

## Eigen Value Analysis of HIV-1 Integrase Inhibitors

Mahindra T. Makhija and Vithal M. Kulkarni\*

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

Received September 27, 2000

A three-dimensional quantitative structure activity relationship using the eigen value analysis (EVA) paradigm applied to 41 HIV-1 integrase inhibitors that inhibit integrase mediated cleavage (3'-processing step) and integration (3'-strand transfer step) *in vitro* was performed. The training set consisted of 35 molecules from five structurally diverse classes: salicylhydrazines, lichen acids, coumarins, quinones, and thiazolothiazepines. Models derived using semiempirical (MOPAC AM1 and PM3) calculated normal-mode frequencies were compared. The predictive ability of each resultant model was evaluated using a test set comprised of six molecules belonging to a different structural class: hydrazides. Models derived using AM1 method showed considerable internal as well as external predictivity ( $r^2_{cv} = 0.806$ ,  $r^2_{pred} = 0.761$  for 3'-processing and  $r^2_{cv} = 0.677$ ,  $r^2_{pred} = 0.591$  for 3'-strand transfer).

### INTRODUCTION

Acquired immunodeficiency syndrome (AIDS) is the most devastating pandemic in recent history of the mankind. It portends an increasing toll in human suffering and is a major hurdle in the economic progress of any developing country. Current approaches for the treatment of AIDS using single agents are plagued by the development of resistance. Combination therapies employing multiple components, each directed against different viral enzymes, may potentially provide an effective means of countering such resistance. The three viral enzymes encoded by the *pol* gene of human immunodeficiency virus (HIV) play key roles in the viral replication cycle. Two of these enzymes, reverse transcriptase and protease, have been the focus of intense research as targets for chemotherapeutic intervention and currently provide the basis for most anti-AIDS therapies.<sup>1,2</sup> HIV integrase (HIV-1 IN) is recognized as an important addition to the list of enzymes whose inhibition may be efficacious in anti-AIDS therapy, since this enzyme is required for viral replication, yet it is not indigenous to the human host.<sup>3–5</sup>

Retroviruses encode the integrase protein at the 3'-end of the *pol* gene. This enzyme, an HIV protease cleavage product of the *gag-pol* fusion protein precursor, catalyzes the integration of a double stranded DNA copy of the RNA genome, synthesized by the reverse transcriptase, into a host chromosome in a two-step reaction. First, integrase cleaves the last two nucleotides from each 3'-end of the linear viral DNA, leaving the terminal dinucleotide CA-3'-OH. This activity is referred to as 3'-processing or dinucleotide cleavage. Second, after transport to the nucleus as a nucleoprotein complex, integrase catalyzes a DNA strand transfer reaction involving a nucleophilic attack from the cleaved 3'-ends to a host chromosome. This process is referred to as strand transfer. Finally, the viral 5'-ends are processed, and the gaps between the viral 5'- and target 3'-ends are repaired.<sup>6,7</sup>

Several classes of inhibitors of HIV-1 IN have been reported to date; none has yet proven to be highly selective for IN and useable for therapeutic development.<sup>8,9</sup> For example, a majority of catechol-containing compounds have been demonstrated to possess potency against the 3'-processing and 3'-strand transfer reactions catalyzed by IN.<sup>9</sup> However, despite such promising initial results, compounds containing this moiety can also cross-link protein,<sup>10</sup> chelate metal,<sup>11,12</sup> and thus generally lack the desired selectivity.<sup>8,9</sup> As opposed to the catechol-containing compounds, non-catechol-containing structures are excellent leads to develop a selective potent IN inhibitor, for they possess considerably less cytotoxicity. Pommier et al. have identified several such non-catechol-containing compounds such as salicylhydrazines,<sup>11</sup> lichen acids,<sup>12</sup> coumarins,<sup>13</sup> quinones,<sup>14</sup> hydrazides,<sup>15</sup> thiazolothiazepines,<sup>16</sup> etc. that inhibit IN function at low micromolar concentrations.

To obtain further insight into the relationship between the structures of the aforementioned classes of compounds and their HIV-1 IN inhibitory activities, three-dimensional quantitative structure activity relationship (3D QSAR) studies have been performed. Perhaps the most well-known of the 3D QSAR techniques is comparative molecular field analysis (CoMFA), developed by Cramer et al. in 1988. CoMFA method revolves around an assumption that the interaction between an inhibitor and its molecular target is primarily noncovalent in nature and shape dependent. Therefore, a QSAR may be derived by sampling the steric and electrostatic fields surrounding a set of ligands and correlating the differences in these fields to biological activity.<sup>17–20</sup> One of the major problems to be overcome when applying CoMFA is that of aligning the structures concerned. Molecular alignment can be relatively straightforward if a set of rigid structural analogues is to be analyzed but becomes increasingly problematic as the diversity of the dataset increases, even if no account is taken of conformational flexibility. One such class is that of HIV-1 IN inhibitors which presents the tremendous diversity of molecular structures. The overall

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

orientation of the structures (taken as a rigid body) within the lattice can also substantially alter modeling statistics in CoMFA. Although, comparative molecular similarity analysis (CoMSIA)<sup>21–24</sup> developed by Klebe et al. is insensitive to the relative orientation of molecules with respect to the lattice, there nonetheless remain substantial difficulties associated with the use of grids and the need for alignment.

There is, therefore, considerable interest in the development of new descriptors of molecular structure that do not require the alignment of molecules but retain the 3D and molecular property information encoded within the molecular fields. Alternative descriptions of molecular fields than those used in CoMFA or molecular surface properties, for example, methods based on autocorrelation vectors,<sup>25</sup> molecular moments,<sup>26</sup> or MS-WHIM descriptors,<sup>27</sup> may provide effective orientation-independent descriptions of molecular structure. In the present paper, a recently described alignment-free descriptor of molecular structure, known as EVA (Eigen-Value descriptor), that is derived from calculated infrared (IR) range vibrational frequencies, has been used to throw some light on the structural requirements of a diverse set of HIV-1 IN inhibitors.<sup>28</sup>

### THE EVA DESCRIPTOR<sup>29</sup>

EVA is a vector descriptor based on EigenValues corresponding to individual molecular vibrational modes. It was developed at Shell Laboratories and has been extensively evaluated by scientists at the University of Sheffield for use in analyzing QSARs.

Normal coordinate analysis (NCA), generally in MOPAC or some other molecular mechanics package, is used to calculate normal modes of vibration for an input structure. These are used to construct a vibrational profile between 0 and 4000 cm<sup>-1</sup> by summing (convoluting) a series of Gaussian distributions, with one centered on each normal mode  $f_i$ . The resulting profile differs from an infrared spectrum in that all bandwidths and intensities are the same for all of the component Gaussians. This common bandwidth is defined by the parameter  $\sigma$  (the resolution or half-width at half-height), which must be provided for each EVA vector; a value of 10 cm<sup>-1</sup> is typical.

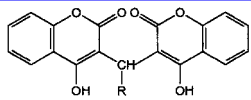
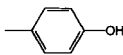
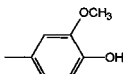
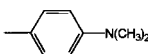
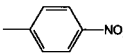
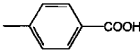


The intensity  $I$  at each frequency  $\nu$  is given by

$$I = \sum_i \frac{e^{-(\nu - f_i)^2 / 2\sigma^2}}{\sigma \cdot \sqrt{2\pi}}$$

EVA vectors are then built by sampling this artificial vibrational profile at intervals across this profile. The distance between sampled frequencies ( $\delta$ ) is also an important EVA parameter. This interval should always be less than  $2\sigma$ . A smaller value of  $\delta$  is generally desirable, however.<sup>30,31</sup>

The EVA descriptor has been shown to be conformationally sensitive and as such can be considered to be a 3D descriptor but with the advantage over CoMFA that structural superposition is not required.<sup>32</sup> However, while EVA removes the need for superposition, the method is sensitive to 3D structure, although not to such an extent as a “true” 3D method such as CoMFA. This reduced sensitivity is a

**Table 1.** Structures and Activities of Coumarins Used in the Training Set (from Ref 13)

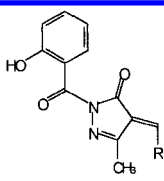
			
Sr.No.	R	3'-Processing IC <sub>50</sub> (mM)	3'-Strand transfer IC <sub>50</sub> (mM)
1		0.134	0.074
2		0.187	0.092
3		0.094	0.069
4		0.054	0.018
5		0.057	0.051
6		0.105	0.148
7		0.054	0.122


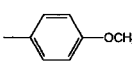
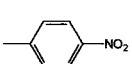

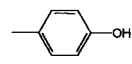


consequence of a Gaussian smearing function to develop the descriptor and as a result EVA might be described as a “2<sup>1/2</sup>D” descriptor.<sup>30</sup> Normally, the range chosen for EVA is 0–4000 cm<sup>-1</sup> to encompass the frequencies of all fundamental molecular vibrations.<sup>31–34</sup> EVA descriptors are somewhat dependent on the conformation used in their calculation, particularly below 1000 cm<sup>-1</sup>, but much less so than is CoMFA. The frequencies below 200 cm<sup>-1</sup> do not contribute to the model significantly, and to prove this most of the calculations were done with a range of 200–4000 cm<sup>-1</sup>. Only one analysis was done using 0–4000 cm<sup>-1</sup> range to know the importance of frequencies below 200 cm<sup>-1</sup> in model derivation. For the purpose of this study, AM1 and PM3 Hamiltonia have been used to calculate normal modes of vibration for an input structure.<sup>33,34</sup>

### MATERIALS AND METHODS

**Data Set for Analysis.** Published in vitro biological data for both 3'-processing and 3'-endjoining (strand transfer) on a series of HIV-1 integrase inhibitors were used for this study.<sup>11–16</sup> The structures and biological activities of the 35 molecules constituting the training set in the QSAR analyses are given in Tables 1–5, respectively.

The test set consists of six molecules (Table 6). These compounds belong to an entirely different structural class and were selected to remove any bias in testing the predictivity of the models.

**Table 2.** Structures and Activities of Salicylhydrazines Used in the Training Set (from Ref 11)


Sr.No.	R	3'-Processing IC <sub>50</sub> (mM)	3'-Strand transfer IC <sub>50</sub> (mM)
8		0.0006	0.0028
9		0.0009	0.0006
10		0.0008	0.0006
11		0.0014	0.0026
12		0.0006	0.0009
13		0.0009	0.0074
14		0.0027	0.0020

All biological activities used in the present study were expressed as

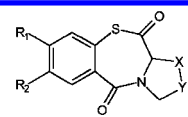
$$\text{pIC}_{50} = -\log \text{IC}_{50}$$



where IC<sub>50</sub> is the millimolar concentration of the inhibitor producing 50% inhibition.

**Computational Approaches.** All molecular modeling techniques, EVA studies described herein were performed on Silicon Graphics INDY R5000 workstation using the SYBYL 6.6 molecular modeling software from Tripos, Inc., St. Louis, MO.<sup>35</sup>

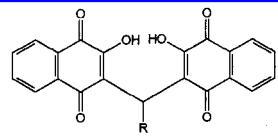
**Molecular Conformation.** The compounds were built from fragments in the SYBYL database. Each structure was fully geometry optimized using the standard Tripos force field<sup>36</sup> with a distance dependent-dielectric function and a 0.001 kcal/mol energy gradient convergence criterion. Partial atomic charges were computed by a semiempirical molecular orbital method using the MOPAC 6.0 program.<sup>37</sup> The charges were computed using the PM3 model Hamiltonian (key-words: 1SCF, RHF, MMOK).<sup>38</sup>


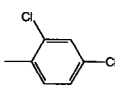
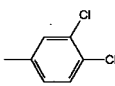
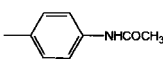

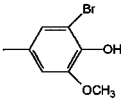
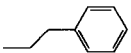
Using the systematic search protocol, rotatable bonds in compounds were searched from 0 to 360° in 10° increments. The low energy conformations of these compounds thus obtained were minimized using Tripos force field and

**Table 3.** Structures and Activities of Thiazolothiazepines Used in Training Set (from Ref 16)


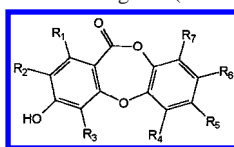
Sr.No.	R <sub>1</sub>	R <sub>2</sub>	-X-Y-	3'-Processing IC <sub>50</sub> (mM)	3'-Strand transfer IC <sub>50</sub> (mM)
15	H	Cl	-S-CH <sub>2</sub> -	0.128	0.090
16	H	Br	-S-CH <sub>2</sub> -	0.058	0.048
17	H	CH <sub>3</sub>	-S-CH <sub>2</sub> -	0.064	0.055
18	H	OCH <sub>3</sub>	-CH <sub>2</sub> -S-	0.215	0.200
19	OCH <sub>3</sub>	OCH <sub>3</sub>	-CH <sub>2</sub> -S-	0.650	0.331
20 <sup>a</sup>			-S-CH <sub>2</sub> -	0.040	0.047
21 <sup>a</sup>			-CH <sub>2</sub> -S-	0.092	0.100

<sup>a</sup> Fused ring system (naphthalene derivative).

**Table 4.** Structures and Activities of Quinones Used in the Training Set (from Ref 14)


Sr.No.	R	3'-Processing IC <sub>50</sub> (mM)	3'-Strand transfer IC <sub>50</sub> (mM)
22		0.068	0.048
23		0.037	0.040
24		0.090	0.052
25		0.086	0.078
26		0.092	0.060
27		0.032	0.020
28		0.083	0.090

subsequently used in the analysis. Prior minimization of structures using Tripos force field is sufficient, because in EVA before normal coordinate analysis, all structures are fully geometry optimized using the default set up i.e.,

**Table 5.** Structures and Activities of Lichen Acids Used in the Training Set (from Ref 12)

no.	R <sub>1</sub>	R <sub>2</sub>	R <sub>3</sub>	R <sub>4</sub>	R <sub>5</sub>	R <sub>6</sub>	R <sub>7</sub>	IC <sub>50</sub> <sup>a</sup>	IC <sub>50</sub> <sup>b</sup>
29	Me	H	CHO	Me	CO <sub>2</sub> H	OH	Me	0.0046	0.0065
30	Me	H	CHO	Me	CO <sub>2</sub> Me	OH	Me	0.0054	0.0044
31	Me	H	CHO	Me	CO <sub>2</sub> H	OH	CH <sub>2</sub> OCOCH:CHCO <sub>2</sub> H	0.0049	0.0046
32	CH <sub>2</sub> COC <sub>5</sub> H <sub>11</sub>	H	H	<i>n</i> -C <sub>5</sub> H <sub>11</sub>	CO <sub>2</sub> H	OH	H	0.0385	0.0309
33	COC <sub>4</sub> H <sub>9</sub>	H	H	<i>n</i> -C <sub>5</sub> H <sub>11</sub>	H	OH	H	0.0510	0.0336
34	Me	H	CHO	CO <sub>2</sub> H	H	OMe	Me	0.0160	0.0088
35	Me	Cl	CHO	Me	H	OMe	Me	0.0024	0.0021

<sup>a</sup> IC<sub>50</sub> (mM) values for 3'-processing activity. <sup>b</sup> IC<sub>50</sub> (mM) values for 3'-strand transfer activity.

ALL\_BONDS\_AND\_ANGLES with stopping criteria of GNORM = 0.05 and SCFCRT = 10<sup>-12</sup> with the corresponding method (AM1 or PM3) used for EVA calculations. This is the default definition file for EVA calculation in SYBYL 6.6.<sup>31</sup>

**Partial Least Squares (PLS) Analysis.** The PLS method has been applied successfully in numerous QSAR studies aiming to rationalize those structural features affecting biological activity. PLS regression seeks a relationship between Y and X, where vector Y is the response or dependent variable and X represents the descriptor data.<sup>39</sup>

PLS analyses were performed following the CoMFA standard implementation in SYBYL. To check statistical significance of the models, cross-validations were done by means of the "leave-one-out" (LOO) procedure using the enhanced version of PLS, the SAMPLS method.<sup>40</sup> The results from cross-validation analysis were expressed as the cross-validated *r*<sup>2</sup> value (*r*<sup>2</sup><sub>cv</sub>). The cross-validated *r*<sup>2</sup> is defined as

$$r_{cv}^2 = 1 - \text{PRESS} / \sum (Y - Y_{\text{mean}})^2$$

where PRESS =  $\sum (Y - Y_{\text{pred}})^2$ .

The optimal number of components was determined by selecting the smallest *s*<sub>press</sub> value. *s*<sub>press</sub> is the root mean Predictive Error Sum of Squares. It is an expected uncertainty in prediction for an individual compound based on the data available from other compounds in the set

$$s_{\text{press}} = (\text{PRESS} / (n - c - 1))^{1/2}$$

where *n* = number of rows and *c* = number of components. Usually the smallest *s*<sub>press</sub> value corresponds to the highest *r*<sup>2</sup><sub>cv</sub> value. The optimal number of components was subsequently used to derive the final QSAR models. Conventional analyses (no cross-validation) were performed without any scaling. The *r*<sup>2</sup><sub>cv</sub>, *s*<sub>press</sub>, *r*<sup>2</sup><sub>conv</sub>, and SE values were computed as defined in SYBYL. SE is the standard error of estimate. It is a measure of the target property uncertainty still unexplained after the QSAR has been derived. In Tables 7–9, Pr<sup>2</sup> = 0 means the probability of obtaining the observed *F*-ratio value (Fischer's significance test) by chance alone, if the target and the explanatory variables themselves are truly uncorrelated. When Pr<sup>2</sup> = 0 is zero, then results are not by chance and are significant.<sup>31</sup> Additionally to perform an even more rigorous statistical test, several runs of cross-validations using five groups were done in which each target

**Table 6.** Structures and Activities of Test Set Molecules (from Ref 15)

Sr.No.	R <sub>1</sub>	R <sub>2</sub>	R <sub>3</sub>	R <sub>4</sub>	3'-Processing pIC <sub>50</sub>	3'-Strand transfer pIC <sub>50</sub>
36		OH	H	H	2.684	3.136
37		OH	H	H	2.173	2.283
38		OH	H	H	1.096	1.420
39 <sup>a</sup>		OH			2.638	2.958
40		OH	H	H	2.040	2.236
41		SH	H	H	1.126	1.269

<sup>a</sup> Fused ring system (naphthalene derivative).

property value is predicted by a model based on about 4/5, or 80% of the available data.

In many QSAR analyses, confidence intervals (mean and standard deviation) for the parameters to be estimated can be calculated by a modern validation method, the bootstrap. The name is derived from the old saying about pulling yourself up by your own bootstraps. The idea is to simulate a statistical sampling procedure by assuming that the original data set is the true population and generating many new datasets from it. These new data sets (called bootstrap samplings) are of the same size as the original data set and are obtained by randomly choosing samples (rows) from the original data, repeated selection of the same row being allowed. The statistical calculation is performed on each of these bootstrap samplings, new values being calculated for each of the parameters to be estimated. The difference between the parameters calculated from the original data set and the average of the parameters calculated from the many

**Table 7.** Summary of EVA Results for 3'-Processing Using 200–4000 cm<sup>-1</sup> Frequency Range

	$\sigma = 3, \delta = 5$		$\sigma = 5, \delta = 5$		$\sigma = 10, \delta = 5$		$\sigma = 25, \delta = 5$	
	AM1	PM3	AM1	PM3	AM1	PM3	AM1	PM3
$r^2_{cv}$	0.806	0.812	0.808	0.835	0.846	0.869	0.864	0.814
compd	3	5	4	5	7	9	9	9
$s_{press}$	0.392	0.399	0.397	0.374	0.375	0.359	0.365	0.428
SEM	0.168	0.062	0.137	0.064	0.088	0.039	0.100	0.121
$r^2_{conv}$	0.964	0.995	0.977	0.995	0.992	0.998	0.990	0.985
$F$ -value	279.08	1275.3	322.37	1194.6	450.54	1794.5	271.75	184.81
$Pr^2 = 0$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
norm coeff of EVA	3.057	3.353	3.256	3.480	3.908	4.083	6.002	5.828
$r^2_{pred}$	0.761	0.521	0.697	0.597	0.716	0.707	0.615	0.508
$r^2_{bs}$	0.975	0.998	0.983	0.995	0.996	0.999	0.996	0.995
SD	0.011	0.001	0.008	0.002	0.002	0.001	0.002	0.003

**Table 8.** Summary of EVA Results for 3'-Strand Transfer Using 200–4000 cm<sup>-1</sup> Frequency Range

	$\sigma = 3, \delta = 5$		$\sigma = 5, \delta = 5$		$\sigma = 10, \delta = 5$		$\sigma = 25, \delta = 5$	
	AM1	PM3	AM1	PM3	AM1	PM3	AM1	PM3
$r^2_{cv}$	0.677	0.745	0.670	0.783	0.766	0.853	0.824	0.783
compd	2	6	3	14	7	15	8	9
$s_{press}$	0.441	0.419	0.452	0.457	0.409	0.386	0.361	0.408
SEM	0.252	0.028	0.252	0.001	0.087	0.001	0.131	0.121
$r^2_{conv}$	0.894	0.999	0.898	0.990	0.989	0.999	0.977	0.981
$F$ -value	135.47	4224.0	90.783	9.6e+06	355.87	766284.6	137.34	143.86
$Pr^2 = 0$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
norm coeff of EVA	2.857	3.755	2.884	4.018	4.317	4.815	5.777	5.928
$r^2_{pred}$	0.591	0.122	0.498	0.169	0.213	0.212	-0.72	0.296
$r^2_{bs}$	0.912	0.999	0.938	0.990	0.995	0.990	0.992	0.994
SD	0.029	0.001	0.022	0.0001	0.003	0.0001	0.005	0.004

bootstrap samplings is a measure of the bias of the original calculation.<sup>31</sup> In the present study, bootstrapping analysis (1000) runs was performed to assess the robustness and statistical confidence of the derived models.

A common test to check the consistency of the models is to scramble the biological data and repeat the model derivation process, allowing detection of possible chance correlations. After randomizing our data set in several distinct ways, in all cases we only observed very low or negative  $r^2_{cv}$  values in the PLS analyses.

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

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

where SD is the sum of the squared deviations between the biological activity of molecules in the test set and the mean biological activity of the training set molecules and PRESS is the sum of the squared deviations between predicted and actual activity values for every molecule in the test set. Like  $r^2_{cv}$ , the predictive  $r^2$  can assume a negative value reflecting a complete lack of predictive ability of the training set for the molecules included in the test set.<sup>41,42</sup>

## RESULTS AND DISCUSSION

**A. 3'-Processing.** The results of EVA analyses are summarized in Table 7. EVA analyses using AM1 method,  $\sigma = 3$ , and  $\delta = 5$ , yielded the best correlation with  $r^2_{cv}$  of 0.806 using three principal components. The conventional  $r^2$  for this analysis was 0.964. This model showed good external predictivity ( $r^2_{pred} = 0.761$ ) for the test set which belongs to an entirely different structural class. Thus, the model is truly predictive and unbiased. The high bootstrapped  $r^2$  value ( $r^2_{bs} = 0.975$ ) and low standard deviation (SD =

**Table 9.** Summary of EVA Results Using  $\sigma = 25, \delta = 5$ , and 0–4000 cm<sup>-1</sup> Frequency Range

	3'-processing		3'-strand transfer	
	AM1	PM3	AM1	PM3
$r^2_{cv}$	0.876	0.796	0.878	0.768
compd	9	10	9	10
$s_{press}$	0.350	0.488	0.306	0.432
SEM	0.082	0.101	0.072	0.103
$r^2_{conv}$	0.993	0.990	0.993	0.987
$F$ -value	408.18	238.50	412.98	177.66
$Pr^2 = 0$	0.000	0.000	0.000	0.000
norm coeff of EVA	6.255	6.078	6.021	6.105
$r^2_{pred}$	0.585	0.544	-0.23	0.415
$r^2_{bs}$	0.997	0.998	0.997	0.997
SD	0.002	0.001	0.001	0.002

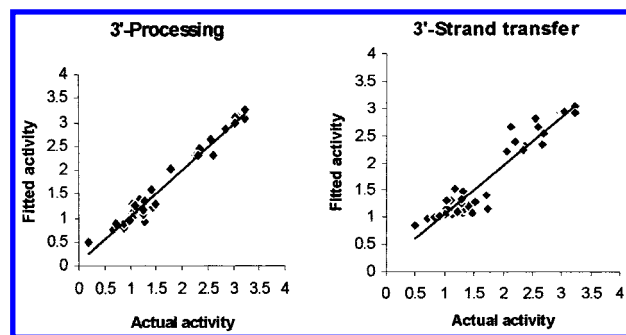
**Table 10.** Results of Analyses with Randomized Biological Activities and Cross-Validation Using Five Groups

	3'-processing <sup>a</sup>		3'-strand transfer <sup>a</sup>	
	$r^2_{cv}{}^b$	$r^2_{cv}{}^c$	$r^2_{cv}{}^b$	$r^2_{cv}{}^c$
mean	-0.11	0.79	-0.29	0.64
SD	0.09	0.03	0.18	0.05
high	0.02	0.84	-0.02	0.73
low	-0.38	0.73	-0.62	0.55

<sup>a</sup> Results using AM1 Hamiltonian ( $\sigma = 3, \delta = 5$ , range = 200–4000 cm<sup>-1</sup>). <sup>b</sup> Cross-validated  $r^2$  with randomized biological activity average of 25 runs. <sup>c</sup> Cross-validation using five groups with optimum number of components average of 25 runs.

0.011) suggest a high degree of confidence in the analysis. The results of cross-validations using five groups and randomized 3'-processing activities are summarized in Table 10. Randomization of the biological activity is the best way to detect inconsistencies and chance correlations in the model. Using 25 runs of randomization, almost in all cases negative  $r^2_{cv}$  values were obtained which proves the stability





**Figure 1.** Fitted vs actual activity values for EVA analysis of the training set. Results are from AM1 method,  $\sigma = 3$ , and  $\delta = 5$ .

of this model. When the same model was derived using PM3 method, internal predictivity remained almost unchanged ( $r^2_{cv} = 0.812$ ), but external predictivity decreased ( $r^2_{pred} = 0.521$ ).

Not much improvement was observed upon increasing the resolution factor ( $\sigma$ ) to 5, 10, and 25, while keeping the sampling interval ( $\delta$ ) to a constant value of 5. The corresponding  $r^2_{cv}$  values obtained with AM1 method were 0.808, 0.846, and 0.864 and those with PM3 method were 0.835, 0.869, and 0.814. The  $r^2_{conv}$  values with AM1 method were 0.977, 0.992, and 0.990 and with PM3 method were 0.995, 0.998, and 0.985 for  $\sigma$  values of 5, 10, and 25, respectively. Moreover, the models become increasingly complex as is evident with higher number of components as the value of  $\sigma$  increases, without any added advantage in terms of external predictivity.

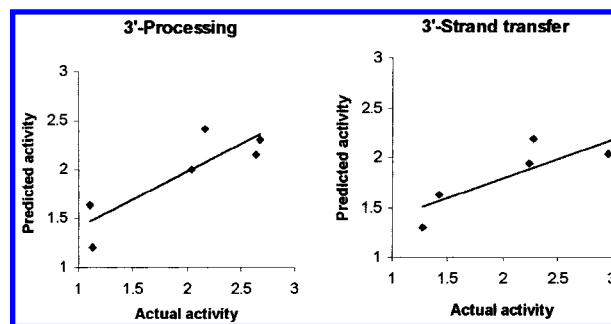
**B. 3'-Strand Transfer.** The results of EVA analyses are summarized in Table 8. Model derived using AM1 method,  $\sigma = 3$  and  $\delta = 5$ , showed good internal predictivity ( $r^2_{cv} = 0.677$ ) and the best external predictivity ( $r^2_{pred} = 0.591$ ). The conventional  $r^2$  for this analysis was 0.894. High  $r^2_{bs}$  (0.912) along with low  $s_{press}$  (0.441) value place further confidence in this model. The results of cross-validations using five groups and randomized 3'-strand transfer activities are summarized in Table 10. With PM3 method,  $r^2_{cv}$  increased to 0.745, but the external predictivity was very poor ( $r^2_{pred} = 0.122$ ).

Like 3'-processing, here also, internal predictivity increased as the value of  $\sigma$  increased to 5, 10, and 25. However, the models became increasingly complex, and the external predictivity for the test set molecules was lost. Similar results were obtained using PM3 Hamiltonian.

The plots of fitted versus actual activity values for the training set molecules and predicted versus actual activity values for the test set molecules are shown in Figures 1 and 2, respectively.

All other five classes of compounds used in the training set were individually selected as test set and remaining compounds for the training set. The model derivation was repeated, and internal as well as external predictivities of these models were checked. The results are impressive and are summarized in Tables 11–14 for 3'-processing and 3'-strand transfer, respectively. Thus, the model is predictive and unbiased.

**Comparison of EVA with CoMSIA.** We have performed the CoMSIA (comparative molecular similarity indices analysis) on the same training and test set ( $r^2_{cv} = 0.821$ ,  $r^2_{conv} = 0.980$ ,  $F$ -value = 279.92,  $r^2_{pred} = 0.608$  for 3'-



**Figure 2.** Predicted vs actual activity values for EVA analysis of the test set. Results are from AM1 method,  $\sigma = 3$ , and  $\delta = 5$ .

**Table 11.** Summary of EVA Results with AM1 Method Using Different Series as Test Sets ( $\sigma = 3$ ,  $\delta = 5$ ; 200–4000  $\text{cm}^{-1}$  for 3'-Processing

sr. no.	test set	$r^2_{cv}$	compd	$r^2_{conv}$	$F$ -value	$r^2_{pred}$
1	coumarins (Table 1)	0.839	6	0.990	5262.4	0.656
2	salicylhydrazines (Table 2)	0.627	6	0.997	1400.3	0.719
3	thiazolothiazepines (Table 3)	0.839	6	0.997	1429.3	0.678
4	quinones (Table 4)	0.822	3	0.966	285.79	0.703
5	lichen acids (Table 5)	0.858	7	0.990	44773.9	0.642

**Table 12.** Summary of EVA Results with AM1 Method Using Different Series as Test Sets ( $\sigma = 3$ ,  $\delta = 5$ ; 200–4000  $\text{cm}^{-1}$  for 3'-Strand Transfer

sr. no.	test set	$r^2_{cv}$	compd	$r^2_{conv}$	$F$ -value	$r^2_{pred}$
1	coumarins (Table 1)	0.735	3	0.946	175.34	0.483
2	salicylhydrazines (Table 2)	0.657	6	0.997	1432.6	0.562
3	thiazolothiazepines (Table 3)	0.773	6	0.996	1260.7	0.503
4	quinones (Table 4)	0.710	3	0.946	176.08	0.526
5	lichen acids (Table 5)	0.752	6	0.999	8496.9	0.585

**Table 13.** Summary of EVA Results with PM3 Method Using Different Series as Test Sets ( $\sigma = 3$ ,  $\delta = 5$ ; 200–4000  $\text{cm}^{-1}$  for 3'-Processing

sr. no.	test set	$r^2_{cv}$	compd	$r^2_{conv}$	$F$ -value	$r^2_{pred}$
1	coumarins (Table 1)	0.833	6	0.999	5787.1	0.565
2	salicylhydrazines (Table 2)	0.634	6	0.998	2334.5	0.694
3	thiazolothiazepines (Table 3)	0.781	6	0.999	5985.4	0.692
4	quinones (Table 4)	0.800	5	0.998	3391.1	0.578
5	lichen acids (Table 5)	0.849	5	0.998	3164.3	0.530

**Table 14.** Summary of EVA Results with PM3 Method Using Different Series as Test Sets ( $\sigma = 3$ ,  $\delta = 5$ ; 200–4000  $\text{cm}^{-1}$  for 3'-Strand Transfer

sr. no.	test set	$r^2_{cv}$	compd	$r^2_{conv}$	$F$ -value	$r^2_{pred}$
1	coumarins (Table 1)	0.696	6	0.998	2562.5	0.410
2	salicylhydrazines (Table 2)	0.634	6	0.999	3726.0	0.483
3	thiazolothiazepines (Table 3)	0.666	6	0.999	5468.8	0.301
4	quinones (Table 4)	0.672	6	0.998	2794.0	0.398
5	lichen acids (Table 5)	0.706	6	0.990	10820.8	0.282

processing and  $r^2_{cv} = 0.759$ ,  $r^2_{conv} = 0.963$ ,  $F$ -value = 144.43,  $r^2_{pred} = 0.660$  for 3'-strand transfer activity). Since this dataset is very diverse and it would be difficult to align it with the conventional alignment methods, we have developed a novel alignment procedure based on molecular

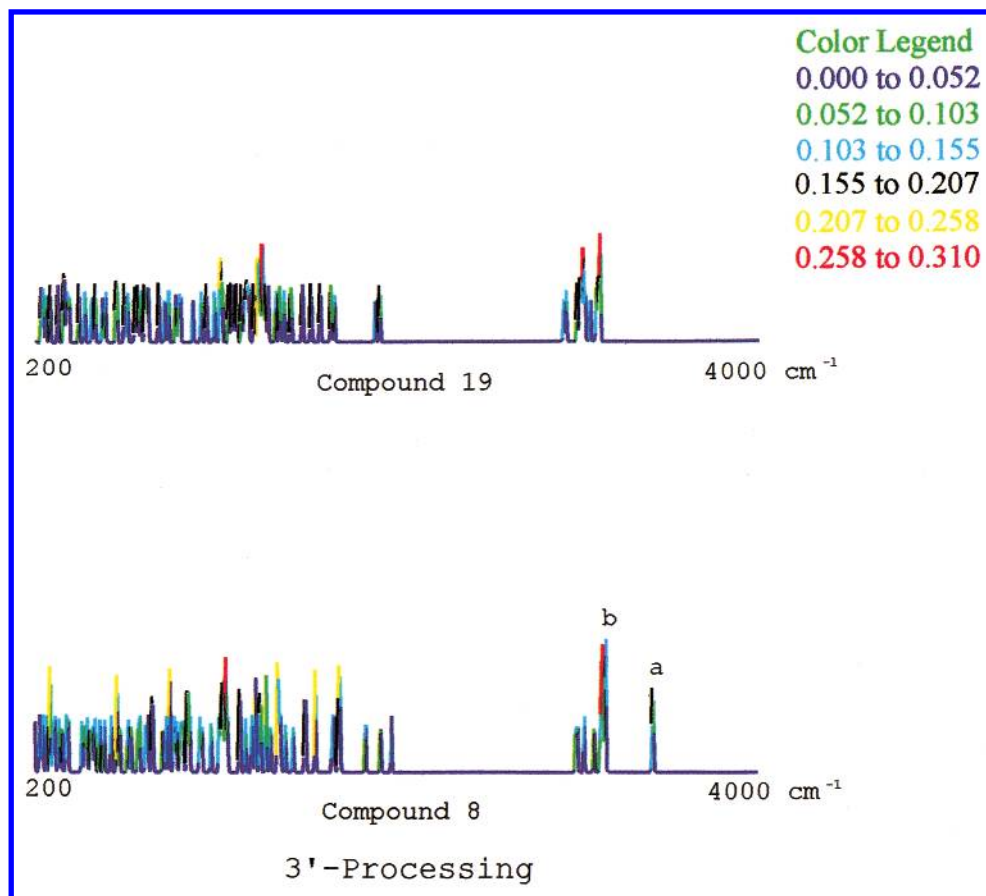


Figure 3. EVA profile of compounds 8 and 19 for 3'-processing activity.

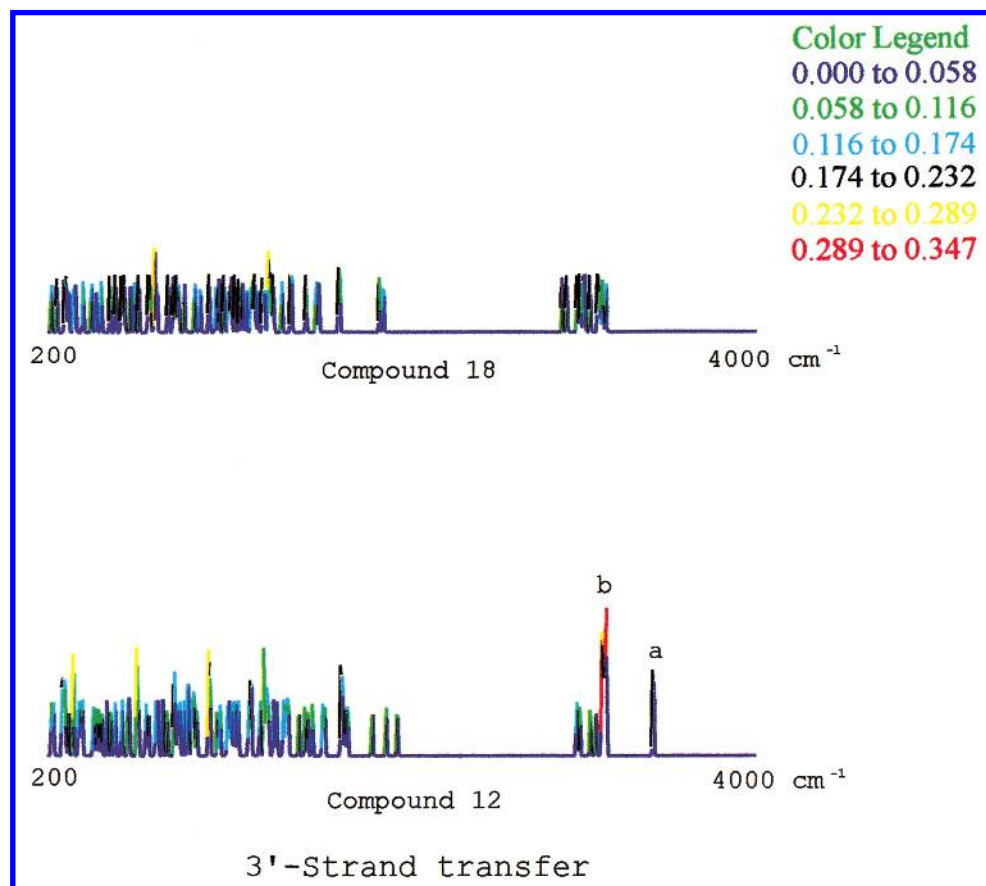


Figure 4. EVA profiles of compounds 12 and 18 for 3'-strand transfer activity.

electrostatic potentials.<sup>43</sup> Thus, the performance of EVA in terms of predictions is almost comparable to that of CoMSIA.

**Graphical Interpretation of Results.** In CoMFA and CoMSIA, 3D isocontour plots are used to visualize those regions of space indicated by the PLS model to have positive or negative correlation with biological activity. While no such 3D visualization is possible with EVA, 2D plots are used here which facilitate interpretation of an EVA QSAR model in similar fashion as the interpretation of an experimental IR spectrum. In EVA, results are presented as a profile across frequencies ranging from the minimum to the maximum specified for each EVA column when it was created.

The EVA profiles of compound 8 (more active) and compound 19 (less active) for 3'-processing and compound 12 (more active) and compound 18 (less active) for 3'-strand transfer are shown for reference in Figures 3 and 4, respectively. Vertical axes in these figures represent actual values for most profiles and run from 0 to 5/ $\sigma$ . Color coding is used only to help visualize differences in height across profiles. Note in particular the two peaks "a" and "b" at the right of the profiles (about 3700 and 3200  $\text{cm}^{-1}$ ). These are the most prominent peaks in the hydrogen bond stretching frequency region for the salicylhydrazine derivatives (compounds 8 and 12), which possess two hydrogen bond donor groups i.e., OH groups. Out of the two peaks, "a" is absent and "b" is attenuated in the EVA profile for less active thiazolothiazepine analogues (compounds 18 and 19), which do not possess any free hydrogen bond donor group. It has been documented that hydrogen bond donor groups such as OH, COOH,  $\text{SO}_3\text{H}$ , etc. are required for potent HIV-1 integrase inhibitory activity<sup>44</sup> and thus, more potency of compounds 8 and 12 can be rationalized on the basis of their EVA profiles. EVA as included in SYBYL 6.6 does not allow proper assignment of vibrational frequencies, and, hence, only a rough scale of 200–4000  $\text{cm}^{-1}$  has been given in Figures 3 and 4 for relative comparison of EVA profiles of an active and inactive compounds. Although Figures 3 and 4 indicate that there is a great deal of vibrational information in the fingerprint region (1–1500  $\text{cm}^{-1}$ ) and considerably less information in the functional group region (1500–4000  $\text{cm}^{-1}$ ), the information in functional group region is specific and can be correlated with the activity in terms of presence or absence of particular functional groups. In case of fingerprint region, most of the group frequencies overlap and are nonspecific, and, hence, it is difficult to correlate them with the activity. The EVA descriptor values over the whole training set are largest at frequencies centered around 1400 and 3700  $\text{cm}^{-1}$ . The fingerprint region is clearly less informative.<sup>31</sup>

## CONCLUSIONS

The EVA method has been applied successfully to rationalize the HIV-1 integrase inhibitory activity of 41 compounds. The present study has shown that EVA can be used to develop good predictive QSAR models for a structurally diverse data set of compounds. This has been achieved without the need to align the structures concerned. The resulting 3D QSAR models show good correlations between EVA and HIV-1 integrase inhibitory activities. The 3D QSAR culminating from the training set yielded a

regression equation with a high degree of statistical significance and performed exceptionally well in predicting the biological activities of compounds in the test set. On the basis of these results, novel molecules can be designed that are predicted to possess improved HIV-1 integrase inhibitory activity.

## ACKNOWLEDGMENT

The authors gratefully acknowledge support for this research from the University Grants Commission (UGC), New Delhi, under its DSA and COSIST projects. MM thanks UGC for the award of senior research fellowship and Dr. David Larson (Support scientist, Tripos, Inc., St. Louis, MO) for providing invaluable information on EVA technique. MM is also grateful to Santosh Kulkarni and Vijay Gokhale for insightful discussions, comments and encouragement.

## REFERENCES AND NOTES

- (1) De Clercq, E. Toward improved anti-HIV chemotherapy. Therapeutic strategies for intervention with HIV infection. *J. Med. Chem.* **1995**, *38*, 2491–2517.
- (2) Hariprasad, V.; Talele, T. T.; Kulkarni, V. M. Design and synthesis of a novel series of nonpeptidic HIV-1 protease inhibitors. *Pharm. Pharmacol. Commun.* **1998**, *4*, 365–372.
- (3) Sakai, H.; Kawamura, M.; Sakuragi, J.; Shibata, R.; Ishimoto, A.; Ono, H.; Ueda, S.; Adachi, A. Integration is essential for efficient gene expression of human immunodeficiency virus type 1. *J. Virol.* **1993**, *7*, 1169–1174.
- (4) Taddeo, B.; Haseltine, W. A.; Farnet, C. M. Integrase mutants of human immunodeficiency virus type 1 with a specific defect in integration. *J. Virol.* **1994**, *68*, 8401–8405.
- (5) Engelman, A.; Englund, G.; Orenstein, J. M.; Martin, M. A.; Craigie, R. Multiple effects of mutations in human immunodeficiency virus type 1 integrase on viral replication. *J. Virol.* **1995**, *69*, 2729–2736.
- (6) Goff, S. P. Genetics of retroviral integration. *Annu. Rev. Genet.* **1992**, *26*, 527–544.
- (7) Craigie, R. Hotspots and Warm spots: Integration Specificity of Retroelements. *Trends Genet.* **1992**, *8*, 187–190.
- (8) Pommier, Y.; Pilon, A.; Bajaj, K.; Mazumder, A.; Neamati, N. HIV-1 integrase as a target for antiviral drugs. *Antiviral Chem. Chemother.* **1997**, *8*, 483–503.
- (9) Neamati, N.; Sunder, S.; Pommier, Y. Design and discovery of HIV-1 integrase inhibitors. *Drug Discovery Today* **1997**, *2*, 487–498.
- (10) Stanwell, C.; Ye, B.; Yuspa, S. H.; Burke, T. R., Jr. Cell protein cross-linking by erbstatin and related compounds. *Biochem. Pharmacol.* **1996**, *52*, 475–480.
- (11) Neamati, N.; Hong, H.; Owen, J. M.; Sunder, S.; Winslow, H. E.; Christensen, J. L.; Zhao, H.; Burke, T. R., Jr.; Milne, G. W. A.; Pommier, Y. Salicylhydrazine-containing inhibitors of HIV-1 integrase: implication for a selective chelation in the integrase active site. *J. Med. Chem.* **1998**, *41*, 3202–3209.
- (12) Neamati, N.; Hong, H.; Mazumder, A.; Wang, S.; Sunder, S.; Nicklaus, M. C.; Milne, G. W. A.; 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.
- (13) Zhao, H.; Neamati, N.; Hong, H.; Mazumder, A.; Wang, S.; Sunder, S.; Milne, G. W. A.; Pommier, Y.; Burke, T. R., Jr. Coumarin-based inhibitor of HIV integrase. *J. Med. Chem.* **1997**, *40*, 242–249.
- (14) Mazumder, A.; Wang, S.; Neamati, N.; Nicklaus, M.; Sunder, S.; Chen, J.; Milne, G. W. A.; Rice, W. G.; Burke, T. R., Jr.; Pommier, Y. Antiretroviral agents as inhibitors of both human immunodeficiency virus type 1 integrase and protease. *J. Med. Chem.* **1996**, *39*, 2472–2481.
- (15) Zhao, H.; Neamati, N.; Sunder, S.; Hong, H.; Wang, S.; Milne, G. W. A.; Pommier, Y.; Burke, T. R., Jr. Hydrazide-containing inhibitors of HIV-1 integrase. *J. Med. Chem.* **1997**, *40*, 937–941.
- (16) 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.
- (17) Cramer, R. D., III; Patterson, D. E.; Bunce, J. D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959–5967.



- (18) Kulkarni, S. S.; Kulkarni, V. M. Three-dimensional quantitative structure activity relationships of interleukin  $1\beta$  converting enzyme inhibitors: a comparative molecular field analysis study. *J. Med. Chem.* **1999**, *42*, 373–380.
- (19) Gokhale, V. M.; Kulkarni, V. M. Comparative molecular field analysis of fungal squalene epoxidase inhibitors. *J. Med. Chem.* **1999**, *42*, 5348–5358.
- (20) Talele, T. T.; Kulkarni, S. S.; Kulkarni, V. M. Development of pharmacophore alignment models as input for comparative molecular field analysis (CoMFA) of azole antifungal agents. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 958–966.
- (21) Klebe, G.; Abraham, U.; Mietzner, T. Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. *J. Med. Chem.* **1994**, *37*, 4130–4146.
- (22) Klebe, G. Comparative molecular similarity indices analysis: CoMSIA. *Perspect Drug Discovery Des.* **1998**, *12*, 87–104.
- (23) Bohm, M.; Sturzebecher, J.; Klebe, G. Three-dimensional quantitative structure activity relationship analyses using comparative molecular field analysis and comparative molecular similarity indices analysis to elucidate selectivity differences of inhibitors binding to trypsin, thrombin and factor Xa. *J. Med. Chem.* **1999**, *42*, 458–477.
- (24) Makhija, M. T.; Kulkarni, V. M. 3D QSAR and molecular modeling of HIV-1 integrase inhibitors. *J. Comput.-Aided Mol. Des.* **2001**. (Communicated).
- (25) Wagener, M.; Sadowski, J.; Gasteiger, J. Autocorrelation of molecular surface properties for modeling corticosteroid binding globulin and cytosolic Ah receptor activity by neural networks. *J. Am. Chem. Soc.* **1995**, *117*, 7769–7775.
- (26) Silverman, B. D.; Platt, D. E. Comparative molecular moment analysis (CoMMA): 3D QSAR without molecular superposition. *J. Med. Chem.* **1996**, *39*, 2129–2140.
- (27) Bravi, G.; Gancia, E.; Mascagni, P.; Pegna, M.; Todeschini, R.; Zaliani, A. MS-WHIM, new 3D theoretical descriptors derived from molecular surface properties: A comparative 3D QSAR study in a series of steroids. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 79–92.
- (28) Ferguson, A. M.; Heritage, T.; Jonathon, P.; Pack, S. E.; Phillips, L.; Rogan, J.; Snaith, P. J. EVA: A new theoretically based molecular descriptor for use in QSAR/QSPR analysis. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 143–152.
- (29) Turner, D. B.; Willett, P.; Ferguson, A. M.; Heritage, T. Evaluation of a novel infrared range vibration-based descriptor (EVA) for QSAR studies. I. General application. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 409–422.
- (30) Turner, D. B.; Willett, P. The EVA spectral descriptor. *Eur. J. Med. Chem.* **2000**, *35*, 367–375.
- (31) SYBYL Ligand-Based Design Manual version 6.6; Tripos, Inc.: St. Louis, MO, 1999.
- (32) Turner, D. B.; Willett, P.; Ferguson, A. M.; Heritage, T. W. Evaluation of a novel molecular vibration-based descriptor (EVA) for QSAR studies: 2. Model validation using a benchmark steroid dataset. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 271–296.
- (33) Ginn, R. M. C.; Turner, D. B.; Willett, P. Similarity searching in files of three-dimensional chemical structures: Evaluation of the EVA descriptor and combination of rankings using data fusion. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 23–37.
- (34) Heritage, T. W.; Ferguson, A. M.; Turner, D. B.; Willett, P. EVA: A novel theoretical descriptor for QSAR studies. *Perspect Drug Discovery Des.* **1998**, 381–398.
- (35) SYBYL Molecular Modeling System, version 6.6; Tripos, Inc.: St. Louis, MO.
- (36) Clark, M.; Cramer, R. D., III.; Van Opdenbosh, N. Validation of the General-Purpose Tripos 5.2 force field. *J. Comput. Chem.* **1989**, *10*, 982–1012.
- (37) MOPAC 6.0 is available from Quantum Chemistry Program Exchange, Indiana University.
- (38) Stewart, J. J. P. MOPAC: a semiempirical molecular orbital program. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 1–103.
- (39) Wold, S.; Albano, C.; Dunn, W. J., III.; Edlund, U.; Esbensen, K.; Geladi, P.; Hellberg, S.; Johansson, E.; Lindberg, W.; Sjostrom, M. Multivariate Data Analysis in Chemistry. In *CHEMOMETRICS: Mathematics and Statistics in Chemistry*; Kowalski, B., Ed.; Reidel: Dordrecht, The Netherlands, 1984.
- (40) Bush, B. L.; Nachbar, R. B., Jr. Sample-distance partial least squares: PLS optimized for many variables, with application to CoMFA. *J. Comput.-Aided Mol. Des.* **1993**, *7*, 587–619.
- (41) Waller, C. L.; Oprea, T. L.; Giolliti, A.; Marshall, G. R. Three-dimensional QSAR of human immunodeficiency virus (1) protease inhibitors. 1. A CoMFA study employing experimentally determined alignment rules. *J. Med. Chem.* **1993**, *36*, 4152–4160.
- (42) Cramer, R. D., III.; Bunce, J. D.; Patterson, D. E. Crossvalidation, bootstrapping, and partial least squares compared with multiple regression in conventional QSAR studies. *Quant. Struct. Act. Relat.* **1988**, *7*, 18–25.
- (43) Makhija, M. T.; Kulkarni, V. M. Molecular electrostatic potentials as input for the alignment of HIV-1 integrase inhibitors in 3D QSAR. *J. Comput.-Aided Mol. Des.* **2001**, in press.
- (44) Goldgur, Y.; Craigie, R.; Cohen, H. G.; Fujiwara, T.; Yoshinaga, T.; Fujishita, T.; Sugimoto, H.; Endo, T.; Murai, H.; Davies, R. D. 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.