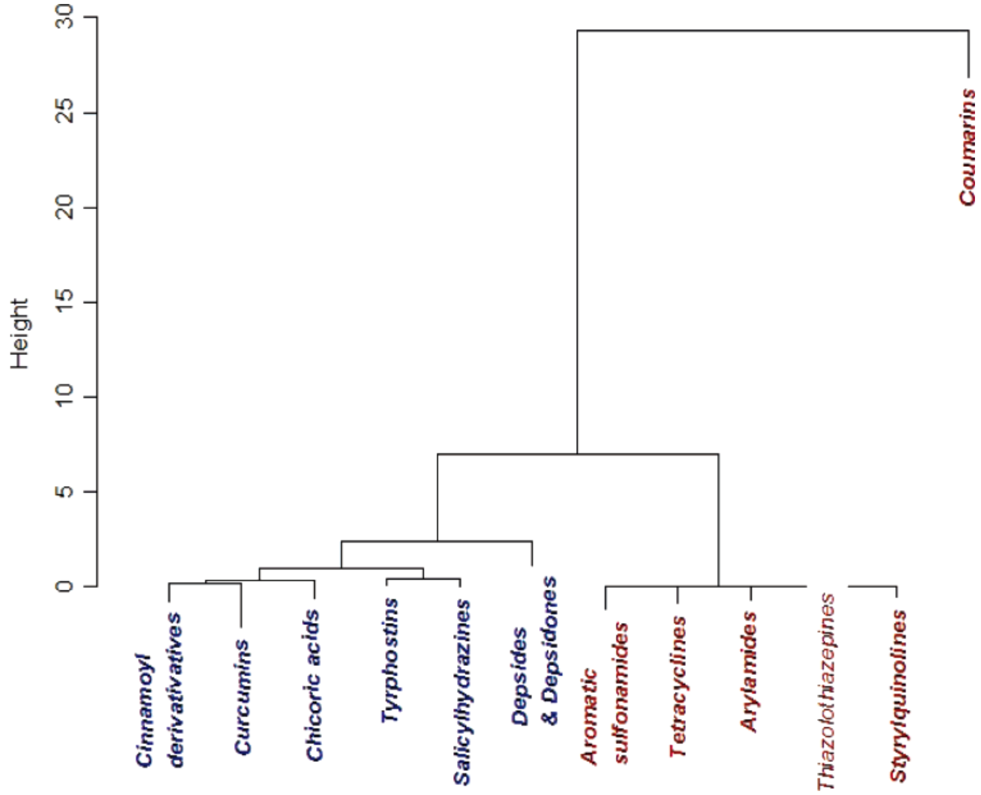


Figure 1. Dendrogram or tree structure of the hierarchical clustering of the data set based on aromatic IPE 4D fingerprints.

Table 18. 4D Fingerprint and Classic Descriptor QSAR Models for the Data Clusters Identified by PAM and Hierarchical Clustering

cluster #	*N*	model	R^2	Q^2
1	77	$pIC_{50} = 6.29 - 15.74\epsilon_7(hs,any) + 21.18\epsilon_6(p+,any) + 0.06Ecoh + 0.02Kappa4 - 3.01\epsilon_2(hs,hbd) - 96.91\epsilon_{30}(np,any) + 1.67\epsilon_1(p+,hbd) - 361.86\epsilon_{13}(any,any)$	0.65	0.60
1A	49	$pIC_{50} = 0.71 - 26.34\epsilon_8(any,any) - 286.42\epsilon_{19}(hs,any) + 0.10Ecoh + 12.97\epsilon_4(hs,p-) + 230.71\epsilon_{19}(np,any) - 2.68\epsilon_2(hbd,hbd) - 33.47\epsilon_8(np,p-)$	0.78	0.71
1B	19	$pIC_{50} = 4.29 + 30.63\epsilon_4(hbd,aro) + 5.75\epsilon_1(p+,hbd) - 5.37\epsilon_1(p+,aro)$	0.95	0.91
1C	9	$pIC_{50} = -4.91 + 208.71\epsilon_{13}(hs,np) + 57.11\epsilon_3(any,any)$	0.98	0.95
2	20	$pIC_{50} = 3.18 + 0.43Chi8 + 103.26\epsilon_{32}(hs,np)$	0.93	0.89
3	51	$pIC_{50} = 5.45 + 4.87\epsilon_2(p+,hbd) + 29.39\epsilon_5(np,aro) + 275.86\epsilon_{39}(np,np) - 147.75\epsilon_{11}(np,np) - 94.56\epsilon_{13}(aro,any) - 9.23\epsilon_3(hs,p-)$	0.88	0.83

IPE decreases pIC_{50}. It is not possible to discern if this key hydrogen-bond donor site involves intramolecular (within the inhibitor) or intermolecular (integrase−inhibitor) hydrogen bonding.

Medium-range eigenvalues of clusters 1A, 1B, and 3 involve nonpolar (recognizing that aromatic groups are nonpolar) IPEs. However, there is high variability in the other IPEs of the medium-range eigenvalues. This could mean that the spatial distribution of nonpolar groups in the 7−9 Å range is a common feature of binding to integrase across the classes of inhibitors in clusters 1A, 1B, and 3, but other interactions specific to the inhibitors of a given cluster are also at play. Finally, only cluster 3 has a long-range eigenvalue with specific IPE types, again nonpolars. This finding would suggest that the QSAR model for inhibitors in cluster 3 is the best-defined model in terms of capturing the entire inhibitor structure. Overall, the spatial distributions of nonpolar groups over the inhibitors of clusters 1A, 1B, and 3 appear important to specifying both the alignment and pose in binding to integrase.

Table 22 indicates that the inhibitors in cluster 1C (depsides and depsidones) and cluster 2 (coumarins) have QSAR models with no eigenvalue descriptors which only involve specific IPE (atom) types. However, both the R^2 and Q^2 values of each of these QSAR models indicate that the models are significant and robust. Two interpretations of the binding features of these inhibitor classes seem plausible. First, these inhibitors may bind in a similar manner to integrase as that of the inhibitors in clusters 1A, 1B, and 3. Second, the inhibitors of clusters 1C and 2 bind distinctly to integrase with respect to one another, and also with respect to the inhibitors of clusters 1A, 1B, and 3. The latter possibility seems more likely on the basis of the markedly high chemical diversities of these inhibitors with respect to one another as well as the other chemical classes of inhibitors. This distinct binding by inhibitors of clusters 1C and 2 could involve different binding sites on integrase, instead of the common site suggested by the QSAR models for the inhibitors of clusters 1A, 1B, and 3. However, markedly different alignments and/or poses at the same integrase binding site as clusters 1A, 1B, and 3 cannot be eliminated from an analysis of the descriptors of the QSAR models.

Overall, a binding "picture" emerges from Table 22, showing that inhibitor classes in clusters 1A, 1B, and 3 bind at a common location on integrase characterized by a

Table 19. Cross-Correlation Descriptor Matrix for Cluster 1A and Cluster 3

	ϵ_8(any,any)	ϵ_{19}(hs,any)	ϵ_4(hs,p−)	ϵ_{19}(np,any)	ϵ_2(hbd,hbd)	ϵ_8(np,p−)
ϵ_8(any,any)	1.000					
ϵ_{19}(hs,any)	0.394	1.000				
ϵ_4(hs,p−)	0.002	0.163	1.000			
ϵ_{19}(np,any)	0.210	0.652	0.012	1.000		
ϵ_2(hbd,hbd)	0.058	0.000	0.021	0.002	1.000	
ϵ_8(np,p-)	0.641	0.241	0.015	0.053	0.049	1.000

	ϵ_2(p+,hbd)	ϵ_5(np,aro)	ϵ_{39}(np,np)	ϵ_{11}(np,np)	ϵ_{13}(aro,any)	ϵ_3(hs,p−)
ϵ_2(p+,hbd)	1.000					
ϵ_5(np,aro)	0.032	1.000				
ϵ_{39}(np,np)	0.016	0.202	1.000			
ϵ_{11}(np,np)	0.003	0.150	0.157	1.000		
ϵ_{13}(aro,any)	0.059	0.016	0.078	0.199	1.000	
ϵ_3(hs,p−)	0.000	0.023	0.098	0.054	0.003	1.000

Table 20. Experimental and Predicted Inhibition Potencies of the Test Compounds in All Inhibitor Structural Classes

		pIC$_{50}$				pIC$_{50}$	
cluster	#	exp.	pred.	cluster	#	exp.	pred.
Cinnamoyl Derivatives				Tetracyclines			
1A	T1	6.70	6.14	3	T35	4.55	5.70
1A	T2[a]	6.05	5.05	3	T36	5.46	5.23
1A	T3	6.00	5.24	3	T37	5.89	5.75
1A	T4	3.50	3.48	3	T38	5.80	6.24
1A	T5	5.52	5.46	Arylamides and naphthalenes			
1A	T6	3.50	3.38	3	T39	6.01	6.65
1A	T7	4.22	4.53	3	T40	4.27	5.19
1A	T8	5.70	4.97	Thiazolothiazepines			
1A	T9	5.10	4.40	3	T41	3.87	3.67
1A	T10	4.17	4.30	3	T42	4.28	3.78
Tyrphostins				3	T43	3.67	3.94
1B	T11	5.72	6.52	3	T44	3.35	3.49
1B	T12	6.10	6.42	3	T45	4.07	4.06
1B	T13	5.48	6.39	3	T46	4.04	3.70
1B	T14	6.35	5.58	Curcumins			
Coumarins				1A	T47[a]	3.85	2.15
2	T15	4.36	4.02	1A	T48	5.05	4.82
2	T16	4.09	4.10	Salicylhydrazines			
2	T17	5.38	4.98	1B	T49	3.74	3.73
2	T18	4.71	4.59	1B	T50	3.61	3.81
2	T19	4.98	4.89	1B	T51	3.90	3.74
2	T20*	3.73	4.59	1B	T52	6.22	5.98
2	T21	4.27	4.41	1B	T53	6.10	6.06
2	T22	4.46	4.63	1B	T54	5.85	6.05
2	T23	5.13	4.69	1B	T55	6.10	6.04
Aromatic Sulfonamides				Styrylquinolines			
3	T24	4.62	4.82	3	T56	5.39	5.85
3	T25	4.12	4.36	3	T57	5.49	5.50
3	T26	3.89	4.00	3	T58	5.43	5.38
3	T27	4.31	4.12	3	T59	6.15	5.66
3	T28	3.71	4.57	3	T60	5.89	5.96
Chicoric Acids				3	T61	5.64	5.56
1A	T29	5.96	5.27	Depsides and Depsidones			
1A	T30	4.61	4.48	1C	T62	5.27	5.52
1A	T31	5.38	4.54	1C	T63	4.29	4.05
1A	T32	5.00	5.80	1C	T64	4.17	3.55
1A	T33	5.55	5.36	1C	T65	5.44	5.30
1A	T34	5.17	5.26				

[a] Outlier eliminated for calculating $R_{\text{prediction}}^2$

Table 21. $R_{\text{prediction}}^2$ of Predictions of the Test Sets Using the QSAR Models of the Five Clusters

QSAR model for cluster	test set clusters				
▼	1A	1B	1C	2	3
1A	**0.67**[a]	−0.11	−7.6	−0.38	−4.26
1B	−2.15	**0.80**	−6.95	−3.65	−3.89
1C			**0.62**		
2	−1.1	−1.25	−1.47	**0.64**[a]	−0.33
3	−0.81	−3.97	−15.84	−4.9	**0.74**

[a] Outliers eliminated for calculation of $R_{\text{prediction}}^2$

common hydrogen-bond donor site across the inhibitors. These inhibitors also adopt similar alignments and poses at this binding site as judged by the common medium-range nonpolar contributions to binding. Differences in the poses and alignments of the inhibitors between these three clusters may arise from differences in the medium-range spatial distributions of a variety of IPEs. There is a lack of information regarding specific IPEs, which is needed to define the binding site(s) of the inhibitors in clusters 1C and 2, even though their QSARs are significant.

These combined cluster and QSAR analyses, using 4D fingerprints and classic QSAR descriptors, of a large set of HIV-1 integrase inhibitors support the emerging view that the 4D fingerprints are, indeed, universal descriptors that can be meaningfully applied to any QSAR/QSPR study. All cluster and QSAR models found in this study are completely, or nearly completely, composed of 4D-fingerprint descriptors. Moreover, none of the QSAR models are due to random chance, and the QSAR models for all clusters, except cluster 1, are significant, predictive, and robust as measured by their high R^2 and Q^2 values (see Table 18). The cluster 1 QSAR model has marginal R^2 and Q^2 values, indicating that it is not particularly predictive. But given that Q^2 is close in value to R^2, the model is likely quite stable, and therefore, it likely identifies significant descriptors reflecting inhibitor−integrase binding qualitatively.

The presence of Ecoh in QSAR models for clusters 1 and 1A is not readily explained. However, it is noted that Ecoh generally increases with an increasing size of a molecule, reflecting the packing energy of the entire molecule with other like molecules. The regression coefficients for Ecoh in both QSAR models for clusters 1 and 1A are positive, so that inhibition potency is predicted to increase with increasing Ecoh. In composite, these observations suggest that Ecoh is a generalized ligand−receptor binding energy term, capturing mainly the size of the inhibitor on pIC$_{50}$. The connectivity descriptors, kappa-4 in the QSAR of cluster 1 and chi-8 in the QSAR of cluster 2, may reflect the efficient representation of a particular bonding topology feature of the classes of inhibitors in these clusters, which otherwise could only be captured by two or more 4D fingerprints.

Table 22. Distribution of Common Key 4D Fingerprint Pharmacophore Groups over the Integrase Inhibitors[a]

cluster	inhibitor class	short-range 3−5 Å	medium-range 7−9 Å	long-range >11 Å
1A	cinnamoyl chicoric curcumins	[−] hydrogen bond donors	[−] nonpolar and polar negative	
1B	tyrphostins salicylhydrazines	[+] hydrogen-bond donors and polar positive [−] aromatic and polar positive	[+] hydrogen bond donors and aromatic	
1C	depsides depsidones			
2	coumarins aromatic sulfonamides			
3	tetracyclines arylamides & naphtalenes thiazolothiazepines styrylquinolines	[+] hydrogen-bond donors and polar positive	[+] nonpolar and aromatic [−] nonpolars	[+] nonpolars
1	members of 1A, 1B, and 1C	[+] hydrogen bond donors and polar positive		

[a] Short-range, medium-range, and long-range refer to the average range of distances between the listed IPEs [pharmacophore atom types] identified in the QSAR models. The sign in brackets next to the IPE type(s) indicates whether the indicated IPEs increase [+] or decrease [−] inhibition potency.

One advantage of the 4D-fingerprint descriptors lies in the difference between a "classical" concept of molecular similarity and that of 4D molecular similarity (4D-MS).[27] The classical concept limits an estimate of molecular similarity to a static structure of the entire molecule. 4D-MS permits similarity comparisons based on all possible combinations of 3D-pharmacophore features inherent to a molecule sampled over its set of energetically accessible conformer states. The importance of such a distinction comes especially into play when the entire molecular structure is not involved in eliciting a biological response, as is true of most ligand−receptor binding modes. In addition, 4D-MS permits the measurement of molecular similarity in multiple ways based on all pair combinations of the eight IPE types. 4D fingerprints also have the major advantage of incorporating three-dimensional shape and conformational information about molecules but yet are independent of alignment constraints. Since an ensemble of conformational states is included in determining the 4D-MS eigenvectors, these descriptors also capture information about molecular flexibility and conformational entropy.

However, a drawback to the 4D-fingerprint descriptors is that they are abstract and do not readily lend themselves to precise interpretation in a QSAR model. Visualizing 4D fingerprints in Euclidean space is difficult despite that they originate from the 3D structures of molecules. Work is in progress to better visualize and interpret the 4D fingerprints. We have used the characterization of distributions of distances between the IPE pairs of the 4D fingerprints in this study to help define plausible spatial pharmacophore features of the clustered sets of integrase inhibitors.

Moreover, in this application, the 4D fingerprints of structurally diverse data sets expressing a common endpoint action have been successfully used in data clustering as well as data fitting. Overall, the use of a combined clustering and QSAR modeling methodology, employing 4D fingerprints and classic QSAR descriptors, may be a useful and reliable strategy in the quantitative analysis of structurally diverse data sets.

ACKNOWLEDGMENT

This work was funded by the National Institutes of Health through the NIH Roadmap for Medical Research, Grant 1 R21 GM075775-01. Information on Novel Preclinical Tools for Predictive ADME-Toxicology can be found at http://grants.nih.gov/grants/guide/rfa-files/RFA-RM-04-023.html. Links to nine initiatives are found at http://nihroadmap.nih.gov/initiatives.asp. This work was also supported in part by The Procter & Gamble Company. Resources of the Laboratory of Molecular Modeling and Design at UIC, and at The Chem21 Group, Inc., were used in performing these studies.

REFERENCES AND NOTES

(1) Beresford, A. P.; Selick, H. E.; Tarbit, M. H. The emerging importance of predictive ADME simulation in drug discovery. *Drug Discovery Today* **2002**, *7*, 109−116.

(2) Van de Waterbeemd, H.; Gifford, E. ADMET *in silico* modelling: Towards prediction paradise? *Nat. Rev. Drug Discovery* **2003**, *2*, 192−204.

(3) Yu, H.; Adedoyin, A. ADME/Tox in drug discovery: integration of experimental and computational technologies. *Drug Discovery Today* **2003**, *8*, 852−861.

(4) Esposito, E. X.; Hopfinger, A. J.; Madura, J. D. Methods for applying the quantitative structure-activity relationship paradigm. In *Methods in Molecular Biology*; Bajorath, J., Ed.; Humana Press: Totowa, NJ, 2004; pp 131−218.

(5) Senese, C. L.; Hopfinger, A. J. A simple clustering technique to improve QSAR model selection and predictivity: application to a receptor independent 4D-QSAR analysis of cyclic urea derived inhibitors of HIV-1 protease. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2180−2192.

(6) Johnston, M. I.; Hoth, D. F. Present status and future prospects for HIV therapies. *Science* **1993**, *26*, 1286−1293.

(7) Brown, P. O. Integration. In *Retroviruses*; Coffin, J. M., Hughes, S. H., Varmus, H. E., Eds.; Cold Spring Harbor Lab. Press: Plainview, New York, 1997; pp 161−203.

(8) Pommier, Y.; Johnson, A. A.; Marchand, C. Integrase inhibitors to treat HIV/AIDS. *Nat. Rev. Drug Discovery* **2005**, *4*, 236−248.

(9) Goldgur, Y.; Craigie, R.; Cohen, G. H.; Fujiwara, T.; Yoshinaga, T.; Fujishita, T.; Sugimoto, H.; Endo, T.; Murai, H.; Davies, D. R. Structure of the HIV-1 integrase catalytic domain complexed with an inhibitor: A platform for antiviral drug design. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 13040−13043.

(10) Barreca, M. L.; Lee, K. W.; Chimirri, A.; Briggs, J. M. Molecular dynamics studies of the wild-type and double mutant HIV-1 Integrase combined with the 5CITEP inhibitor: Mechanism for inhibition and drug resistance. *Biophys. J.* **2003**, *84*, 1450−1463.

(11) Marchand, C.; Zhang, X.; Pais, G. C. G.; Cowansage, K.; Neamati, N.; Burke, T. R.; Pommier, Y. Structural determinants for HIV-1 integrase inhibition by β-diketo acids. *J. Biol. Chem.* **2002**, *277*, 12596−12603.

(12) Artico, M.; Di Santo, R.; Costi, R.; Novellino, E.; Greco, G.; Massa, S.; Tramontano, E.; Marongiu, M. E.; De Montis, A.; Colla, P. L. Geometrically and conformationally restrained cinnamoyl compounds as inhibitors of HIV-1 integrase: synthesis, biological evaluation and molecular modeling. *J. Med. Chem.* **1998**, *41*, 3948−3960.

(13) Burke, T. R., Jr.; Fesen, M. R.; Mazumdar, A.; Wang, J.; Carothers, A. M.; Grunberger, D.; Driscoll, J.; Kohn, K.; Pommier, Y. Hydroxylated aromatic inhibitors of HIV Integrase. *J. Med. Chem.* **1995**, *38*, 4171−4178.

(14) Mazumdar, A.; Gazit, A.; Levitzki, A.; Nicklaus, M.; Yung, J.; Kohlhagen, G.; Pommier, Y. Effect of tyrphostins, protein kinase inhibitors on Human Immunodeficiency virus type 1 integrase. *Biochemistry* **1995**, *35*, 15111−15122.

(15) Zhao, H.; Neamati, N.; Hong, H.; Mazumdar, A.; Wang, S.; Sunder, S.; Milne, G. W.; Pommier, Y.; Burke, T. R., Jr. Coumarin-based inhibitors of HIV integrase. *J. Med. Chem*. **1997**, *40*, 242−249.

(16) Nicklaus, M. C.; Neamati, N.; Hong, H.; Mazumdar, A.; Sunder, S.; Chen, J.; Milne, G. W.; Pommier, Y. J. HIV-1 integrase pharmacophore: discovery of inhibitors through three-dimensional database searching. *J. Med. Chem*. **1997**, *40*, 920−929.

(17) Lin, Z.; Neamati, N.; Zhao, H.; Kiryu, Y.; Turpin, J. A.; Aberham, C.; Strebel, K.; Kohn, K.; Witvrouw, M.; Pannecouque, C.; Debyser, Z.; De Clercq, E.; Rice, W. G.; Pommier, Y.; Burke, T. R., Jr. Chicoric acid analogues as HIV-1 inetgrase inhibitors. *J. Med. Chem*. **1999**, *42*, 1401−1414.

(18) Neamati, N.; Hong, H.; Sunder, S.; Milne, G. W. A.; Pommier, Y. Potent inhibitors of human immunodeficiency virus type 1 integrase: identification of a novel four- point pharmacophore and tetracyclines as novel inhibitors. *Mol. Pharmacol.* **1997**, *52*, 1041−1055.

(19) Zhao, H.; Neamati, N.; Mazumdar, A.; Sunder, S.; Pommier, Y.; Burke, T. R., Jr. Arylamide inhibitors of HIV-1 integrase. *J. Med. Chem.* **1997**, *40*, 1186−1194.

(20) 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. Thiazolothiazepine inhibitors of HIV-1 integrase. *J. Med. Chem.* **1999**, *42*, 3334−3341.

(21) Mazumdar, A.; Neamati, N.; Sunder, S.; Schulz, J.; Pertz, H.; Eich, E.; Pommier, Y. Curcumin analogs with altered potencies against HIV-1 integrase as probes for biochemical mechanisms of drug action. *J. Med. Chem.* **1997**, *40*, 3057−3063.

(22) 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. Salicylhydrazine-containing inhibitors of HIV-1 integrase: implication for a selective chelation in the integrase active site. *J. Med. Chem.* **1998**, *41*, 3202−3209.

(23) Zouhiri, F.; Mouscadet, J. F.; Mekouar, K.; Desmaele, D.; Savoure, D.; Leh, H.; Subra, F.; Le Bret, M.; Auclair, C.; d'Angelo, J. Structure-activity relationships and binding mode of styrylquinolines as potent inhibitors of HIV-1 integrase and replication of HIV-1 in cell culture. *J. Med. Chem.* **2000**, *43*, 1533−1540.

(24) Neamati, N.; Hong, H.; Mazumdar, A.; Wang, S.; Sunder, S.; Nicklaus, M. C.; Milne, G. W.; Proksa, B.; Pommier, Y. Depsides and depsidones as inhibitors of HIV-1 integrase: discovery of novel inhibitors through 3D database searching. *J. Med. Chem.* **1997**, *40*, 942−951.

(25) *HyperChem*, HyperChem Program Release 6.0 for MS Windows; Hybercube, Inc.: Gainesville, FL, 1996.

(26) *4D-QSAR and 4D-QSAR-MS*, 4D-QSAR molecular similarity program version 3.0; The Chem21 Group Inc.: Lake Forest, IL, 2001.

(27) Duca, J. S.; Hopfinger, A. J. Estimation of molecular similarity based on 4D- QSAR analysis: Formalism and validation. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1367−1387.

(28) Doherty, D. C. *Molsim Version 3.2 User's Guide*; The Chem21 Group Inc.: Lake Forest, IL, 2001; pp 18−26.

(29) *ClogP Daylight Chemical Information Software*, version 4.51; Daylight Chemical Information Inc.: Los Altos, CA, 1998.

(30) Iyer, M.; Mishra, R.; Han, Y.; Hopfinger, A. J. Predicting blood-brain barrier partitioning of organic molecules using membrane-interaction QSAR analysis. *Pharm. Res.* **2002**, *19*, 1611−1620.

(31) Kaufman, L.; Rousseeuw, P. J. *Finding Groups in Data: An Introduction to Cluster Analysis*; Wiley-Interscience, New York, 1990; pp 102−156.

(32) Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53−65.

(33) Pan, D.; Iyer, M.; Liu, J.; Li, Y.; Hopfinger, A. J. Constructing optimum blood- brain barrier QSAR models using a combination of 4D-molecular similarity measures and cluster analysis. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2083−2098.

(34) Yuan, H.; Parrill, A. L. QSAR studies of HIV-1 integrase inhibition. *Bioorg. Med. Chem.* **2002**, *10*, 4169−4183.

(35) *S-plus-6 for Windows*, ver. 1.0; Insightful Corp.: Seattle, WA, 2001.

(36) Rogers, D.; Hopfinger, A. J. Applications of genetic function approximation to quantitative structure-activity relationships and quantitative structure-property relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854−866.

(37) Rogers, D. *WOLF 6.2 GFA Program*, ver.2.1; Accelrys: San Diego, CA, 1994.

(38) Hall, L. H.; Kier, L. B. The molecular connectivity chi indexes and kappa shape indexes in structure-property modeling. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH: New York, 1991; Vol. 2, pp 367−422.

(39) Fernández, M.; Caballero, J. Bayesian-regularized genetic neural networks applied to the modeling of non-peptide antagonists for the human luteinizing hormone-releasing hormone receptor. *J. Mol. Graphics Modell.* **2006**, *25*, 410−422.

CI700153G