Available online at www.sciencedirect.com
SCIENCE DIRECT
Bioorganic & Medicinal Chemistry Letters
ELSEVIER
Bioorganic & Medicinal Chemistry Letters 16 (2006) 4467–4474

Comparative QSAR modeling of CCR5 receptor binding affinity of substituted 1-(3,3-diphenylpropyl)-piperidinyl amides and ureas

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 6 April 2006; revised 6 June 2006; accepted 12 June 2006
Available online 27 June 2006

Abstract—The present QSAR study attempts to explore the structural and physicochemical requirements of substituted 1-(3,3-diphenylpropyl)-piperidinyl amides and ureas for CCR5 binding affinity using linear free energy-related (LFER) 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 descriptor like partition coefficient ($\log P_{\text{calcd}}$) was also tried as an additional descriptor. Statistical techniques like stepwise regression, multiple linear regression with factor analysis as the data preprocessing step (FA-MLR), partial least squares with factor analysis as the preprocessing step (FA-PLS), principal component regression analysis (PCRA), multiple linear regression with genetic function approximation (GFA-MLR), and genetic partial least squares (G/PLS) were applied to identify the structural and physicochemical requirements for the CCR5 binding affinity. The generated equations were statistically validated using leave-one-out technique. The quality of equations obtained from stepwise regression, FA-MLR, FA-PLS, and PCRA is of acceptable statistical range (explained variance ranging from 71.9% to 80.4%, while predicted variance ranging from 67.4% to 77.0%). The GFA-derived models show high intercorrelation among predictor variables used in the equations while the G/PLS model shows lowest statistical quality among all types of models. The best models were also subjected to leave-25%-out crossvalidation.

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 postentry 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 transmembrane subunit (gp41). Each of the stages can serve as a target for the HIV entry.

Postentry 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

Abbreviations: QSAR, quantitative structure–activity relationships; AIDS, acquired immunodeficiency syndrome; HIV, human immunodeficiency virus; LFER, linear free energy related; G/PLS, genetic partial least squares; GFA, genetic function approximation; FA, factor analysis.

Keywords: QSAR; Hansch analysis; LFER; CCR5; Piperidinyl amide; Urea.

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

0960-894X/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmcl.2006.06.031

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 group of compounds, for example, 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 QSAR modeling of CCR5 binding affinity data of substituted 1-(3,3-diphenylpropyl)-piperidinyl amides and ureas.[15,16]

The CCR5 binding affinity data reported by Burrows et al.[15,16] have been used as the model dataset for the present QSAR study: the affinity (50% inhibitory concentration) data [IC_{50} (μM) and IC_{50} (nM)] of substituted 1-(3,3-diphenylpropyl)-piperidinyl amides and ureas (Table 1) for ^{125}I-labeled RANTES (regulated on activation normal T-cell expressed and secreted) to Chinese hamster ovary (CHO) cells expressing human CCR5 have been converted to the logarithmic scale [pIC_{50} (mM)] and then used for subsequent QSAR analyses as the response variable. There are five regions of structural variations in the compounds: one is the R^1-position (showing limited substitution pattern), second one is the X position (showing limited structural variations), and the remaining are R^2, R^3, and R^4 positions of the phenyl rings (showing diverse substitution pattern) (Table 1). This paper uses classical LFER approach using substituent constants;[17–19] thus, compounds containing the common scaffold of [1-(3,3-diphenylpropyl)piperidin-4-yl]-2-phenylacetamides and [1-(3,3-diphenylpropyl)piperidin-4-yl]-*N*-benzylureas were only considered for the present analysis. The objective of the work was to find out the contribution pattern of the phenyl ring substituents. The binding affinity data were subjected to classical QSAR analysis using linear free energy-related (LFER) model of Hansch[17–19] with lipophilicity (π), electronic (Hammett σ), and steric (molar refractivity mr and STERIMOL *L*, *B*1, and *B*5) parameters of the phenyl ring substituents along with appropriate dummy parameters as descriptors. Various indicator variables used in the study have been defined in Table 2. The values of the physiochemical substituent constants (Table S1 in Supplementary material) were taken from the literature.[17] Hydrophobic whole molecular descriptor (partition coefficient $\log P_{calcd}$) was also tried as predictor variable. SMILES were generated from the structures using the JME molecular editor (http://www.molinspiration.com/jme/) and then $\log P$ values were calculated using the ALOGPS 2.1 software [Virtual Computational Chemistry Laboratory (VCC-LAB); http://vcclab.org/lab/alogps]. The calculated $\log P$ ($\log P_{calcd}$) values for all the compounds are given in Table 1.

For the development of equations, six methods were used: (1) stepwise regression,[20] (2) multiple linear regression with factor analysis[21,22] as the data pre-processing step for variable selection (FA-MLR), (3) partial least squares[23,24] with factor analysis as preprocessing step (FA-PLS), (4) principal component regression analysis (PCRA),[22] (5) multiple linear regression with genetic function approximation (GFA-MLR),[24,25] and (6) genetic partial least squares (G/PLS).[26] The details of the methods are given Supplementary material.

The stepwise regression, factor analysis (FA), and principal component regression analysis were performed using the statistical software SPSS.[27] PLS was performed using statistical software MINITAB.[28] Model extraction from the data using genetic function approximation (GFA) and genetic partial least squares G/PLS was done using QSAR+ environment of Cerius^2 software.[29]

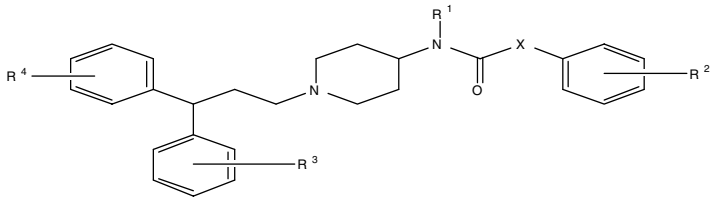
The statistical qualities of the MLR equations[30] were judged by the parameters like explained variance (R_a^2), correlation coefficient (R), standard error of estimate (s), and variance ratio (F) at specified degrees of freedom (df). All accepted MLR equations have regression coefficients and F ratios significant at 95% and 99% levels, respectively, if not stated otherwise. The generated QSAR equations were validated by predicted residual sum of squares (PRESS) (leave-one-out or LOO),[31,32] crossvalidation R^2 (Q^2), standard deviation based on PRESS (S_{PRESS}) and standard deviation of error of prediction (SDEP). Finally, leave-25%-out crossvalidation was applied on selected equations.

Stepwise regression. Using stepping criterion based on F value ($F = 3$ for inclusion; $F = 2.9$ for exclusion), the following best equation was derived with eight variables.

$$\begin{aligned} pIC_{50} = & -0.863(\pm0.820)\sigma_{R2\_m} + 1.230(\pm0.355)\sigma_{R2\_p} \\ & + 0.737(\pm0.614)\sigma_{R3\_p} + 1.179(\pm0.469)mr_{R3\_p} \\ & - 0.356(\pm0.142)mr^2_{R3\_p} + 0.791(\pm0.371)I_{NHCH2} \\ & - 1.279(\pm0.670)I_{BRANCH} - 1.206(\pm0.666)I_{R4\_4S} \\ & + 2.986(\pm0.185) \\ n = 79,\ & R_a^2 = 0.719,\ R^2 = 0.747,\ R = 0.865, \\ F = 25.9(df8, 70),\ & s = 0.455,\ SDEP = 0.487, \\ S_{PRESS} = 0.517,\ & Q^2 = 0.674,\ PRESS = 18.706 \end{aligned} \quad (1)$$

The 95% confidence intervals of the regression coefficients are mentioned within parentheses. Eq. 1 can explain and predict 71.9% and 67.4%, respectively, of the variance of the binding affinity data. The calculated binding affinity values according to Eq. 1 are given in Table 1. Various indicator variables used in this study have been defined in Table 2. The regression

Table 1. Structural features, calculated partition coefficients, and CCR5 binding affinity of 1-(3,3-diphenylpropyl)-piperidinyl amides and ureas



Compound	X	R^1	R^2	R^3	R^4	log P_{calcd}	pIC_{50}				
							Obsd[a]	Calcd[b]	Calcd[c]	Calcd[d]	Calcd[e]
1	—	CH_3	4-F	—	—	5.27	2.143	3.174	3.081	3.172	2.867
2	—	CH_3	3-NO_2	—	—	5.26	2.292	2.488	2.815	2.158	2.309
3	$CH(CH_3)$	CH_3	—	—	—	5.84	2.167	1.822	1.766	1.933	3.068
4	CH_2	CH_3	—	—	—	5.51	3.114	3.101	2.854	3.170	2.836
5	CH_2	CH_3	2-Cl	—	—	6.08	2.444	3.101	2.743	2.665	2.379
6	CH_2	CH_3	3-Cl	—	—	6.09	2.658	2.781	2.739	2.894	2.655
7	CH_2	CH_3	4-Cl	—	—	6.09	3.097	3.383	3.409	3.174	3.194
8	CH_2	CH_3	3,4-Cl	—	—	6.65	3.108	3.064	3.023	2.899	3.142
9	CH_2	CH_3	2,4-Cl	—	—	6.70	2.585	3.383	2.975	2.670	2.776
10	CH_2	CH_3	2-F	—	—	5.48	2.721	3.101	2.853	2.665	2.488
11	CH_2	CH_3	3-F	—	—	5.50	2.854	2.807	2.854	2.937	2.602
12	CH_2	CH_3	4-F	—	—	5.55	3.180	3.174	3.119	3.172	3.002
13	CH_2	CH_3	3,4-F	—	—	5.75	3.161	2.881	3.105	2.939	2.823
14	CH_2	CH_3	3-OMe	—	—	5.57	3.167	2.997	2.856	3.141	2.876
15	CH_2	CH_3	4-OCH_3	—	—	5.66	3.237	2.769	3.110	3.106	3.008
16	CH_2	CH_3	3,4-OMe	—	—	5.66	3.187	2.665	3.110	3.078	3.023
17	CH_2	CH_3	3,5-OMe	—	—	5.64	2.569	2.894	2.854	3.053	2.803
18	CH_2	CH_3	2,4,5-OMe	—	—	5.62	2.959	2.665	3.113	2.574	2.721
19	CH_2	CH_3	4-Br	—	—	6.08	3.237	3.383	3.477	3.144	3.294
20	CH_2	CH_3	4-Benzyloxy	—	—	6.82	2.456	2.818	2.351	2.771	2.547
21	CH_2	CH_3	4-Phenyl	—	—	6.82	2.638	3.088	2.853	2.771	2.778
22	CH_2	CH_3	4-CF_3	—	—	6.23	3.432	3.765	3.599	3.471	3.585
23	CH_2	CH_3	4-OCF_3	—	—	6.54	3.538	3.531	3.424	3.218	3.231
24	CH_2	CH_3	4-NHCOMe	—	—	5.32	3.167	3.101	3.500	3.365	3.093
25	CH_2	CH_3	4-CN	—	—	5.24	4.222	3.912	4.051	3.931	3.770
26	CH_2	CH_3	4-SO_2NH_2	—	—	4.54	4.041	3.838	3.547	4.095	4.077
27	CH_2	CH_3	4-$SO_2N(Me)_2$	—	—	4.77	4.337	3.900	3.889	3.959	4.036
28	CH_2	CH_3	4-SMe	—	—	5.75	3.252	3.101	3.426	3.047	3.154
29	CH_2	CH_3	4-CO_2Me	—	—	5.41	3.201	3.654	3.939	3.539	3.361
30	CH_2	CH_3	4-OH	—	—	5.39	3.328	2.646	2.698	3.241	3.053
31	CH_2	CH_3	4-NO_2	—	—	5.38	3.824	4.060	4.050	4.063	3.821
32	CH_2	Ethyl	4-OCF_3	—	—	6.66	3.509	3.531	3.318	3.218	3.430
33	CH_2	Ethyl	4-CN	—	—	5.59	4.180	3.912	4.097	3.931	3.954
34	CH_2	Ethyl	4-SO_2NH_2	—	—	5.36	4.420	3.838	3.983	4.095	4.354
35	CH_2	Ethyl	4-SO_2Me	—	—	5.53	4.119	3.986	4.138	4.236	4.326
36	CH_2	Ethyl	4-NO_2	—	—	5.72	3.959	4.060	4.055	4.063	4.054
37	CH_2	*c*-propyl	4-SO_2NH_2	—	—	5.09	4.481	3.838	3.903	4.095	4.163
38	CH_2	*c*-propyl	4-SO_2Me	—	—	5.29	4.292	3.986	4.106	4.236	4.164
39	CH_2	*c*-propyl	4-NO_2	—	—	6.04	3.509	4.060	3.969	4.063	3.689
40	CH_2	Allyl	4-OCF_3	—	—	6.79	3.456	3.531	3.188	3.218	3.247
41	CH_2	Allyl	4-SO_2Me	—	—	5.61	4.432	3.986	4.138	4.236	4.248
42	CH_2	Allyl	4-NO_2	—	—	6.11	3.745	4.060	3.939	4.063	3.813
43	NH	CH_3	3-CN	—	—	5.09	2.310	2.617	2.757	2.539	2.517
44	NH	CH_3	3-CH_3	—	—	5.69	2.229	3.161	2.850	3.159	2.952
45	NH	Allyl	—	—	—	5.83	2.678	3.101	2.826	3.170	2.980
46	NH	CH_3	3,4-Cl	—	—	6.84	3.432	3.064	2.830	2.899	3.449
47	NH	CH_3	4-F	—	—	5.95	3.721	3.174	3.056	3.172	3.359
48	NH	Ethyl	4-CH_3	—	—	6.02	3.495	2.891	2.865	3.001	3.640
49	NH–CH_2	CH_3	—	—	—	4.96	4.000	3.892	3.575	3.836	3.507
50	NH–CH_2	Ethyl	—	—	—	5.34	4.208	3.892	3.712	3.836	3.759
51	NH–$CH(CH_3)$	Ethyl	—	—	—	5.70	2.268	2.613	2.669	2.599	3.577
52	NH–CH_2	Allyl	3-CH_3	—	—	5.80	3.620	3.952	3.711	3.826	3.553
53	NH–CH_2	Allyl	4-OCH_3	—	—	5.71	3.959	3.560	3.983	3.773	3.656
54	NH–CH_2	Ethyl	3-CH_3	—	—	5.56	4.456	3.952	3.734	3.826	3.958
55	NH–CH_2	Ethyl	4-OCH_3	—	—	5.38	3.377	3.560	3.977	3.773	3.673
56	NH–CH_2	Ethyl	4-SO_2CH_3	—	—	4.92	4.310	4.777	4.836	4.903	4.808
57	CH_2	Ethyl	—	4-F	4-F	5.23	3.108	2.813	2.876	2.905	2.827

(*continued on next page*)

Table 1 (*continued*)

Compound	X	R^1	R^2	R^3	R^4	log P_{calcd}	pIC_{50}				
							Obsd[a]	Calcd[b]	Calcd[c]	Calcd[d]	Calcd[e]
58	CH_2	Ethyl	—	4-F	—	5.15	3.509	4.019	4.101	4.353	4.342
59	CH_2	Ethyl	—	4-Cl	—	5.98	5.071	4.621	4.819	4.558	5.007
60	CH_2	Ethyl	—	4-Cl	4-Cl	6.39	3.119	3.414	3.351	3.110	3.044
61	CH_2	Ethyl	—	3-Cl	—	5.98	4.237	3.986	4.066	4.236	4.019
62	CH_2	Ethyl	—	3,4-Cl_2	—	6.41	4.108	4.621	4.587	4.558	4.401
63	CH_2	Ethyl	—	4-CH_3	—	5.74	4.553	4.295	4.506	4.304	4.517
64	CH_2	Ethyl	—	4-CF_3	—	5.79	5.638	4.770	5.036	5.098	5.522
65	CH_2	Ethyl	—	4-CO_2CH_3	—	5.06	5.149	5.132	5.390	5.097	5.101
66	CH_2	Ethyl	—	4-$CONH_2$	—	3.95	3.585	4.951	4.158	5.004	4.584
67	CH_2	Ethyl	—	4-OCH_3	—	5.55	5.201	4.382	4.629	4.472	4.543
68	CH_2	Ethyl	—	4-Ph	—	6.73	4.071	4.562	4.277	4.129	4.200
69	CH_2	Ethyl	—	4-SCH_3	—	5.71	4.921	4.821	5.141	4.671	4.774
70	CH_2	Ethyl	—	4-SO_2CH_3	—	4.55	5.770	5.345	5.300	5.296	5.471
71	CH_2	Ethyl	—	4-NH_2	—	5.03	3.699	3.918	3.964	3.853	3.679
72	CH_2	Ethyl	—	4-$NHCOCH_3$	—	4.61	4.585	4.838	4.776	4.727	4.477
73	CH_2	Ethyl	—	4-NHCOPh	—	5.95	3.886	3.548	3.610	3.679	3.681
74	CH_2	Ethyl	—	4-$NHSO_2CH_3$	—	4.62	4.745	4.861	4.814	4.916	4.552
75	CH_2	Ethyl	—	4-$NHSO_2Ph$	—	5.54	3.131	3.233	3.295	3.604	3.414
76	CH_2	Ethyl	—	4-Cl	3-F	5.64	4.721	4.621	4.890	4.558	5.205
77	CH_2	Ethyl	—	4-NH_2	3-F	4.69	3.979	3.918	3.756	3.853	4.050
78	CH_2	Ethyl	—	4-$NHCOCH_3$	3-F	4.66	5.102	4.838	4.816	4.727	4.894
79	CH_2	Ethyl	—	4-$NHSO_2CH_3$	3-F	4.67	5.041	4.861	4.854	4.916	4.908

[a] Taken from Refs. 15 and 16.
[b] From Eq. 1.
[c] From Eq. 3.
[d] From Eq. 4.
[e] From Eq. 5.

Table 2. Definitions of indicator parameters

Parameter	Definition
I_{CH2}	Indicator variable having value 1 if methylene group is present at X position, value 0 otherwise.
I_{NH}	Indicator variable having value 1 if NH group is present at X position, value 0 otherwise
I_{NHCH}	Indicator variable having value 1 if NH–CH_2 group is present at X position, value 0 otherwise
I_{BRANCH}	Indicator variable having value 1 if branching is present at X position, value 0 otherwise
I_{ETHYL}	Indicator variable having value 1 if ethyl group is present at R^1-position, value 0 otherwise
I_{ALLYL}	Indicator variable having value 1 if allyl group is present at R^1-position, value 0 otherwise
$I_{C\text{-}PRO}$	Indicator variable having value 1 if *c*-propyl group is present at R^1-position, value 0 otherwise
I_{R2_2S}	Indicator variable having value 1 if any substituent is present at 2nd position at R^2, value 0 otherwise
I_{R4_3S}	Indicator variable having value 1 if any substituent is present at 3rd position at R^4, value 0 otherwise
I_{R4_4S}	Indicator variable having value 1 if any substituent is present at 4th position at R^4, value 0 otherwise

coefficients of σ_{R2_m} and σ_{R3_p} are significant at 96.1% and 98.1% levels, respectively. The negative coefficient of σ_{R2_m} indicate that presence of electron-withdrawing substituents (like NO_2, CN or OCH_3) at the *meta*-position of the phenyl ring (R^2) is not favorable for the binding affinity, while the positive coefficients of σ_{R2_p} and σ_{R3_p} indicate that presence of electron withdrawing substituents (like NO_2, SO_2NH_2, SO_2CH_3, and $SO_2N(CH_3)_2$) at the *para*-position of the phenyl ring (R^2) and substituents (like CF_3, SO_2CH_3) at the *para*-position of the phenyl ring (R^3) is conducive for the binding affinity. The molar refractivity (mr) of the *para*-substituents at R^3-position shows a parabolic relation with the binding affinity in the above equation. This indicates that the binding affinity increases with increase in volume of the *para*-substituents up to some critical value (optimum mr value being 1.656 according to Eq. 1) after which the binding affinity decreases. The positive coefficient of I_{NHCH2} indicates that presence of $NHCH_2$ group at the X position is favorable for the binding affinity. The negative coefficient of I_{BRANCH} denotes that the presence of branching at the X position is detrimental for the binding affinity. The positive coefficient of I_{R4_4S} indicates that the presence of substituents at the *para*-position of the phenyl ring (R^4) is favorable for the binding affinity. The intercorrelation (*r*) matrix among the predictor variables used in Eq. 1 is given in Table S2 in Supplementary material section. Leave-25%-out crossvalidation was also applied on Eq. 1; the average regression coefficients for the different variables and the corresponding standard deviations for the four cycles are shown in Table 3.

FA-MLR. From the factor analysis on the data matrix consisting of the CCR5 binding affinity data, physiochemical parameters, and indicator variables, it was observed that 19 factors could explain the data matrix

Table 3. Results of leave-25%-out crossvalidation applied on Eqs. 1, 3, 4, and 5 model equation, $pC = \Sigma\beta_i x_i + \alpha$

Eq. no.	No. of cycles	Average regression coefficients (±standard deviations)	Q^2 statistic (average pres)
1	4[a]	$-0.829(\pm0.240)\sigma_{R2\_m} + 1.207(\pm0.209)\sigma_{R2\_p} + 0.755(\pm0.246)\sigma_{R3\_p}$ $+1.181(\pm0.192)mr_{R3\_p} - 0.351(\pm0.062)mr^2_{R3\_p} + 0.799(\pm0.177)I_{NHCH}$ $-1.270(\pm0.090)I_{BRANCH} - 1.248(\pm0.093)I_{R4\_4S} + 2.979(\pm0.086)$	0.570 (0.439)
3	4[a]	$4.583(\pm0.082)\log P_{calcd} - 0.413(\pm0.007)[\log P_{calcd}]^2 + 0.980(\pm0.122)\sigma_{R2\_p}$ $+0.599(\pm0.035)L_{R2\_p} - 0.053(\pm0.004)L^2_{R2\_p} + 0.829(\pm0.135)\sigma_{R3\_p}$ $+1.392(\pm0.161)mr_{R3\_p} - 0.411(\pm0.053)mr^2_{R3\_p} + 0.880(\pm0.187)I_{NHCH}$ $-1.025(\pm0.116)I_{BRANCH} - 1.251(\pm0.039)I_{R4\_4S} - 11.019(\pm0.155)$	0.665 (0.389)
4	4[a]	$-2.040(\pm0.304)\sigma^2_{R2\_m} + 0.456(\pm0.047)\sigma_{R2\_p} + 0.680(\pm0.216)\sigma^2_{R2\_p}$ $-0.190(\pm0.031)\pi_{R2\_p} - 0.079(\pm0.031)mr^2_{R3\_p} + 0.861(\pm0.155)\sigma_{R3\_p}$ $+0.676(\pm0.184)I_{NHCH} - 1.237(\pm0.077)I_{BRANCH} - 0.494(\pm0.109)I_{R2\_2S}$ $-1.368(\pm0.095)I_{R4\_4S} + 2.972(\pm0.050)$	0.600 (0.406)
5	4[a]	$-0.150(\pm0.038)fs1 + 0.159(\pm0.010)fs2 + 0.141(\pm0.036)fs3 + 0.561(\pm0.042)fs4$ $-0.158(\pm0.020)fs5 - 0.153(\pm0.019)fs6 + 0.173(\pm0.018)fs8 - 0.154(\pm0.056)fs9$ $+0.235(\pm0.014)fs10 - 0.212(\pm0.016)fs12 + 0.132(\pm0.030)fs14 - 0.125(\pm0.018)fs18$ $+3.651(\pm0.019)$	0.760 (0.307)

Q^2 denotes crossvalidated R^2. Average pres means average of absolute values of predicted residuals.
[a] Compounds were deleted in 4 cycles in the following manner: (**1**, **5**, **9**, …, **77**), (**2**, **6**, **10**, …, **78**), (**3**, **7**, **11**, …, **79**), (**4**, **8**, **12**, …, **76**).

to the extent of 95.06%. The binding affinity is highly loaded with factors 5 (highly loaded in σ_{R2_p}, $\sigma^2_{R2\_p}$), factors 10 (highly loaded in σ_{R3_p}), 16 (highly loaded in I_{BRANCH}), and 12 (highly loaded in I_{R4_4S}), moderately loaded with factors 8 (highly loaded in $\sigma^2_{R3\_p}$), 2 (highly loaded in mr_{R3_p}, $mr^2_{R3\_p}$, L_{R3_p}, $L^2_{R3\_p}$, $B5_{R3\_p}$, $B5^2_{R3\_p}$), 5 (highly loaded in $\log P_{calcd}$, $[\log P_{calcd}]^2$), 9 (highly loaded in I_{NHCH2}), 1 (highly loaded in mr_{R2_m}, $mr^2_{R2\_m}$, L_{R2_m}, $L^2_{R2\_m}$, $B1_{R2\_m}$, $B2^2_{R2\_m}$, $B5_{R2\_m}$, $B5^2_{R2\_m}$), 6 (highly loaded in σ_{R2_m}), 3 (highly loaded in mr_{R2_p}, $mr^2_{R2\_p}$, L_{R2_p}, $L^2_{R2\_p}$, $B5_{R2\_p}$, $B5^2_{R2\_p}$), 14 (highly loaded in I_{R4_3S}), and 18 (highly loaded in I_{R2_2S}). The binding affinity is poorly loaded with other factors (factors 7, 11, 13, 15, 17, and 19). Based on the results of the factor analysis (table not shown), the following equation was derived with ten variables.

$$\begin{aligned} pIC_{50} = {} & 3.709(\pm2.315)\log P_{calcd} - 0.332(\pm0.204) \\ & \times [\log P_{calcd}]^2 - 1.752(\pm1.389)\sigma^2_{R2\_m} \\ & + 1.238(\pm0.331)\sigma_{R2\_p} + 0.813(\pm0.573)\sigma_{R3\_p} \\ & + 1.391(\pm0.483)mr_{R3\_p} - 0.415(\pm0.144)mr^2_{R3\_p} \\ & + 0.764(\pm0.349)I_{NHCH2} - 1.346(\pm0.625)I_{BRANCH} \\ & - 1.229(\pm0.623)I_{R4\_4S} - 7.287(\pm6.615) \\ n = 79,\ & R^2_a = 0.757,\ R^2 = 0.788,\ R = 0.888, \\ F = 25.3(&df10, 68),\ s = 0.423,\ SDEP = 0.468, \\ S_{PRESS} = 0.505,\ & Q^2 = 0.698,\ PRESS = 17.344 \end{aligned} \quad (2)$$

Eq. 2 can explain and predict 75.7% and 69.8%, respectively, of the variance of the CCR5 binding affinity data. The parameter $\log P_{calcd}$ shows parabolic relation with the binding affinity. The binding affinity increases with increase in $\log P_{calcd}$ value up to a critical level (optimum $\log P_{calcd}$ value being 5.586) after which the affinity decreases. The molar refractivity of R^3 substituents (mr_{R3_p}) also shows parabolic relation with the binding affinity. This indicates that the binding affinity increases with increase in volume of the *para*-substituents (at R^3) up to some critical value (optimum mr value being 1.676) after which it decreases.

$$\begin{aligned} pIC_{50} = {} & 4.794(\pm2.367)\log P_{calcd} - 0.430(\pm0.210) \\ & \times [\log P_{calcd}]^2 + 1.013(\pm0.361)\sigma_{R2\_p} \\ & + 0.579(\pm0.397)L_{R2\_p} - 0.050(\pm0.046)L^2_{R2\_p} \\ & + 0.840(\pm0.547)\sigma_{R3\_p} + 1.416(\pm0.459)mr_{R3\_p} \\ & - 0.423(\pm0.138)mr^2_{R3\_p} + 0.878(\pm0.329)I_{NHCH2} \\ & - 1.058(\pm0.609)I_{BRANCH} - 1.250(\pm0.595)I_{R4\_4S} \\ & - 11.609(\pm6.770) \\ n = 79,\ & R^2_a = 0.779,\ R^2 = 0.810,\ R = 0.900, \\ F = 25.9(&df11, 67),\ s = 0.403,\ SDEP = 0.462, \\ S_{PRESS} = 0.501,\ & Q^2 = 0.706,\ PRESS = 16.845 \end{aligned} \quad (3)$$

Eq. 3 can explain and predict 77.9% and 70.6%, respectively, of the variance of the CCR5 binding affinity data. The optimum value of $\log P_{calcd}$ according to Eq. 3 is 5.574. The regression coefficient of $L^2_{R2\_p}$ is significant at 96.6% level. The STERIMOL parameter (L) of the *para*-substituents at R^3-position shows a parabolic relation with the CCR5 binding affinity. This indicates that the binding affinity increases with increase of length of the *para*-substituents up to some critical value (optimum L value being 5.79) after which it decreases. The calculated binding affinity values according to Eq. 3 are given in Table 1. The intercorrelation (r) matrix among the predictor variables used in Eqs. 2 and 3 is given in Table S2 in Supplementary material. Leave-25%-out crossvalidation was also applied on Eq. 3; the average regression coefficients for the different variables and the corresponding standard deviations for the four cycles are shown in Table 3. The regression coefficients of the variables in Eq. 3 at four cycles in leave-25%-out crossvalidation and predictive r^2 values for different test sets at four iterations are shown in Table 4.

FA-PLS. The number of optimum components was three to obtain the final equation (optimized by

Table 4. Prediction of binding affinity of test set compounds in four crossvalidation cycles (leave-25%-out) based on the descriptor set of Eq. 3

Cycle	Test set compounds	Training set compounds	Regression coefficients											r^2pred[a]
			Intercept	$\log P_{calcd}$	$[\log P_{calcd}]^2$	σ_{R2_p}	L_{R2_p}	$L^2_{R2\_p}$	mr_{R3_p}	$mr^2_{R3\_p}$	σ_{R3_p}	I_{R4_p}	I_{NHCH}	I_{BRANCH}	
1	**1**, **5**, **9**, **13**, **17**, **21**, **25**, **29**, **33**, **37**, **41**, **45**, **49**, **53**, **57**, **61**, **65**, **69**, **73**, **77**	Rest of the compounds ($n = 59$)	−11.154	4.696	−0.423	0.808	0.554	−0.048	1.478	−0.444	0.879		0.746	−1.079	0.725
2	**2**, **6**, **10**, **14**, **18**, **22**, **26**, **30**, **34**, **38**, **42**, **46**, **50**, **54**, **58**, **62**, **66**, **70**, **74**, **78**	Rest of the compounds ($n = 59$)	−11.136	4.584	−0.412	1.093	0.613	−0.052	1.370	−0.415	0.864	−1.279	0.693	−0.892	0.697
3	**3**, **7**, **11**, **15**, **19**, **23**, **27**, **31**, **35**, **39**, **43**, **47**, **51**, **55**, **59**, **63**, **67**, **71**, **75**, **79**	Rest of the compounds ($n = 59$)	−10.827	4.504	−0.406	0.991	0.636	−0.057	1.176	−0.335	0.940	−1.223	1.049		0.620
4	**4**, **8**, **12**, **16**, **20**, **24**, **28**, **32**, **36**, **40**, **44**, **48**, **52**, **56**, **60**, **64**, **68**, **72**, **76**	Rest of the compounds ($n = 60$)	−10.960	4.550	−0.409	1.030	0.594	−0.056	1.545	−0.451	0.633		1.032	−1.104	0.619

[a] $r^2_{pred} = 1 - \frac{\sum (Y_{pred(test)} - Y_{(test)})^2}{\sum (Y_{(test)} - \overline{Y}_{training\ set})^2}$, $Y_{pred(test)}$ and $Y_{(test)}$ indicate predicted and observed binding affinity values, respectively, of the test set compounds and $\overline{Y}_{training}$ indicates mean binding affinity value of the

crossvalidation). Based on the standardized regression coefficients, the following variables were selected for the final equation.

$$\begin{aligned} pIC_{50} = & -2.008\sigma^2_{R2\_m} + 0.461\sigma_{R2\_p} + 0.786\sigma^2_{R2\_p} \\ & - 0.201\pi_{R2\_p} - 0.071mr^2_{R3\_p} + 0.890\sigma_{R3\_p} \\ & + 0.163B5_{R3\_p} + 0.032B5^2_{R3\_p} - 0.007L^2_{R3\_p} \\ & + 0.666I_{NHCH2} - 1.237I_{BRANCH} - 0.504I_{R2\_2S} \\ & - 1.448I_{R4\_4S} + 3.006 \end{aligned}$$

$$n = 79,\ R^2_a = 0.757,\ R^2 = 0.797,\ R = 0.893,$$
$$F = 98.5, \mathrm{SDEP} = 0.437,\ S_{PRESS} = 0.448,$$
$$Q^2 = 0.737,\ \mathrm{PRESS} = 15.068 \quad (4)$$

Eq. 4 could explain 75.7% of the variance and predict 73.7% of the variance. The negative coefficient lipophilic substituent constant (π_{R2_p}) of the *para*-substituents at the R^2-position shows that the binding affinity decreases with increase in lipophilicity of *para*-substituents. The negative coefficient of $\sigma^2_{R2\_m}$ indicates that presence of electron withdrawing substituents at the *meta*-position of the phenyl ring (R^2) is not favorable for the binding affinity. The positive coefficients of σ_{R2_p} and $\sigma^2_{R2\_p}$ indicate that presence of electron withdrawing substituents at the *para*-position of the phenyl ring (at R^2) is conducive for the binding affinity. The occurrence of positive coefficients for both first and second order terms of σ_{R2_p} indicates an inverted parabolic relation with a critical value of σ_{R2_p} of less than zero (−0.29). The calculated binding affinity values according to Eq. 4 have been presented in Table 1. Leave-25%-out crossvalidation was also applied on Eq. 4; the average regression coefficients for the different variables and the corresponding standard deviations for the four cycles are shown in Table 3. The statistical quality of the FA-PLS model is better than that of the G/PLS model which indicates that factor analysis may be a better method for variable selection for the subsequent PLS analysis than a genetic (random) method.

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.

$$\begin{aligned} pIC_{50} = & -0.156(\pm0.086)fs1 + 0.158(\pm0.086)fs2 \\ & + 0.141(\pm0.086)fs3 + 0.559(\pm0.086)fs4 \\ & - 0.157(\pm0.086)fs5 - 0.154(\pm0.086)fs6 \\ & + 0.170(\pm0.086)fs8 - 0.157(\pm0.086)fs9 \\ & + 0.234(\pm0.086)fs10 - 0.207(\pm0.086)fs12 \\ & + 0.132(\pm0.086)fs14 - 0.121(\pm0.086)fs18 \\ & + 3.651(\pm0.086) \end{aligned}$$

$$n = 79,\ R^2_a = 0.804,\ R^2 = 0.834,\ R = 0.913,$$
$$F = 27.6(\mathrm{df}12, 66),\ s = 0.380, \mathrm{SDEP} = 0.408,$$
$$S_{PRESS} = 0.447,\ Q^2 = 0.770, \mathrm{PRESS} = 13.174 \quad (5)$$

Eq. 5 shows good equation statistics (80.4% explained variance) and crossvalidation parameters (77.0% predicted variance). Of all the eight models, PCRA-derived model has the best predicted variance for the CCR5 binding affinity data. The variables (factor scores) used in Eq. 5 are perfectly orthogonal to each other. As factor scores are used instead of selected descriptors in a 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 binding affinity can be obtained. Factor score 1 indicates the importance of molar refractivity (mr), length (*L*), and width (*B5*) of *meta*-substituents at R^2. Factor score 2 indicates the importance of molar refractivity (mr), length (*L*) and width (*B5*) of *para*-substituents at R^3. Factor score 3 indicates the importance of molar refractivity (mr), length (*L*) and width (*B5*) of *para*-substituents at R^2. Factor score 4 indicates the importance of lipophilicity (π), electronic (Hammett σ) parameters, and width (*B5*) of *para*-substituents at R^2. Factor score 5 indicates the importance of the lipophilicity ($\log P_{\text{calcd}}$) of the entire molecules, hydrophobicity parameter of the *para*-substituents at R^2. Factor score 6 signifies the importance of the electronic (Hammett σ) parameters of the *meta*-substituents at R^2, while factor score 8 and factor score 10 signify the importance of the electronic (Hammett σ) parameters of the *para*-substituents at R^3. Factor score 9 signifies the importance of $NHCH_2$ substituents at X, while factor score 12 signifies the importance of *para*-substituents at R^4-position. Factor score 14 signifies the importance of *meta*-substituents at R^4-position, while factor score 18 signifies the importance of *ortho* substituents at R^2-position. The calculated binding affinity values according to Eq. 5 have been presented in Table 1. Leave-25%-out crossvalidation was also applied on Eq. 5; the average regression coefficients for the different variables and the corresponding standard deviations for the four cycles are shown in Table 3.

The GFA-derived models could explain and predict up to 78.6% and 73.9%, respectively, of the variance of the CCR5 receptor binding affinity data but the GFA-generated models have high intercorrelation among the X variables (up to 0.835). The G/PLS-derived model could explain and predict 70.8% and 65.9%, respectively, of the variance of the CCR5 receptor binding affinity data, which show the lowest statistical quality among all models derived. The genetic methods (GFA-MLR and G/PLS) could not produce good reliable models (equations not shown); this may be due to the random selection of the descriptors. Leave-25%-out crossvalidation was applied on Eqs. 1, 3, 4, and 5 and the results are shown in Table 3. Crossvalidation statistics indicate robustness of the formulated models.

The present QSAR study has explored the structural and physicochemical requirements of 1-(3,3-diphenylpropyl)-piperidinyl amides and ureas for the CCR5 binding affinity using linear free energy-related (LFER) model of Hansch. The quality of models obtained from stepwise regression, FA-MLR and FA-PLS are of comparable range (explained variance ranging from 71.9% to 77.9% while predicted variance ranging from 67.4% to 73.7%). The best model came from PCRA technique with explained and predicted variances of 80.4% and 77.0%, respectively. The genetic methods (GFA-MLR and G/PLS) could not produce good reliable models; this may be due to the random selection of the descriptors. However, the rational selection of descriptors through factor analysis could generate statistically valid models in case of FA-MLR and FA-PLS.

The presence of electron withdrawing substituents (like NO_2, CN, and OCH_3) at the *meta*-position of the phenyl ring (R^2) is detrimental for the binding affinity, while presence of electron-withdrawing substituents (like NO_2, SO_2NH_2, SO_2CH_3, and $SO_2N(CH_3)_2$) at the *para*-position of the phenyl ring (R^2) and substituents (like CF_3, SO_2CH_3) at the *para*-position of the phenyl ring (R^3) is conducive for the binding affinity. The presence of electron withdrawing substituents (like CF_3, SO_2CH_3) at the *para*-position of the phenyl ring (R^3) is conducive for the CCR5 receptor binding affinity. Molar refractivity of the *para*-substituents at R^3-position shows a parabolic relation with the binding affinity, which implies the importance of the volume of the *para*-substituents at R^3. The coefficients of STERIMOL parameter (*B5*) of the *para*-substituents at R^3-position show the importance of the volume of the *para*-substituents at R^3. The calculated $\log P$ also shows a parabolic relation with the binding affinity, which implies the importance of the lipophilicity of the whole molecule. The range of the optimum values of $\log P_{\text{calcd}}$ is between 5.574 and 5.586 as found from two different models. The lipophilicity of the *para*-substituents at R^2-position is detrimental to the binding affinity. STERIMOL parameter (*L*) of the *para*-substituents at R^2-position shows a parabolic relation with the binding affinity, which shows the importance of the length of the *para* substituents at R^2. The presence of substituents at the *para*-position of the phenyl ring (at R^4) is detrimental for the binding affinity. The presence of substituents at the *ortho*-position of the phenyl ring (at R^2) is also detrimental for the binding affinity. The presence of $-NH-CH_2-$ group at the X position is conducive to the binding affinity. The presence of branching at the X position is not conducive for the binding affinity.

Acknowledgments

One of the authors (J.T.L.) thanks the AICTE, New Delhi, for a fellowship. K.R. thanks the DST, Govt. of India, New Delhi, for financial assistance under Fast Track scheme for Young Scientists (2005).

Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bmcl.2006.06.031.

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 (London)* **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. Roy, K.; Leonard, J. T. *Bioorg. Med. Chem.* **2004**, *12*, 745.
11. Leonard, J. T.; Roy, K. *QSAR Comb. Sci.* **2004**, *23*, 387.
12. Roy, K.; Leonard, J. T. *Bioorg. Med. Chem.* **2005**, *13*, 2967.
13. Roy, K.; Leonard, J. T. *J. Chem. Inf. Model.* **2005**, *45*, 1352.
14. Leonard, J. T.; Roy, K. *Bioorg. Med. Chem.* **2006**, *14*, 1039.
15. Burrows, J. N.; Cumming, J. G.; Fillery, S. M.; Hamlin, G. A.; Hudson, J. A.; Jackson, R. J.; McLaughlin, S.; Shaw, J. S. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 25.
16. Cumming, J. G.; Cooper, A. E.; Grime, K.; Logan, C. J.; McLaughlin, S.; Oldfield, J.; Shaw, J. S.; Tucker, H.; Winter, J.; Whittaker, D. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 5012.
17. Kubinyi, H. In *Burger's Medicinal Chemistry and Drug Discovery*; 5th ed.; Wolff, M. E., Ed.; John Wiley: New York, 1995; Vol. I, p 507.
18. Hansch, C.; Leo, A.; Hoekman, D. *Exploring QSAR. Hydrophobic, Electronic and Steric Constants*; American Chemical Society: Washington, DC, 1995.
19. Hansch, C.; Fujita, T. *J. Am. Chem. Soc.* **1964**, *86*, 1616.
20. Darlington, R. B. *Regression and Linear Models*; McGraw-Hill: New York, 1990.
21. Franke, R. *Theoretical Drug Design Methods*; Elsevier: Amsterdam, 1984, p 184.
22. Franke, R.; Gruska, A. In *Chemometric Methods in Molecular Design*; van de Waterbeemd, H., Ed.; VCH: Weinheim, 1995; p 113.
23. Wold, S. In *Chemometric Methods in Molecular Design*; van de Waterbeemd, H., Ed.; VCH: Weinheim, 1995; p 195.
24. Fan, Y.; Shi, L. M.; Kohn, K. W.; Pommier, Y.; Weinstein, J. N. *J. Med. Chem.* **2001**, *44*, 3254.
25. Rogers, D.; Hopfinger, A. J. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854.
26. Dunn, W. J., III; Rogers, D. In *Genetic Algorithms in Molecular Modeling*; Devillers, J., Ed.; Academic Press: London, 1996; p 109.
27. SPSS is a statistical software of SPSS Inc., Chicago, IL, USA.
28. MINITAB is a statistical software of Minitab Inc., State College, PA, USA.
29. Cerius^2 version 4.8 is a product of Accelrys, Inc., San Diego, USA.
30. Snedecor, G. W.; Cochran, W. G. In *Statistical Methods*; Oxford and IBH: New Delhi, 1967; p 381.
31. Wold, S.; Eriksson, L. In *Chemometric Methods in Molecular Design*; van de Waterbeemd, H., Ed.; VCH: Weinheim, 1995; p 312.
32. Debnath, A. K. In *Combinatorial Library Design and Evaluation*; Ghose, A. K., Viswanadhan, V. N., Eds.; Marcel Dekker: New York, 2001; p 73.