Digest Journal of Nanomaterials and Biostructures Vol. 4, No. 4, December 2009, p. 823 – 834

QSAR STUDIES ON NOVEL ANTI-HIV AGENTS USING FA-MLR, FA-PLS AND PCRA TECHNIQUES

R.VEERASAMY[a*], S. RAVICHANDRAN [a], A. JAIN [b], H. RAJAK [c], R. K. AGRAWAL [b]

[a]*Faculty of Pharmacy, AIMST University,Semeling – 08100, Kedah, Malaysia*
[b]*Pharmaceutical Chemistry Research Laboratory, Dept. of Pharmaceutical Sciences, Dr. Hari Singh Gour University, Sagar (M.P) - 470 003, India*
[c]*Institute of Pharmaceutical Sciences, Guru Ghasidas University, Bilaspur-495 009 (CG), India*

In the present work, quantitative structure activity relationship studies were performed to explore the structural and physicochemical requirements of phenyl ethyl thiourea (PET) derivatives for anti-HIV activity. QSAR models have been developed using steric, electronic and thermodynamic descriptors. Statistical techniques like multiple linear regression with factor analysis as the data preprocessing step (FA-MLR), principal component regression analysis (PCRA), and partial least squares with factor analysis as the data preprocessing step (FA-PLS) analysis were applied to identify the structural and physicochemical requirements for anti-HIV activity. The generated equations were statistically validated using leave-one-out technique and the best models were also subjected to leave-25% out cross-validation. The quality of fit and predictive ability of equations obtained from FA-MLR, PCRA, and FA-PLS is of acceptable statistical range (explained variance ranging from 80.7% to 94.0%, while predicted variance ranging from 75.9% to 93.4%). The robustness of the best models was checked by Y–randomization test and identified as good predictive models. The coefficient of density and van der Waal's energy shows that the activity increases with increase in density and van der Waals energy of molecules. The coefficient of molar refractivity shows that the activity decreases with increase in volume and critical pressure of the molecules is detrimental to activity. The information generated from the present study may be useful in the design of more potent PET derivatives as anti HIV agents.

(Received August 4, 2009; accepted October 7, 2009)

Keywords: QSAR; FA-MLR; FA-PLS; PCRA; anti-HIV; PET derivatives

1. Introduction

HIV- 1 (Human Immunodeficiency Virus Type-1) is the pathogenic retrovirus and causative agent of AIDS or AIDS- related complex (ARC) [1,2]. When viral RNA is translated into a polypeptide sequence, it is assembled in a long polypeptide chain, which includes several individual proteins namely, reverse transcriptase, protease, integrase, etc. Before these enzymes become functional, they must be cut from the longer polypeptide chain.

Acquired immune deficiency syndrome (AIDS) is a formidable pandemic that is still wreaking havoc world wide. The catastrophic potential of this virally caused disease may not have been fully realized. The causative moiety of the disease is human immunodeficiency virus (HIV), which is a retrovirus of the lentivirus family [3]. The three viral enzymes; reverse transcriptase, protease and integrase encoded by the gag and gag-pol genes of HIV play an important role in the virus replication cycle. Among them, viral protease catalyzes the formation of viral functional

*Corresponding author: phravi75@rediffmail.com

enzymes and proteins necessary for its survival. The viral particles at this stage are called virions. The virus particles after the protease action have all the necessary constituents of mature virus and are capable of invading other T4 cells and repeating the life cycle of proviral DNA from viral RNA, the key stage in viral replication. Its central role in virus maturation makes protease is a prime target for anti-HIV-therapy [4].

QSAR analyses of HIV-1 reverse transcriptase inhibitors [5], HIV-1 protease inhibitors [6,7] and HIV-1 integrase inhibitors [8] and gp 120 envelope glycoprotein [9] were reported. The present group of authors has developed a few quantitative structure-activity relationship models to predict anti-HIV activity of different group of compounds [10-22]. In continuation of such efforts, in this article, we have performed QSAR analysis for anti-HIV activity of PET derivatives [23,24]using modeling software WIN CAChe 6.1 and statistical software STATISTICA 6.

The purpose of the present study is to investigate the physico-chemical parameters responsible for the anti-HIV activity of PET derivatives, to explore the correlation between them and is expected to get more information for designing novel PET derivatives with potent anti-HIV activity.

There is high structural diversity and a sufficient range of the biological activity in the selected series of PET derivatives. So we have selected this series of compounds for our QSAR studies. We carried out QSAR analysis using different statistical techniques and established QSAR models to guide further structural optimization and predict the biological potency of clinical drug candidates.

2. Experimental

Materials and Methods

All of the Molecular Modeling studies, reported herein were performed using Win CAChe 6.1 (Product of Fujitsu private limited, Japan, http://www.cachesoftware.com /contacts/japan.shtml) modeling software, Molecular modeling pro 6.1.0, trial version (ChemSW, Inc., www.chemsw.com) and the QSAR models were executed with STATISTICA 6 (Softstat, Inc., Tulsa, USA) software.

Biological data

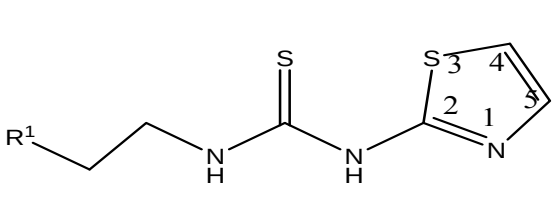
In the present work we have taken 71 PET compounds and their anti-HIV activity from the reported work [23,24]. Many of these compounds inhibited wild type HIV-1 with ED_{50} values between 0.001 μM and 0.005 μM in MT-4 cells. One of these thiourea derivatives troviridine showed good anti-HIV activity (0.02 μM, in clinical trial) with low cytotoxicity for MT4 cells (Cantrell et al., 1996). There is high structural diversity and a sufficient range of the biological activity in the selected series of PET derivatives. It insists as to select these series of compounds for our QSAR studies. All the anti-HIV activities used in the present study were expressed as $pED_{50} = -\log_{10} ED_{50}$. Where ED_{50} is the micro molar concentration of the compounds producing 50% reduction in the cytopathic effect caused by the virus is stated as the means of at least two experiments. ED_{50} values were assessed by XTT assays [25]. The compounds which did not show confirmed anti-HIV activity in the above cited literature have not been taken for our study.

Optimization of molecules structure

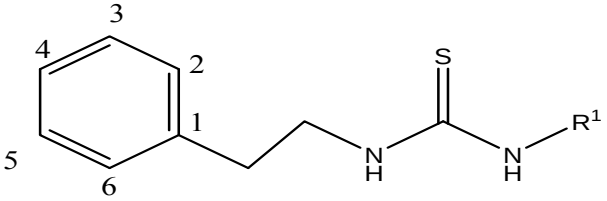
From the structures of 71 PET analogues, sixty compounds constituted as a training set and eleven compounds were used in the test set. All 71 PET compounds were built on workspace of Win CAChe 6.1 (molecular modeling software, a product of Fujitsu private limited, Japan) and energy minimization of the molecules was done using Allinger's MM2 force field followed by semi empirical PM3 method available in MOPAC module with RMS gradient 0.001Å. The stable conformations of the molecules were selected automatically by the software when the molecules subjected for optimization. Most stable structure for each compound was generated and used for calculating various physico-chemical descriptors like thermodynamic, steric and electronic values of descriptors. Some of the descriptors were calculated using the above

optimized structure of the compounds by modeling software Molecular modeling pro 6.1.0, trial version (ChemSW, Inc., www.chemsw.com).

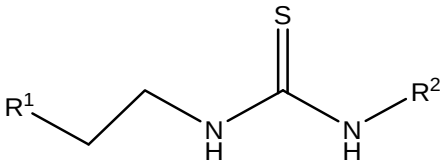
Table 1. Structures of PET analogs and their anti-HIV activity in MT-4 cell lines



Compound 1- 25



Compound 26-38



Compound 39-46

Comp No	R^1	R^2	pED_{50} (μM) Experimental[b]
1	Phenyl	-	-0.1139
2	2-fluorophenyl	-	1
3[a]	3-fluorophenyl	-	0.6021
4	4-fluorophenyl	-	-0.5185
5[a]	2-methoxyphenyl	-	0.3979
6	3-methoxyphenyl	-	0.2218
7	4-methoxyphenyl	-	-0.7404
8	2-methylphenyl	-	0.0227
9	2-nitrophenyl	-	-0.0414
10	2-hydroxyphenyl	-	-0.602
11	2-chlorophenyl	-	0.3979
12	3-ethoxyphenyl	-	0.8239
13	3-propoxyphenyl	-	-0.3424
14[a]	3-isopropoxyphenyl	-	0.3979
15	3-phenoxyphenyl	-	-0.4471
16	2,6-dimethoxyphenyl	-	1.0457
17	2,5-dimethoxyphenyl	-	0.3979
18	3-bromo-6-methoxyphenyl	-	1.301
19	2-fluoro-6-methoxyphenyl	-	0.5229
20[a]	2-ethoxy-6-fluorophenyl	-	0.6989
21	2,6-difluorophenyl	-	1.6989
22	2-chloro-6-fluorophenyl	-	1.301
23	2-pyridyl	-	-0.1139
24	3-pyridyl	-	-0.8062
25	2-furyl	-	-0.716
26	4-methylthiazol-2-yl	-	0.3979
27	4-ethylthiazol-2-yl	-	0.1549
28	4-propylthiazol-2-yl	-	-0.2041
29	4-isopropylthiazol-2-yl	-	-0.1139
30[a]	4-butylthiazol-2-yl	-	-0.1139
31	4-cyanothiazol-2-yl	-	0.6989
32	4-(trifluoro methyl)thiazol-2-yl	-	0.301
33	4-(ethoxy carbonyl)thiazol-2-yl	-	0.301
34	5-chlorothiazol-2-yl	-	-0.4314
35	1,3,4-thiazol-2-yl	-	-0.7243

36	2-pyridyl	-	0.6989
37	5-bromo-2-pyridyl	-	1.301
38	5-methyl-2-pyridyl	-	0.8239
39[a]	2,6-difluorophenyl	4-cyano thiazoly-2-yl	1.5229
40	2,6-difluorophenyl	5-bromo-2-pyridyl	2
41	2,6-difluorophenyl	5-methyl-2-pyridyl	2
42[a]	2-ethoxy-6-fluorophenyl	5-methyl-2-pyridyl	0.6989
43	2-ethoxy-6-fluorophenyl	5-bromo-2-pyridyl	1.6989
44	2-pyridyl	5-methyl-2-pyridyl	0.5228
45	2-pyridyl	5-bromo-2-pyridyl	1.6989
46	2,6-difluorophenyl	4-ethylthiazol-2-yl	1.0969

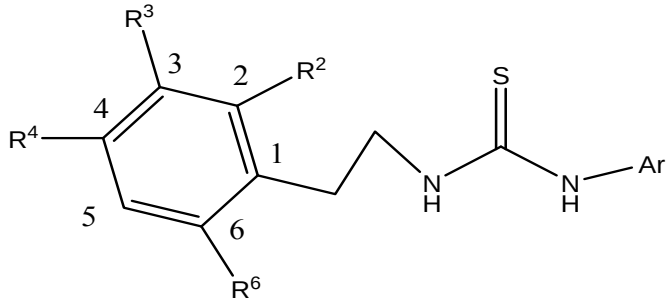
a - test set compounds

b - the experimental ED_{50} values (in micro molar) were converted into $-\log ED_{50}$ (pED50, in micro molar).

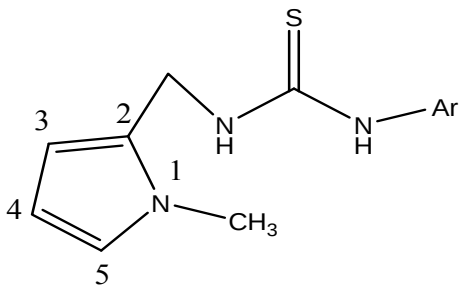
Descriptors calculation

The physicochemical properties were calculated on project leader file of the modeling software Win CAChe 6.1. In present study the calculated descriptors were conformational minimum energies (CME), Zero-order connectivity index (CI0), First-order connectivity index (CI1), Second-order connectivity index (CI2), dipole moment (DM), total energy at its current geometry after optimization of structure (TE), heat of formation at its current geometry after optimization of structure (HF), dipole vector X (DVX), dipole vector Y (DVY), dipole vector Z (DVZ), Highest occupied molecular orbital (HOMO), lowest unoccupied molecular orbital (LUMO), octanol-water partition coefficient (logP), squared partition coefficient ($logP^2$), molar refractivity (MR), shape index order 1 (SI1), shape index order 2 (SI2), Zero-order valance connectivity index (VI0), First-order valance connectivity index (VI1), Second-order valance connectivity index (VI2) and solvent accessible surface area (SAS). We have not included ionization potential (IP) and electron affinity for developing QSAR models because these are highly inter-correlated with HOMO and LUMO, respectively.

Table 2. Structures of PET analogs and their anti-HIV activity in MT-4 cell lines



Compound 47-69



Compound 70 & 71

Comp No	R^2	R^3	R^4	R^6	Ar	pED_{50} (μM)
						Experimental[b]
47	F	$(CO)N(Me)_2$	H	F	5-bromo-2-pyridyl	1.0969
48	F	CH_2NAc	H	F	5-bromo-2-pyridyl	0.0457
49	F	CN	H	F	5-chloro-2-pyridyl	2.2218
50	F	$N(Me)_2$	H	F	5-chloro-2-pyridyl	1.3979
51[a]	F	$N(Me)_2$	H	F	5-bromo-2-pyridyl	1.3979
52	F	OCH_3	H	F	5-bromo-2-pyridyl	1.8239
53[a]	F	OC_2H_5	H	F	5-bromo-2-pyridyl	2.2218

54	F	CH_2OCH_3	H	F	5-bromo-2-pyridyl	2.2218
55	Cl	OC_2H_5	H	F	5-bromo-2-pyridyl	2.1549
56	Cl	OC_2H_5	H	F	5-chloro-2-pyridyl	2.0969
57	Cl	OC_2H_5	H	F	5-iodo-2-pyridyl	1.8239
58	Cl	OC_2H_5	H	F	5-cyano-2-pyridyl	2.5229
59	H	OCH_3	H	OCH_3	5-chloro-2-pyridyl	1.3979
60	H	OC_2H_5	H	OC_2H_5	5-bromo-2-pyridyl	1.7447
61	F	H	H	OC_2H_5	5-bromo-2-pyridyl	1.886
62	F	F	H	OC_2H_5	5-bromo-2-pyridyl	2.1549
63[a]	F	F	H	OCH_3	5-bromo-2-pyridyl	2.1589
64	F	OCH_3	H	OCH_3	5-chloro-2-pyridyl	2.6989
65	F	OC_2H_5	H	OCH_3	5-chloro-2-pyridyl	2.301
66	OCH_3	OCH_3	H	F	5-bromo-2-pyridyl	1.9208
67	F	H	$N(Me)_2$	F	5-bromo-2-pyridyl	1.7447
68	F	CN	$N(Me)_2$	H	5-bromo-2-pyridyl	1.886
69	Cl	OC_2H_5	Cl	F	5-bromo-2-pyridyl	1.886
70	-	-	-	-	5-cyano-2-pyridyl	0.959
71[a]	-	-	-	-	5-chloro-2-pyridyl	0.602

a - test set compounds

b - the experimental ED_{50} values (in micro molar) were converted into $-\log ED_{50}$ (pED50, in micro molar).

QSAR models development and validation

In present study, we have used 50 physicochemical descriptors calculated by Win CAChe 6.1 and Molecular modeling pro 6.1.0 (the complete descriptors data set of all compounds will be provided on request). All the calculated descriptors were considered as independent variable and biological activity as dependent variable. STATISTICA 6 (StatSoft, Inc., Tulsa, USA) software was used to generate QSAR models by different statistical techniques.

In the present study, we used three statistical methods to develop the models: (1) multiple linear regression with factor analysis as the data pre-processing step for variable selection (FA-MLR), (2) principal component regression analysis (PCRA), and (3) partial least squares with factor analysis (FA-PLS).

In case of FA-MLR, factor analysis (FA) was used as the data-preprocessing step to identify the important predictor variables contributing to the response variable and to avoid collinearities among them even though classical approach of multiple linear regression technique was used as the final statistical tool for developing QSAR relations. In a typical factor analysis procedure, the data matrix is first standardized, correlation matrix and subsequently reduced correlation matrix are constructed, eigenvalue problem is then solved and the factor pattern can be obtained from the corresponding eigenvectors. 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 dataset(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 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 case of PCRA, factor scores (as obtained from FA) are used as the predictor variables. PCRA has an advantage that collinearities among X variables are not a disturbing factor and that the number of variables included in the analysis may exceed the number of observations. In PCRA, all descriptors are assumed to be important while the aim of factor analysis is to identify

relevant descriptors. PLS is a generalization of regression, which can handle data with strongly correlated and/or noisy or numerous X variables. The linear PLS model finds 'new variables' (latent variables) 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 of testing this significance. 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. In case of PLS analysis on the present dataset, factor loading table obtained from factor analysis was used for primary variable screening. From the factor loading table, variables with high loading (>0.7) in such factors where the activity shows high or moderate loading were selected for the PLS regression.

Statistical measures used were n-number of compounds in regression, *r*-correlation coefficient, r^2-squared correlation coefficient, *F*- test (Fischer's value) for statistical significance, SD- standard deviation, q^2- cross validated correlation coefficient and correlation matrix to show correlation among the parameters. The squared correlation coefficient (or coefficient of multiple determination) r^2 is a relative measure of fit by the regression equation. Correspondingly, it represents the part of the variation in the observed data that is explained by the regression. The correlation coefficient values closer to 1.0 represent the better fit of the regression. The F-test reflects the ratio of the variance explained by the model and the variance due to the error in the regression. High values of the F-test indicate that the model is statistically significant. Standard deviation is measured by the error mean square, which expresses the variation of the residuals or the variation about the regression line. Thus standard deviation is an absolute measure of quality of fit and should have a low value for the regression to be significant.

$q^2 = 1 -$ PRESS

$$\sum_{i=1}^{N} (y_i - y_m)^2$$

$$\text{PRESS} = \sum_{i=1}^{N} (y_{pred,i} - y_i)^2$$

Where y_i is the activity for training set compounds, y_m is the mean observed value, corresponding to the mean of the values for each cross-validation group, and $y_{pred,i}$ is the predicted activity for y_i. The predictive ability of the selected model was also confirmed by external R^2 and R^2CVext [24].

$$R^2\text{CVext} = 1 - \frac{\sum_{i=1}^{test} (y_{exp} - y_{pred})^2}{\sum_{i=1}^{test} (y_{exp} - \bar{y}_{tr})^2}$$

Where $\bar{y}_{tr}$ is the averaged value for the dependent variable for the training set.

Furthermore Tropsha et al. considered a QSAR model predictive, if the following conditions are satisfied:

$$r^2_{CVext} > 0.5,\ r^2 > 0.6,$$

$r^2 - r^2_o / r^2 < 0.1$, $r^2 - r'^2_o / r^2 < 0.1$ and $0.85 \leq k \leq 1.15$ or $0.85 \leq k' \leq 1.15$

Mathematical definitions of r^2_o, r'^2_o, k and k' are based on regression of the observed activities against predicted activities and the opposite (regression of the predicted activities against observed activities). The definitions are given clearly by Tropsha et al., and are not discussed here.

The robustness of a QSAR model was checked by Y – randomization test. In this technique, new QSAR models were developed by shuffling the dependent variable vector randomly and keeping the original independent variable as such. The new QSAR models are expected to have low r^2 and q^2 values. If the opposite happens then an acceptable QSAR model can not be obtained for the specific modeling method and data.

3. Results and Discussion

FA-MLR

It was observed that 8 factors could explain the data matrix to the extent of 98.5%, from the factor analysis on the data matrix consisting of the anti-HIV activity data and physiochemical parameters. Table 4 shows that the biological activity is highly loaded with factors 3 (highly loaded in VDW), 2 (highly loaded in EF, HF and GF), and 1 (highly loaded in MR, CI3, VI1, VI2, CT and EV), moderately loaded with factors 6 (highly loaded in TE), 5 (highly loaded in logP), 4 and 7 (highly loaded in HP and D respectively), and poorly loaded with factor 8 (considerably loaded in CP). Based on the factor analysis the following equation was derived with six variables.

pED_{50} = 4.918 (± 1.536) – 0.177 (± 0.029) MR + 0.342 (± 0.134) logP – 0.117 (± 0.022) CP + 0.092 (± 0.024) EV + 2.89 (± 0.706) D + 0.233 (± 0.038) VDW (1)

n = 60, r = 0.898, r^2 = 0.807, r^2a = 0.785, SEE = 0.464, F (6, 53) = 36.86, P <0.001, r^2_{CV} = 0.748, S_{PRESS} =0.530, PRESS = 14.9, SDEP = 0.502.

Eq. (1) could explain 80.7% of the variance and predict 74.8% of the variance. The negative coefficient of MR and CP showed that, the volume and high critical pressure is detrimental to the activity, respectively. The biological activity increase when the logP and VDW energy of the molecule increase. We have included EV in this model even it's highly inter-correlated with other descriptors, because it's not having multicolinearity problem (variance inflation factor (VIF) is less than ten). It's indicate that the volume of the substitutent should be less mean while the lipophilicity of the substituent would be more for increasing the anti-HIV activity of PET derivatives. Leave 25% out crossvalidation also carried out for this model and the data is shown in Table 6. The predictive ability of the selected model was also confirmed by external r^2_{CVext} method. According to

Tropsha et al., the proposed QSAR model is predictive as it satisfies all the conditions like $r^2_{CVext} > 0.5$, $r^2 > 0.6$, $r^2 - r^2_o / r^2 < 0.1$, $r^2 - r'^2_o / r^2 < 0.1$ and $0.85 \leq k \leq 1.15$ or $0.85 \leq k' \leq 1.15$ but this model satisfy the following criteria r^2_{CVext} = 0.508 > 0.5, r^2 =0.732 > 0.6, $r^2 – r'^2_0 / r^2$ = 0.024 < 0.1, k = 0.8864 < 1.15 but > 0.85 and k' = 0.9867 < 1.15 but > 0.85, except the condition $r^2 - r^2_0 / r^2$ < 0.1.The value of $r^2 - r^2_0 / r^2$ is 0.2143. So this QSAR model is not predictive as its not satisfy all the conditions reported by Tropsha et al [26].

pED_{50} = 5.099 (± 1.511) – 0.174 (± 0.028) MR + 0.278 (± 0.137) logP – 0.142(± 0.026) CP + 0.079 (± 0.025) EV + 0.004 (± 0.002) HF + 3.937 (± 0.917) D + 0.239 (± 0.038) VDW (2)

n = 60, r = 0.904, r^2 = 0.817, r^2a = 0.793, SEE = 0.455, F (7, 52) = 33.25, P <0.001, r^2_{CV} = 0.762, S_{PRESS} =0.520, PRESS = 14.0, SDEP = 0.488.

Eq. (2) with seven variables could explain 81.7% of the variance and predict 76.2% of the variance. There is significant improvement in statistical quality when one extra variable (HF), Eq. (2), is included additionally to the variables in Eq. (1). The positive contribution of the HF on the biological activity indicated that heat of formation is responsible for the anti-HIV activity of the PET compounds. The inter-correlation (r) matrix among the predictor variables is given in Table 5. The predictive ability of the selected model was also confirmed by external r^2_{CVext} method. According to Tropsha et al., the proposed QSAR model is predictive as it satisfies all the conditions r^2_{CVext} = 0.591 > 0.5, r^2 =0.750 > 0.6, $r^2 - r^2_0 / r^2$ = 0.0968 < 0.1, $r^2 – r'^2_0 / r^2$ = 0.0172 < 0.1, k = 0.8956 < 1.15 but > 0.85 and k' = 0.9645 < 1.15 but > 0.85. The robustness of this model was checked by Y – randomization test. The low r^2 and r^2_{CV} values indicate that the good results in our original model are not due to a chance correlation or structural dependency of the training set.

FA-PLS

The number of optimum components was found to be 6 to obtain the final equation. Based on the standardized regression coefficients, the following variables were selected for the final equation:

pED_{50} = 0.193 * HP – 0.024 * MR – 0.544 * VI2 + 0.169* CI4 – 0.097 * CP + 0.251 * logP + 2.654 * D + 0.213 * VDW (3)

n = 60, r = 0.907, r^2 = 0.823, r^2a = 0.803, SEE = 0.444, F = 41.0, P <0.001

r^2_{CV} = 0.778, S_{PRESS} =0.497, PRESS = 13.1, SDEP = 0.471.

Eq. (3) could explain 82.3% of the variance and predict 77.8% of the variance. The positive coefficient of the HP and CI4 indicates that the increase in the value of hanse polarity and connectivity index 4 of PET compounds is conducive to the activity. The calculated values according to Eq. (3) are presented in Table 3. The predictive ability of the selected model was also confirmed by external r^2_{CVext} method. According to Tropsha et al., the proposed QSAR model is predictive as it satisfies all the conditions r^2_{CVext} = 0.539 > 0.5, r^2 =0.731 > 0.6, r^2 - r^2_0/ r^2 = 0.0742 < 0.1, r^2 – r'^2_0/ r^2 = 0.0203 < 0.1, k = 0.8845 < 1.15 but > 0.85 and k' = 0.9583 < 1.15 but > 0.85. The robustness of this model was checked by Y – randomization test (data not given). The low r^2 and r^2_{CV} values indicate that the good results in our original model are not due to a chance correlation or structural dependency of the training set.

PCRA

When factor scores were used as the predictor parameters in a multiple regression equation using forward selection method (PCRA), the following equation was obtained:

pED_{50} = 0.878 (± 0.033) + 0.324 (± 0.034) fs1 + 0.354 (± 0.034) fs2 + 0.777 (± 0.034) fs3 + 0.210 (± 0.034) fs5 + 0.248 (± 0.034) fs6 (4)

n = 60, r = 0.969, r^2 = 0.940, r^2a = 0.934, SEE = 0.257, F = 167.8, P <0.001

r^2_{CV} = 0.920, S_{PRESS} =0.301, PRESS = 4.7, SDEP = 0.283.

Eq. (4) could explain 94.0% of the variance and predict 92.0% of the variance. The variables (factor scores) used in Eq. (4) are perfectly orthogonal to each other. As factor scores are used, instead of selected descriptors, in MLR equation in PCRA and any one factor-score contains information from different descriptors, loss of information is thus avoided and the quality of PCRA equation is better than those derived from FA-MLR. From the factor scores used, significance of the original variables for modeling the activity can be obtained. Factor score 1 indicates the importance of molar refractivity (MR), length and critical temperature of the entire molecule. Factor score 2 indicates the importance of enthalpy of formation and Gibbs energy of formation of the entire molecules. Factor score 3 indicates the importance of vander waals energy of the entire molecules. Factor score 5 signifies the importance of lipophilicity (logP) of the entire molecules, while factor score 6 indicates the importance of torsion energy of the entire molecules. The robustness of this model was checked by Y – randomization test (Data not given). The low r^2 and r^2_{CV} values indicate that the good results in our original model are not due to a chance correlation or structural dependency of the training set.

Table 3. Observed, calculated and predicted (LOO) anti-HIV activity data of PET derivatives.

Comp. No	Obsd	Calcd[a]	Calcd[b]	Calcd[c]
Training Set				
1	-0.1139	-0.2794	-0.2199	-0.0795
2	1.0000	0.1455	0.2452	0.7287
4	-0.5185	0.1273	0.1456	-0.3179
6	0.2218	0.1831	0.2413	0.2785
7	-0.7404	0.1783	0.2250	-0.3640
8	0.0227	-0.0078	-0.1106	0.0811
9	-0.0413	0.1073	0.2592	-0.4183
10	-0.602	-0.4381	-1.0758	-0.9869
11	0.3979	0.1657	0.1628	0.2183
12	0.8239	0.1012	0.2555	0.7908
13	-0.3424	0.0391	0.1565	0.0683
15	-0.4471	-0.3375	-0.2595	-0.0268
16	1.0457	0.6895	0.9188	1.0649
17	0.3979	0.6758	0.8206	0.6016

18	1.301	0.8813	0.7626	0.8711
19	0.5229	0.9351	1.0764	0.7382
21	1.6989	0.8337	0.9612	1.3796
22	1.301	0.8851	0.9013	1.0379
23	-0.1139	-0.7370	-0.39225	-0.2644
24	-0.8062	-0.5379	-0.2378	-0.5967
25	-0.716	-0.6140	-0.5438	-0.9963
26	-0.3979	-0.0833	-0.40866	-0.3889
27	0.1549	0.1389	-0.0109	0.2500
28	-0.2041	0.0402	-0.1034	0.0401
29	-0.113	-0.1694	-0.5118	-0.0374
31	0.6989	0.5930	0.4388	0.2375
32	0.301	0.6895	0.3901	0.2559
33	0.301	0.1644	0.3640	0.3807
34	-0.4314	-0.0462	-0.0585	-0.4348
35	-0.7243	-0.2661	0.0097	-0.5020
36	0.6989	0.9365	0.9016	1.1781
37	1.301	1.2628	1.0264	1.2246
38	0.8239	1.0952	0.8493	1.2451
40	2.0000	1.9980	1.8411	1.8150
41	2.0000	1.6848	1.5846	2.1349
43	0.6989	1.9799	1.8932	1.0522
44	0.5228	0.8345	0.7565	0.8806
45	1.6989	0.9324	0.8547	1.2918
46	1.0969	0.7804	0.8840	1.0822
47	1.0969	1.2961	1.3411	1.0335
48	0.0457	0.3715	0.3314	-0.2643
49	2.2218	2.6464	2.4533	2.0518
50	1.3979	1.7414	1.7485	1.7580
52	1.8239	2.1412	2.1024	1.8120
54	2.2218	1.8350	1.9352	2.1564
55	2.1549	1.9236	1.9311	2.0668
56	2.0969	1.8822	1.9872	2.2270
57	1.8239	2.2395	2.3605	1.7952
58	2.5229	2.0370	2.0744	2.3551
59	1.3979	1.8133	1.8463	1.773
60	1.7447	1.2975	1.5735	1.9741
61	1.886	1.6678	1.6432	1.8819
62	2.1549	1.7387	1.8296	2.0643
64	2.6989	2.0994	2.1661	2.5130
65	2.301	1.8018	2.02450	2.3199
66	1.9208	2.1538	2.2152	2.0003
67	1.7447	1.7028	1.5309	1.6880
68	1.886	1.9507	1.5984	1.5551
69	1.886	2.1902	2.1586	1.8508
70	0.959	0.6103	0.8575	0.5769
Test Set				
3	0.6021	0.1059	-0.2073	
5	0.3979	0.2543	0.1004	
14	0.3979	-0.6168	-0.6252	
20	0.6989	0.8097	1.0191	
30	-0.1139	0.0410	0.0619	
39	1.5229	2.1048	1.8716	
42	0.6989	1.0930	1.3993	
51	1.3979	1.7719	1.2919	

53	2.2218	1.8331	2.1549
63	2.1589	1.9598	1.9546
71	0.602	0.1614	0.3609

Obsd – observed activity, Calcd – calculated activity, Pred – predicted activity, [a] using Eq. (2), [b] using Eq. (3) and [c] using Eq. (4).

Table 4. Factor loadings of the variables after VARIMAX rotation.

Variable	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Communality**
pED_{50}	0.324	0.353	0.777	0.134	0.21	0.247	0.112	-0.02	0.971
HP	-0.059	-0.114	0.024	**0.976**	-0.075	0.02	0.122	-0.01	0.991
MR	**0.917**	0.201	0.171	-0.132	0.185	0.12	0.091	0.093	0.996
PC	**0.874**	0.313	0.216	-0.137	0.178	0.152	0.063	0.093	0.997
CI1	**0.845**	0.416	0.219	-0.024	0.145	0.186	0.007	0.03	0.993
CI2	**0.777**	0.476	0.152	0.040	0.229	0.225	0.047	0.112	0.975
CI3	**0.792**	0.451	0.256	0.07	0.163	0.21	0.028	0.042	0.975
VI1	**0.857**	0.304	0.100	-0.221	0.261	-0.027	0.142	0.083	0.984
VI2	**0.796**	0.280	0.001	-0.164	0.349	-0.007	0.240	0.248	0.981
KP	**0.871**	0.322	0.246	-0.137	0.051	0.163	-0.04	-0.02	0.974
CI4	**0.813**	0.447	0.230	0.041	0.185	0.13	0.019	0.04	0.971
LogP	0.480	0.238	0.264	-0.146	**0.751**	0.168	0.133	0.004	0.990
CT	**0.930**	0.008	0.086	0.158	0.039	0.144	0.221	-0.154	0.993
CP	**-0.660**	-0.396	-0.308	0.109	-0.251	-0.261	0.301	-0.23	0.975
EF	-0.331	**-0.906**	-0.165	0.084	-0.08	-0.117	-0.09	-0.031	0.997
GF	-0.178	**-0.949**	-0.133	0.020	-0.12	-0.104	-0.133	0.003	0.995
EV	**0.94**	0.070	0.089	0.124	0.003	0.14	0.180	-0.14	0.987
HF	-0.308	**-0.907**	-0.139	0.149	-0.075	-0.12	-0.091	-0.03	0.990
D	0.324	0.42	0.140	0.252	0.144	0.128	**0.766**	0.007	0.990
TE	-0.378	-0.320	-0.302	-0.039	-0.135	**-0.79**	-0.114	-0.002	0.996
VDW	**0.657**	0.274	**0.550**	-0.268	0.084	0.22	0.082	0.11	0.957
Variance	0.466	0.213	0.076	0.064	0.055	0.054	0.045	0.010	0.983

Table 5. Inter-correlation matrix for anti-HIV activity and important physicochemical variables.

	pED_{50}	**D**	**MR**	**CP**	**VDW**	**HF**	**logP**	**EV**
pED_{50}	1							
D	0.5400	1						
MR	0.5606	0.4880	1					
CP	-0.6638	-0.2445	-0.8267	1				
VDW	0.7478	0.4424	0.8475	-0.8073	1			
HF	-0.5623	-0.5575	-0.5504	0.6506	-0.6193	1		
logP	0.6250	0.4904	0.7273	-0.6991	0.6973	-0.5161	1	
EV	0.4841	0.5307	0.8265	-0.6210	0.6793	-0.3809	0.5211	1

Table 6. Results of leave-25%-out cross-validation.

Equation	No. of cycles[a]	Average regression coefficient	q^2 (Average pres)
2	4	5.008 (± 0.033) - 0.175 (± 0.033) MR + 0.296 (± 0.168) logP - 0.138 (± 0.033) CP + 0.080 (± 0.030) EV + 0.003 (± 002) HF+ 3.878 (± 1.148) D + 0.238 (± 0.045) VDW	0.746 (0.258)
3	4	0.153 (± 0.094) HP - 0.055 (± 0.035) MR - 0.447 (± 0.381) VI2 + 0.165 (± 0.360) CI4 - 0.075 (± 0.034) CP + 0.304 (± 0.161) logP + 2.743 (± 1.247) VDW	0.760 (0.242)
4	4	0.876 (± 0.038) + 0.324 (± 0.038) fs1 + 0.353 (± 0.038) fs2 + 0.776 (± 0.038) fs3 + 0.215 (± 0.038) fs5 + 0.244 (± 0.038) fs6	0.910 (0.089)

Average pres means average of absolute values of predicted residuals.
[a] Compounds were deleted in 4 cycles in the following manner: (1, 5, 9,…, 57), (2, 6, 10,……, 58), (3, 7, 11,……, 59) and (4, 8, 12,…, 60)

4. Conclusions

The structural and physicochemical requirements of PET derivatives for anti-HIV activity have been explored by the present QSAR study. The best QSAR model is obtained from PCRA (Eq. (4)) technique with explained and predicted variance of 94.0% and 92.0%, respectively. The quality of model came from stepwise regression, FA-MLR and FA-PLS (Eqs. (2 and 3)) are of comparable range with explained variance 81.7%, 82.3% and predicted variance 76.2%, 77.8%, respectively. All the developed QSAR models are having the following four descriptors MR, D, VDW and CP indicates that these variables are more important to explain the anti-HIV activity of PET compounds. The negative coefficient of MR and CP indicates that these parameters are detrimental to activity when they are increased. The positive coefficient of D and VDW indicates that these parameters are conducive to activity when they are increased. So we have concluded the QSAR study of PET compounds with the volume of substituent should be less mean while the density and lipophilicity of subtituents should be high for their anti-HIV activity. The information generated from the present study may be useful in the design of more potent PET derivatives as anti HIV agents.

Acknowledgement

We would like to thank IGEMBA yahoo group members for providing the required research articles for the present study and extent our thanks to Prof. V. K. Mourya for providing computational facilities to carry out this work.

References

[1] R. C. Gallo, S. Z. Salahuddin, M. Popovic, G. M. Shearer, M. Kaplan, B. F. Haynes, T. J. Palker, R. Redfield, J. Oleske, B. Safai, Science **224**, 500 (1984).
[2] E. D. Clercq, J. Med. Chem. **38**, 2491 (1995).
[3] J. Milton, M. J. Slater, A. J. Bird, D. Spinks, G. Scott, C. E. Price, S. Downing, D. V. S. Green, Madar, R. Bethell, D. K. Stammers, Bioorg. Med. Chem. Lett. **8**, 2623 (1998).
[4] S. D. Young, S. F. Britcher, L. O. Tran, L. S. Payne, W. C. Lumma, T. A. Lyle, J. R. Huff, P. S. Anderson, D. B. Olsen, S. S. Carroll, Antimicrob. Agents. Chemother. **39**, 2602 (1995).
[5] P. Pungpo, S. Hannongbua, J. Mol. Graphics & Modell. **18**, 581 (2000).

[6] A. C. Nair, P. Jayatilleke, X. Wang, S. Miertus, W. J. Welsh, J. Med. Chem. **45**, 973 (2002).
[7] P. R. N. Jayatilleke, A. C. Nair, R. Zauhar, W. J. Welsh, J. Med. Chem. **43**, 4446 (2000).

[8] K. Raghavan, J. K. Buolamwini, M. R. Fesen, Y. Pommier, K. W. Kohn, J. Med. Chem. **38**, 890 (1995).
[9] A. K. Debnath, S. Jiang, N. Strick, K. Lin, P. Haberfield, J. Med. Chem. **37**, 1099 (1994).
[10] V. Ravichandran, R. K. Agrawal, Bioorg. Med. Chem. Lett. **17**, 2197 (2007).
[11] V. Ravichandran, V. K. Mourya, R. K. Agrawal, Arkivoc **XIV**, 204 (2007).
[12] V. Ravichandran, V. K. Mourya, R. K. Agrawal, Internet Electron. J. Mol. Des. **6**, 363 (2007).
[13] V. Ravichandran, B. R. Prashanthakumar, S. Sankar, R. K. Agrawal, Med. Chem. Res. **17**, 1 (2008).
[14] V. Ravichandran, P. K. Jain, V. K. Mourya, R. K. Agrawal, Med. Chem. Res. **16**, 342 (2007).
[15] V. Ravichandran, V. K. Mourya, R. K. Agrawal, Digest J. Nanomat. Biostruct. **3**, 9 (2008).
[16] V. Ravichandran, V. K. Mourya, R. K. Agrawal, Indian J. Pharm. Edu. & Res. **42**, 40 (2008).
[17] V. Ravichandran, B. R. Prashanthakumar, S. Sankar, R. K. Agrawal, Eurp. J. Med. Chem. **44**, 1180 (2009).
[18] V. Ravichandran, A. Jain, V. K. Mourya, R. K. Agrawal, Chem. Pap. **62**, 596 (2008).
[19] V. Ravichandran, V. K. Mourya, R. K. Agrawal, J. Enzym. Inhib. Med. Chem. (Inpress) (2008).
[20] V. Ravichandran, V. K. Mourya, R. K. Agrawal, Digest J. Nanomat. Biostruct. **3**, 19 (2008).
[21] K. K. Sahu, V. Ravichandran, V. K. Mourya, R. K. Agrawal, Med. Chem. Res. **15**, 418 (2007).
[22] K. K. Sahu, V. Ravichandran, V. K. Mourya, R. K. Agrawal, Acta. Chim. Slov. **55**, 138 (2008).
[23] F. W. Bell, A. S. Cantrell, M. Hogberg, S. R. Jaskunas, N. G. Johansson, C. L .Jordan, M. D. Kinnic, P. Lind, J. M. Morin, R. Noreen, B. Oberg, J. A. Palkowitz, C. A. Parrish, P. Pranc, C. Sahlberg, R. J. Ternansky, R. T. Vasileff, L. Vrang, S. J. West, H. Zhang, X. X. Zhou, J. Med. Chem. **38**, 4929 (1995).
[24] A. S. Cantrell, P. Engelhardt, M. Hogberg, S. R. Jaskunas, N. G. Johansson, C. L. Jordan, M. D. Kinnic, P. Lind, J. M. Morin, R. Noreen, B. Oberg, J. A. Palkowitz, C. A. Parrish, P. Pranc, C. Sahlberg, R. J. Ternansky, R. T. Vasileff, L. Vrang, S. J. West, H. Zhang, X. X. J. Med. Chem. **39**, 4261 (1996).
[25] O. S. Weislow, R. Kiser, D. L. Fine, J. Bader, R. H. Shoemaker, M. R. Boyd, J. Nat. Cancer Inst. **81**, 577 (1989).
[26] A. Tropsha, P. Gramatica, V. K. Gombar, QSAR & Combin Sci. **22**, 69 (2003).