Available online at www.sciencedirect.com
SCIENCE DIRECT®
Bioorganic & Medicinal Chemistry 13 (2005) 2967–2973
Bioorganic & Medicinal Chemistry

QSAR by LFER model of cytotoxicity data of anti-HIV 5-phenyl-1-phenylamino-1*H*-imidazole derivatives using principal component factor analysis and genetic function approximation

Kunal Roy* and J. Thomas Leonard

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

Received 9 November 2004; revised 27 January 2005; accepted 2 February 2005

Abstract—Cytotoxicity data of anti-HIV 5-phenyl-1-phenylamino-1*H*-imidazole derivatives were subjected to quantitative structure–activity relationship (QSAR) study using linear free energy related (LFER) model of Hansch using electronic (Hammett σ), hydrophobicity (π) and steric (molar refractivity and STERIMOL *L*, *B*1, *B*2, *B*3 and *B*4) parameters of phenyl ring substituents of the compounds, along with appropriate indicator variables. Principal component 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. The generated multiple linear regression (MLR) equations were statistically validated using leave-one-out technique. Genetic function approximation (GFA) was also used on the same data set to develop QSAR equations, which produced the same best equation as obtained with FA-MLR. The final equation is of acceptable statistical quality (explained variance 80.2%) and predictive potential (leave-one-out predicted variance 74%). The analysis explores the structural and physicochemical contributions of the compounds for cytotoxicity. A thiol substituent at 2 position of the imidazole nucleus decreases cytotoxicity when compared to the corresponding unsubstituted congener. Presence of hydrogen bond donor group at *meta* position of the phenyl ring present at 5 position of the imidazole nucleus also reduces cytotoxicity. Additionally, absence of any substituent at 2 and 3 positions of the phenyl ring of 1-phenylamino fragment reduces the cytotoxicity. The negative coefficient of σ_p indicates that presence of electron-withdrawing substituents at the *para* position of the phenyl ring of the 1-phenylamino fragment is not favourable for the cytotoxicity. Again, lipophilicity of *meta* substituents of the 5-phenyl ring increases cytotoxicity. The coefficients of molar refractivity (MR_m) and STERIMOL parameters for *meta* substituents (L_m, $B1_m$ and $B4_m$) of the phenyl ring of 1-phenylamino fragment indicate that the length, width and overall size of *meta* substituents are conducive factors for the cytotoxicity.

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 global HIV/AIDS epidemic[2] has claimed more than 3 million lives in 2003, and an estimated 5 million people acquired the HIV in the same year, bringing to 40 million (including 2.5 million children under 15 years) the number of people globally living with HIV/AIDS. The pandemic spread of this disease has prompted an unprecedented scientific and clinical effort to understand and combat it.

The replicative cycle of HIV can be divided into entry and post entry steps.[3,4] Entry of the HIV into a target

Abbreviations: Quantitative structure–activity relationships (QSAR); Acquired immunodeficiency syndrome (AIDS); Human immunodeficiency virus (HIV); Linear free energy related (LFER); Molar refractivity (MR); Genetic function approximation (GFA).

Keywords: QSAR; Hansch analysis; LFER; Anti-HIV activity; 5-Phenyl-1-phenylamino-1*H*-imidazole; Cytotoxicity.

* Corresponding author. Tel.: +91 33 2414 6676; fax: +91 33 2414 6677; 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.2005.02.003

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, gp120-co-receptor binding inhibitors and fusion inhibitors.

Post entry steps[5] 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. Integrase inhibitors are in clinical trials. 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 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.[6,7] 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.[8] A new generation of antiviral drugs intended to counter HIV-1 entry into susceptible cells is now under development. These compounds generally referred to as fusion or entry inhibitors (which include chemokine-receptor inhibitor, CD4-receptor inhibitor, membrane-fusion inhibitor and other attachment inhibitors) are expected to have different toxicity and resistance profiles than the existing reverse transcriptase and protease inhibitors.[9] 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.[10,11]

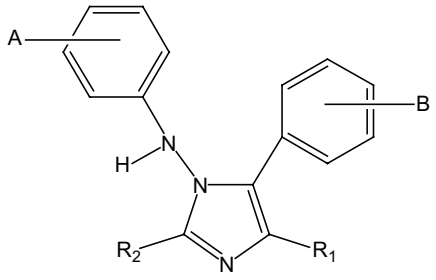
The present group of authors has developed a few quantitative structure–activity relationship (QSAR) models for anti-HIV activities of different group of compounds, for example, 2-amino-6-arylsulfonylbenzonitriles,[12,13] alkenyldiarylmethanes[14] and benzylpyrazoles.[15] In continuation of such efforts, the present paper deals with QSAR modeling of anti-HIV 5-phenyl-1-phenylamino-1*H*-imidazole derivatives.

2. Materials and methods

Cytotoxicity data reported by Lagoja et al.[16] have been used for the present QSAR study: CC_{50} (cytotoxicity data: the 50% cytotoxic concentration to reduce MT-4 cell viability). The activity data [CC_{50} (μM)] of 5-phenyl-1-phenylamino-1*H*-imidazole derivatives (Table 1) have been converted to the logarithmic scale [−log CC (mM)] and then used for subsequent QSAR analyses as the response variables. There are four regions of structural variations in the compounds: one is the 1-phenylamino fragment and the second one is the 5-phenyl ring on the imidazole nucleus, both showing diverse substitution pattern; again, the third one is the R_2 (thiol/hydrogen) substitution and the fourth one is R_1 substitution both of which have limited structural variations (Table 1). This paper uses classical LFER approach using substituent constants;[17,18] thus, compounds containing the common scaffold of 5-phenyl-1-phenylamino-1*H*-imidazole were only considered for the present analysis. The objective of the work was to find out the contribution pattern of the phenyl and imidazole ring substituents. Some of the compounds reported in the original papers were excluded in the present study because of their nongraded quantitative activity data or presence of uncommon structural features. The activity data (Table 1) were subjected to QSAR analyses using linear free energy related (LFER) model of Hansch[17,18] using lipophilicity (π), electronic parameter (Hammett σ) and steric (molar refractivity MR and STERIMOL *L*, *B*1, *B*2, *B*3 and *B*4) parameters of the phenyl ring substituents along with appropriate indicator variables (defined in Table 2). The values of the physiochemical substituent constants (Table 3) were taken from the literature.[19]

Though classical approach of multiple regression technique was used as the final statistical tool for developing QSAR relations, factor analysis (FA)[20,21] 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, correlation matrix and subsequently reduced correlation matrix are constructed. An eigen value problem is then solved and the factor pattern can be obtained from the corresponding eigen vectors. The principal objectives of factor analysis are to display multidimen-

Table 1. Structural features, observed and calculated cytotoxicity data of anti-HIV 5-phenyl-1-phenylamino-1*H*-imidazole derivatives



Sl. no	A	B	R_1	R_2	Cytotoxicity data				
					Obs.[a]	Calc.[b]	Res.[b]	Calc.[c]	Res.[c]
1	3-Cl	H	CH_3	SH	1.176	1.291	−0.115	1.301	−0.125
2	2-Cl	H	CH_3	SH	0.791	0.802	−0.011	0.796	−0.005
3	4-F	H	CH_3	SH	0.657	1.073	−0.416	1.061	−0.404
4	4-Cl	H	CH_3	SH	0.872	0.780	0.092	0.774	0.098
5	3-Cl	3-Br	CH_3	SH	1.357	1.512	−0.155	1.518	−0.161
6	3-Cl	4-Br	CH_3	SH	1.478	1.291	0.187	1.301	0.177
7	3-Cl	3-Cl	CH_3	SH	1.509	1.473	0.036	1.480	0.029
8	3-Cl	4-Cl	CH_3	SH	1.389	1.291	0.098	1.301	0.088
9	3-Cl	4-OCH_3	CH_3	SH	1.412	1.291	0.121	1.301	0.111
10	H	H	CH_3	SH	0.718	0.802	−0.084	0.796	−0.078
11	3-Br	H	CH_3	SH	1.354	1.312	0.042	1.381	−0.027
12	2,5-Cl	H	CH_3	SH	0.832	0.916	−0.084	0.935	−0.103
13	3-NO_2	H	CH_3	SH	1.316	1.382	−0.066	1.338	−0.022
14	3-F	H	CH_3	SH	0.966	1.227	−0.261	1.159	−0.193
15	3-CH_3	H	CH_3	SH	1.054	1.305	−0.251	1.291	−0.237
16	3-CH_3	H	$(CH_3)_2CH$	SH	1.370	1.305	0.065	1.291	0.079
17	3-Cl	H	C_2H_5	SH	1.409	1.291	0.118	1.301	0.108
18	3-CH_3	H	C_2H_5	SH	1.271	1.305	−0.034	1.291	−0.020
19	3-Cl	H	C_6H_5	SH	1.440	1.291	0.149	1.301	0.139
20	3-OCH_3	H	CH_3	SH	1.420	1.305	0.115	1.353	0.067
21	3-Cl	3-CN	CH_3	SH	1.003	1.144	−0.141	1.158	−0.155
22	3-CH_3	3-CN	CH_3	SH	1.341	1.159	0.182	1.147	0.194
23	3-Cl	3-$OCOCH_3$	CH_3	SH	1.172	1.126	0.046	1.140	0.032
24	3-Cl	3-COOH	CH_3	SH	0.921	0.697	0.224	0.706	0.215
25	3-CH_3	3-COOH	CH_3	SH	0.728	0.711	0.017	0.695	0.033
26	4-C_2H_5	H	CH_3	SH	1.463	1.435	0.028	1.415	0.048
27	4-CH_3S	H	CH_3	SH	1.275	1.177	0.098	1.162	0.113
28	3-Cl	H	CH_3	H	1.757	1.593	0.164	1.601	0.156
29	3-Cl	3-Br	CH_3	H	1.785	1.814	−0.029	1.817	−0.032
30	3-Cl	3-Cl	CH_3	H	1.923	1.776	0.147	1.780	0.143
31	H	H	CH_3	H	1.282	1.104	0.178	1.096	0.186
32	3-CH_3	H	CH_3	H	1.568	1.607	−0.039	1.590	−0.022
33	4-F	H	CH_3	H	1.511	1.375	0.136	1.361	0.150
34	3-CH_3	H	C_2H_5	H	1.555	1.607	−0.052	1.590	−0.035
35	3,5-CH_3	H	CH_3	H	1.690	1.736	−0.046	1.719	−0.029
36	3-OCH_3	H	CH_3	H	1.463	1.607	−0.144	1.652	−0.189
37	3-Cl	3-CN	CH_3	H	1.434	1.447	−0.013	1.457	−0.023
38	3-CH_3	3-CN	CH_3	H	1.350	1.461	−0.111	1.447	−0.097
39	3-Cl	3-$CONH_2$	CH_3	H	0.584	0.698	−0.114	0.711	−0.127
40	3-CH_3	3-$CONH_2$	CH_3	H	0.790	0.712	0.078	0.700	0.090
41	3-Cl	3-$OCOCH_3$	CH_3	H	1.480	1.429	0.051	1.440	0.040
42	3-Cl	3-COOH	CH_3	H	0.794	0.999	−0.205	1.005	−0.211

[a] Obs. = Observed (Ref. 16), Calc. = Calculated, Res. = Residual = Obs. − Calc.
[b] From Eq. 1.
[c] From Eq. 4.

sional data in a space of lower dimensionality with minimum loss of information (explaining >95% of the variance of the data matrix) and to extract the basic features behind the data with ultimate goal of interpretation and/or prediction. Factor analysis was performed on the data set(s) containing biological activity and all descriptor variables, which were to be considered. The factors were extracted by principal component method and then rotated by VARIMAX rotation (a kind of rotation, which is used in principal component analysis so that

Table 2. Definitions of indicator and physicochemical parameters

Parameter	Definition
$I_{R1\_CH_3}$	Indicator variable having value 1 if methyl group is present at the 4 position of the imidazole nucleus, value 0 otherwise
I_{R2_SH}	Indicator variable having value 1 if thiol group is present at the 2 position of the imidazole nucleus, value 0 otherwise
$I_{m\_B\_Hydrogen\_Donor}$	Indicator variable having value 1 if any hydrogen donor substituent is present at the *meta* position of the 5-phenyl nucleus, value 0 otherwise
$I_{23\_A\_NIL}$	Indicator variable having value 1 if no substituent is present at 2 and 3 positions of the *N*-phenylamino fragment, value 0 otherwise
$B4_{m\_A}/B1_{m\_A}/L_{m\_A}/MR_{m\_A}$	*B*4/*B*1/*L*/MR value of *meta* substituent (A) present on the phenyl ring of 1-phenylamino fragment
σ_{p_A}	Hammett σ constant of *para* substituent (A) present on the phenyl ring of 1-phenylamino fragment
π_{m_B}	π value of *meta* substituent (B) present on 5-phenyl ring of the imidazole nucleus

Table 3. Values of physicochemical parameters (substituent constants)[a]

Substituents	Substituent constants								
R	π	MR[b]	σ_m	σ_p	*L*	*B*1	*B*2	*B*3	*B*4
H	0	0.103	0	0	2.06	1.00	1.00	1.00	1.00
Cl	0.71	0.603	0.37	0.23	3.52	1.80	1.80	1.80	1.80
Br	0.86	0.888	0.39	0.23	3.83	1.95	1.95	1.95	1.95
NO_2	−0.28	0.736	0.71	0.78	3.44	1.70	1.70	2.44	2.44
F	0.14	0.092	0.34	0.06	2.65	1.35	1.35	1.35	1.35
CH_3	0.56	0.565	−0.07	−0.17	3.00	1.52	2.04	1.90	1.90
OCH_3	−0.02	0.787	0.12	−0.27	3.98	1.35	2.87	1.90	1.90
C_2H_5	1.02	1.030	−0.07	−0.15	4.11	1.52	2.97	1.90	1.90
SCH_3	0.61	1.382	0.15	0	4.30	1.70	3.26	1.90	1.90
CN	−0.57	0.633	0.56	0.66	4.23	1.60	1.60	1.60	1.60
$OCOCH_3$	−0.64	1.247	0.39	0.31	4.87	1.35	3.68	1.90	1.90
COOH	−0.32	0.693	0.37	0.45	3.91	1.60	1.60	2.36	2.66
$CONH_2$	−1.49	0.981	0.28	0.36	4.06	1.60	1.60	2.42	3.07

[a] Taken from Ref. 19.
[b] MR values scaled to a factor of 0.1 as usual.

the axes are rotated to a position in which the sum of the *vari*ances of the loadings is the *max*imum 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 nonzero loadings in such factors where biological activity also has nonzero loading were considered important in explaining variance of the activity. Further, variables with nonzero loadings in different factors were combined in a multivariate equation.

Genetic function approximation (GFA) technique[22,23] was used to generate a population of equations rather than one single equation for correlation between biological activity and physicochemical properties. GFA involves the combination of multivariate adaptive regression splines (MARS) algorithm with genetic algorithm to evolve population of equations that best fit the training set data. It provides an error measure, called the lack of fit (LOF) score that automatically penalizes models with too many features. It also inspires the use of splines as a powerful tool for nonlinear modeling. GFA is done as follows: (i) an initial population of equations is generated by random choice of descriptors; (ii) pairs from the population of equations are chosen at random and 'crossovers' are performed and progeny equations are generated; (iii) it is better at discovering combinations of features that take advantage of correlations between multiple features; (iv) the fitness of each progeny equation is assessed by LOF measure; (v) it can use a larger variety of equation term types in construction of its models; (vi) if the fitness of new progeny equation is better, then it is preserved. The model with proper balance of all statistical terms will be used to explain variance in the biological activity. A distinctive feature of GFA is that it produces a population of models (e.g., 100), instead of generating a single model, as do most other statistical methods. The range of variations in this population gives added information on the quality fit and importance of the descriptors.

The factor analysis and multiple regression were performed using the statistical software SPSS.[24] The statistical qualities of the equations[25] 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 equations have regression coefficients and *F* ratios significant at 95% and 99% levels, respectively, if not stated otherwise. PRESS (leave-one-out)[26,27] and bootstrap

statistics were calculated using the QSAR+ module of Cerius^2 software[28] and the reported parameters are *cross-validation* R^2 (Q^2), *predicted residual sum of squares* (PRESS), *standard deviation based on* PRESS (S_{PRESS}), *standard deviation of error of prediction* (SDEP) and *bootstrap* r^2 (bsr^2). Genetic function approximation analysis was done using GFA module of Cerius^2 software.[28]

3. Results and discussion

From the factor analysis on the data matrix consisting of cytotoxicity data, physiochemical parameters and indicator variables, it was observed that 12 factors could explain the data matrix to the extent of 96.1%. Based on the factor analysis the following equation was derived with six variables.

$$\begin{aligned}\mathrm{pC} &= 0.143(\pm0.138)B4_{m\_\mathrm{A}} - 1.722(\pm1.108)\sigma_{p\_\mathrm{A}} \\ &\quad + 0.257(\pm0.124)\pi_{m\_\mathrm{B}} - 0.302(\pm0.104)I_{\mathrm{R2\_SH}} \\ &\quad - 0.512(\pm0.177)I_{m\_\mathrm{B\_Hydrogen\_Donor}} \\ &\quad - 0.375(\pm0.181)I_{23\_\mathrm{A\_NIL}} + 1.193(\pm0.400) \\ n &= 42, \quad R^2_{\mathrm{a}} = 0.793, \quad R^2 = 0.823, \quad R = 0.907, \\ F &= 27.1(\mathrm{df}\ 6, 35), \quad s = 0.152, \\ Q^2 &= 0.737, \quad \mathrm{SDEP} = 0.187, \quad S_{\mathrm{PRESS}} = 0.210, \\ \mathrm{PRESS} &= 1.196, \quad \mathrm{bsr}^2(\pm\mathrm{sd}) = 0.824(\pm0.003)\end{aligned} \tag{1}$$

The 95% confidence intervals of the regression coefficients are shown within parentheses. Eq. 1 could explain 79.3% of the variance and predict 73.7% of the variance. Compound **3** acts as an outlier for Eq. 1, which requires further introspection. The positive coefficient of STERIMOL parameter $B4_{m\_\mathrm{A}}$ for the *meta* substituents (A) of the phenyl ring of 1-phenylamino fragment indicates that the width of *meta* substituents is a conducive factor for the cytotoxicity. The negative coefficient of σ_{p_A} indicates that presence of electron-withdrawing substituents at the *para* position of the phenyl ring of the 1-phenylamino fragment is not favourable for the cytotoxicity. Again, lipophilicity of the *meta* substituents (B) of the 5-phenyl ring is conducive for the cytotoxicity as evidenced from the positive coefficient of π_{m_B}. Presence of the thiol group at 2 position of the imidazole nucleus is conducive for reducing the cytotoxicity as evidenced from the negative coefficient of $I_{\mathrm{R2\_SH}}$. Furthermore, the negative coefficient of $I_{m\_\mathrm{B\_Hydrogen\_Donor}}$ indicates that presence of hydrogen bond donor group at the *meta* position of the 5-phenyl nucleus (B) reduces the cytotoxicity. The negative coefficient of $I_{23\_\mathrm{A\_NIL}}$ implies that absence of the substituents at the 2 and 3 positions (A) of the phenyl ring of 1-phenylamino fragment reduces the cytotoxicity.

When the $B4_{m\_\mathrm{A}}$ in Eq. 1 is replaced with L_{m_A} or $B1_{m\_\mathrm{A}}$, there is a marginal improvement in the statistical quality. It signifies the importance of the shape parameters of *meta* substituents (A) of the phenyl ring of 1-phenylamino fragment for the cytotoxicity.

$$\begin{aligned}\mathrm{pC} &= 0.095(\pm0.089)L_{m\_\mathrm{A}} - 1.715(\pm1.104)\sigma_{p\_\mathrm{A}} \\ &\quad + 0.245(\pm0.124)\pi_{m\_\mathrm{B}} - 0.305(\pm0.102)I_{\mathrm{R2\_SH}} \\ &\quad - 0.519(\pm0.177)I_{m\_\mathrm{B\_Hydrogen\_Donor}} \\ &\quad - 0.378(\pm0.177)I_{23\_\mathrm{A\_NIL}} + 1.087(\pm0.485) \\ n &= 42, \quad R^2_{\mathrm{a}} = 0.794, \quad R^2 = 0.824, \quad R = 0.908, \\ F &= 27.4(\mathrm{df}\ 6, 35), \quad s = 0.151, \\ Q^2 &= 0.731, \quad \mathrm{SDEP} = 0.171, \quad S_{\mathrm{PRESS}} = 0.187, \\ \mathrm{PRESS} &= 1.224, \quad \mathrm{bsr}^2(\pm\mathrm{sd}) = 0.825(\pm0.003)\end{aligned} \tag{2}$$

$$\begin{aligned}\mathrm{pC} &= 0.182(\pm0.168)B1_{m\_\mathrm{A}} - 1.721(\pm1.100)\sigma_{p\_\mathrm{A}} \\ &\quad + 0.246(\pm0.124)\pi_{m\_\mathrm{B}} - 0.309(\pm0.102)I_{\mathrm{R2\_SH}} \\ &\quad - 0.528(\pm0.177)I_{m\_\mathrm{B\_Hydrogen\_Donor}} \\ &\quad - 0.382(\pm0.175)I_{23\_\mathrm{A\_NIL}} + 1.119(\pm0.449) \\ n &= 42, \quad R^2_{\mathrm{a}} = 0.795, \quad R^2 = 0.825, \quad R = 0.908, \\ F &= 27.5(\mathrm{df}\ 6, 35), \quad s = 0.151, \\ Q^2 &= 0.733, \quad \mathrm{SDEP} = 0.170, \quad S_{\mathrm{PRESS}} = 0.186, \\ \mathrm{PRESS} &= 1.215, \quad \mathrm{bsr}^2(\pm\mathrm{sd}) = 0.826(\pm0.003)\end{aligned} \tag{3}$$

Again, when molar refractivity of *meta* substituents (A) of the phenyl ring of 1-*phenylamino* fragment is used instead of STERIMOL parameter, there is further rise in statistical quality.

$$\begin{aligned}\mathrm{pC} &= 0.278(\pm0.223)\mathrm{MR}_{m\_\mathrm{A}} - 1.685(\pm1.080)\sigma_{p\_\mathrm{A}} \\ &\quad + 0.252(\pm0.122)\pi_{m\_\mathrm{B}} - 0.300(\pm0.099)I_{\mathrm{R2\_SH}} \\ &\quad - 0.515(\pm0.173)I_{m\_\mathrm{B\_Hydrogen\_Donor}} \\ &\quad - 0.366(\pm0.175)I_{23\_\mathrm{A\_NIL}} + 1.404(\pm0.177) \\ n &= 42, \quad R^2_{\mathrm{a}} = 0.802, \quad R^2 = 0.831, \quad R = 0.912, \\ F &= 28.7(\mathrm{df}\ 6, 35), \quad s = 0.148, \\ Q^2 &= 0.740, \quad \mathrm{SDEP} = 0.168, \quad S_{\mathrm{PRESS}} = 0.184, \\ \mathrm{PRESS} &= 1.183, \quad \mathrm{bsr}^2(\pm\mathrm{sd}) = 0.832(\pm0.003)\end{aligned} \tag{4}$$

Eq. 4 could explain 80.2% variance and predict 74.0% variance. Compound **3** acts as an outlier for Eq. 4. The positive coefficient of molar refractivity (MR) of *meta* substituents of the phenyl nucleus of *1-phenylamino* moiety shows that the volume of *meta* substituents (A) increases the cytotoxicity.

Each of Eqs. 1–4 involve six predictor variables for 42 data points and thus maintains the recommended ratio of number of predictor variables to number of data point of 1:5. Furthermore, the leave-one-out Q^2 value of each these equations is more than the recommended cut-off value of 0.5.[29,30]

Genetic function approximation (GFA) was also used on the same dataset to develop QSAR equations using

Table 4. Intercorrelation (r) matrix for physiochemical parameters and indicator variables

	MR_{m_A}	σ_{p_A}	π_{m_B}	I_{R2_SH}	$I_{m\_B\_Hydrogen\_Donor}$	$I_{23\_A\_NIL}$	$B1_{m\_A}$	L_{m_A}
σ_{p_A}	−0.191							
π_{m_B}	−0.059	0.024						
I_{R2_SH}	−0.161	0.013	0.211					
$I_{m\_B\_Hydrogen\_Donor}$	0.112	−0.040	−0.546	−0.186				
$I_{23\_A\_NIL}$	−0.399	−0.035	0.073	0.073	−0.119			
$B1_{m\_A}$	0.849	−0.193	−0.035	−0.100	0.147	−0.374		
L_{m_A}	0.937	−0.196	−0.008	−0.122	0.101	−0.391	0.892	
$B4_{m\_A}$	0.941	−0.194	−0.093	−0.166	0.126	−0.417	0.816	0.838

GFA module of Cerius[2], which produced the same best equation [Eq. 4] as obtained with FA-MLR (multiple linear regression equation based on factor analysis).

The intercorrelation (r) matrix among the predictor variables used in Eqs. 1–4 is given in Table 4. The calculated and residual cytotoxicity values according to Eqs. 1 and 4 are given in Table 1.

4. Conclusions

The present QSAR study has explored the structural and physicochemical contributions of 5-phenyl-1-phenylamino-1*H*-imidazole derivatives for cytotoxicity using linear free energy related (LFER) model of Hansch. A thiol substituent at 2 position of the imidazole nucleus decreases cytotoxicity when compared to the corresponding unsubstituted congener. Presence of hydrogen bond donor group at *meta* position of the phenyl ring present at 5 position of the imidazole nucleus also reduces cytotoxicity. Additionally, absence of any substituent at 2 and 3 positions of the phenyl ring of 1-phenylamino fragment reduces the cytotoxicity. The presence of electron-withdrawing substituents at the *para* position of the phenyl ring of 1-phenylamino fragment is not favourable for the cytotoxicity. Lipophilicity of *meta* substituents of the 5-phenyl ring increases cytotoxicity. The coefficients of molar refractivity (MR_m) and STERIMOL parameters for *meta* substituents (L_m, $B1_m$ and $B4_m$) of the phenyl ring of *1*-phenylamino fragment indicate that the length, width and overall size of *meta* substituents on the phenyl ring of 1-phenylamino fragment are conducive factors for the cytotoxicity.

Acknowledgements

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. *Cur. Pharm. Des.* **2002**, *8*, 615.
2. http://www.unaids.org/Unaids/EN/Resources/Publications.
3. Jiang, S.; Zhao, Q.; Debanth, A. K. *Curr. Pharm. Des.* **2002**, *8*, 563.
4. Sanders, R. W.; Dankers, M. M.; Busser, E.; Caffrey, M.; Moore, J. P.; Berkhout, B. *Retrovirology* **2004**, *1*, 3.
5. Mager, P. P. *Med. Res. Rev.* **2001**, *21*, 348.
6. Farber, J. M.; Berger, E. A. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 1749.
7. Richman, D. D. *Nature* **2001**, *410*, 995.
8. Kazmierski, W.; Bifulco, N.; Yang, H.; Boone, L.; DeAnda, F.; Watson, C.; Kenakin, T. *Bioorg. Med. Chem.* **2003**, *11*, 2663.
9. Tsamis, F.; Gavrilov, S.; Kajumo, F.; Seibert, C.; Kuhmann, S.; Ketas, T.; Trkola, A.; Palani, A.; Clader, J. W.; Tagat, J. R.; McCombie, S.; Baroudy, B.; Moore, J. P.; Sakmar, T. P.; Dragic, T. *J. Virol.* **2003**, *77*, 5201.
10. 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.
11. Stevens, M.; Pannecouque, C.; DeClercq, E.; Balzarini, J. *Antimicrob. Agents Chemother.* **2003**, *47*, 3109.
12. Leonard, J. T.; Roy, K. *QSAR Comb. Sci.* **2004**, *23*, 23.
13. Roy, K.; Leonard, J. T. *Bioorg. Med. Chem.* **2004**, *12*, 745.
14. Leonard, J. T.; Roy, K. *Drug Des. Discov.* **2003**, *18*, 165.
15. Leonard, J. T.; Roy, K. *QSAR Comb. Sci.* **2004**, *23*, 387.
16. Lagoja, I. M.; Pannecouque, C.; Van Aerschot, A.; Witvrouw, M.; Debyser, Z.; Balzarini, J.; Hardewijn, P.; De Clercq, E. *J. Med. Chem.* **2003**, *46*, 1546.
17. Hansch, C.; Fujita, T. *J. Am. Chem. Soc.* **1964**, *86*, 1616.
18. Hansch, C.; Leo, A.; Hoekman, D. *Exploring QSAR. Hydrophobic, Electronic and Steric Constants*; American Chemical Society: Washington, DC, 1995.
19. Kubinyi, H. In *Burger's Medicinal Chemistry and Drug Discovery*; Wolff, M. E., Ed.; 5th ed.; John Wiley & Sons: New York, 1995; Vol. 1, p 507.
20. Franke, R. *Theoretical Drug Design Methods*; Elsevier: Amsterdam, 1984; p 184.
21. Franke, R.; Gruska, A. In *Chemometric Methods in Molecular Design*; van de Waterbeemd, H., Ed.; VCH: Weinheim, 1995; p 113.
22. Rogers, D.; Hopfinger, A. J. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854.
23. Fan, Y.; Shi, L. M.; Kohn, K. W.; Pommier, Y.; Weinstein, J. N. *J. Med. Chem.* **2001**, *44*, 3254.
24. SPSS is a statistical software of SPSS Inc., IL, USA.
25. Snedecor, G. W.; Cochran, W. G. *Statistical Methods*; Oxford and IBH: New Delhi, 1967; p 381.
26. Wold, S.; Eriksson, L. In *Chemometric Methods in Molecular Design*; van de Waterbeemd, H., Ed.; VCH: Weinheim, 1995; p 312.

27. Debnath, A. K. In *Combinatorial Library Design and Evaluation*; Ghose, A. K., Viswanadhan, V. N., Eds.; Marcel Dekker: New York, 2001; p 73.
28. Cerius^2 version 4.8 is a product of Accelrys, Inc., San Diego, USA.
29. Walker, J. D.; Jaworska, J.; Comber, M. H.; Schultz, T. W.; Dearden, J. C. *Environ. Toxicol. Chem.* **2003**, *22*, 1653.
30. Eriksson, L.; Jaworska, J.; Worth, A. P.; Cronin, M. T. D.; McDowell, R. M.; Gramatica, P. *Environ. Health Perspect.* **2003**, *111*, 1361.