Available online at www.sciencedirect.com
SCIENCE DIRECT®
Bioorganic & Medicinal Chemistry
Bioorganic & Medicinal Chemistry 14 (2006) 1039–1046

QSAR by LFER model of HIV protease inhibitor mannitol derivatives using FA-MLR, PCRA, and PLS techniques

J. Thomas Leonard and Kunal Roy*

Drug Theoretics and Cheminformatics Lab, Division of Medicinal and Pharmaceutical Chemistry, Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700 032, India

Received 5 August 2005; revised 7 September 2005; accepted 8 September 2005
Available online 5 October 2005

Abstract—The present quantitative structure–activity relationship (QSAR) study attempts to explore the structural and physicochemical requirements of mannitol derivatives for HIV protease inhibitory activity using linear free energy related model of Hansch. QSAR models have been developed using electronic (Hammett σ), hydrophobicity (π), and steric (molar refractivity and STERIMOL L, $B1$, and $B5$) parameters of phenyl ring substituents of the compounds along with appropriate dummy variables. Whole molecular descriptors like partition coefficient ($\log P_{calcd}$) and molar refractivity (MR) were also tried as additional descriptors. Statistical techniques like stepwise regression, multiple linear regression with factor analysis as the data preprocessing step (FA-MLR), principal component regression analysis (PCRA), and partial least squares (PLS) analysis were applied to identify the structural and physicochemical requirements for HIV protease inhibitory activity. The generated equations were statistically validated using leave-one-out technique. The quality of equations obtained from stepwise, FA-MLR, PCRA, and PLS are of acceptable statistical range (explained variance ranging from 74.0% to 80.5%, while predicted variance ranges from 70.3% to 77.1%). The coefficient of molar refractivity shows that the activity decreases with increase in volume. Lipophilicity of the *para* substituents at Y position is conducive to the activity while lipophilicity of the *para* substituents at X position is detrimental to the activity. The coefficients of molar refractivity (mr_{Y_p}) and STERIMOL parameters for *para* substituents at X and Y positions ($B1_{X\_p}$ and $B5_{Y\_p}$) of the phenyl rings indicate that the width of the substituents at X position and the overall size of *para* substituents at Y position are the detrimental factors for the activity. The fluoro substituent at *ortho* position (Y) decreases the activity when compared to the corresponding unsubstituted congener. Presence of hydrogen bond donor groups at *para* position (Y) also reduces the activity. Additionally, presence of substituent at *ortho* position (X) and the presence of substituent at *para* position (Y) are conducive for the activity. Presence of fluorine at X and Y positions also increases the activity.

1. Introduction

Acquired immunodeficiency syndrome (AIDS) is the most 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 macrophages, and results in opportunistic infections, neurological and neoplastic diseases 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 consists of 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 (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; (3) and 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 inhibitors. The antiviral agents that inhibit HIV entry to the target cells are denoted as HIV entry inhibitors, which consist of three categories: gp120-CD4 binding inhibitors, gp 120-co-receptor binding inhibitors, and fusion inhibitors.

Abbreviations: QSAR, Quantitative structure–activity relationships; AIDS, Acquired immuno deficiency syndrome; HIV, Human immunodeficiency virus; LFER, Linear free energy related; MR, Molar refractivity; PLS, Partial least squares.

Keywords: QSAR; Hansch analysis; LFER; HIV-protease; Mannitol derivatives.

* Corresponding author. Tel.: +91 33 2867 0786; fax: +91 33 2837 1078; e-mail: kunalroy_in@yahoo.com
URL: http://www.geocities.com/kunalroy_in

0968-0896/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2005.09.022

Post-entry steps[4] require the viral reverse transcriptase (RT), integrase, and protease (PR) 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] Reverse transcriptase inhibitors were the first agents approved for the treatment of HIV-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. Appropriate combinations of these drugs (referred to as highly active antiretroviral therapy or HAART) markedly suppress viral replication in most treated persons, leading to significant restoration of immune system function. HAART is responsible for dramatic reductions in HIV-associated morbidity and mortality.[5,6] However, the quest for improved therapies continues, because of problems that seriously limit the current HAART regimens, including toxic side effects, viral persistence, and difficulties in adhering to treatment, high cost, and the emergence of drug-resistant escape variants.[7]

Among various methods of anti-HIV activity screening, some important methods are cytoprotection assay, integration enzyme assay, multinuclear-activation galactosidase indicator (MAGI) assay, RT inhibition assay, HIV attachment assay, fusion assay, cytotoxicity assay, time-of-addition assay, inhibition of HIV-1 transactivation, etc.[8,9]

The present group of authors has developed a few quantitative structure–activity relationship (QSAR) models for anti-HIV activities of different group of compounds, for example, 2-amino-6-arylsulfonylbenzonitriles,[10,11] alkenyldiarylmethanes,[12] benzylpyrazoles,[13] thiazolidin-4-ones,[14] and imidazoles.[15] In continuation of such efforts, the present paper deals with QSAR modeling of HIV protease inhibitory activity of mannitol derivatives.[16]

2. Results and discussion

2.1. Stepwise regression

Using the stepping criteria based on F value ($F = 4$ for inclusion; $F = 3.99$ for exclusion), the following best equation was derived with four variables.

$$\begin{aligned} pC = {} & -0.311(\pm 0.249)\pi_{X\_p} - 0.961(\pm 0.721)\sigma^2_{Y\_p} \\ & -0.804(\pm 0.272)I_{Y\_Hbond\_do\_P} \\ & +0.523(\pm 0.170)N_{X\_O} + 2.600(\pm 0.114) \\ n = 35,\ & R^2_a = 0.764,\ R^2 = 0.792,\ R = 0.890, \\ F = 28.5(& df\ 4, 30),\ s = 0.247,\ \mathrm{SDEP} = 0.261, \\ S_{\mathrm{PRESS}} = 0.282,\ & Q^2 = 0.730,\ \mathrm{PRESS} = 2.390. \end{aligned} \quad (1)$$

The 95% confidence intervals of the regression coefficients are mentioned within parentheses. All regression coefficients are significant at 95% level. Eq. 1 could explain 76.4% of the variance and predict 73.0% of the variance. The calculated values according to Eq. 1 are presented in Table 1. The negative coefficient of the lipophilic substituent constant (π_{X_p}) of the *para* substituents of X position shows that the activity decreases with increase in lipophilicity of *para* substituents at X position. The negative coefficient of $\sigma^2_{Y\_p}$ indicates that presence of highly electron-withdrawing substituents (like CN or CF_3) at the *para* position (Y) is not favorable for the binding affinity. Presence of hydrogen bond donor groups (like CH_2OH or CH_2NH_2) at the *para* position of Y decreases the activity as evident from the negative coefficient of $I_{Y\_Hbond\_do\_P}$. Additionally, presence of substituents at *ortho* position of X is conducive to the activity as evident from the positive coefficient of N_{X_O}.

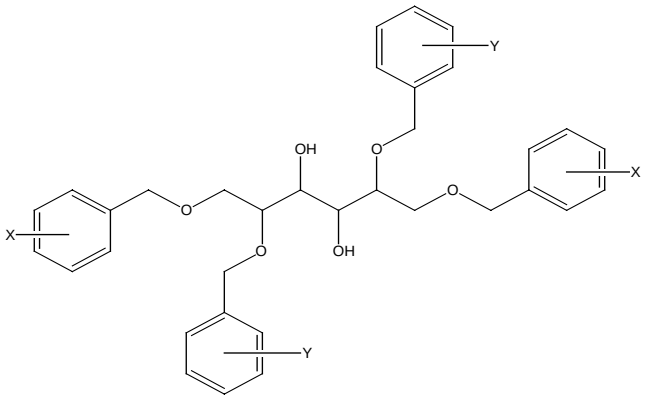
2.2. FA-MLR

From the factor analysis on the data matrix consisting of the protease inhibitory data, physiochemical parameters, and indicator and integer variables, it was observed that 12 factors could explain the data matrix to the extent of 95.9%. Table 2 shows that the biological activity is highly loaded with factors 4 (highly loaded in N_{X_F}, N_{X_O}, and $I_{X\_di\_F}$) and 1 (highly loaded in steric factors MR, MR^2, mr_{Y_p}, $mr^2_{Y\_p}$, L_{Y_p}, $B1_{Y\_p}$, and $B5_{Y\_p}$), moderately loaded with factors 7 (highly loaded in π_{X_p}), 3 (highly loaded in $\log P_{\mathrm{calcd}}$, $[\log P_{\mathrm{calcd}}]^2$, and π_{Y_p}), 9 (considerably loaded in $I_{Y\_H\_bond\_do\_P}$), and 12 (not considerably loaded in any variables). The biological activity is poorly loaded with other factors (factors 2, 5, 6, 8, 10, and 11). Based on the factor analysis the following equation was derived with seven variables.

$$\begin{aligned} pC = {} & -0.016(\pm 0.012)MR + 0.510(\pm 0.181)N_{X\_O} \\ & -0.702(\pm 0.292)I_{Y\_Hbond\_do\_P} + 5.118(\pm 1.964) \\ n = 35,\ & R^2_a = 0.740,\ R^2 = 0.763,\ R = 0.873, \\ F = 33.2(& df\ 3, 31),\ s = 0.260,\ \mathrm{SDEP} = 0.274, \\ S_{\mathrm{PRESS}} = 0.291,\ & Q^2 = 0.703,\ \mathrm{PRESS} = 2.621. \end{aligned} \quad (2)$$

Eq. 2 could explain 74.0% of the variance and predict 70.3% of the variance. The negative coefficient of molar refractivity (MR) showed that the volume is detrimental to the activity (considering MR a measure of volume and neglecting its polarizability component). When MR was replaced with $B1_{X\_P}$ and $\sigma^2_{Y\_P}$ in Eq. 2, there was further improvement in the statistical quality:

Table 1. Structural features, observed and calculated data of HIV-protease inhibitory activity of mannitol derivatives



Compound	X	Y	log P_{calcd}	MR	Obsd[a]	Calcd[b]	Calcd[c]	Calcd[d]
1	H	H	4.50	155.46	2.620	2.600	2.760	2.619
2	4-CH_3	4-CH_3	5.30	179.05	2.155	2.399	2.277	2.299
3	4-CF_3	4-CF_3	6.15	181.48	2.222	2.047	2.227	2.102
4	4-F	4-F	4.46	157.08	2.796	2.553	2.726	2.942
5	4-CN	4-CN	3.03	179.86	2.509	2.359	2.260	2.445
6	H	4-F	4.52	156.27	3.155	2.597	2.743	2.855
7	H	4-Cl	5.64	164.67	2.699	2.550	2.571	2.589
8	H	4-Br	5.53	170.84	2.678	2.550	2.445	2.515
9	H	3-F	4.54	156.27	2.745	2.600	2.743	2.770
10	H	2-F	4.55	156.27	2.432	2.600	2.490	2.551
11	H	2,6-F_2	4.62	157.08	2.194	2.600	2.221	2.294
12	H	2,4-F_2	4.57	157.08	2.585	2.597	2.474	2.598
13	H	4-CF_3	5.47	168.47	2.000	2.320	2.493	2.514
14	H	4-CH_3	4.82	167.26	2.432	2.573	2.518	2.540
15	H	4-CN	3.98	167.66	2.268	2.182	2.510	2.423
16	H	4-$COOCH_3$	4.27	179.96	2.076	2.406	2.258	2.066
17	H	4-COOH	3.69	169.09	1.585	1.602	1.766	1.863
18	H	4-CH_2OH	3.24	169.99	2.000	1.794	1.748	1.760
19	H	4-CH_2NH_2	3.07	173.70	1.569	1.785	1.672	1.595
20	H	4-$CONH_2$	2.96	173.88	1.699	1.672	1.668	1.570
21	4-F	H	4.52	156.27	2.602	2.557	2.743	2.706
22	3-F	H	4.53	156.27	2.824	2.600	2.743	2.809
23	2-F	H	4.54	156.27	3.222	3.123	3.206	3.056
24	2,6-F_2	H	4.66	157.08	3.481	3.646	3.653	3.492
25	4-Br	H	5.52	170.84	2.222	2.333	2.445	2.219
26	4-Cl	H	5.62	164.67	2.284	2.380	2.571	2.286
27	4-CF_3	H	5.47	168.47	2.301	2.327	2.493	2.207
28	4-CH_3	H	4.82	167.26	2.523	2.426	2.518	2.378
29	4-CN	H	3.96	167.66	2.367	2.777	2.510	2.640
30	4-$COOCH_3$	H	4.24	179.96	2.699	2.603	2.258	2.495
31	4-COOH	H	3.65	169.09	2.301	2.607	2.481	2.505
32	4-CH_2OH	H	3.25	169.99	2.921	2.793	2.462	2.668
33	2-F	2-CH_3	4.91	168.07	3.097	3.123	2.965	3.056
34	2,6-F_2	4-F	4.60	157.89	3.481	3.642	3.637	3.728
35	2-F	4-F	4.57	157.08	3.699	3.120	3.190	3.292

[a] Obsd = Observed (Ref. 16), Calcd = Calculated.
[b] From Eq. 1.
[c] From Eq. 5.
[d] From Eq. 7.

$$pC = -0.326(\pm0.290)B1_{X\_p} - 0.990(\pm0.728)\sigma^2_{Y\_p} + 0.455(\pm0.182)N_{X\_O} - 0.237(\pm0.231)I_{Y\_F\_O} - 0.908(\pm0.292)I_{Y\_Hbond\_do\_P} + 3.032(\pm0.413) \quad (3)$$

$$n = 35,\ R^2_a = 0.764,\ R^2 = 0.798,\ R = 0.894,$$
$$F = 23.0(df\ 5, 29), s = 0.247, \mathrm{SDEP} = 0.269,$$
$$S_{\mathrm{PRESS}} = 0.295,\ Q^2 = 0.713,\ \mathrm{PRESS} = 2.528.$$

Eq. 3 could explain 76.4% of the variance and predict 71.3% of the variance. The regression coefficient of $I_{Y\_F\_O}$ is significant at 95.4% level. Presence of a negative $B1_{X\_p}$ steric term suggests that the width of the *para* substituent (X) is not conducive to the activity; thus, groups like Br and CF_3 at X position (*para*) may cause steric hindrance for the binding to the active site.

Table 2. Factor loadings of the variables [data matrix: HIV protease inhibitory activity, physicochemical, indicator, and integer variables] after VARIMAX rotation

	1	2	3	4	5	6	7	8	9	10	11	12	Communality
pC	−0.54	−0.09	0.26	0.55	−0.12	−0.09	−0.28	−0.06	0.23	−0.16	0.19	−0.23	0.931
$\log P_{\text{calcd}}$	−0.12	−0.03	**0.89**	0.04	0.01	−0.03	0.41	−0.02	−0.02	0.03	−0.05	−0.08	0.988
$\log P^2_{\text{calc}}$	−0.07	0.00	**0.88**	0.02	0.00	−0.01	0.44	0.01	−0.02	0.09	−0.03	−0.07	0.989
MR	**0.70**	**0.61**	−0.03	−0.16	−0.16	0.15	0.07	0.12	−0.05	0.10	0.06	−0.12	0.980
MR^2	**0.70**	**0.61**	−0.03	−0.16	−0.16	0.15	0.07	0.12	−0.04	0.09	0.06	−0.12	0.981
π_{X_P}	0.01	0.13	0.34	−0.05	−0.05	−0.09	**0.89**	−0.10	−0.02	0.02	−0.04	0.00	0.954
$\pi^2_{X\_P}$	−0.04	0.48	0.19	−0.12	−0.08	0.06	0.64	0.37	−0.08	−0.01	0.18	0.03	0.877
mr_{X_P}	−0.12	**0.95**	0.01	−0.11	−0.08	−0.04	0.08	0.15	−0.08	−0.04	−0.02	−0.03	0.986
$mr^2_{X\_P}$	−0.13	**0.94**	−0.01	−0.06	−0.06	−0.06	−0.04	0.02	−0.05	−0.02	−0.16	−0.08	0.955
σ_{X_P}	−0.06	0.50	−0.01	−0.05	−0.05	0.15	0.05	**0.82**	−0.03	0.01	−0.04	0.00	0.955
$\sigma^2_{X\_P}$	0.00	0.50	−0.02	−0.07	−0.05	0.16	−0.01	**0.83**	−0.02	0.01	0.07	−0.02	0.973
L_{X_P}	−0.16	**0.87**	−0.05	−0.13	−0.10	0.02	0.00	0.38	−0.07	−0.06	0.10	0.04	0.974
$B1_{X\_P}$	−0.13	**0.72**	0.12	−0.14	−0.11	0.01	0.46	0.38	−0.07	−0.04	0.18	0.08	0.983
$B5_{X\_P}$	−0.12	**0.90**	0.08	−0.12	−0.09	−0.06	0.16	0.13	−0.06	0.00	0.10	0.08	0.915
π_{Y_P}	−0.03	0.03	**0.96**	0.00	0.00	−0.02	−0.09	−0.03	0.09	−0.03	0.04	0.18	0.972
$\pi^2_{Y\_P}$	0.62	−0.10	−0.03	−0.13	−0.10	0.12	0.03	0.02	0.08	**0.72**	−0.02	0.00	0.958
mr_{Y_P}	**0.93**	−0.15	−0.11	−0.16	−0.13	0.17	−0.06	−0.05	−0.03	0.05	−0.05	−0.05	0.986
$mr^2_{Y\_P}$	**0.91**	−0.16	−0.15	−0.12	−0.09	0.09	−0.06	−0.06	−0.11	−0.11	−0.04	−0.13	0.959
σ_{Y_P}	0.39	−0.04	0.02	−0.05	−0.04	**0.88**	−0.04	0.13	0.05	0.02	−0.03	0.02	0.956
$\sigma^2_{Y\_P}$	0.41	0.03	−0.03	−0.09	−0.07	**0.88**	−0.02	0.14	0.06	0.05	0.02	−0.03	0.977
L_{Y_P}	**0.83**	−0.18	−0.21	−0.09	−0.09	0.39	−0.06	−0.01	0.21	0.01	−0.04	0.02	0.976
$B1_{Y\_P}$	**0.73**	−0.16	0.29	−0.07	−0.08	0.36	−0.07	0.00	0.34	0.23	−0.01	0.16	0.991
$B5_{Y\_P}$	**0.90**	−0.17	−0.12	−0.11	−0.09	0.11	0.05	−0.08	0.11	0.15	−0.04	0.16	0.956
N_{X_F}	−0.28	−0.25	0.00	**0.84**	−0.12	−0.06	−0.04	−0.08	0.00	−0.04	0.26	0.01	0.944
N_{Y_F}	−0.28	−0.25	0.01	0.06	**0.76**	−0.09	−0.04	−0.05	0.45	−0.13	−0.06	−0.03	0.965
N_{X_O}	−0.15	−0.16	0.02	**0.94**	−0.06	−0.05	−0.04	−0.05	0.01	0.01	0.04	−0.12	0.952
$I_{Y\_F\_O}$	−0.14	−0.14	0.00	−0.11	**0.93**	−0.04	−0.02	−0.03	−0.08	0.02	−0.08	−0.08	0.945
$I_{Y\_Hbond\_do\_P}$	0.47	−0.16	−0.66	−0.09	−0.08	−0.14	0.13	−0.10	0.05	0.29	−0.16	0.28	0.924
I_{Y_P}	0.67	−0.22	0.05	0.02	−0.01	0.23	−0.07	−0.06	0.60	0.08	0.04	0.17	0.965
I_X	−0.38	0.51	0.06	0.38	−0.24	−0.07	0.19	0.20	−0.08	−0.07	0.52	−0.03	0.974
I_Y	0.52	−0.38	0.04	−0.04	0.26	0.18	−0.08	−0.08	0.60	0.11	−0.07	−0.16	0.936
$I_{X\_di\_F}$	−0.04	−0.06	0.01	**0.94**	0.01	−0.01	0.00	0.00	−0.04	−0.03	−0.19	0.13	0.944
$I_{Y\_di\_F}$	−0.07	−0.09	0.02	−0.07	**0.95**	−0.02	−0.04	−0.04	−0.01	−0.01	0.05	0.09	0.931
% variance	0.289	0.243	0.110	0.096	0.064	0.042	0.035	0.024	0.019	0.015	0.012	0.010	0.959

Note. Factor loadings more than 0.7 are shown in bold face.

$$\begin{aligned} pC = {} & -0.393(\pm 0.299)B1_{X\_p} + 0.257(\pm 0.198)\pi_{Y\_p} \\ & -0.409(\pm 0.135)B5_{Y\_p} + 0.431(\pm 0.186)N_{X\_O} \\ & -0.274(\pm 0.237)I_{Y\_F\_O} + 3.592(\pm 0.532) \\ n = 35,\ & R_a^2 = 0.761,\ R^2 = 0.796,\ R = 0.892, \\ F = 22.7(df\ & 5,29), s = 0.249, \text{SDEP} = 0.267, \\ S_{\text{PRESS}} = 0.293,\ & Q^2 = 0.718,\ \text{PRESS} = 2.489. \end{aligned} \tag{4}$$

Eq. 4 with five variables could explain 76.1% of the variance and predict 71.8% of the variance. The presence of a negative $B5_{Y\_p}$ steric term suggests that the width of the *para* substituent (Y) is not conducive to the activity; thus, groups like $COOCH_3$, CH_2NH_2 and $CONH_2$ at Y (*para*) position may cause steric hindrance for the binding to the active site.

When an additional term N_{X_O} (highly loaded with factor 4) was included in Eq. 2, there was further improvement in the statistical quality of the predicted variance:

$$\begin{aligned} pC = {} & -0.020(\pm 0.012)MR + 0.464(\pm 0.176)N_{X\_O} \\ & - 0.253(\pm 0.231)I_{Y\_F\_O} \\ & - 0.715(\pm 0.274)I_{Y\_Hbond\_do\_P} \\ & + 5.941(\pm 1.997) \\ n = 35,\ & R_a^2 = 0.770,\ R^2 = 0.797,\ R = 0.893, \\ F = 29.4(df\ & 4,30), s = 0.244,\ \text{SDEP} = 0.255, \\ S_{\text{PRESS}} = 0.276,\ & Q^2 = 0.741,\ \text{PRESS} = 2.281. \end{aligned} \tag{5}$$

Eq. 5 could explain 77.0% of the variance and predict 74.1% of the variance. The regression coefficient of $I_{Y\_F\_O}$ is significant at 96.7% level. The statistics

Table 3. Intercorrelation (r) matrix for physiochemical, indicator, and integer variables

	π_{X_P}	$B1_{X\_P}$	π_{Y_P}	$\sigma^2_{Y\_P}$	$B5_{Y\_P}$	N_{X_O}	$I_{Y\_F\_O}$	$I_{Y\_Hbond\_do\_P}$
MR	0.130	0.458	−0.056	0.487	0.564	−0.333	−0.316	0.249
π_{X_P}	1.000	0.502	0.248	−0.107	0.004	−0.092	−0.069	−0.087
$B1_{X\_P}$	0.502	1.000	0.109	0.028	−0.249	−0.271	−0.204	−0.255
π_{Y_P}	0.248	0.109	1.000	−0.055	−0.109	0.021	0.000	−0.592
$\sigma^2_{Y\_P}$	−0.107	0.028	−0.055	1.000	0.479	−0.191	−0.145	0.071
$B5_{Y\_P}$	0.004	−0.249	−0.109	0.479	1.000	−0.218	−0.188	0.641
N_{X_O}	−0.092	−0.271	0.021	−0.191	−0.218	1.000	−0.110	−0.137
$I_{Y\_F\_O}$	−0.069	−0.204	0.000	−0.145	−0.188	−0.110	1.000	−0.103

(R^2 and R_a^2) of Eqs. 3–5 are of same quality, but the predicted variance (Q^2) of the Eq. 5 is much superior to other models. Eq. 5 involves four descriptors for 35 data points and thus maintains the recommended ratio of number of descriptors to number of data points of 1:5.[17,18] The calculated activity values according to Eq. 5 is given in Table 1. The intercorrelation (r) matrix among the predictor variables used in Eqs. 1–5 is given in Table 3.

2.3. 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:

$$\begin{aligned} pC = & -0.273(\pm0.080)fs1 + 0.132(\pm0.080)fs3 \\ & +0.278(\pm0.080)fs4 - 0.141(\pm0.080)fs7 \\ & +0.117(\pm0.080)fs9 - 0.119(\pm0.080)fs12 \\ & +2.527(\pm0.078) \qquad (6) \\ n = 35,\ R_a^2 = 0.805,\ R^2 = 0.839,\ R = 0.916, & \\ F = 24.3(df\ 6,28),\ s = 0.225,\ \mathrm{SDEP} = 0.259, & \\ S_{\mathrm{PRESS}} = 0.289,\ Q^2 = 0.734,\ \mathrm{PRESS} = 2.341. & \end{aligned}$$

Eq. 6 shows excellent equation statistics (80.5% explained variance) and crossvalidation parameters (73.4% predicted variance). The variables (factor scores) used in Eq. 6 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 importance of molar refractivity (MR) of the entire molecules, and length (L) and width ($B1$ and $B5$) of substituents at X position. Factor score 3 indicates importance of the lipophilicity ($\log P_{\mathrm{calcd}}$) of the entire molecules and the π value of the Y substituents, while factor score 4 signifies importance of fluorine at X position. Factor score 7 signifies the importance of lipophilicity of the substituents at X, while factor score 9 signifies importance of the substituents at Y position. Factor score 12 signifies the importance of the hydrogen bond donor groups at Y.

2.4. PLS

The number of optimum components was found to be 3 to obtain the final equation (optimized by crossvalidation). Based on the standardized regression coefficients, the following variables were selected for the final equation:

$$\begin{aligned} pC = & -0.246\pi_{\mathrm{X\_p}} - 0.198B1_{\mathrm{X\_p}} + 0.148\pi_{\mathrm{Y\_p}} \\ & -0.252mr_{\mathrm{Y\_p}} - 0.156B5_{\mathrm{Y\_p}} + 0.191N_{\mathrm{X\_F}} \\ & +0.151N_{\mathrm{Y\_F}} + 0.246N_{\mathrm{X\_O}} + 0.116I_{\mathrm{Y\_P}} \\ & -0.461I_{\mathrm{Y\_Hbond\_do\_P}} - 0.189I_{\mathrm{Y\_di\_F}} \qquad (7) \\ & -0.219I_{\mathrm{Y\_F\_O}} + 2.997 \\ n = 35,\ R_a^2 = 0.791,\ R^2 = 0.865,\ R = 0.873, & \\ Q^2 = 0.771,\ \mathrm{PRESS} = 2.022. & \end{aligned}$$

The negative coefficient of the molar refractivity substituent constant ($\mathrm{mr_{Y\_p}}$) of the *para* substituents (Y) shows that the increase in the size of the *para* substituents at Y position is not conducive to the activity. The positive coefficient of $N_{\mathrm{Y\_F}}$ and $N_{\mathrm{X\_F}}$ indicates that the presence of fluorine at X and Y positions favors the activity. Presence of hydrogen bond donor groups at the *para* position of Y decreases the activity. The positive coefficient of $I_{\mathrm{Y\_P}}$ indicates that presence of a substituent at *para* position (Y) favors activity. The negative coefficients of $I_{\mathrm{Y\_di\_F}}$ and $I_{\mathrm{Y\_F\_O}}$ indicate that the presence of two fluoro substituents and also presence of fluorine at *ortho* position (Y) decrease activity. Of all the models, PLS derived Eq. 7 has the best predicted variance (77.1%) for the activity data. The calculated values according to Eq. 7 are presented in Table 1.

Leave-33%-out crossvalidation was applied on Eqs. 1, 5, and 7 and the results are shown in Table 4. Crossvalidation statistics indicate robustness of the formulated models.

3. Conclusions

The present QSAR study has explored the structural and physicochemical requirements of mannitol derivatives for HIV protease inhibitory activity using linear free energy related (LFER) model of Hansch. The coefficients of molar refractivity for *para* substituents of Y ($\mathrm{mr_{Y\_p}}$) and

Table 4. Results of leave-33%-out cross-validation applied on Eqs. 1, 5 and 7 Model equation, $pC = \Sigma\beta_i x_i + \alpha$

Equation	No. of cycles	Average regression coefficients (±standard deviations)	Q^2 statistic (Average Pres)
1	3[a]	$-0.317(\pm0.023)\pi_{\mathrm{X\_p}} - 1.004(\pm0.286)\sigma^2_{\mathrm{Y\_p}} + 0.533(\pm0.103)N_{\mathrm{X\_O}}$ $-0.814(\pm0.047)I_{\mathrm{Y\_Hbond\_do\_P}} + 2.601(\pm0.055)$	0.716 (0.220)
5	3[a]	$-0.021(\pm0.003)\mathrm{MR} + 0.469(\pm0.046)N_{\mathrm{X\_O}} - 0.245(\pm0.033)I_{\mathrm{Y\_F\_O}}$ $-0.703(\pm0.094)I_{\mathrm{Y\_Hbond\_do\_P}} + 5.957(\pm0.396)$	0.741 (0.204)
7	3[a,b]	$-0.241(\pm0.048)\pi_{\mathrm{X\_p}} - 0.170(\pm0.025)B1_{\mathrm{X\_p}} + 0.130(\pm0.029)\pi_{\mathrm{Y\_p}}$ $-0.274(\pm0.115)mr_{\mathrm{Y\_p}} - 0.146(\pm0.027)B5_{\mathrm{Y\_p}} + 0.216(\pm0.047)N_{\mathrm{X\_F}}$ $+0.103(\pm0.082)N_{\mathrm{Y\_F}} + 0.267(\pm0.034)N_{\mathrm{X\_O}} - 0.218(\pm0.073)I_{\mathrm{Y\_F\_O}}$ $-0.428(\pm0.127)I_{\mathrm{Y\_Hbond\_do\_P}} + 0.084(\pm0.115)I_{\mathrm{Y\_P}}$ $-0.108I_{\mathrm{Y\_di\_F}} + 2.968(\pm0.058)$	0.744 (0.196)

Q^2 denotes cross-validated R^2. Average Pres means the average of absolute values of predicted residuals.

[a] Compounds were deleted in 3 cycles in the following manner: (1, 4, 7, …, 34), (2, 5, 8, …, 35), (3, 6, 9, …, 30).

[b] Average and standard deviation were not calculated for $I_{\mathrm{Y\_di\_F}}$, since $I_{\mathrm{Y\_di\_F}}$ appeared only once in three cycles.

STERIMOL parameters for *para* substituents at X ($B1_{X\_p}$) and at Y ($B5_{Y\_p}$) of the phenyl ring of benzyloxy fragments indicate that the widths of the *para* substituents at X and Y positions are detrimental factors for the activity. The coefficient of molar refractivity (MR) also shows that the volume of compounds is not conducive for the activity. The lipophilicity of the *para* substituents of Y is conducive for the activity, while the lipophilicity of *para* substituents of X is detrimental for the HIV protease inhibitory activity. Furthermore, presence of fluoro substituent at *ortho* position and hydrogen bond donor groups at *para* position of Y is detrimental to the activity. Again, presence of two fluoro substituents at Y is also detrimental to the activity. Additionally, presence of substituents at *ortho* position of X and at *para* position of Y is conducive to the activity, while presence of fluorine at X and Y positions favors the activity.

4. Materials and methods

HIV protease inhibitory data reported by Bouzide et al.[16] have been used for the present QSAR study. The activity data [$IC_{50}(\mu M)$ determined with Matayoshi fluorometric assay] of 1,2,5,6-tetra-*O*-benzyl-D-mannitol derivatives (Table 1) have been converted to the logarithmic scale [pC(mM)] and then used for subsequent QSAR analyses as the response variables. There are two regions of structural variations in the compounds: one is at X position and the other at Y position (Table 1). This paper uses the classical LFER approach using substituent constants (Table 5) and whole molecular physicochemical descriptors.[19,20] The objective of the work was to find out the contribution pattern of the phenyl ring substituents to the protease inhibitory activity. Compounds with non-graded quantitative activity data reported in the original work have been excluded in the present study. The activity data (Table 1) were subjected to QSAR analyses using linear free energy related (LFER) model of Hansch[19,20] with lipophilicity (π), electronic parameter (Hammett σ), and steric (molar refractivity MR and STERIMOL *L*, *B*1, and *B*5) parameters of the phenyl ring substituents along with appropriate indicator and integer variables (defined in Table 6). The values of the physiochemical substituent constants (Table 5) were taken from the literature.[21] Hydrophobic and steric whole molecular descriptors (partition coefficient $\log P_{calcd}$ and molar refractivity MR) were also tried as predictor variables. SMILES were generated from the structures using the JME molecular editor (http://www.molinspiration.com/jme/) and then $\log P_{calcd}$ values were calculated using the ALOGPS 2.1 software [Virtual Computational Chemistry Laboratory (VCC-LAB); http://vcclab.org/lab/alogps]. The software Chem Draw Ultra ver 5.0[22] was used for the calculation of MR values (Ghose and Crippen's fragmentation method[23]). The calculated $\log P_{calcd}$ and MR values for all compounds are given in Table 1.

Table 5. Values of physicochemical parameters (substituent constants)[a]

	π	mr[b]	σ_p	*L*	*B*1	*B*5
H	0	0.103	0	2.06	1.00	1.00
CH_3	0.56	0.56	−0.17	2.87	1.52	2.04
CF_3	0.88	0.50	0.54	3.30	1.99	2.61
CN	−0.57	0.63	0.66	4.23	1.60	1.60
F	0.14	0.09	0.06	2.65	1.35	1.35
Cl	0.71	0.60	0.23	3.52	1.80	1.80
Br	0.86	0.89	0.23	3.82	1.95	1.95
$COOCH_3$	−0.01	1.29	0.45	4.73	1.64	3.36
COOH	−0.02	0.69	0.45	3.91	1.60	2.66
CH_2OH	−0.62	0.72	−0.05	3.97	1.52	2.70
CH_2NH_2	−1.04	0.91	−0.11	4.02	1.52	3.05
$CONH_2$	−1.07	0.98	0.36	4.06	1.50	3.07

[a] Taken from Ref. 21.
[b] mr values scaled to a factor of 0.1 as usual.

For the development of equations, four methods were used: (1) stepwise regression, (2) multiple linear regression with factor analysis as the data pre-processing step for variable selection (FA-MLR), (3) principal component regression analysis (PCRA), and (4) partial least squares (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

Table 6. Definitions of indicator, integer, and physicochemical parameters

Parameter	Definition
$\log P_{calcd}$	Calculated $\log P$ values for whole molecules
MR	Calculated MR values for whole molecules
π_{X_P}	π value of *para* substituents present at X
σ_{X_P}	Hammett σ constant of *para* substituents present at X
mr_{X_P}	Molar refractivity value of *para* substituents present at X
π_{X_P}	π value of *para* substituents present at Y
σ_{Y_P}	Hammett σ constant of *para* substituents present at Y
mr_{Y_P}	Molar refractivity value of *para* substituent present at Y
N_{X_F}	Number of fluorine atoms present at X
N_{Y_F}	Number of fluorine atoms present at Y
N_{X_O}	Number of substituents present at *ortho* position of X
$I_{Y\_F\_O}$	Indicator variable having value 1 if fluorine is present at *ortho* position of Y, value 0 otherwise
$I_{Y\_Hbond\_do\_P}$	Indicator variable having value 1 if hydrogen bond donor is present at *para* position of Y, value 0 otherwise
I_{Y_P}	Indicator variable having value 1 any substituent is present at *para* position of Y, value 0 otherwise
I_X	Indicator variable having value 1 if any substituent is present in X, value 0 otherwise
I_Y	Indicator variable having value 1 if any substituent is present in Y, value 0 otherwise
$I_{X\_di\_F}$	Indicator variable having value 1 if two fluorine atoms are present at X, value 0 otherwise
$I_{Y\_di\_F}$	Indicator variable having value 1 if two fluorine atoms are present at Y, value 0 otherwise

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, $F = 4$ for inclusion; $F = 3.99$ 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 have 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 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.

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)[25,26] 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. An 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 (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.

Along with FA-MLR, PCRA was also tried for the dataset. In PCRA,[26] 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.[20] 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.[27] 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. Crossvalidation is a practical and reliable method of testing this significance.[28] 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.[29] 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 (rotated component matrix), variables with high loading (>0.7) in such factors where the activity shows high or moderate loading 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, factor analysis (FA) and principal component regression analysis (PCRA) were performed using the statistical software SPSS while PLS was performed with MINITAB.[30] The statistical qualities of the MLR equations[31] 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). All accepted MLR equations have regression coefficients and F ratios significant at 95% and 99% levels, respectively, if not stated otherwise. The generated QSAR equations were validated by *PRESS (leave-one-out or LOO)*,[32,33] *cross-validation* R^2 (Q^2), and *predicted residual sum of squares* (PRESS) *standard deviation based on PRESS* (S_{PRESS}) and *standard deviation of error of prediction* (SDEP). Finally, leave-33%-out crossvalidation was applied on selected equations.

Acknowledgments

One of the authors (J.T.L.) thanks the AICTE, New Delhi for a QIP fellowship. K.R. thanks the AICTE, New Delhi for a financial grant under the Career Award for Young Teachers scheme.

References and notes

1. Campiani, G.; Ramunno, A.; Maga, G.; Nacci, V.; Fattorusso, C.; Catalanotti, B.; Morelli, E.; Novellino, E. *Curr. Pharm. Des.* **2002**, *8*, 615.

2. Jiang, S.; Zhao, Q.; Debanth, A. K. *Curr. Pharm. Des.* **2002**, *8*, 563.
3. Sanders, R. W.; Dankers, M. M.; Busser, E.; Caffrey, M.; Moore, J. P.; Berkhout, B. *Retrovirology* **2004**, *1*, 3.
4. Mager, P. P. *Med. Res. Rev.* **2001**, *21*, 348.
5. Farber, J. M.; Berger, E. A. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 1749.
6. Richman, D. D. *Nature* **2001**, *410*, 995.
7. Kazmierski, W.; Bifulco, N.; Yang, H.; Boone, L.; DeAnda, F.; Watson, C.; Kenakin, T. *Bioorg. Med. Chem.* **2003**, *11*, 2663.
8. Xu, G.; Kannan, A.; Hartman, T. L.; Wargo, H.; Watson, K.; Turpin, J. A.; Buckheit, R. W., Jr.; Johnson, A. A.; Pommier, Y.; Cushman, M. *Bioorg. Med. Chem.* **2002**, *10*, 2807.
9. Stevens, M.; Pannecouque, C.; DeClercq, E.; Balzarini, J. *Antimicrob. Agents Chemother.* **2003**, *47*, 3109.
10. Leonard, J. T.; Roy, K. *QSAR Comb. Sci.* **2004**, *23*, 23.
11. Roy, K.; Leonard, J. T. *Bioorg. Med. Chem.* **2004**, *12*, 745.
12. Leonard, J. T.; Roy, K. *Drug Des. Dicov.* **2003**, *18*, 165.
13. Leonard, J. T.; Roy, K. *QSAR Comb. Sci.* **2004**, *23*, 387.
14. Roy, K.; Leonard, J. T. *QSAR Comb. Sci.* **2005**, *24*, 579.
15. Roy, K.; Leonard, J. T. *Bioorg. Med. Chem.* **2005**, *13*, 2967.
16. Bouzide, A.; Sauvé, G.; Sévigny, G.; Yelle, J. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 3601.
17. Walker, J. D.; Jaworska, J.; Comber, M. H.; Schultz, T. W.; Dearden, J. C. *Environ. Toxicol. Chem.* **2003**, *22*, 1653.
18. Eriksson, L.; Jaworska, J.; Worth, A. P.; Cronin, M. T. D.; McDowell, R. M.; Gramatica, P. *Environ. Health Perspect.* **2003**, *111*, 1361.
19. Hansch, C.; Fujita, T. *J. Am. Chem. Soc.* **1964**, *86*, 1616.
20. Kubinyi, H. In Wolff, M. E. (Ed.), *Burger's Medicinal Chemistry and Drug Discovery*; 5th ed.; Vol. I, Wiley: New York, 1995; p 507.
21. Hansch, C.; Leo, A.; Hoekman, D. *Exploring QSAR. Hydrophobic, electronic and steric constants*; American Chemical Society: Washington, DC, 1995.
22. CHEM DRAW ULTRA VERSION 5.0 is a program of CambridgeSoft Corporation, USA.
23. Ghose, A. K.; Crippen, G. M. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 21.
24. Darlington, R. B. *Regression and linear models*; McGraw-Hill: New York, 1990.
25. Franke, R. *Theoretical Drug Design Methods*; Elsevier: Amsterdam, 1984; p 184.
26. Franke, R.; Gruska, A. In *Chemometric Methods in Molecular Design*; van de Waterbeemd, H., Ed.; VCH: Weinheim, 1995, p 113.
27. Wold, S. *Chemometric Methods in Molecular Design*; VCH: Weinheim, 1995.
28. Fan, Y.; Shi, L. M.; Kohn, K. W.; Pommier, Y.; Weinstein, J. N. *J. Med. Chem.* **2001**, *44*, 3254.
29. Kulkarni, S. S.; Kulkarni, V. M. *J. Med. Chem.* **1999**, *42*, 373.
30. SPSS is a statistical software of SPSS Inc., USA; MINITAB is a statistical software of Minitab Inc., USA.
31. Snedecor, G. W.; Cochran, W. G. In *Statistical Methods*; Oxford and IBH Publishing Co. Pvt. Ltd: New Delhi, 1967; p 381.
32. Wold, S.; Eriksson, L. In *Chemometric Methods in Molecular Design*; van de Waterbeemd, H., Ed.; VCH: Weinheim, 1995, p 312.
33. Debnath, A. K. In *Combinatorial Library design and Evaluation*; Ghose, A. K., Viswanadhan, V. N., Eds.; Marcel Dekker Inc.: New York, 2001, p 73.