

Original article

Exploring molecular shape analysis of styrylquinoline derivatives as HIV-1 integrase inhibitors

J. Thomas Leonard, Kunal Roy*

Drug Theoretics and Cheminformatics Lab, Division of Medicinal and Pharmaceutical Chemistry, Department of Pharmaceutical Technology, Jadavpur University, Raja S C Mullick Road, Kolkata, West Bengal 700 032, India

Received 5 August 2006; received in revised form 7 February 2007; accepted 26 February 2007

Available online 14 March 2007

Abstract

HIV-1 integrase inhibitory activity data of styrylquinoline derivatives have been subjected to 3D-QSAR study by molecular shape analysis (MSA) technique using Cerius² version 4.8 software (Accelrys). For the selection of test set compounds, initially a QSAR analysis was done based on topological and structural descriptors and *K*-means clustering technique was used to classify the entire data set ($n = 36$). Clusters were formed from the factor scores of the whole data set comprising of topological and structural descriptors without the biological activity, and based on the clusters, the data set was divided into training and test sets ($n = 26$ and $n = 10$, respectively) so that all clusters are properly represented in both training and test sets. In the molecular shape analysis, the major steps were (1) generation of conformers and energy minimization; (2) hypothesizing an active conformer (global minimum of the most active compound); (3) selecting a candidate shape reference compound (based on active conformation); (4) performing pair-wise molecular superimposition using maximum common subgroup [MCSG] method; (5) measuring molecular shape commonality using MSA descriptors; (6) determination of other molecular features by calculating spatial and conformational parameters; (7) selection of conformers; (8) generation of QSAR equations by standard statistical techniques. The best model obtained from stepwise regression and GFA techniques shows 51.6% predicted variance (leave-one-out) and 57.3% explained variance. In case of FA–PLS regression, the best relation shows 54.0% predicted variance and 57.9% explained variance. The R^2_{pred} and R^2_{test} values for the GFA derived model are 0.611 and 0.664, respectively, while the best FA–PLS model has R^2_{pred} and R^2_{test} values of 0.602 and 0.656, respectively. These models show the importance of Jurs descriptors (total polar surface area, relative polar surface area, relative hydrophobic surface area, relative positive charge), fraction area of the molecular shadow in the XZ plane (ShadowXZfrac), common overlap steric volume and the ratio of common overlap steric volume to volume of individual molecules. Statistically reliable MSA models obtained from this study suggest that this technique could be useful to design potent HIV-1 integrase inhibitors.

© 2007 Elsevier Masson SAS. All rights reserved.

Keywords: QSAR; MSA; Styrylquinoline; Anti-HIV; Integrase; Inhibitory activity; GFA; G/PLS; FA

1. Introduction

Acquired immunodeficiency syndrome (AIDS) is a fatal disorder for which no complete and successful chemotherapy has

been developed so far. Human immunodeficiency virus subtype 1 (HIV-1), a retrovirus of the lentivirus family, has been found to be prevalent in causing this disease. HIV-1 produces a progressive immunosuppression by destruction of CD4+ T lymphocytes (“helper” cells, which lead attack against infections), and results in opportunistic infections and death [1].

The replicative cycle of HIV can be divided into entry and post entry steps [2,3]. Entry of the HIV into a target cell takes place in three vital steps: (1) the trimeric HIV-1 envelope glycoprotein complex mediated viral entry into susceptible target cells: the surface subunit (gp120) attaches to the receptor

Abbreviations: QSAR, Quantitative structure–activity relationships; GFA, Genetic function approximation; PLS, Partial least squares; FA, Factor analysis; G/PLS, Genetic partial least squares.

* Corresponding author. Tel.: +91 9831594140.

E-mail address: kunalroy_in@yahoo.com (K. Roy).

URL: http://www.geocities.com/kunalroy_in

(CD4); (2) gp120-co-receptor (CXCR4 or CCR5) interaction, which results in the exposure of a co-receptor-binding domain in gp120 on the cell surface; and (3) subsequent conformational changes within the Env complex which lead to membrane fusion mediated by the trans-membrane subunit (gp41). Each of the stages can serve as a target for the HIV entry.

Post entry steps [4] require the viral reverse transcriptase (RT), integrase (IN) and protease (PR) enzymes to complete the viral replication cycle. The virally encoded RT enzyme mediates reverse transcription. RT is a heterodimeric (p51 and p66 subunits) and multifunctional enzyme presenting both RNA and DNA polymerase and RNaseH activities, being responsible for the conversion of the single stranded viral RNA into the double stranded proviral DNA [1]. The viral integrase enzyme is required for the integration of proviral DNA into the host genome before replication. When the infected cell synthesizes new protein, integrated proviral DNA is also translated into the protein building blocks of new viral progeny. Subsequent expression of the virus by the host cells produces the gag and gag-pol proteins Pr44 and Pr160 of HIV–DNA that are processed by the HIV-encoded PR into functional proteins and enzymes. The viral components then assemble on the cell surface and bud out as immature viral particles. The final maturation of newly formed viruses requires the HIV-1 protease to make up an infectious virion. The inhibition of the key enzymes, HIV-1 reverse transcriptase and HIV-1 protease, provides the most attractive target for the anti-HIV drug development [5–7].

Among various methods of anti-HIV activity screening, some important methods are cytoprotection assay, integration enzyme assay, RT inhibition assay, HIV attachment assay, fusion assay, etc [8,9].

The present group of authors has developed a few quantitative structure–activity relationship (QSAR) models for anti-HIV activities of different groups of compounds, e.g., 2-amino-6-arylsulfonylbenzonitriles [10], benzylpyrazoles [11], imidazoles [12], phenylpropylamines [13] and mannitol [14] derivatives. In continuation of such efforts, the present paper deals with molecular shape analysis of styrylquinoline derivatives [15,16] as potent inhibitors of HIV-1 integrase. Previously, HIV-1 integrase inhibitory activity of styrylquinoline derivatives had been analyzed through CoMFA and docking studies [17]. Electrostatic potentials on the molecular surfaces [18] have also been used to model the HIV-1 integrase inhibitory activity.

2. Materials and methods

Anti-HIV-1 integrase inhibitory activity data reported by Mekouar et al. [15,16] have been used for the present QSAR study: the affinity data [IC_{50} (μM)] of styrylquinoline derivatives (Table 1) to inhibit 50% of HIV-1 IN in 3'-processing have been converted to the logarithmic scale [pIC_{50} (M)] and then used for subsequent QSAR analyses as the response variable. Sodium salt compounds were excluded in the present study [19].

All computational experiments were conducted within QSAR+ environment of Cerius² 4.8 version [19] from Accelrys (San Diego, USA) on a Silicon Graphics O2 workstation running under the IRIX 6.5 operating system. It is a standard

practice in QSAR analyses to test the predictive potential of a developed model by applying it on an external data set. Sufficient number of compounds of external data set being often unavailable, it is usual practice to divide the original data set into a training set (based on which the model will be generated) and a test set (which will be used for prediction purpose). For the division of the data set into training and test sets, attempt was made to adopt some rational strategy. A molecule that is structurally very similar to the training set molecules will be predicted well because the model has captured features that are common to the training set molecules and is able to find them in the new molecule. We thus attempted to classify the compounds based on their structural features. Initially the data set was subjected to QSAR analysis based on topological and structural descriptors. Various topological indices calculated are Balaban J, connectivity indices, kappa shape indices, E-state indices and structural parameters [Rotlbonds (number of rotatable bonds), hydrogen bond acceptors and hydrogen bond donors]. The values for the topological and structural descriptors for the compounds have been generated by QSAR+ and Descriptor+ modules of the Cerius² version 4.8 software [19].

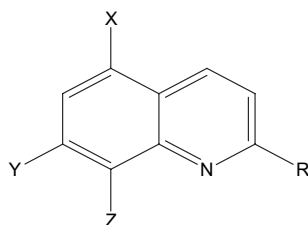
The *E-State index* developed by Kier and Hall [20] is an atom level descriptor encoding both the electronic character and topological environment of each skeletal atom in a molecule. It is derived from chemical graph theoretic approach and has two basic components: (1) intrinsic topological and electronic state of an atom and (2) effect of the environment influencing the atom, considering differences in the intrinsic topological states of different atoms and topological distance among them, which determine the magnitude of the interactions. Every other atom has an effect on a specific atom depending on the difference in electronegativity or electron-richness between them and also on their relative distance.

The data set was divided into training (75% of the total data set) and test (25% of the total data set) sets based on the chemical features, for which we have used *K*-means clustering technique [21]. *K*-means clustering is a non-hierarchical method, which differs from hierarchical clustering in that the data are partitioned without any hierarchy. It expresses the final cluster membership for each case only. The idea consists of clustering a series of compounds into several statistically representative classes of chemicals. At the end of the analysis the data will be split between *K* clusters. As a result of the *K*-means clustering analysis, one can examine the means for each cluster on each dimension to assess how distinct the *K* clusters are. This procedure ensures that any chemical classes (as determined by the clusters derived from the *K*-means clustering technique) will be represented in both the series of compounds i.e. training and test sets. In this technique clusters were formed (Table 2) using the factor scores of the topological and structural descriptors without the biological activity using SPSS [22]. After division of the data set into training and test sets, QSAR with topological and structural parameters was repeated based on the training set data to see the prediction potential of the developed model. Finally, the 3D-QSAR analysis was performed.

Molecular shape analysis (MSA) was used as the 3D-QSAR technique. Molecular shape analysis [23] is a formalism

Table 1

Structural features, observed and calculated HIV-1 integrase inhibitory activity data of styrylquinoline derivatives



Cpd. no.	Structural features				pIC ₅₀		
	R	X	Y	Z	Obs. ^a	Calc. ^b	Calc. ^c
<i>Training set</i>							
2	Styr-1-yl	H	CO ₂ H	OH	5.28	4.899	4.858
3	–CH=CH-Furan-2-yl	H	CO ₂ H	OH	5.72	5.071	5.102
4	–CH=CH-Thiaphen-3-yl	H	CO ₂ H	OH	5.47	5.046	4.975
5	–CH=CH-Pyridin-3-yl	H	CO ₂ H	OH	5.39	4.798	4.920
8	4-NHCOCH ₃ -Styr-1-yl	H	CO ₂ H	OH	5.85	5.520	5.392
9	4-OH-Styr-1-yl	H	CO ₂ H	OH	5.80	5.489	5.477
11	2,4-(OH) ₂ -Styr-1-yl	H	CO ₂ H	OH	5.43	5.767	5.832
12	3,4-(OH) ₂ -Styr-1-yl	H	CO ₂ H	OH	5.62	5.881	5.927
14	3-OH,4-OMe-Styr-1-yl	H	CO ₂ H	OH	6.05	5.699	5.579
15	2,3,4-(OH) ₃ -Styr-1-yl	H	CO ₂ H	OH	6.52	6.760	6.674
16	3,4-(OH) ₂ ,5-OMe-Styr-1-yl	H	CO ₂ H	OH	6.15	5.430	5.561
17	3,5-(OMe) ₂ , 4-OH-Styr-1-yl	H	CO ₂ H	OH	5.31	5.495	5.404
18	3,5-(Br) ₂ , 4-OH-Styr-1-yl	H	CO ₂ H	OH	5.89	5.584	5.424
19	3,4-(OH) ₂ , 5-I-Styr-1-yl	H	CO ₂ H	OH	5.40	5.548	5.576
20	3,4-(OH) ₂ -Styr-1-yl	H	CO ₂ Me	OH	4.00	4.976	5.029
23	–CH ₃	H	H	OH	4.00	3.841	3.952
24	–CH ₃	H	H	–OCO-3,4-(OMe) ₂ -Styr-1-yl	4.00	3.926	3.895
25	–CH ₃	H	H	–OCO-3,4-(OH) ₂ -Styr-1-yl	4.00	4.435	4.527
27	Styr-1-yl	H	H	OH	4.00	4.314	4.122
28	–CH=CH-(8-OH)Quinolin-2-yl	H	H	OH	4.00	4.855	4.691
30	3,4-(OH) ₂ -Styr-1-yl	H	H	NO ₂	4.00	4.973	5.145
31	3,4-(OH) ₂ -Styr-1-yl	H	H	NH ₂	4.00	4.722	4.682
33	3,4-(OH) ₂ -Styr-1-yl	H	H	OH	5.13	4.809	4.914
34	3,4-(OH) ₂ -Styr-1-yl	H	3,4-(OH) ₂ -styr-1-yl	OH	5.66	5.485	5.425
35	3,4-(OH) ₂ -Styr-1-yl	H	CN	OH	5.52	4.643	4.726
36	3-CO ₂ H,4-OH-Styr-1-yl	H	CO ₂ H	OH	5.57	5.773	5.951
<i>Test set</i>							
1	–CH ₃	H	CO ₂ H	OH	4.00	4.583	4.920
6	4-NO ₂ -Styr-1-yl	H	CO ₂ H	OH	5.92	5.469	5.563
7	4-NH ₂ -Styr-1-yl	H	CO ₂ H	OH	5.46	4.871	4.897
10	3,5-(OH) ₂ -Styr-1-yl	H	CO ₂ H	OH	5.49	5.825	5.911
13	3-Me, 4-OH-Styr-1-yl	H	CO ₂ H	OH	5.55	5.754	5.600
21	3,4-(OH) ₂ -Styr-1-yl	Cl	Cl	OH	4.00	4.821	4.834
22	3,4-(OH) ₂ -Styr-1-yl	H	CO ₂ H	OH	5.64	5.588	5.701
26	Styr-1-yl	H	H	OAc	4.00	4.232	4.028
29	3,4-(OH) ₂ -Styr-1-yl	H	H	H	4.00	4.426	4.469
32	3,4-(OAc) ₂ -Styr-1-yl	H	H	OAc	4.00	5.035	4.839

^a Refs. [15,16]; Obs. = observed; Calc. = calculated.^b From Eq. (14).^c From Eq. (19).

that deals with quantitative characterization, representation and manipulation of molecular shape in the construction of a QSAR. The overall aim of molecular shape analysis is to identify the biologically relevant conformation without knowledge of the receptor geometry and in a quantitative fashion explain the activity of a series of congeners. The major steps of molecular shape analysis were (1) generation of conformers and energy minimization; (2) hypothesizing an active conformer

(global minimum of the most active compound); (3) selecting a candidate shape reference compound (based on active conformation); (4) performing pair-wise molecular superimposition using maximum common subgroup [MCSG] method; (5) measuring molecular shape commonality using MSA descriptors; (6) determination of other molecular features by calculating spatial, electronic and conformational parameters; (7) selection of conformers; (8) generation of QSAR equations by different

Table 2
Stepwise clustering of the compounds for selection of test set members

Cluster	Number of compounds in clusters/sub-clusters			Compounds (Sl nos.) in each clusters							Number of compounds in test set
	Cluster	Subcluster level 1	Subcluster level 2								
1	4			4	28	29	31				1
2	32	5		1	3	22	23	24			2
		3		7	8	36					1
		4		5	18	19	21				1
		3		6	30	32					2
		17	5	2	9	26	27	33			1
			12	10	11	12	13	14	15		2
				16	17	20	25	34	35		

statistical tools. A complete list of descriptors used in MSA is given in Table 3 (along with MSA descriptors, spatial, thermodynamic and structural parameters were also considered). Multiple conformations of each molecule were generated using the Boltzmann jump as a conformational search method. The upper limit of the number of conformations per molecule was 150. Each conformer was subjected to an energy minimization procedure (open force field) to generate the lowest energy conformation for each structure. The lowest energy conformer of the most active compound **15** for HIV-1 integrase inhibitory activity was selected as a shape reference to which all the structures of compounds in the study were aligned through pair-wise superpositioning. The method used for performing the alignment was maximum common subgroup [MCSG] [19]. This method looks at molecules as points and lines, and uses the techniques of graph theory to identify patterns. It finds the largest subset of atoms in the shape reference compound that is shared by all the structures in the study table and uses this subset for alignment. A rigid fit of atom pairings was performed to superimpose each structure so that it overlays the shape reference compound.

For the development of MSA models, five statistical methods were used: (1) stepwise regression, (2) genetic function approximation (GFA), (3) multiple linear regression with factor analysis as the preprocessing step for variable selection (FA–MLR), (4) partial least squares regression with factor analysis as the preprocessing step for variable selection (FA–PLS) and (5) genetic partial least squares (G/PLS).

In stepwise regression [24], a multiple-term linear equation was built step-by-step. The basic procedures involve (1) identifying an initial model, (2) iteratively “stepping,” that is, repeatedly altering the model at the previous step by adding or removing a predictor variable in accordance with the “stepping criteria” (in our case $F = 3$ for inclusion; $F = 2.9$ for exclusion for the forward selection method) and (3) terminating the search when stepping is no longer possible given the stepping criteria, or when a specified maximum number of steps has been reached. Specifically, at each step all variables are reviewed and evaluated to determine which one will contribute most to the equation. That variable will then be included in the model, and the process starts again. A limitation of the stepwise

Table 3
List of descriptors used in molecular shape analysis

Sl. no.	Spatial parameters		Molecular shape analysis (MSA) parameters		Thermodynamic parameters	Structural parameters
1	Vm	JursDPSA-3	DIFFV		AlogP98	Rotlbonds
2	Radius of Gyration	JursFPSA-1	COSV		MolRef	H bond donors
3	Density	JursFPSA-2	Fo			H bond acceptors
4	PMI	JursFPSA-3	NCOSV			
5	Area	JursFNSA-1	ShapeRMS			
6	ShadowXY	JursFNSA-2	SRVol			
7	ShadowYZ	JursFNSA-3				
8	ShadowXZ	JursWPSA-1				
9	ShadowXYfrac	JursWPSA-2				
10	ShadowYZfrac	JursWPSA-3				
11	ShadowXZfrac	JursWNSA-1				
12	Xlength	JursWNSA-2				
13	Ylength	JursWNSA-3				
14	Zlength	JursRPCG				
15	Shadow η	JursRNCG				
16	JursPPSA-1	JursRPCS				
17	JursPPSA-2	JursRNCS				
18	JursPPSA-3	JursTPSA				
19	JursPNSA-1	JursTASA				
20	JursPNSA-2	JursRPASA				
21	JursPNSA-3	JursRASA				
22	JursDPSA-1	JursSASA				
23	JursDPSA-2					

regression search approach is that it presumes there is a single “best” subset of X variables and seeks to identify it. There is often no unique “best” subset, and all possible regression models with a similar number of X variables as in the stepwise regression solution should be fitted subsequently to study whether some other subsets of X variables might be better.

Genetic function approximation (GFA) technique [25,26] was used to generate a population of equations rather than a single equation for correlation between biological activity and physicochemical properties. GFA involves the combination of multivariate adaptive regression splines (MARS) algorithm with genetic algorithm to evolve population of equations that best fit the training set data. It provides an error measure, called the lack of fit (LOF) score that automatically penalizes models with too many features. It also inspires the use of splines as a powerful tool for non-linear modeling. The model with proper balance of all statistical terms will be used to explain the variance of the biological activity. A distinctive feature of GFA is that it produces a population of models (e.g., 100), instead of generating a single model, as most other statistical methods do. The range of variations in this population gives added information on the quality fit and importance of the descriptors.

In case of FA–MLR, though classical approach of multiple regression technique was used as the final statistical tool for developing QSAR relations, factor analysis (FA) [27,28] was used as the data-preprocessing step to identify the important predictor variables contributing to the response variable and to avoid collinearities among them. In a typical factor analysis procedure, the data matrix is first standardized, and correlation matrix and subsequently reduced correlation matrix are constructed. The eigen value problem is then solved and the factor pattern can be obtained from the corresponding eigen vectors. The principal objectives of factor analysis are to display multidimensional data in a space of lower dimensionality with minimum loss of information (explaining >95% of the variance of the data matrix) and to extract the basic features behind the data with ultimate goal of interpretation and/or prediction. Factor analysis was performed on the data set(s) containing biological activity and all descriptor variables, which were to be considered. The factors were extracted by principal component method and then rotated by VARIMAX rotation (a kind of rotation which is used in principal component analysis so that the axes are rotated to a position in which the sum of the variances of the loadings is the maximum possible) to obtain Thurston’s simple structure. The simple structure is characterized by the property that as many variables as possible fall on the coordinate axes when presented in common factor space, so that largest possible number of factor loadings becomes zero. This is done to obtain a numerically comprehensive picture of the relatedness of the variables. Only variables with non-zero loadings in such factors where biological activity also has non-zero loading were considered important in explaining variance of the activity. Further, variables with non-zero loadings in different factors were combined in a multivariate equation.

In FA–PLS, factor analysis was used to identify the important predictor variables contributing to the response variable and PLS regression was used as the final statistical tool for developing

QSAR relations, while in G/PLS, genetic method was used to identify the important predictor variables. PLS is a generalization of regression, which can handle data with strongly correlated and/or noisy or numerous X variables [29]. It gives a reduced solution, which is statistically more robust than MLR. The linear PLS model finds “new variables” (latent variables or X scores) which are linear combinations of the original variables. To avoid overfitting, a strict test for the significance of each consecutive PLS component is necessary and then stopping when the components are non-significant. Cross-validation is a practical and reliable method for testing this significance. Application of PLS thus allows the construction of larger QSAR equations while still avoiding overfitting and eliminating most variables. PLS is normally used in combination with cross-validation to obtain the optimum number of components. This ensures that the QSAR equations are selected based on their ability to predict the data rather than to fit the data [30]. In case of FA–PLS, variables with high loading (>0.7) in such factors where the inhibitory activity shows high or moderate loading in the factor loading table (rotated component matrix) were selected for the PLS regression. Based on the standardized regression coefficients, the variables with smaller coefficients were removed from the PLS regression, until there is no further improvement in Q^2 value, irrespective of the components.

The stepwise regression and factor analysis (FA) were performed using the statistical software SPSS [22]. PLS was performed using statistical software MINITAB [31]. Genetic function approximation (GFA) and G/PLS were done using QSAR+ environment of Cerius² software [19].

The statistical qualities of the equations [32] were judged by the parameters like *explained variance* (R_a^2), *correlation coefficient* (R), *standard error of estimate* (s), and *variance ratio* (F) at specified *degrees of freedom* (df), *root mean square error* ($RMSE$) and *average of absolute values of residuals* ($AVRES$). All accepted equations have regression coefficients and F ratios significant at 95% and 99% levels, respectively, if not stated otherwise. All the generated models were validated by PRESS (leave-one-out) [33,34], *cross-validation* R^2 (Q^2), *predicted residual sum of squares* ($PRESS$), *standard deviation based on PRESS* (S_{PRESS}), *standard deviation of error of prediction* ($SDEP$) and *bootstrap r^2* (bsr^2). Definitions of some of the statistical terms are given below.

Coefficient of determination R^2 : This is the most commonly used term to describe the goodness of fit of data for a regression model. This statistic is defined in the following equation:

$$R = \sqrt{1 - \frac{\sum (Y_{\text{Calc}} - Y)^2}{\sum (Y - \bar{Y})^2}} \quad (1)$$

In Eq. (1), Y_{Calc} and Y indicate calculated and observed activity values, respectively, and \bar{Y} indicates mean activity value.

Explained variance R_a^2 : Explained variance of the training set without validation may be defined as follows:

$$R_a^2 = \frac{(n-1)R^2 - p}{n-p-1} \quad (2)$$

In Eq. (2), R^2 is squared correlation coefficient, p is number of predictor variables and n is number of compounds.

Variance ratio (F): It gives an indication about the stability of the regression coefficients.

$$F = \frac{\frac{\sum (Y_{\text{Calc}} - \bar{Y})^2}{p}}{\frac{\sum (Y_{\text{Calc}} - Y)^2}{n-p-1}}, \text{ df} = p, n-p-1 \quad (3)$$

where df is degree of freedom.

Standard error of estimate (s): This is defined as:

$$s = \sqrt{\frac{\sum (Y_{\text{Calc}} - Y)^2}{n-p-1}} \quad (4)$$

AVRES: Average of absolute values of residuals is defined as:

$$\text{AVRES} = \frac{|Y_{\text{Calc}} - Y|}{n} \quad (5)$$

In Eq. (5), Y_{Calc} and Y indicate calculated and observed activity values, respectively, and n indicates number of compounds.

RMSE: Squared root mean error is defined as:

$$\text{RMSE} = \sqrt{\frac{\sum (Y_{\text{Calc}} - Y)^2}{n}} \quad (6)$$

All the generated QSAR models were validated by leave-one-out method [33,34] and cross-validation R^2 (Q^2), predicted residual sum of squares (PRESS), standard deviation based on PRESS (S_{PRESS}) and standard deviation of error of prediction (SDEP) and bootstrap r^2 (bsr^2) values were reported.

Cross-validation R^2 (Q^2): It measures predictive R^2 (leave-one-out) and part of the variance explained in the validation data.

$$Q^2 = 1 - \frac{\sum (Y_{\text{pred}} - Y)^2}{\sum (Y - \bar{Y})^2} \quad (7)$$

In Eq. (7), Y_{pred} and Y indicate predicted and observed activity values, respectively, and \bar{Y} indicates mean activity value.

PRESS: It is the predicted residual sum of squares, the difference between predicted and the calculated values [33,34].

$$\text{PRESS} = \sum (Y_{\text{pred}} - Y)^2 \quad (8)$$

Standard deviation of error of prediction (SDEP): SDEP is a measure of prediction of error [33,34].

$$\text{SDEP} = \sqrt{\frac{\text{PRESS}}{n}} \quad (9)$$

S_{PRESS} : Standard deviation based on PRESS is defined as:

$$S_{\text{PRESS}} = \sqrt{\frac{\text{PRESS}}{n-p-1}} \quad (10)$$

Bootstrap r^2 : This is the average squared correlation coefficient calculated during the validation procedure (leave-one-out). The models derived on training set compounds were also validated

through the external validation using the parameters like R_{pred}^2 and R_{test}^2 .

R_{pred}^2 : The predictive R^2 was based only on molecules present in the test set and is defined as:

$$R_{\text{pred}}^2 = 1 - \frac{\sum (Y_{\text{pred}(\text{test})} - Y_{(\text{test})})^2}{\sum (Y_{(\text{test})} - \bar{Y}_{(\text{training})})^2} \quad (11)$$

In Eq. (11), $Y_{\text{pred}(\text{test})}$ and $Y_{(\text{test})}$ indicate predicted and observed activity values, respectively, of the test set compounds and $\bar{Y}_{(\text{training})}$ indicates mean activity value of the training set.

R_{test}^2 is the squared correlation coefficient (R^2) between the observed and predicted data of the test set.

Finally, randomization test at 99% confidence level was carried out for the selected models.



The acceptability criteria of a valid QSAR model include a Q^2 value of more than 0.5 and a difference of Q^2 and R^2 value being less than 0.3 [35]. The external validation is a more reliable way to establish a predictive QSAR model [36]. When the data set is divided into training and test sets and a model is generated based on the training set compounds, the predictive R^2 value should be more than 0.5.

3. Results and discussion

3.1. QSAR of the whole data set using topological and structural descriptors

Considering all the 36 compounds factor analysis was performed using the topological and structural descriptors. Eight factors could explain 96.53% of variance. Based on the factor analysis the following best equation was developed.

$$\begin{aligned} \text{pIC}_{50} = & -0.778 (\pm 0.477) \text{S}_{\text{aaaC}} - 0.226 (\pm 0.155) \text{S}_{\text{sCH}_3} \\ & + 0.190 (\pm 0.122) \text{Rotlbonds} + 4.991 (\pm 1.100) \\ n = 36, R_a^2 = 0.604, R^2 = 0.638, R = 0.799, \\ s = 0.528, F = 18.8 (\text{df } 3, 32), \\ Q^2 = 0.519, \text{PRESS} = 11.861, \text{SDEP} = 0.574, \\ S_{\text{PRESS}} = 0.609 \end{aligned} \quad (12)$$

The 95% confidence intervals of the regression coefficients are mentioned within parentheses. Eq. (12) could explain 60.4% of the variance and predict 51.9% of the variance. The negative coefficients of S_{aaaC} (E-state value of fragment ) and S_{sCH_3} (E-state value of fragment $-\text{CH}_3$) show that the activity decreases with increase in the E-state values of  fragment and methyl group. The number of rotatable bonds (i.e., flexibility in the structure) is conducive for the HIV-1 integrase inhibitory activity.

3.2. QSAR of training set compounds using topological and structural descriptors

After division (Table 1) of the data set into training ($n = 26$) and test ($n = 10$) sets using K -means clustering technique, the training set compounds were subjected to factor analysis.

From the factor analysis on the data matrix consisting of HIV-1 integrase inhibitory data, topological and structural descriptors, it was observed that seven factors could explain the data matrix to the extent of 96.38%.

$$\begin{aligned} \text{pIC}_{50} = & 0.035 (\pm 0.021) \text{S}_{\text{sOH}} - 0.594 (\pm 0.360) \text{S}_{\text{dssC}} \\ & + 3.751 (\pm 0.540) \\ n = & 26, R_a^2 = 0.598, R^2 = 0.630, R = 0.794, \\ F = & 19.6 (\text{df } 2, 23), s = 0.526, \text{SDEP} = 0.569, \\ S_{\text{PRESS}} = & 0.605, \\ Q^2 = & 0.510, \text{PRESS} = 8.415, \text{AVRES} = 0.419, \\ \text{RMSE} = & 0.494, \\ n_{\text{test}} = & 10, R_{\text{pred}}^2 = 0.598, \text{AVRES}_{\text{test}} = 0.455, \\ \text{RMSE}_{\text{test}} = & 0.559 \end{aligned} \quad (13)$$

Eq. (13) could explain 59.8% of the variance and predict 51.0% of the variance. The positive coefficient of S_{sOH} shows that the inhibitory activity increases with increase in the E-state value of hydroxyl group. While the negative coefficient of S_{dssC} shows that the inhibitory activity decreases with increase in the E-state values of >= fragment. The predictive R^2 (R_{pred}^2) value for the test set was found to be 0.598.

3.3. Molecular shape analysis

The view of aligned training set molecules is shown in Fig. 1. The values of important descriptors used in the MSA models are given in Table 4.

3.3.1. Stepwise regression and GFA

Both stepwise regression [using stepping criteria based on F value ($F = 3$ for inclusion; $F = 2.9$ for exclusion)] and GFA [50,000 iterations, initial equation length 4, and other defaults] techniques resulted in the same best equation Eq. (14) as noted below.

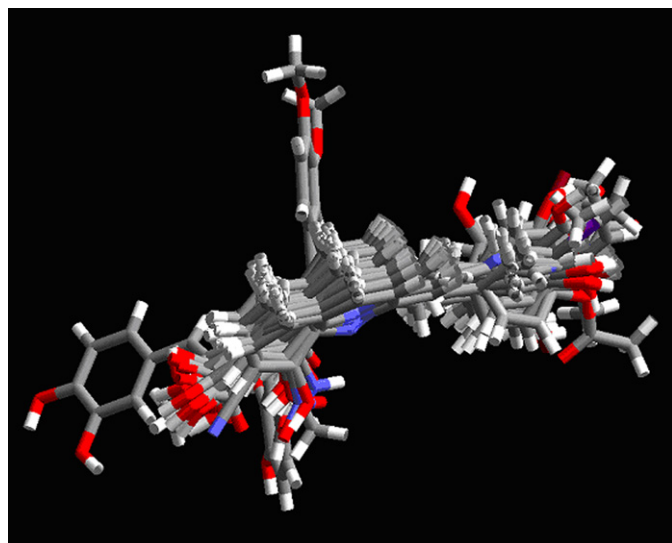


Fig. 1. View of aligned training set compounds in molecular shape analysis.

$$\begin{aligned} \text{pIC}_{50} = & 0.006 (\pm 0.004) \text{JursTPSA} + 0.012 (\pm 0.008) \text{COSV} \\ & + 1.974 (\pm 1.338) \\ n = & 26, \text{LOF} = 0.363, R_a^2 = 0.573, R^2 = 0.607, R = 0.779, \\ F = & 17.8 (\text{df } 2, 23), s = 0.542, \text{SDEP} = 0.565, \\ S_{\text{PRESS}} = & 0.601, \\ Q^2 = & 0.516, \text{PRESS} = 8.320, \text{bsr}^2 (\pm \text{sd}) = 0.608 (\pm 0.011), \\ \text{AVRES} = & 0.435, \text{RMSE} = 0.509, \\ n_{\text{test}} = & 10, R_{\text{pred}}^2 = 0.611, \text{AVRES}_{\text{test}} = 0.473, \\ \text{RMSE}_{\text{test}} = & 0.550 \end{aligned} \quad (14)$$

The 95% confidence intervals of the regression coefficients are shown within parentheses. The positive coefficient of JursTPSA indicates that the total polar surface area (TPSA) (sum of solvent-accessible surface areas of atoms with absolute value of partial charges) is conducive for inhibitory activity. Compounds with high TPSA values [e.g., $\text{R} = 3,4\text{-(OH)}_2\text{-styr-1-yl}$ (compound **12**), $3,4\text{-(OH)}_2,5\text{-OMe-styr-1-yl}$ (compound **16**)] have more inhibitory activity than the corresponding methyl congeners. This indicates that presence of polar substituents will be favorable possibly because of dipole–dipole interactions with the active site. The positive coefficient of COSV indicates that the common overlap steric volume (COSV) (common volume between each individual molecule and the molecule selected as the reference compound) is conducive for the inhibitory activity. Compounds with a high COSV value [e.g., $\text{R} = 3,4\text{-(OH)}_2\text{-styr-1-yl}$ (compound **12**), $3\text{-OH},4\text{-OMe-styr-1-yl}$ (compound **14**)] have more inhibitory activity than the corresponding methyl congeners. It is easy to understand that the compounds, which are similar in shape and size to the reference compound, will be similarly active as the reference compound. The calculated inhibitory activity values and the predicted values of the test set compounds according to Eq. (14) are given in Table 1. The predictive R^2 (R_{pred}^2) is found to be 0.611. The R_{test}^2 for Eq. (14) is 0.664, which is better than the R_{test}^2 value (0.633) reported for the best CoMFA model in Ref. [17].

3.3.2. FA–MLR

From the factor analysis on the data matrix consisting of HIV-1 integrase inhibitory data and molecular shape parameters (along with spatial and thermodynamic descriptors), it was observed that eight factors could explain the data matrix to the extent of 96.22% (Table 5). Based on the factor analysis the following equation was derived with two variables.

$$\begin{aligned} \text{pIC}_{50} = & -4.081 (\pm 2.681) \text{JursRSA} + 0.012 (\pm 0.008) \text{COSV} \\ & + 5.915 (\pm 2.937) \\ n = & 26, R_a^2 = 0.567, R^2 = 0.602, R = 0.776, \\ F = & 17.4 (\text{df } 2, 23), s = 0.546, \text{SDEP} = 0.575, S_{\text{PRESS}} = 0.611, \\ Q^2 = & 0.501, \text{PRESS} = 8.586, \text{AVRES} = 0.437, \text{RMSE} = 0.513, \\ n_{\text{test}} = & 10, R_{\text{pred}}^2 = 0.566, \text{AVRES}_{\text{test}} = 0.489, \text{RMSE}_{\text{test}} = 0.581 \end{aligned} \quad (15)$$

Eq. (15) could explain 56.7% of the variance and predict 50.1% of the variance. The negative coefficient of JursRSA

Table 4
The values of selected descriptors used in this analysis

Cpd. no.	S_aaaC	COSV	Fo	NCOSV	ShapeRMS	JursTPSA	JursRPCG	JursRPSA	JursRASA	ShadowXZfrac	ShadowYlength
<i>Training set</i>											
2	0.981	171.899	0.665	86.625	0.572	142.757	0.164	0.268	0.732	0.746	9.287
3	0.934	174.085	0.725	66.104	1.032	165.801	0.152	0.322	0.678	0.725	8.590
4	0.993	185.440	0.744	63.790	0.493	141.054	0.175	0.273	0.727	0.715	9.268
5	0.953	147.052	0.578	107.161	1.112	172.591	0.160	0.326	0.674	0.657	8.684
8	0.900	195.433	0.636	111.619	1.063	197.011	0.129	0.321	0.679	0.628	9.785
9	0.909	191.276	0.714	76.710	1.125	199.722	0.145	0.364	0.636	0.714	9.728
11	0.800	189.295	0.688	86.021	1.001	246.940	0.129	0.439	0.561	0.650	9.210
12	0.822	195.607	0.710	79.964	0.751	253.292	0.127	0.447	0.553	0.711	9.101
14	0.862	204.941	0.699	88.193	1.509	207.534	0.128	0.349	0.651	0.686	9.806
15	0.734	249.333	0.877	34.932	0.000	292.472	0.113	0.511	0.489	0.685	9.508
16	0.784	156.344	0.519	144.655	1.570	254.681	0.113	0.425	0.575	0.669	10.862
17	0.834	184.142	0.579	134.155	0.612	213.852	0.114	0.336	0.664	0.587	10.484
18	0.880	203.115	0.668	100.751	0.875	192.903	0.145	0.316	0.684	0.602	9.263
19	0.821	176.144	0.587	123.720	0.624	236.877	0.127	0.394	0.606	0.676	10.632
20	0.975	151.407	0.517	141.642	1.390	192.649	0.129	0.324	0.676	0.699	10.478
23	1.651	125.104	0.844	23.198	0.019	62.853	0.253	0.181	0.819	0.697	8.208
24	1.588	119.926	0.377	198.462	0.043	85.676	0.152	0.139	0.861	0.498	9.871
25	1.514	123.682	0.435	160.672	0.014	158.713	0.151	0.266	0.734	0.670	10.135
27	1.578	168.212	0.728	62.964	0.412	57.778	0.176	0.118	0.882	0.731	9.021
28	2.925	182.881	0.656	95.914	1.423	115.761	0.126	0.208	0.792	0.825	9.030
30	1.020	146.644	0.556	116.885	0.840	200.850	0.256	0.373	0.627	0.646	9.255
31	1.766	165.695	0.659	85.891	0.480	126.524	0.121	0.242	0.758	0.743	9.243
33	1.419	151.860	0.612	96.274	0.875	165.557	0.124	0.319	0.681	0.687	8.733
34	1.211	162.536	0.446	201.547	1.272	252.036	0.082	0.354	0.646	0.546	11.627
35	1.098	141.800	0.534	123.518	0.899	158.004	0.115	0.288	0.712	0.650	8.736
36	0.742	165.072	0.560	129.880	0.885	292.522	0.114	0.493	0.507	0.681	9.576
<i>Test set</i>											
1	1.054	140.838	0.800	35.275	0.117	150.432	0.208	0.380	0.620	0.726	8.113
6	0.820	169.813	0.600	113.405	1.058	236.114	0.231	0.411	0.589	0.756	9.913
7	0.946	157.117	0.581	113.466	1.099	165.643	0.141	0.295	0.705	0.687	8.904
10	0.806	190.210	0.690	85.418	0.565	254.437	0.130	0.456	0.544	0.714	9.719
13	0.872	213.503	0.729	79.555	1.377	200.493	0.128	0.340	0.660	0.712	9.445
21	0.922	157.414	0.571	118.356	0.873	157.293	0.121	0.282	0.718	0.648	9.215
22	0.907	173.525	0.615	108.678	0.542	247.934	0.128	0.436	0.564	0.616	9.203
26	1.628	161.613	0.604	106.141	1.188	57.094	0.184	0.103	0.897	0.685	10.805
29	2.051	143.891	0.600	96.117	0.839	120.124	0.148	0.237	0.763	0.726	7.724
32	1.412	174.521	0.488	182.757	1.487	159.379	0.114	0.226	0.774	0.634	10.754

indicates that the relative hydrophobic surface area (RASA), which is the ratio between the total hydrophobic surface area and the total molecular solvent-accessible surface area, is detrimental for inhibitory activity. Compounds having high JursRASA values [e.g., R = styr-1-yl (compound **2**), $-\text{CH}=\text{CH}$ -thiaphen-3-yl (compound **4**)] have less inhibitory activity than the substituted styryl congeners. We have discussed above that polar substituents are preferred for the inhibitory activity and thus hydrophobic substituents will be detrimental. The predictive R^2 (R^2_{pred}) value for the test set was found to be 0.566.

$$\begin{aligned} \text{pIC}_{50} &= 4.081 (\pm 2.681) \text{JursRPSA} + 0.012 (\pm 0.008) \text{COSV} \\ &\quad + 1.835 (\pm 1.351) \\ n &= 26, R_a^2 = 0.567, R^2 = 0.602, R = 0.776, \\ F &= 17.4 (\text{df } 2, 23), s = 0.546, \text{SDEP} = 0.575, \text{S}_{\text{PRESS}} = 0.611, \\ Q^2 &= 0.500, \text{PRESS} = 8.586, \text{AVRES} = 0.437, \text{RMSE} = 0.513 \\ n_{\text{test}} &= 10, R_{\text{pred}}^2 = 0.565, \text{AVRES}_{\text{test}} = 0.490, \text{RMSE}_{\text{test}} = 0.582 \end{aligned} \quad (16)$$

Eq. (16) could explain 56.7% of the variance and predict 50.0% of the variance. The positive coefficient of JursRPSA indicates that the relative polar surface area (RPSA), which is the ratio between the total polar surface area and the total molecular solvent-accessible surface area, is conducive for inhibitory activity. Compounds with high JursRPSA values [e.g., R = 2,4-(OH)₂-styr-1-yl (compound **11**), 3,4-(OH)₂-styr-1-yl (compound **12**), 3,4-(OH)₂,5-OMe-styr-1-yl (compound **16**)] have more inhibitory activity than the corresponding methyl congeners. The predictive R^2 (R^2_{pred}) value for the test set was found to be 0.565.

$$\begin{aligned} \text{pIC}_{50} &= 0.007 (\pm 0.004) \text{JursTPSA} \\ &\quad - 4.785 (\pm 4.292) \text{ShadowXZfrac} + 3.895 (\pm 2.323) \text{Fo} \\ &\quad + 0.504 (\pm 0.541) \text{ShapeRMS} + 4.195 (\pm 2.483) \\ n &= 26, R_a^2 = 0.617, R^2 = 0.678, R = 0.823, \\ F &= 11.0 (\text{df } 4, 21), s = 0.513, \text{SDEP} = 0.542, \text{S}_{\text{PRESS}} = 0.603, \\ Q^2 &= 0.556, \text{PRESS} = 7.628, \text{AVRES} = 0.363, \text{RMSE} = 0.461 \\ n_{\text{test}} &= 10, R_{\text{pred}}^2 = 0.509, \text{AVRES}_{\text{test}} = 0.514, \text{RMSE}_{\text{test}} = 0.618 \end{aligned} \quad (17)$$

Table 5

Factor loadings of the variables (MSA, spatial and thermodynamic parameters) after VARIMAX rotation

	F1	F2	F3	F4	F5	F6	F7	F8	Communality
pIC ₅₀	0.148	0.542	0.325	0.089	0.493	−0.102	0.338	0.337	0.911
V _m	0.932	0.184	0.182	0.214	−0.009	−0.111	0.044	0.005	0.996
DIFFV	0.932	0.184	0.182	0.214	−0.009	−0.111	0.044	0.005	0.996
COSV	0.200	0.322	0.332	−0.326	0.776	−0.072	−0.041	0.022	0.969
F _o	−0.570	0.098	0.091	−0.362	0.695	0.128	−0.082	0.006	0.980
NCOSV	0.715	−0.049	−0.058	0.415	−0.531	−0.052	0.067	−0.010	0.979
ShapeRMS	0.354	0.202	0.059	−0.337	−0.079	0.036	0.797	0.000	0.925
Radius of gyration	0.887	0.311	0.175	−0.078	−0.033	−0.003	0.188	0.093	0.965
JursSASA	0.931	0.200	0.188	0.207	−0.034	−0.091	0.053	0.033	0.997
JursPPSA-1	0.778	−0.094	−0.552	0.252	−0.078	−0.084	0.017	0.016	0.996
JursPNSA-1	0.457	0.362	0.807	0.022	0.033	−0.036	0.053	0.027	0.997
JursDPSA-1	0.217	−0.305	−0.908	0.155	−0.074	−0.032	−0.024	−0.007	0.996
JursPPSA-2	0.840	0.419	−0.225	0.236	0.014	−0.040	0.064	−0.031	0.994
JursPNSA-2	−0.583	−0.611	−0.509	−0.115	−0.066	−0.027	−0.029	0.026	0.993
JursDPSA-2	0.789	0.557	0.128	0.197	0.042	−0.009	0.052	−0.031	0.994
JursPPSA-3	0.679	0.672	0.011	−0.092	0.027	0.079	0.049	0.239	0.988
JursPNSA-3	−0.369	−0.811	−0.400	−0.080	0.013	0.057	−0.065	0.106	0.980
JursDPSA-3	0.490	0.810	0.294	0.028	−0.001	−0.015	0.063	0.002	0.988
JursFPSA-1	0.006	−0.346	−0.917	0.142	−0.049	0.063	−0.035	−0.005	0.988
JursFNSA-1	−0.006	0.346	0.917	−0.142	0.049	−0.063	0.035	0.005	0.988
JursFPSA-2	0.751	0.512	−0.341	0.159	0.010	−0.132	0.063	−0.045	0.990
JursFNSA-2	−0.428	−0.690	−0.570	−0.016	−0.065	0.066	−0.017	0.047	0.995
JursFPSA-3	0.190	0.812	−0.133	−0.341	0.058	0.109	0.014	0.343	0.963
JursFNSA-3	−0.082	−0.869	−0.401	0.003	0.028	0.113	−0.059	0.142	0.960
JursWPSA-1	0.900	0.020	−0.300	0.299	−0.057	−0.040	0.035	0.015	0.997
JursWNSA-1	0.649	0.327	0.662	0.143	0.030	0.044	0.060	0.036	0.995
JursWPSA-2	0.864	0.353	−0.169	0.292	0.019	0.050	0.065	−0.020	0.991
JursWNSA-2	−0.670	−0.531	−0.444	−0.194	−0.065	−0.121	−0.040	0.007	0.986
JursWPSA-3	0.807	0.533	0.051	0.066	0.021	0.131	0.061	0.169	0.992
JursWNSA-3	−0.545	−0.721	−0.368	−0.159	−0.003	−0.031	−0.067	0.074	0.989
JursRPCG	−0.715	−0.243	−0.005	0.138	−0.124	0.040	−0.019	−0.578	0.941
JursRNCG	−0.760	−0.486	−0.219	−0.012	0.019	0.314	−0.056	−0.040	0.966
JursRPCS	−0.489	−0.461	−0.258	−0.127	0.003	0.582	−0.093	0.157	0.907
JursRNCS	−0.716	−0.100	0.012	−0.115	0.010	0.549	0.134	−0.294	0.942
JursTPSA	0.378	0.867	0.276	0.026	0.091	−0.007	0.043	0.055	0.984
JursTASA	0.620	−0.723	−0.093	0.201	−0.136	−0.093	0.011	−0.024	0.984
JursRPSA	0.154	0.932	0.266	−0.046	0.096	−0.062	0.037	0.052	0.983
JursRSA	−0.154	−0.932	−0.266	0.046	−0.096	0.062	−0.037	−0.052	0.983
ShadowXY	0.878	0.334	0.238	−0.117	0.099	−0.028	0.144	−0.002	0.983
ShadowXZ	0.772	0.154	0.189	0.421	−0.257	−0.080	0.006	0.141	0.925
ShadowYZ	0.604	−0.024	0.017	0.711	−0.228	−0.166	−0.178	−0.024	0.982
ShadowXYfrac	−0.584	−0.159	−0.035	−0.164	0.401	−0.392	−0.134	−0.031	0.728
ShadowXZfrac	−0.357	−0.047	−0.042	−0.910	0.039	0.052	−0.032	−0.011	0.964
ShadowYZfrac	−0.372	−0.009	0.131	−0.811	−0.044	−0.359	−0.105	0.083	0.963
Shadow η	0.041	0.056	0.078	−0.950	0.172	−0.010	0.182	0.025	0.976
Xlength	0.780	0.313	0.295	−0.207	−0.045	−0.018	0.256	0.165	0.932
Ylength	0.800	0.213	0.013	0.222	−0.083	0.271	−0.015	−0.184	0.850
Zlength	0.452	−0.112	−0.023	0.821	−0.183	−0.180	−0.140	0.021	0.978
Area	0.927	0.153	0.181	0.252	0.006	−0.097	0.002	0.021	0.990
Density	0.099	0.084	0.865	0.120	0.102	−0.054	−0.030	−0.004	0.794
PMI	0.725	0.246	0.541	0.183	0.071	0.156	0.118	0.077	0.961
AlogP ₉₈	0.627	−0.417	0.430	0.049	−0.131	0.188	−0.082	−0.077	0.819
MolRef	0.932	0.096	0.282	0.151	−0.040	−0.048	0.031	0.015	0.985
% Variance	0.393	0.214	0.147	0.103	0.040	0.027	0.021	0.018	0.962

Eq. (17) could explain 61.7% of the variance and predict 55.6% of the variance. The regression coefficients of ShapeRMS and ShadowXZfrac are significant at 93.4% and 96.9% levels, respectively. An increase in the fraction area of the molecular shadow in the XZ plane (ShadowXZfrac) is detrimental for the inhibitory activity. Compounds with high ShadowXZfrac

values [e.g., R = styr-1-yl (compound **2**), −CH=CH−(8-OH)-quinolin-2-yl (compound **28**)] have less inhibitory activity than the substituted styryl congeners. Root mean square (ShapeRMS) deviation between the individual molecule and the shape reference compound is conducive for the inhibitory activity. Compounds with high ShapeRMS values [e.g.,

R = 3-OH,4-OMe-styr-1-yl (compound **14**), 3,4-(OH)₂,5-OMe-styr-1-yl (compound **16**) have more inhibitory activity than the other substituted styryl congeners. The intercorrelation (*r*) matrix among the predictor variables used in Eqs. (14)–(17) is given in Table 6. The predictive R^2 (R^2_{pred}) value for the test set was found to be 0.509.

On introduction of structural descriptors along with MSA, spatial and thermodynamic descriptors, there is significant improvement in cross-validation statistics (Q^2 rises up to 0.658), but it has a negative impact on external validation, as R^2_{pred} values are unacceptably low.

3.3.3. FA–PLS

The optimum number of components was found to be one to obtain both the models Eqs. (18) and (19) (optimized by cross-validation). Based on the standardized regression coefficients, the variables were selected for the respective models.

$$\begin{aligned} \text{pIC}_{50} &= 0.003\text{JursTPSA} + 2.173\text{JursRPSA} - 3.835\text{JursRPCG} \\ &\quad + 0.007\text{COSV} + 0.683\text{Fo} + 2.753 \\ n &= 26, R^2_a = 0.616, R^2 = 0.631, R = 0.794, \\ Q^2 &= 0.571, \text{PRESS} = 7.370, \text{SDEP} = 0.532, S_{\text{PRESS}} = 0.554, \\ \text{AVRES} &= 0.432, \text{RMSE} = 0.494, \\ n_{\text{test}} &= 10, R^2_{\text{pred}} = 0.527, \text{AVRES}_{\text{test}} = 0.505, \text{RMSE}_{\text{test}} = 0.607 \end{aligned} \quad (18)$$

$$\begin{aligned} \text{pIC}_{50} &= 0.004\text{JursTPSA} + 2.506\text{JursRPSA} + 0.008\text{COSV} \\ &\quad + 2.248 \\ n &= 26, R^2_a = 0.579, R^2 = 0.596, R = 0.772, \\ Q^2 &= 0.540, \text{PRESS} = 7.907, \text{SDEP} = 0.551, S_{\text{PRESS}} = 0.574, \\ \text{AVRES} &= 0.439, \text{RMSE} = 0.517, \\ n_{\text{test}} &= 10, R^2_{\text{pred}} = 0.602, \text{AVRES}_{\text{test}} = 0.454, \text{RMSE}_{\text{test}} = 0.556 \end{aligned} \quad (19)$$

The negative coefficient of JursRPCG indicates that the relative positive charge (RPCG), the charge of the most positive atom divided by the total positive charge, is detrimental for inhibitory activity. Compounds with high RPCG values [e.g., R = CH₃ (compound **23**) and Z = NO₂ (compound **30**)] have less inhibitory activity than the substituted styryl (at R) and hydroxyl (at Z) congeners. The positive coefficient of Fo indicates that the ratio between common overlap steric volume and volume of individual molecule (Fo), is conducive for inhibitory activity. Compounds with high Fo values [e.g.,

R = 2,3,4-(OH)₃-styr-1-yl (compound **15**)] have more inhibitory activity than the other substituted styryl congeners. This is in agreement with the positive coefficients of COSV found in Eqs. (18) and (19). The calculated inhibitory activity values and the predicted values of the test set compounds according to Eq. (19) are given in Table 1. The predictive R^2 (R^2_{pred}) values for the test set were found to be 0.527 and 0.602 for Eq. (18) and (19), respectively. The R^2_{test} for Eq. (19) is 0.656, which is better than the R^2_{test} value (0.633) reported for the best CoMFA model in Ref. [17]. The requirement of polar surface area as evident from this study is in accordance with the CoMFA contour map with electrostatic field [17]. The results of randomization test (99% confidence level) applied on all the equations are shown in Table 7, which justifies acceptability of the models.

3.3.4. G/PLS

Although G/PLS could generate models with good internal validation ($Q^2 > 0.5$) statistics, but all the models failed in external validation (R^2_{pred}) on the test set compounds.

Both internal and external validations can assess the predictive ability of a model generated. The prediction of the developed models for the test set is not only based on the selection of the training and test sets [21] but also based on the selection of the variables and the statistical method used (Table 8). In this case the test set compounds selected from the *K*-means clusters could be predicted well for many models, but not for all the models generated though internal validation (Q^2) statistics were good (>0.5) for all the models. Particularly the models developed through G/PLS could not predict well for the test set. So the predictive ability of the models for the test compounds also depends on the variables and the statistical method used. Although the test compounds were selected through *K*-means clustering technique there is no guarantee that all the models ($Q^2 > 0.5$) would also predict well for the test set compounds.

4. Conclusions

The present molecular shape analyses explore the spatial and shape requirements for the inhibitory activity of styrylquinoline derivatives for HIV-1 integrase inhibitory data. The quality of models obtained from stepwise regression, GFA–MLR, FA–MLR and FA–PLS is of comparable range (explained variance ranging from 56.7% to 61.7% while predicted variance ranging from 50.0% to 57.0%). However, the best

Table 6
Intercorrelation (*r*) matrix for MSA and spatial parameters for Eqs. (14)–(17)

	COSV	JursTPSA	JursRPSA	JursRASA	Fo	ShapeRMS
JursTPSA	0.505	1.000	0.968	−0.968	−0.059	0.353
JursRPSA	0.513	0.968	1.000	−1.000	0.100	0.305
JursRASA	−0.513	−0.968	−1.000	1.000	−0.100	−0.305
Fo	0.619	−0.059	0.100	−0.100	1.000	−0.175
ShapeRMS	0.161	0.353	0.305	−0.305	−0.175	1.000
ShadowXZfrac	0.217	−0.201	−0.061	0.061	0.557	0.130

Table 7
Results of randomization test applied on the developed models (99% confidence level)

Equation no.	(13)	(14)	(15)	(16)	(17)	(18)	(19)
QSAR method	Topological	MSA	MSA	MSA	MSA	MSA	MSA
Modeling technique	FA–MLR	Stepwise/GFA	FA–MLR	FA–MLR	FA–MLR	FA–PLS	FA–PLS
R from non-random model	0.794	0.779	0.776	0.776	0.823	0.794	0.772
No. of random trials	99	99	99	99	99	99	99
No. of random R s less than non-random R	99	99	99	99	99	99	99
No. of random R s more than non-random R	0	0	0	0	0	0	0
Mean value of R from random trials \pm SD	0.258 (\pm 0.113)	0.270 (\pm 0.118)	0.253 (\pm 0.114)	0.247 (\pm 0.123)	0.382 (\pm 0.118)	0.099 (\pm 0.183)	0.075 (\pm 0.140)

external validation statistics were obtained with stepwise regression and GFA derived model with R^2_{pred} and R^2_{test} being 0.611 and 0.664, respectively, while the FA–PLS derived model has R^2_{pred} and R^2_{test} values of 0.602 and 0.656,

respectively. The R^2_{test} for the FA–PLS derived model is better than the R^2_{test} value (0.633) reported for the best CoMFA model in Ref. [17]. The models derived from G/PLS were not so good in external validation statistics.

Table 8
Predictive ability of the developed models based on the descriptors and statistical methods

Sl. no.	Descriptors	R^2	Q^2	R^2_{pred}	R^2_{test}
<i>FA–MLR (topological descriptors)</i>					
1	S_sOH, S_dssC	0.630	0.510	0.598	0.602
<i>FA–MLR (MSA, spatial and thermodynamic descriptors)</i>					
1	COSV, JursTPSA	0.607	0.516	0.611	0.664
2	COSV, JursRASA	0.602	0.501	0.566	0.630
3	COSV, JursRPSA	0.602	0.500	0.565	0.631
4	JursRPSA, COSV, ShadowXZfrac	0.647	0.531	0.550	0.567
5	ShadowXZfrac, JursTPSA, Fo, ShapeRMS	0.678	0.556	0.509	0.536
<i>FA–MLR (MSA, spatial, thermodynamic and structural descriptors)</i>					
1	ShadowXZfrac, RPCG, JursTPSA, Fo, ShapeRMS	0.734	0.539	0.276	0.236
2	Fo, Rotlbonds	0.683	0.608	0.287	0.292
3	ShadowXZfrac, Rotlbonds, Fo, ShapeRMS	0.742	0.633	0.181	0.211
4	NCOSV, Shadow η , Rotlbonds, ShapeRMS	0.751	0.654	0.181	0.214
5	NCOSV, Rotlbonds	0.696	0.627	0.345	0.342
6	Fo, JursDPSA-1, Rotlbonds	0.720	0.643	0.411	0.388
7	JursRPSA, JursRPCG, Fo, ShadowXZfrac	0.701	0.544	0.153	0.183
8	JursRPCS, JursRPCG, Fo, ShadowXZfrac	0.684	0.658	−0.098	0.045
<i>FA–PLS (MSA, spatial and thermodynamic descriptors)</i>					
1	COSV, Fo, JursTPSA, JursRPSA, JursRPCG, ShapeRMS	0.631	0.571	0.527	0.536
2	COSV, JursTPSA, JursRPSA	0.596	0.540	0.602	0.656
3	COSV, JursTPSA, JursRPSA, ShapeRMS	0.604	0.529	0.594	0.699
4	COSV, Fo, JursTPSA, JursRPSA, JursRPCG, ShadowXZfrac, AlogP98	0.653	0.560	0.543	0.519
<i>G/PLS (MSA, spatial and thermodynamic descriptors)</i>					
1	JursRPCG, JursRPCS, JursTASA, ShadowYZ, ShadowXZfrac	0.741	0.632	−0.273	0.093
2	JursRPCG, JursRPCS, JursTASA, JursFPSA-3, ShadowYZ, ShadowXZfrac	0.794	0.618	−0.650	0.003

Spatial (Jurs) parameters played major role in the developed MSA models: the total polar surface area (TPSA) and relative polar surface area (RPSA) are conducive for inhibitory activity, while the relative hydrophobic surface area (RASA) and relative positive charge (RPCG) are detrimental to the activity. It appears that surface area of the ligands and corresponding charge distribution are important for the binding affinity. Because of possible presence of dipole–dipole interactions with the active site, hydrophobic substituents are not preferred. The requirement of polar surface area as evident from this study is in accordance with the previously reported CoMFA contour map with electrostatic field [17]. An increase in the fraction area of the molecular shadow in the XZ plane (ShadowXZfrac) is also an important criterion for the inhibitory activity. MSA parameters like the common overlap steric volume (COSV), the ratio of common overlap steric volume and the volume of individual molecule (Fo) and root mean square (ShapeRMS) deviation between the individual molecule and the shape reference compound are also very essential parameters in this study. All the developed models were validated by the external validation set. Statistically reliable 3D-QSAR models obtained from this study suggest that this technique (MSA) could be useful to design potent HIV-1 integrase inhibitors.

Acknowledgements

One of the authors (JTL) thanks the AICTE, New Delhi for a fellowship. KR thanks the Department of Science and Technology (DST), Government of India, New Delhi for a financial grant under the Fast Track Scheme for Young Scientists.

References

- [1] G. Campiani, A. Ramunno, G. Maga, V. Nacci, C. Fattorusso, B. Catalanotti, E. Morelli, E. Novellino, *Curr. Pharm. Des.* 8 (2002) 615–657.
- [2] S. Jiang, Q. Zhao, A.K. Debanth, *Curr. Pharm. Des.* 8 (2002) 563–580.
- [3] R.W. Sanders, M.M. Dankers, E. Busser, M. Caffrey, J.P. Moore, B. Berkhout, *Retrovirology* 1 (2004) 3–13.
- [4] P.P. Mager, *Med. Res. Rev.* 21 (2001) 348–353.
- [5] J.M. Farber, E.A. Berger, *Proc. Natl. Acad. Sci. U.S.A.* 99 (2002) 1749–1751.
- [6] D.D. Richman, *Nature* 410 (2001) 995–1001.
- [7] W. Kazmierski, N. Bifulco, H. Yang, L. Boone, F. DeAnda, C. Watson, T. Kenakin, *Bioorg. Med. Chem.* 11 (2003) 2663–2676.
- [8] G. Xu, A. Kannan, T.L. Hartman, H. Wargo, K. Watson, J.A. Turpin, R.W. Buckheit Jr., A.A. Johnson, Y. Pommier, M. Cushman, *Bioorg. Med. Chem.* 10 (2002) 2807–2816.
- [9] M. Stevens, C. Pannecouque, E. DeClercq, J. Balzarini, *Antimicrob. Agents Chemother.* 47 (2003) 3109–3116.
- [10] K. Roy, J.T. Leonard, *Bioorg. Med. Chem.* 12 (2004) 745–754.
- [11] J.T. Leonard, K. Roy, *QSAR Comb. Sci.* 23 (2004) 387–398.
- [12] K. Roy, J.T. Leonard, *Bioorg. Med. Chem.* 13 (2005) 2967–2973.
- [13] K. Roy, J.T. Leonard, *J. Chem. Inf. Model* 45 (2005) 1352–1368.
- [14] J.T. Leonard, K. Roy, *Bioorg. Med. Chem.* 14 (2006) 1039–1046.
- [15] K. Mekouar, J.F. Mouscadet, D. Desmaële, F. Subra, H. Leh, D. Savoure, C. Auclair, J. d'Angelo, *J. Med. Chem.* 41 (1998) 2846–2857.
- [16] F. Zouhiri, J.F. Mouscadet, K. Mekouar, D. Desmaële, D. Savouré, H. Leh, F. Subra, M. Le Bret, C. Auclair, J. d'Angelo, *J. Med. Chem.* 43 (2000) 1533–1540.
- [17] X.H. Ma, X.Y. Zhang, J.J. Tan, W.Z. Chen, C.X. Wang, *Acta Pharmacol. Sin.* 25 (2004) 950–958.
- [18] J. Polanski, F. Zouhiri, L. Jeanson, D. Desmaële, J. d'Angelo, J.-F. Mouscadet, R. Gieleciak, J. Gasteiger, M. Le Bret, *J. Med. Chem.* 45 (2002) 4647–4654.
- [19] Cerius² version 4.8 is a product of Accelrys Inc, San Diego, USA, <http://www.accelrys.com/cerius2>.
- [20] L.B. Kier, L.H. Hall, *Molecular Structure Description: The Electrotopological State*, Academic Press, San Diego, 1999.
- [21] J.T. Leonard, K. Roy, *QSAR Comb. Sci.* 25 (2006) 235–251.
- [22] SPSS is a statistical software of SPSS Inc, USA.
- [23] A.J. Hopfinger, J.S. Tokarsi, in: P.S. Charifson (Ed.), *Three-Dimensional Quantitative Structure–Activity Relationship Analysis: Practical Applications of Computer-aided Drug Design*, Marcel Dekker Inc., New York, 1997, pp. 105–164.
- [24] R.B. Darlington, *Regression and Linear Models*, McGraw-Hill, New York, 1990.
- [25] D. Rogers, A.J. Hopfinger, *J. Chem. Inf. Comput. Sci.* 34 (1994) 854–866.
- [26] Y. Fan, L.M. Shi, K.W. Kohn, Y. Pommier, J.N. Weinstein, *J. Med. Chem.* 44 (2001) 3254–3263.
- [27] R. Franke, *Theoretical Drug Design Methods*, Elsevier, Amsterdam, 1984, p. 184.
- [28] R. Franke, A. Gruska, in: H. van de Waterbeemd (Ed.), *Chemometric Methods in Molecular Design*, VCH, Weinheim, 1995, p. 113.
- [29] S. Wold, in: H. van de Waterbeemd (Ed.), *Chemometric Methods in Molecular Design*, VCH, Weinheim, 1995, pp. 195–218.
- [30] S.S. Kulkarni, V.M. Kulkarni, *J. Med. Chem.* 42 (1999) 373–380.
- [31] Minitab is a statistical software of Minitab Inc, USA.
- [32] G.W. Snedecor, W.G. Cochran, *Statistical Methods*, Oxford & IBH Publishing Co. Pvt. Ltd, New Delhi, 1967, p. 381.
- [33] S. Wold, L. Eriksson, in: H. van de Waterbeemd (Ed.), *Chemometric Methods in Molecular Design*, VCH, Weinheim, 1995, pp. 312–317.
- [34] A.K. Debnath, in: A.K. Ghose, V.N. Viswanadhan (Eds.), *Combinatorial Library Design and Evaluation*, Marcel Dekker Inc, New York, 2001, pp. 73–129.
- [35] L. Eriksson, J. Jaworska, A.P. Worth, M.T.D. Cronin, R.M. McDowell, P. Gramatica, *Environ. Health Perspect.* 111 (2003) 1361–1375.
- [36] R. Guha, P.C. Jurs, *J. Chem. Inf. Model* 45 (2005) 65–73.