Journal of Enzyme Inhibition and Medicinal Chemistry, 2011; 26(5): 696–705
© 2011 Informa UK, Ltd.
ISSN 1475-6366 print/ISSN 1475-6374 online
DOI: 10.3109/14756366.2010.548328

informa
healthcare

RESEARCH ARTICLE

Consensus features of CP-MLR and GA in modeling HIV-1 RT inhibitory activity of 4-benzyl/benzoylpyridin-2-one analogues

Shreekant Deshpande[1], Rinki Singh[1], Mohammad Goodarzi[2], Seturam B. Katti[1], and Yenamandra S. Prabhakar[1]

[1]Medicinal and Process Chemistry Division, Central Drug Research Institute, CSIR, Lucknow, India, and [2]Department of Chemistry, Faculty of Sciences and Young Researchers Club, Islamic Azad University, Arak Branch, Arak, Markazi, Iran

Abstract

The HIV-1 reverse transcriptase (RT) inhibitory activity of benzyl/benzoylpyridinones is modeled with molecular features identified in combinatorial protocol in multiple linear regression (CP-MLR) and genetic algorithm (GA). Among the features, nDB and LogP are found to be the most influential descriptors to modulate the activity. Although the coefficient of nDB suggested in favor of benzylpyridinones skeleton, the coefficient of LogP suggested the favorability of hydrophilic nature in compounds for better activity. The partial least squares analysis of the descriptors common to CP-MLR and GA has displayed their predictivity over the total descriptors identified in both the approaches. The back-propagation artificial neural networks model from the five most significant common descriptors (nDB, T(O..O), MATS8e, LogP, and BELp4) has explained 93.2% variance in the HIV-1 RT activity of the training set compounds and showed a test set r^2 of 0.89. The results suggest that the descriptors have the ability to identify the patterns in the compounds to predict potential analogues.

Keywords: Benzyl/benzoylpyridinones, HIV-1 RT inhibitors, QSAR, CPMLR, GA, ANN

Introduction

HIV-1 reverse transcriptase (HIV-1 RT) is a key enzyme in the progression of infection by HIV retrovirus. It has been widely explored as a drug target[1,2]. Two classes of compounds have gained attention as potential inhibitors of this enzyme. They are termed as nucleoside/nucleotide reverse transcriptase inhibitors (NRTIs) and non-nucleoside reverse transcriptase inhibitors (NNRTIs)[3]. Between them, NNRTIs have received a great deal of attention because of low toxicity and favorable pharmacokinetic properties[4,5]. The low toxicity of NNRTIs is attributed to their interaction with an allosteric site on the enzyme[6]. Since the introduction of NNRTIs, >30 different structure classes are shown to bind with the allosteric site of the enzyme and elicit the desired response[7]. Also, some of these compounds have been put to clinical use[8]. The flexibility of HIV-1 RT in accommodating diverse NNRTIs has been subjugated by the development of quick resistance to different compounds[9,10]. This has necessitated continued efforts to discover new ways to modify the chemical space of compounds and/or alternative targets for the HIV chemotherapy.

In medicinal chemistry, structure–activity relationships pave way to notional insight of the activity (or receptor space) against the chemical space. The application of quantification protocols to this paradigm fine tunes the notional insight of the activity in terms of properties of the chemical space and gives an opportunity to understand and modulate the variations around the scaffold on a broad canvas of diverse structures. Among the different NNRTIs, pyridinone derivatives[11] represent simple structure space. They bear some structure space resemblance with the 7-chloro-1-(2,6-difluorophenyl)-1H,3H-thiazolo[3,4-a] benzimidazole (7-Cl-TBZ) and thiazolidinones (Figure 1).

Address for Correspondence: Yenamandra S. Prabhakar, Central Drug Research Institute, Medicinal Process Chemistry Division, Lucknow, India. E-mail: yenpra@yahoo.com

(Received 07 August 2010; revised 24 November 2010; accepted 13 December 2010)

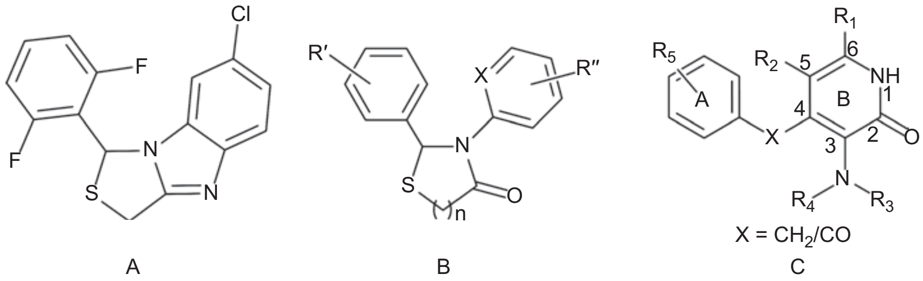

Figure 1. (A) 7-Chloro-1-(2,6-difluorophenyl)-1H,3H-thiazolo[3,4-a] benzimidazole (7-Cl-TBZ), (B) 2,3-diaryl-thiazolidin-4-ones, and (C) 4-benzyl/benzoyl-pyridin-2-ones benzylpyridinones.

Since all the NNRTIs are reported to interact with the HIV-1 RT allosteric site, it is of interest to investigate the important structural features of pyridinones for the HIV-1 RT inhibitory activity. Earlier, we had investigated the quantitative structure–activity relationships (QSARs) of HIV-1 RT inhibitory activity of 2,3-diaryl thiazolidin-4-one class of NNRTIs with different physicochemical and topological indices[12–14]. These studies while confirming the importance of compounds attaining "butterfly-like" conformation for the activity also indicated the prospects of 3-heteroaryl moiety (of thiazolidinones) in modulating the activity. Also, the QSAR of HIV-1 RT inhibitory activity of 2-arylsulfonyl-6-substituted benzonitriles was investigated using Fujita-Ban and Hansch approaches[15]. This has led to suggest the importance of sulfonyl and amine moieties for the activity. In this milieu to explore the scope of chemical space of 4-benzyl/benzoyl-3-dimethylaminopyridin-2(1H)-ones (Figure 1C; Table 1)[16] as HIV-1 RT inhibitors, an attempt has made to rationalize their activity with 0D-2D descriptors from DRAGON software[17].

Feature selection procedures are essential components of modeling studies wherever the number of descriptors involved is very large. It is known in modeling studies that different feature selection approaches show different "bias" in the selection of features from a pool of descriptors to model the phenomenon. Earlier, a hybrid-genetic algorithm (GA)-based descriptor optimization was used in QSAR to model the HIV protease inhibition of tipranavir analogs[18]. Apart from this, weights and biases of neural network were also used in developing highly significant QSAR models from descriptor pools[19]. Additionally, the descriptors consensus to different feature selection approaches may be more promising to pursue in modeling and lead optimization studies. With this in view, two feature selection approaches namely combinatorial protocol in multiple linear regression (CP-MLR) and GA have been used to identify the descriptors for modeling the activity of 4-benzyl/benzoyl-3-dimethylaminopyridin-2(1H)-ones. In this, CP-MLR is a filter-based feature selection procedure[20–22]. It involves a systematic search for the identification of influential features to model the activity. In contrast to CP-MLR, the GA is a stochastic procedure[23]. Being multi-model approaches, both CP-MLR and GA identify different structural features across molecular frame to explain the activity and provide a holistic view to the structure–activity relations[24]. As both these approaches involve different search algorithms, the consensus features evolved from them may be highly significant to model the activity. Furthermore, in QSAR studies, artificial neural networks (ANN) have a special place to develop highly significant predictive models[25–28]. The consensus features of the CP-MLR and GA may serve as good input variables for the ANN to develop predictive models. The results are presented below.

Materials and methods

Chemical structure database and biological activity

The study has involved a series of 55 4-benzyl/benzoyl-pyridin-2-ones (Figure 1C) from the literature (hereafter referred as benzylpyridinones) along with their anti-HIV activity (concentration to achieve 50% inhibition (IC$_{50}$) of wild-type HIV-1 RT in LAI cell line) (Table 1)[16]. For modeling study, the activity has been expressed in the form of logarithm of inverse of inhibitory concentration (–logIC$_{50}$). Adopting the standard procedure, the structure files of the compounds were generated in the ChemDraw[29]. In DRAGON software[17], these structures have resulted in 475 descriptors representing the 0D to 2D characteristics of the molecules. Here, all those descriptors showing a correlation of less than 0.1 with the dependent variable (descriptor vs. activity $r < 0.1$) and descriptor–descriptor intercorrelation ≥0.9 ($r \geq 0.9$) were excluded. It has resulted in 99 descriptors for the investigation. Apart from these, LogP of compounds calculated from the Chem3D Ultra[29] is also incorporated as a descriptor. This makes total descriptors used in analysis as 100. Before proceeding with model development, using the descriptors in single linkage hierarchical cluster analysis[30], all 55 compounds were partitioned into training (35 compounds) and test (20 compounds) sets. Only the training set compounds were used for the development of models.

Descriptors consensus to different feature selection approaches may be more promising to pursue in modeling and lead optimization studies. With this view, two different feature selection approaches namely CP-MLR (a filter directed approach)[20] and GA (a stochastic approach)[23] have been separately used to identify potential features to model the HIV-1 RT inhibitory activity of benzylpyridinones. The descriptors surfaced from

Table 1. Observed and predicted HIV-1 RT inhibitory activity 4-benzyl/benzoylpyridin-2-ones (Figure 1C).

						$-\log IC_{50}$				
							Pred[†]			
S. No.*	R_1	R_2	R_3	R_4	R_5	Obs	Eq. (9)	Eq. (11)	PLS	ANN
1[#]	Me	Et	Me	Me	3,5-diMe	8.10	7.66	8.12	8.31	8.07
2	Me	Et	Me	Me	3,5-diMe	8.40	7.88	8.14	8.48	8.04
3	Me	Me	Me	Me	3,5-diMe	8.30	7.86	8.07	8.30	8.06
4	Et	Me	Me	Me	3,5-diMe	8.00	7.90	7.52	8.00	7.80
5[#]	Me	i-Bu	Me	Me	3,5-diMe	7.30	6.85	7.52	7.56	8.09
6[#]	i-Pr	Me	Me	Me	3,5-diMe	6.10	7.62	6.99	7.05	6.22
7[§]	Me	n-Pr	Me	Me	3,5-diMe	7.30	7.27	7.17	7.10	7.70
8[§]	H	H	Me	Me	3,5-diMe	5.10	5.55	6.20	6.53	5.22
9	Me	H	Me	Me	3,5-diMe	6.40	6.21	7.99	6.40	6.10
10	$-(CH_2)_4-$[‡]		Me	Me	3,5-diMe	8.00	7.93	7.38	8.00	8.01
11	Me	Et	Me	Me	3-Me	8.60	8.26	8.60	8.60	8.07
12[§]	Me	Me	Me	Me	3-Me	7.80	8.38	8.52	8.64	8.04
13	Et	Me	Me	Me	3-Me	7.00	8.48	8.05	7.00	7.23
14[§]	$-(CH_2)_4-$[‡]		Me	Me	3-Me	7.20	8.37	7.92	7.68	7.83
15	$-(CH_2)_3-$[‡]		Me	Me	3-Me	7.80	8.07	8.01	7.80	7.67
16	Me	$(CH_2)_2OMe$	Me	Me	3-Me	8.40	8.67	8.43	8.40	8.46
17[§]	Me	$(CH_2)_3OMe$	Me	Me	3-Me	8.10	8.31	8.78	8.81	8.35
18	Me	$(CH_2)_3OMe$	Me	Me	3,5-diMe	8.70	7.90	8.38	8.70	8.61
19[#]	MeOH	Et	Me	Me	3,5-diMe	9.00	9.13	8.52	9.16	9.07
20	Me	Et	H	CHO	3,5-diMe	7.10	6.21	5.67	7.10	7.10
21[§]	Me	Et	H	Me	3,5-diMe	8.00	8.10	8.13	8.32	8.05
22	Me	Et	Me	Et	3,5-diMe	8.10	7.70	8.04	8.10	8.01
23	Me	Et	Me	*n*-Pr	3,5-diMe	7.80	7.01	7.32	7.80	7.95
24	Me	Et	Me	$CH(Me)CH_2OMe$	3,5-diMe	8.22	8.18	8.05	8.22	8.25
25[#]	Me	Et	Me	$(CH_2)_3SMe$	3,5-diMe	7.60	7.54	7.46	7.42	8.01
26[§]	Me	Et	Me	CH_2CH_2 OMe	3,5-diMe	8.70	8.02	8.11	8.23	8.44
27	Me	Et	Me	$(CH_2)_5OH$	3,5-diMe	8.40	8.30	7.40	8.40	8.23
28	Me	Et	H	COMe	3,5-diMe	6.40	6.38	6.25	6.40	6.38
29	Me	Et	H	COEt	3,5-diMe	5.40	5.83	5.66	5.40	5.09
30	Me	Et	H	COC_3H_7	3,5-diMe	4.00	5.50	5.18	4.00	4.20
31[#]	Me	Et	H	Et	3,5-diMe	7.80	7.71	7.69	7.66	7.94
32	Me	Et	H	*n*-Pr	3,5-diMe	7.70	7.19	7.08	7.70	7.86
33[§]	Me	Et	H	*n*-Bu	3,5-diMe	6.90	6.82	6.31	6.42	6.30
34	Me	Et	Et	Et	3,5-diMe	7.80	7.57	7.88	7.80	7.84
35	Me	Et	*n*-Bu	*n*-Bu	3,5-diMe	4.30	5.98	5.73	4.30	4.34
36	Me	Et	H	Benzyl	3,5-diMe	6.60	6.27	6.06	6.60	6.51
37	Me	Et	Benzyl	Benzyl	3,5-diMe	4.00	5.15	5.34	4.00	4.04
38[#,‡]	Me	Et			3,5-diMe	6.80	8.19	8.38	8.20	7.67
39[¶]	Me	Et			3,5-diMe	6.20	4.44	4.85	6.20	6.19
40[§,‖]	Me	Et			3,5-diMe	7.90	7.31	6.95	6.92	7.80
41	Me	Et	Me	$(CH_2)_2OH$	3-Me	8.30	8.71	9.22	8.30	8.36
42	Me	Et	Me	$(CH_2)_3OH$	3-Me	8.52	8.70	8.61	8.52	8.80
43[#]	Me	Et	Me	$(CH_2)_5OH$	3-Me	8.00	8.79	7.83	7.96	8.43
44	Me	Et	Me	$(CH_2)_2OMe$	3-Me	9.00	8.43	8.47	9.00	8.86
45	Me	Et	Me	$(CH_2)_2OEt$	3-Me	7.89	8.40	8.76	7.89	8.20
46	Me	Et	Me	CH_2CN	3-Me	8.40	8.44	8.94	8.40	8.10
47	Me	Et	Me	$(CH_2)_2CN$	3-Me	7.80	8.34	8.73	7.80	8.08
48[#]	Me	Et	Me	$(CH_2)_3CN$	3-Me	8.30	8.24	9.08	8.51	8.11
49[#]	Me	Et	H	NH-CS-NHEt	3-Me	4.60	5.54	5.92	5.31	5.19
50	Me	Et	H	NH-CS-NHPh	3-Me	5.50	5.86	5.18	5.50	5.43
51	Me	Et	H	NH-CS-NHCOPh	3-Me	5.20	3.39	5.23	5.20	5.40
52[§]	Me	Et	H	$NH-CS-NH_2$	3-Me	6.50	5.87	6.52	5.69	6.69
53	Me	Et	Me	$(CH_2)_2OCH_3$	3-Me	9.00	8.34	8.49	9.00	8.96

Table 1. continued on next page

Table 1. Continued.

						$-\log IC_{50}$				
							Pred[†]			
S. No.[*]	R_1	R_2	R_3	R_4	R_5	Obs	Eq. (9)	Eq. (11)	PLS	ANN
54	Me	Et	Me	Et	3-CHCHCN	9.00	8.91	8.70	9.00	8.09
55	Me	Et	Me	$(CH_2)_2OCH_3$	3-CHCHCN	9.00	9.23	8.64	9.00	8.83

[*] Compounds **2**, **11**–**19** and **53** are 4-benzoylpyridin-2-ones and rest of them are 4-benzylpyridin-2-ones.

[†] Equation 9 is from CP-MLR; Equation 11 is from GA; PLS predicted activities are from PLS model of 11 common descriptors of CPMLR and GA (Table 4); ANN predicted activities are from BP-ANN model developed using 5 input features (Table 5).

[#,§] Test set compounds for CP-MLR, GA and PLS models; in case of BP-ANN model, compounds identified with "#" and "§", respectively correspond to validation and test sets.

[‡] Morpholin-4-yl in place of $-NR_3R_4$ (Figure 1C).

[¶] Piperidin-1yl in place of $-NR_3R_4$ (Figure 1C).

[‖] Pyrrol-1-yl in place of $-NR_3R_4$ (Figure 1C).

these feature selection approaches were pooled together and utilized in partial least squares (PLS) analysis[31,32] to develop single-window structure–activity models. In PLS, the normalized regression coefficients of descriptors provide estimate of each descriptor's fraction contribution to the explained activity. Hence, it is used to rank the descriptors' significance in the PLS model. The high-ranked descriptors of PLS analysis were used in back-propagation ANN[25–28] to develop the predictive models. Since 35 compounds were considered for training the QSAR models, equations containing up to five descriptors were explored (ratio of number of molecules to number of descriptors is >1:5). The computational procedure is briefly described.

Feature selections

CP-MLR

CP-MLR is a filter-based feature selection procedure[20–22]. The thrust of this procedure is in the embedded "filters". Briefly, filter-1 seeds the variables by limiting inter-parameter correlations to predefined level (default value 0.3); filter-2 controls the seeds through t values of variables' coefficients in regression (default threshold value ≥2.0); filter-3 provides comparability of equations of seeds with different numbers of variables in terms of square-root of adjusted multiple correlation coefficient of regression equation, r-bar (default value 0.74); and filter-4 estimates the consistency of the equation in terms of cross-validated r^2 or Q^2 with leave-one-out (LOO) cross-validation as a default option (default threshold value $0.3 \leq Q^2 \leq 1.0$). In CP-MLR, the filters operate in tandem and process the seeds (a string of variables as a bundle) leading to their selection or rejection. Since the principle of combinatorics work in the formation of seeds, the number of seeds result from a set of variables are much more than the individual variables participating in their (seeds) formation. The limits of number of descriptors per seed are the model search perimeter. The models were reassessed for the chance correlations through 100 simulation runs with the randomized biological response[12–14,33] and were also validated with test set compounds.

The selection in CP-MLR proceeded with an initial threshold of filter-1 as 0.3 and subsequently liberated it to 0.79 to boost the formation of different seeds. Considering the degree of correlation of individual descriptors of the dataset with the activity, the search was started with two-variable seeds and an initial filter-3 value of 0.71. The information-rich descriptors were collected by successively incrementing the number of variables per seed as well as the threshold of filter-3 to the optimum r-bar value of the preceding generation.

Genetic algorithm

The GA variable subset selection routine as implemented in MOBY DIGS[23,34] was used for the selection of GA features. It has proceeded with an initial population of 100 solutions (chromosomes) with maximum allowed variables in a solution as five. The fitness for each chromosome was calculated based on LOO cross-validation (Q^2). The reproduction/mutation trade-off (T) value was set to 0.5. Based on the T value, the crossover and mutation values of GA were automatically fixed *in situ* in the computation. The optimum solutions were identified at the end of 100 generations of GA evolution process (selection, crossover and mutation).

Back-propagation ANN

The training set (35 compounds) of CP-MLR/GA analysis was considered as such for the training set of ANN. The test set (20 compounds) of CP-MLR/GA analysis was randomly divided into ANNs validation (10 compounds) and test (10 compounds) sets. The compounds from the training set were used for the model generation whereas the compounds from the validation set were used to stop the overtraining of network. And the compounds from the test set were used to verify the predictivity of the generated model. Coinciding with the number of descriptors in individual feature selection models, for ANN also five descriptors were considered in the input. Before training the networks, the input and output values were normalized with autoscaling of all data. The initial weights were selected randomly between (−0.3) and (0.3). In a standard evaluation procedure with different numbers of hidden layer nodes, the optimum number of nodes for hidden layer was found as four[25–28,35]. The optimization of number of nodes necessary for hidden layer has proceeded by starting with two hidden nodes followed by training the network for best possible output (minimum root mean square error of prediction as a fitness function for training and validation sets). The process has been repeated with incremented hidden layer nodes followed by training the network for assessing output. Using this trial-and-error procedure, the optimum number of hidden nodes necessary for minimizing the error in output is estimated as four. The goal of training the network is to minimize the output errors by changing the weights between the layers. Equation 1 gives the changes in the values of the weights in the network in the optimization of the output.

$$\Delta w_{ij,n} = F_n + \alpha\Delta w_{ij,n-1} \tag{1}$$

In this, Δw_{ij} is the change in the weight factor for each network node, α is the momentum factor, and F is a weight update function, which indicates how weights are changed during the learning process. The weights of hidden layer were optimized using the second derivative optimization method namely Levenberg–Marquardt algorithm[36,37].

Levenberg–Marquardt algorithm

In this algorithm, the update function, F_n, is calculated using equations.

$$F_0 = -g_0 \tag{2}$$

$$g = J^T e \tag{3}$$

$$F_n = -\left[J^T J \times \mu I\right]^{-1} J^T \times e \tag{4}$$

where g is gradient and J is the Jacobian matrix that contains first derivatives of the network errors with respect to the weights, and e is a vector of network errors. The parameter μ is multiplied by some factor (λ) whenever a step would result in an increased e and when a step reduces e, μ is divided by λ.

Statistical parameters

In training the network, the over fitting of data was controlled by comparing the root-mean-square errors (RMSEs) of training and validation sets. It measures the goodness of the output and is useful for the comparison of the target values. The training of the network for the prediction of target value was stopped when the RMSE of the validation set began to increase while that of training set continues to decrease. The goodness-of-fit of activity of the test set compounds was used to further validate the developed models. The predictive ability of the constructed models were assessed using different statistical measures namely, the training, validation and test sets' correlation coefficients (r^2), and corresponding root mean square error of prediction (RMSEP), relative standard error of prediction (RSEP) and mean absolute error (MAE) values. They are calculated using the following equations.

$$r^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_{\text{pred}} - y_{\text{obs}}\right)^2}{\sum_{i=1}^{n}\left(y_{\text{obs}} - y_{\text{mean}}\right)^2} \tag{5}$$

$$\text{RMSEP} = \sqrt{\frac{\sum_{i=1}^{n}\left(y_{\text{pred}} - y_{\text{obs}}\right)^2}{n}} \tag{6}$$

$$\text{RSEP}(\%) = 100\sqrt{\frac{\sum_{i=1}^{n}\left(y_{\text{pred}} - y_{\text{obs}}\right)^2}{\sum_{i=1}^{n}\left(y_{\text{obs}}\right)^2}} \tag{7}$$

$$\text{MAE}(\%) = \frac{100}{n}\sqrt{\sum_{i=1}^{n}\left|\left(y_{\text{pred}} - y_{\text{obs}}\right)\right|} \tag{8}$$

where y_{obs} is the observed activity, y_{mean} is the mean of observed activity and y_{pred} is the predicted activity of the compound in the sample, and n is the number of samples in the concerned set. The ANN computations were carried out using the MATLAB 7.6 for windows[38].

Results and discussion

In CP-MLR, at the end of a search, 18 descriptors (Table 2) were identified as significant ones to model the HIV-1 RT inhibitory activity of benzylpyridinones (Table 1). They are constituent features of several overlapping five-parameter models surfaced for the activity of the compounds (Table 3). Many of these models have explained >78% variance ($r^2 \geq 0.78$) in the activity of training set compounds. They have also accounted for >50% variance ($r^2_t \geq 0.50$) in the activity of test set compounds. Equation 9 is a regression model from among them.

$$\begin{aligned}-\text{LogIC}_{50} = {} & 23.039 - 2.208(0.272)\text{nDB} \\ & - 57.934(28.607)\text{X2A} + 0.157(0.038)\text{VRA2} \\ & + 59.300(17.037)\text{JGI4} - 0.855(0.115)\text{LogP}\end{aligned}$$

$$n = 35,\ r^2 = 0.832, s = 0.650, Q^2 = 0.700,$$

$$Q^2_{\text{G5}} = 0.732,\ F = 28.54$$

$$r^2_{\ t} = 0.605,\ r^2_{\text{Yrand}}(\text{max}) = 0.113(0.325) \tag{9}$$

$$\begin{aligned}-\text{LogIC50} = {} & 9.011 - 4.416(0.545)\text{nDB}_\text{S} \\ & - 1.332(0.658)\text{X2A}_\text{S} + 2.849(0.690)\text{VRA2}_\text{S} \\ & + 2.728(0.783)\text{JGI4}_\text{S} - 4.852(0.655)\text{LogP}_\text{S}\end{aligned} \tag{9s}$$

In this and in all other regression equations, n is the number of compounds, r^2 is the squared correlation coefficient, Q^2 and Q^2_{G5} are cross-validated R^2 from LOO and leave group of five out, respectively, *s* is the standard error of the estimate and *F* is the *F*-ratio between the variances of calculated and observed activities. The values given in the parentheses are the standard errors of the regression coefficients. In the randomization study involving 100 simulations per model, none of the identified models has shown any chance correlation. Furthermore, the models were validated through a test set of 20 analogues listed in Table 1. The predictions of all the test set compounds are within the reasonable limits of their actual values (Table 1). Equation 9s is a derivative of Equation 9, derived using the scaled *X* (X_S) in place of *X* as shown.

$$X_s = \frac{X - X_{MIN}}{X_{MAX} - X_{MIN}} \qquad (10)$$

where X_{MIN} and X_{MAX} are minimum and maximum values of the training set feature *X*. This transforms the descriptor values between "0" and "1", and provides an opportunity for direct comparison of the regression coefficients within the equation. The scaled descriptors are identified with subscript "S" suffixed to the abbreviated names.

Furthermore, the analysis of molecular features in GA has resulted in 14 descriptors as important ones to explain the activity of the compounds (Table 2). They are part of several overlapping five-descriptor models (Table 3) emerged from this approach. These models have explained >83% variance ($r^2 \geq 0.83$) in the activity of training set compounds and showed test set r^2 values ≥0.50. Equation 11 is one among them. Equation 11s is a variant of Equation 11 derived using scaled descriptors.

$$\begin{aligned}-\text{LogIC}_{50} &= 4.257 - 1.494(0.274)\text{nDB}\\ &+ 1.456(0.536)\text{GGI4} + 5.708(1.407)\text{MATS8e}\\ &+ 3.877(1.761)\text{GATS4p} - 0.761(0.101)\text{LogP}\end{aligned}$$

$$n = 35,\ r^2 = 0.839, s = 0.635, Q^2 = 0.735,$$

$$Q^2_{G5} = 0.751, F = 30.17$$

$$\text{r}^2_t = 0.564,\ \text{r}^2_{\text{Yrand}}(\text{max}) = 0.137(0.375) \qquad (11)$$

$$\begin{aligned}-\text{LogIC50} &= 6.878 - 2.989(0.549)\text{nDB}_S\\ &- 1.851(0.682)\text{GGI4}_S + 1.958(0.483)\text{MATS8e}_S\\ &+ 1.101(0.500)\text{GATS4p}_S - 4.318(0.571)\text{LogP}_S\end{aligned} \qquad (11s)$$

The activity predictions from this and other GA equations are within the acceptable limits of their actual values (Table 1). Jointly, the QSAR equations from CP-MLR and

Table 2. Information content of descriptors identified from CP-MLR and GA approaches.

SNo	Descriptor	Class*	FS[†]		Information content[#]
1	nDB	Const	C	G	Number of double bonds
2	X2A	Topo	C		Average connectivity index chi-2
3	PW3		C		Ratio of path/walk 3 - Randic shape index
4	PW4		C		Ratio of path/walk 4 - Randic shape index
5	BIC5			G	Bond information content of neighborhood symmetry of 5 order
6	VRA2		C	G	Average Randic-type eigenvector-based index from adjacency matrix
7	T(N..O)		C		Sum of topological distances between N & O
8	T(O..O)		C	G	Sum of topological distances between O & O
9	BEHp6	BCUT		G	6th Highest eigenvalue of Burden matrix weighted by atomic polarizabilities
10	BELp4		C	G	4th Lowest eigenvalue of Burden matrix weighted by atomic polarizabilities
11	GGI4	Galvez	C	G	Topological charge index of order 4
12	GGI6		C	G	Topological charge index of order 6
13	JGI2		C		Mean topological charge index of order 2
14	JGI4		C	G	Mean topological charge index of order 4
15	JGI6		C		Mean topological charge index of order 6
16	MATS8e	2D-Auto	C	G	Moran autocorrelation of lag 8 weighted by atomic Sanderson electronegativities
17	GATS4e			G	Geary autocorrelation of lag 4 weighted by atomic Sanderson electronegativities
18	GATS8e		C		Geary autocorrelation of lag 8 weighted by atomic Sanderson electronegativities
19	GATS4p		C	G	Geary autocorrelation of lag 4 weighted by atomic polarizabilities
20	H-046	ACF	C	G	H attached to C0(sp3) with no X attached to next C
21	LogP	PROP	C	G	Octanol-water partition coefficient

*Descriptor class: Const, constitutional; Topo, topological; BCUT, BCUT; Galvez, Galvez topological charge indices; 2D-AUTO: 2D autocorrelations; ACF, atom-centered fragments; PROP; proprieties.

[†] Feature selection approach involved in descriptor identification, C for CP-MLR and G for GA.

[#] See Ref. (17).

Table 3. Five parameter models for HIV-1 RT inhibitory activity of 4-benzyl/benzoylpyridin-2-ones (Table 1) from CP-MLR and GA along with statistics.

			Normal stat.[#]			Cross-valid. Stat					Y-Rand r^2_{Yrand}
S. No.	Model*	FS[†]	R^2	s	F	Q^2	Q^2_{G5}	SPRESS	SDEP	Test r^2_t	(max)
1	1, 6, 10, 14, 20	C;G	0.848	0.617	32.31	0.752	0.786	0.788	0.717	0.512	0.109 (0.351)
2	1, 6, 14, 16, 21	C	0.846	0.620	31.98	0.731	0.738	0.820	0.747	0.532	0.124 (0.331)
3	1, 6, 14, 18, 21	C	0.841	0.563	30.63	0.717	0.726	0.841	0.765	0.518	0.149 (0.387)
4	1, 3, 6, 14, 21	C	0.839	0.633	30.40	0.730	0.735	0.822	0.748	0.576	0.132 (0.383)
5	1, 6, 14, 15, 21	C	0.834	0.644	29.23	0.712	0.707	0.849	0.773	0.548	0.133 (0.383)
6	1, 8, 11, 12, 19	C;G	0.834	0.646	28.92	0.730	0.741	0.822	0.749	0.533	0.078 (0.364)
7	1, 4, 7, 13, 21	C	0.832	0.650	28.57	0.743	0.696	0.801	0.729	0.512	0.082 (0.289)
8[§]	1, 2, 6, 14, 21	C	0.832	0.650	28.54	0.700	0.732	0.867	0.789	0.605	0.113 (0.325)
9	1, 5, 14, 16, 17	G	0.830	0.653	28.25	0.753	0.725	0.786	0.715	0.513	0.135 (0.393)
10[‡]	1, 11, 16, 19, 21	G	0.839	0.635	30.17	0.735	0.751	0.815	0.742	0.564	0.137 (0.375)
11	1, 6, 10, 14, 20	C;G	0.848	0.617	32.31	0.752	0.774	0.788	0.717	0.512	0.143 (0.399)
12	1, 8, 11, 12, 19	C;G	0.834	0.646	28.92	0.730	0.756	0.822	0.748	0.533	0.132 (0.323)

*The number corresponds to the descriptor serial number given in (Table 2).

[†] Feature selection approach involved in model formation, C for CP-MLR and G for GA.

[#] In all the models, the number of observations is 35; all statistical abbreviations represent their standard meaning.

[§] Equation 9 in discussion.

[‡] Equation 11 in discussion.

GA approaches have led to 21 descriptors as information rich features to model the activity (Table 2). They have come from six descriptor classes namely, constitutional, topological, BCUT, Galvez, 2D-Autocorrelations and atom-centered fragments. The physical meaning of these descriptors in terms of structural features is described in Table 2. They provide composite property map of the compounds for the HIV-1 RT inhibitory activity. Several of these descriptors have shown their significance in the QSAR models of HIV-1 RT inhibitory activity of thiazolidin-4-ones[12–14]. From this list of descriptors, LogP and nDB are found to be the most influential to modulate the activity of these compounds. Among the 21 descriptors, 11 are common to both CP-MLR and GA approaches. Table 3 shows some CP-MLR and GA models emerged from the descriptors for the activity. Also, both feature selection approaches have shared some common models between them (Table 3).

The 21 descriptors of CP-MLR and GA, and the 11 common descriptors of both these approaches are further analyzed in PLS to facilitate the development of single-window structure–activity models comprising these features. For PLS analysis, the descriptors have been autoscaled (zero mean and unit SD) to give each one of them equal weight in the study. In the cross-validation procedure of the PLS analysis[31,32], four components are found to be the optimum to explain the activity of the compounds. The PLS model from the 21 descriptors of CP-MLR and GA has explained 89.0% variance ($r^2 = 0.890$, $Q^2 = 0.848$, $s = 0.515$, $F = 60.91$) in the HIV-1 RT inhibitory activity of the training set compounds and showed a test set r^2 value 0.569. On the other hand, the PLS model from the 11 common descriptors of CP-MLR and GA has explained 88.8% variance ($r^2 = 0.888$, $Q^2 = 0.834$, $s = 0.520$, $F = 59.46$) in the HIV-1 RT inhibitory activity of the training set compounds and showed 0.607 as test set r^2 value. As the PLS, models emerged from 21 and 11 descriptors have shown almost same level of statistical significance, under principle of parsimony the later may be regarded as better model to explain the activity. The MLR-like PLS coefficients of these two feature sets are shown in Table 4. The plot of fraction contribution of these descriptors to the activity is shown in Figure 2. In both PLS equations, the descriptors nDB, LogP, T(O..O), MATS8e and BELp4 are found to be significant to modulate the activity. Here, nDB accounts for the non-conjugate double bonds, including functional groups, in the molecule. In these compounds, they are due to carbonyl and thionyl functions. In a majority of the analogues of this dataset, it can be attributed to the variation in the bridge carbon between A and B rings (Figure 1C). In regression as well as PLS models, the coefficient of nDB suggested in favor of lesser number (or absence) of these double bonds for better activity. It may be viewed as that between A and B rings, a CH2 bridge is more favorable than a carbonyl bridge (Figure 1C) for the activity. Also in all models, the sign of regression coefficient LogP is negative. Even though LogP is a parameter for hydrophobicity, it suggests the molecular polarity as well. In these analogues, the negative coefficient of LogP may be viewed as favorability of hydrophilic or polar compounds for better activity. The earlier modeling study on these analogues has suggested that –NH–CO-portion of pyridinone moiety offers polar interactions with the receptor[16,39]. This may be satisfying one polar interaction site of the enzyme. The descriptor T(O..O), sum of topological distances between oxygen atoms, has participated in the models with positive regression coefficient. This suggested that in these compounds increasing separation between oxygen atoms as well as their number favor the activity. This may be viewed as the importance of electronegative oxygen in different parts of the structure for the activity. Also, the 2D-autocorrelation descriptor MATS8e with positive regression coefficient suggested the importance of lag 8 autocorrelation weighted by atomic electronegativities for the activity. In these analogues, a small value for BELp4, the 4th lowest eigenvalue of Burden matrix weighted by atomic polarizabilities, would be beneficial for the activity. Galvez topological charge index of orders 4 and 6 (JGI4, GGI4, and GGI6) and Geary autocorrelation of lag 4 weighted by atomic polarizabilities (GATS4p) are the other charge and polarizability indices showed significance in the models (Tables 2 and 3). All these descriptors signify the importance of specific path lengths (in the molecules) weighted by atomic charges and polarizabilities for the activity.

ANN is a powerful tool to identify the patterns in the data. Also, ANN models are difficult to interpret due to the complex computations embedded in the neural networks in deriving the models. In this background, application of well-selected features to ANN input leads to meaningful outputs in terms of rationale behind the input variables[40]. In this scenario, the features from the selection approaches suggest the direction of modification of the chemical space for the activity modulation. Since the number of descriptors in each model of CP-MLR and GA approaches is five, for ANN also five descriptors are considered as input features. In PLS analysis, among the 11 common features of CP-MLR and GA, nDB, T(O..O), BELp4, MATS8e and LogP are more significant ones. Hence, they have been used as input for the development of BP-ANN model for the activity. The architecture and network parameters of ANN are shown in Table 5. In model development, the over fitting of training set has been controlled by the RMSE values of training and test set compounds. The training of the network for the prediction of target value ($-\log EC_{50}$) has been stopped when the RMSE of the validation set has began to increase while that of training set continues to decrease. The developed model has been further evaluated for the goodness-of-fit with the test set. The statistics of ANN model are shown in Table 5. In ANN, these descriptors have well explained the HIV-1 RT activity of the compounds (training, validation and test sets r^2 are 0.932, 0.925, and 0. 890, respectively) (Table 5). The plots of observed versus ANN predicted activities are shown in Figure 3. Attempts are also made to develop ANN model with 11 common features of CP-MLR and GA as input features. This has resulted in excellent predictions for training set but showed relatively less significant predictions in case of validation and test sets (training,

Table 4. MLR-like PLS models from the combined as well as common descriptors of CP-MLR and GA approaches (Table 2) for the HIV-1 RT inhibitory activity ($-\log IC_{50}$) of 4-benzyl/benzoylpyridin-2-ones (Table 1).

		MLR-like coeff (f.c)*	
S. No.	Descriptor	$(\text{CP-MLR}) \cup (\text{GA})$[†]	$(\text{CP-MLR}) \cap (\text{GA})$[#]
1	nDB	−1.164 (−0.180)	−1.612 (−0.257)
2	X2A	−6.139 (−0.007)	
3	PW3	0.187 (0.0005)	
4	PW4	9.942 (0.019)	
5	BIC5	−0.022 (0.0002)	
6	VRA2	0.008 (0.009)	0.025 (0.030)
7	T(N..O)	0.022 (0.051)	
8	T(O..O)	0.050 (0.076)	0.071 (0.109)
9	BEHp6	2.687 (0.053)	
10	BELp4	−2.607 (−0.063)	−3.712 (−0.093)
11	GGI4	0.639 (0.044)	0.990 (0.070)
12	GGI6	−0.980 (−0.037)	−1.541 (−0.060)
13	JGI2	2.209 (0.007)	
14	JGI4	17.480 (0.059)	24.782 (0.086)
15	JGI6	−8.924 (−0.011)	
16	MATS8e	2.252 (0.050)	4.338 (0.099)
17	GATS4e	−2.611 (−0.087)	
18	GATS8e	−1.155 (−0.088)	
19	GATS4p	2.783 (0.056)	4.267 (0.088)
20	H-046	−0.023 (−0.024)	−0.013 (−0.015)
21	LogP	−0.215 (−0.078)	−0.255 (−0.095)
	Constant	3.402	7.986
	Statistics		
	N	35	35
	r^2	0.890	0.888
	S	0.515	0.520
	F	60.91	59.46
	Q^2	0.848	0.834
	Q^2_{G5}	0.862	0.843
	SPRESS	0.605	0.633
	SEDP	0.564	0.586
	r^2_t	0.569	0.607
	r^2_{Yrand}(max)	0.116 (0.383)	0.080 (0.345)

*Coefficients of MLR-like PLS equation in terms of descriptors for their original values; f.c is fraction contribution of regression coefficient, computed from the normalized regression coefficients obtained from the autoscaled (zero mean and unit SD) data.

[†]Combined descriptors of CP-MLR and GA.

[#]Descriptors common to CP-MLR and GA.

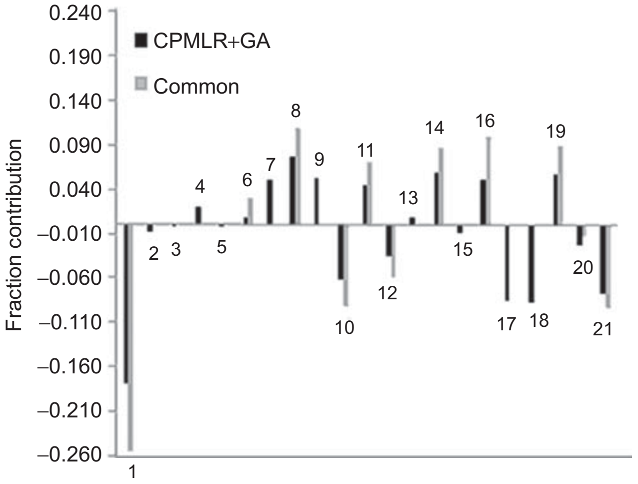

Figure 2. Plots of fraction contribution of MLR-like PLS coefficients (normalized) of the combined and common descriptors of CP-MLR and GA for the HIV-1 RT inhibitory activity of 4-benzyl/benzoyl-pyridin-2-ones; the numbers on the bars refer to the descriptors' numbers (Table 2).

Table 5. ANN Architecture and goodness of fit of HIV-1 RT inhibitory activity of 4-benzyl/benzoylpyridin-2-ones (Table 1) in training, validation and test sets with five most significant features from PLS in BP-ANN model*.

ANN architecture and parameters			
Layer	Nodes	Training parameters	
Input	5 + 1(bias)	Learning rate (μ)	0.611
Hidden	4	Momentum (α)	0.661
Output	1	Transfer function	Sigmoid
		Optimization algorithm	Levenberg–Marquardt
		Iterations (λ)	19

ANN statistics of HIV-1 RT model					
Sample	Sample size	r^2	RMSEP	RSEP (%)	MAE (%)
Train	35	0.931	0.377	4.999	7.907
Valid	10	0.925	0.465	6.234	19.083
Test	10	0.890	0.342	4.609	16.875

*ANN input features are nDB, T(O..O), BELp4, MATS8e and LogP.

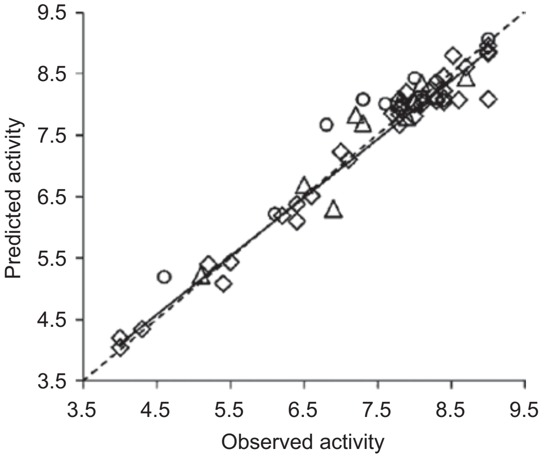

Figure 3. The plots of observed versus predicted activities of 4-benzyl/benzoyl-pyridin-2-ones' training (open diamonds), validation (open citrcles) and test (open triangles) sets from BP-ANN. The solid line indicates the best fit. The dashed line passing through the origin, making an angle of 45° with the axis, bisects the plot area.

validation and test sets r^2 are 0.989, 0.774, and 0.681, respectively). There may be several reasons for this kind of behavior: one is too many input variables for a relatively small dataset. However, in view of the magnitude of test set r^2 value (0.681) the 11 descriptors ANN model still qualifies as predictive model. The results clearly suggested that these descriptors have the ability to identify the patterns in the data and predict the activity of potential analogues.

Conclusions

The feature selection approaches CP-MLR and GA have led to the identification of 21 descriptors to model the HIV-1 RT inhibitory activity of benzylpyridinones. Several of these descriptors have shown significance in explaining the HIV-1 RT inhibitory activity of thiazolidin-4-ones as well. Among the 21 descriptors identified in this exploration, 11 are common to both CP-MLR and GA approaches. Of all the descriptors, LogP and nDB are found to be the most influential to modulate the activity of the benzylpyridinones. In regression as well as PLS models the coefficient of nDB suggested in favor of a CH_2 bridge in between A and B rings of these analogues (Figure 1C) for the activity. The regression coefficient of LogP suggested the favorability of hydrophilic or polar compounds for better activity. In BP-ANN, the five most significant descriptors of PLS analysis (nDB, T(O..O), MATS8e, LogP and BELp4) have explained 93.2% variance in the HIV-1 RT activity of the training set compounds and showed a test set r^2 of 0.890. These results suggest that the descriptors emerged from this study have the ability to identify the patterns in the compounds and can predict the activity of potential analogues.

Acknowledgement

One of the authors (S.D.) thanks CSIR, New Delhi, India, for the financial support in the form of Senior Research Fellowship. CDRI Communication No.7902.

Declaration of interest

The authors report no conflicts of interest in this work.

References

1. Pauwels R. New non-nucleoside reverse transcriptase inhibitors (NNRTIs) in development for the treatment of HIV infections. Curr Opin Pharmacol 2004;4:437–446.
2. Clercq ED. The design of drugs for HIV and HCV. Nat Rev Drug Discov 2007;6:1001–1018.
3. Safadi YE, Boudou VV, Marquet R. HIV-1 reverse transcriptase inhibitors. Appl Microbiol Biotechnol 2007;75:723–737.
4. Clercq ED. Non-nucleoside reverse transcriptase inhibitors (NNRTIs): past, present and future. Chem Biodivers 2004;1:44–64.
5. Prajapati DG, Ramajayam R, Yadav MR, Giridhar R. The search for potent, small molecule NNRTIs: A review. Bioorg Med Chem 2009;17:5744–5762.
6. Esnouf R, Ren J, Ross C, Jones Y, Stammers D, Stuart D. Mechanism of inhibition of HIV-1 reverse transcriptase by non-nucleoside inhibitors. Nat Struct Biol 1995;2:303–308.

7. Ragno R, Frasca S, Manetti F, Brizzi A, Massa S. HIV-reverse transcriptase inhibition: inclusion of ligand-induced fit by cross-docking studies. J Med Chem 2005;48:200–212.
8. Clercq ED. Emerging antiHIV drugs. Expert Opin Emerg Drugs 2005;10: 241–274.
9. Buckheit RW, Fliakas-Boltz V, Yeagy-Bargo S, Weislow O, Mayers DL, Boyer PL, Hughes SH, Pan BC, Chu SH, Bader JP. Resistance to 1-[(2-hydroxyethoxy)methyl]-6-(phenylthio)thymine derivatives is generated by mutations at multiple sites in the HIV-1 reverse transcriptase. Virology 1995;210:186–193.
10. Richman D, Shih CK, Lowy I, Rose J, Prodanovich P, Goff S, Griffin J. Human immunodeficiency virus type 1 mutants resistant to nonnucleoside inhibitors of reverse transcriptase arise in tissue culture. Proc Natl Acad Sci (USA) 1991;88: 11241–11245.
11. Franco JLM, Mayorga KM, Gordiano CJ, Castillo R. Pyridin-2(1H)-ones: a promising class of HIV-1 non-nucleoside reverse transcriptase inhibitors. Chem Med Chem 2007;2:1141–1147.
12. Prabhakar YS, Solomon VR, Rawal RK, Gupta MK, Katti SB. CP-MLR/PLS Directed Structure–Activity Modeling of the HIV-1 RT Inhibitory Activity of 2,3-Diaryl-1,3-thiazolidin-4-ones. QSAR Comb Sci 2004;23:234–244.
13. Prabhakar YS, Rawal RK, Gupta MK, Solomon VR, Katti SB. Topological descriptors in modeling the HIV inhibitory activity of 2-aryl-3-pyridyl-thiazolidin-4-ones. Comb Chem High T Scr 2005;5:431–437.
14. Rawal RK, Prabhakar YS, Katti SB. Molecular surface features in modeling the HIV-1 RT inhibitory activity of 2-(2,6-disubstituted phenyl)-3-(substituted pyrimidin-2-yl)-thiazolidin-4-ones. QSAR Comb Sci 2007;26:398–406.
15. Sharma BK, Kumar R, Singh P. Quantitative structure–activity relationship study of 2-arylsulfonyl-6-substituted benzonitriles as non-nucleoside reverse transcriptase inhibitors of HIV-1. J Enzyme Inhib Med Chem 2002;17:219–225.
16. Benjahad A, Croisy M, Monneret C, Bisagni E, Mabire D, Coupa S et al. 4-Benzyl and 4-benzoyl-3-dimethylaminopyridin-2(1H)-ones: in vitro evaluation of new C-3-amino-substituted and C-5,6-alkyl-substituted analogues against clinically important HIV mutant strains. J Med Chem 2005;48:1948–1964.
17. DRAGON software version 5.0-2005. By Todeschini R, Consonni V, Mauri A, Pavan M. Milano, Italy. http://disat. unimib.it/chm/Dragon.htm.
18. Reddy AS, Kumar S, Garg R. Hybrid-genetic algorithm based descriptor optimization and QSAR models for predicting the biological activity of Tipranavir analogs for HIV protease inhibition. J Mol Graph Model 2010;28:852–862.
19. Guha R, Stanton DT, Jurs PC. Interpreting computational neural network quantitative structure–activity relationship models: a detailed interpretation of the weights and biases. J Chem Inf Model 2005;45:1109–1121.
20. Prabhakar YS. A combinatorial approach to the variable selection in multiple linear regression analysis of Selwood et al. data set -- a case study. QSAR Comb Sci 2003;22:583–595.
21. Prabhakar YS, Gupta MK, Roy N, Venkateswarlu Y. A high dimensional QSAR study on the aldose reductase inhibitory activity of some flavones: topological descriptors in modeling the activity. J Chem Inf Model 2006;46:86–92.
22. Saquib M, Gupta MK, Sagar R, Prabhakar YS, Shaw AK, Kumar R, Maulik PR, Gaikwad AN, Sinha S, Srivastava AK, Chaturvedi V, Srivastava R, Srivastava BS. C-3 Alkyl/Arylalkyl-2,3-dideoxy Hex-2-enopyranosides as antitubercular agents: synthesis, biological evaluation, and QSAR study. J Med Chem 2007;50:2942–2950.
23. Pavan M, Mauri A, Todeschini R. Total ranking models by the genetic algorithm variable subset selection (GA-VSS) approach for environmental priority settings. Anal Bioanal Chem 2004;380:430–444.
24. Deshpande S, Gupta MK, Prabhakar YS. Multi-model environment as a rational approach for drug design: an experience with CP-MLR. IUP J Chem 2010; 3:1–33.
25. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by backpropagating errors. Nature, 1986;323:533–536.
26. Gasteiger J, Zupan J. Neural networks in chemistry. Angew Chem Intl Ed Engl 1993;32:503–527.
27. Bishop CM. Neural Networks for Pattern Recognition. Clarendon Press: Oxford, 1995. P 116–163 and 253–294.
28. Graupe D. Principles of Artificial Neural Networks, 2nd ed. World Scientific Publishing Co.: Singapore; 2007. P 59–111.
29. ChemDraw Ultra 6.0 and Chem 3D Ultra, Cambridge Soft Corporation, Cambridge, USA.
30. SYSTAT, Version 7.0: SPSS Inc., 444 North Michigan Avenue, Chicago, IL 60611.
31. Wold S. Cross validatory estimation of the number of components in factor and principal components analysis. Technometrics 1978;20:397–406.
32. Stahle L, Wold S, in: Eillis GP, West WB. ggEds., Progress in Medicinal Chemistry, vol. 25, Elsevier Science Publishers, B.V. Amsterdam, 1988. P 291–338 (Chapter 6).
33. So SS, Karplus M. Three-dimensional quantitative structure–activity relationships from molecular similarity matrices and genetic neural networks. 1. Method and validations. J Med Chem 1997;40:4347–4359.
34. Todeschini R, Consonni V, Pavan M. MOBYDIGS software (Version 1.2) for Windows, Talete Srl, Milan, Italy, 2002. http://www.talete.mi.it/mobydigs.htm.
35. Marini F, Bucci R, Magrì AL, Magrì AD. Artificial neural networks in chemometrics: History, examples and perspectives. Microchem J 2008;88:178–185.
36. Marquardt DW. An algorithm for leastsquares estimation of nonlinear parameters. J Soc Ind Appl Math 1963;11:431–441.
37. Hagan MT, Menhaj MB. Training Feedforward Networks with the Marquardt Algorithm. IEEE T Neural Net 1994;5:989–993.
38. MATLAB, Version 7.6: <http://www.mathworks.com/products/matlab/>.
39. Carlsson J, Boukharta L, Aqvist J. Combining docking, molecular dynamics and the linear interaction energy method to predict binding modes and affinities for non-nucleoside inhibitors to HIV-1 reverse transcriptase. J Med Chem 2008;51:2648–2656.
40. Goodarzi M, Deshpande S, Murugesan V, Katti SB, Prabhakar YS. Is Feature Selection Essential for ANN Modeling? QSAR Comb Sci 2009; 28: 1487–1499.