Journal of Applied Chemistry

JAC
http://chemistry.journals.semnan.ac.ir

Prediction of the anti-HIV activities of PETT analogs as non-nucleoside HIV-1 reverse transcriptase inhibitors by linear and non-linear QSAR models

Mansour Arab Chamjangali*, Ghadamali Bagherian, Motahhareh Ashrafi, and Amir Hossein Amin

College of Chemistry, Shahrood University of Technology, Shahrood, Iran

Article history:
Received: 1/Apr/2014
Received in revised form: 15/May/2014
Accepted: 1/Jun/2014

Abstract

Quantitative structure-activity relationship (QSAR) models were constructed in order to predict the anti-HIV activity of a set of phenethyl thiazole thiourea (PETT) analogs by calculated descriptors. Molecular descriptors calculated by Dragon software were subjected to variable reduction using the stepwise regression. The variables were then used as inputs for QSAR model generation using multiple linear regression (MLR) and artificial neural network (ANN). Validation study of the MLR and ANN models was performed using the test set and leave-one-out techniques. The results obtained for prediction of the test set by the MLR and ANN models showed squared correlation coefficients of 0.766 and 0.913, respectively.

Keywords: Bayesian regularized algorithm; Anti-HIV; Thiourea; Inhibitors; Artificial neural network (ANN).

1. Introduction

Acquired immunodeficiency syndrome (AIDS) caused by the human immunodeficiency virus type 1(HIV-1) infects many people every day, and millions of people have died from this disease [1, 2]. According to the joint united nation program on HIV/AIDS (UNAIDS), about 22 million people died from AIDS, and over 39 million people lived with HIV/AIDS at the end of 2008[3]. The essential step in the life cycle of HIV-1 is the reversed transcription of the viral RNA genome to produce a double-strand DNA copy, so this process is referred to as reverse transcriptase [4]. Reverse transcriptase (RT) is the prime target for the development of drugs for the HIV/AIDS therapy [5]. RT inhibitors can be divided into two major categories: nucleoside reverse transcriptase inhibitors (NRTIs) and non-nucleoside reverse transcriptase inhibitors (NNRTIs)[5, 6]. NR.TIs cause DNA chain termination when they are incorporated into a growing DNA strand.6 NNRTIs directly block the RT enzyme by binding to a pocket adjacent to the catalytic site of the enzyme.5 In this study, NNRTIs gained the greatest importance because of their specificity and low toxicity [7].Although the therapeutic efficiency of NNRTIs is severely limited by the emergence of HIV-1 drug-resistance, their use in combination therapy has been encouraging and has revived interest in the search

[*] ***Corresponding Author: Email:*** *marab@shahroodut.ac.ir*

for new, selective, and potent NNRTIs. Therefore, in the past 15 years, more than 50 structurally diverse NNRTIs have been described [8].

Computational methods, namely QSAR, have been developed as effective approaches in facilitating new drug discovery. By using these methods, the biological activity of chosen molecules can be estimated before the experimental test. Consequently, they are simple and cheap, and accelerate to design molecules with the desired biological activity [9].

The main steps in QSAR studies can be summarized as structure entry and optimization, descriptor calculations, descriptor selections, model construction, and validation of the proposed model [10, 11]. QSAR model can be formulated based on experimentally derived descriptors or theoretically calculated descriptors.1 The latter group of descriptors can be determined solely by computational methods, and no laboratory measurements are needed. Thus this saves time, material, space, and equipment, and it is available for any molecule, real or hypothetic [12].

There are several major techniques that can be applied for QSAR modeling such as multiple linear regression (MLR) and partial least square (PLS) for inspection of the linear relationship between biological activity and molecular descriptor [13, 14, 15] and artificial neural network (ANN) and support vector machine (SVM) to analyze non-linear relationship between the activity of interest and molecular descriptor [16-19]. In the present contribution, we attempted to establish structure-activity for a set of PETT analogs by means of the ANN and MLR methods. To the best of our knowledge, this is the first report on modeling of these derivatives by the ANN and MLR methods.

2. Theory

Artificial neural networks (ANNs) are computational simulations of biological neural networks [20] that learn through experience with appropriate learning exemplar by detecting the patterns and relationships in data, not from pre-programming [21]. The fundamental component of a neural network is neuron. Neurons are essentially computation units that consist of weighted input, transfer function, and one output [11, 21]. These processing elements can be one of the three different kinds. Neurons in the input layer receive their values from independent variables. In turn, the hidden neuron collects values from the precedent neuron, giving a result that passes to a successor neuron. Finally, neuron in the output layer takes values from other units, and corresponds to different dependent variables [22]. There are different types of network architecture but the most common ANN architecture that is useful for QSAR studies and drug research is multiple-layer feed forward with back propagation learning rule. In this method, a set of input values are propagated in forward direction through the net such that an output is calculated for each node based on the current weight. For a given set of data, the calculated error is the difference between ANN output value and the experimental output [23]. These errors are then used as inputs to feedback connection from which adjustments are made to the synaptic weight layer by the layer in backward direction[6].The goal of training (weight adjustment of) the network is to minimize the output error. There are many back propagation training algorithms with different computational, storage requirement, and different speed available [17]. In this study, we used the Bayesian-regularization training algorithm due to the ability of this algorithm in solving some of the main weaknesses of back propagation ANNs. This is faster than the standard back propagation neural networks. In addition, in this method, the concerns about over fitting and overtraining are eliminated so that the definitive and reproducible model is attained [3].

Overfitting problem or poor generalization capability happens when a neural network overlearns during the training period. In such situation, the error on the training set is small, while for a new set of data, which is presented to the network, the error is large. In other word, the network has memorized the training examples but it has not learned to generalize the new situation [24]. The Bayesian regularization approach modifies the objective function (Eq. 1) by adding a term, MSW, which is the mean of the sum of squares of the network weights:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 \qquad (1)$$

$$F = \alpha MSE + \beta MSW \qquad (2)$$

$$MSW = \frac{1}{n}\sum_{j=1}^{n} w_j^{\ 2} \qquad (3)$$

In these equations, $\hat{y}_i$ is the network output for compound i, y_i is the target value for compound i, N is the number of compounds, and α and β are the parameters which are to be optimized in the Bayesian framework of Mackey [25]. Using this performance function causes the network to have smaller weights and bias, and this forces the network to be smoother and less likely to overfit weights. BR takes place within the Levenberg-Marquardt algorithm performed in Matlab environment. BR ANN has been successfully used in QSAR studies [26-28].

3. Methods and calculations

3.1. Data set

The data sets for IC50 activity of 49 PETT analogs were chosen from the reference[28] Concentration of the compound required to achieve 50% protection of MT-cell from HIV-1 induced cytopatothogency (IC_{50}) was converted to pIC_{50} ([-log (IC50 $\times 10^{-6}$)]) and then used as the dependent variable in the subsequent QSAR studies. Basic skeleton of these compounds is presented in Figure 1, and the various substituents along with the experimental pIC_{50} data for the compounds are listed in Table 1.

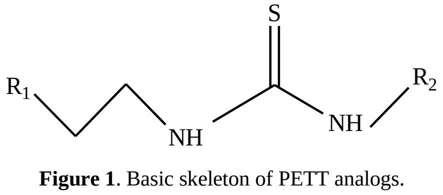


Figure 1. Basic skeleton of PETT analogs.

3.2. Structure entry and descriptor generations

The success of QSAR modeling and accuracy of the results firstly depends on the exact calculation of the values of molecular descriptors. These numerical values computed for compounds after a low-energy conformation were obtained by standard optimization chemical methods (e.g. ab initio, semi-empirical method). For this requirement, the 2D structures of the molecules were drawn by Hyperchem 7. Software (http://www.hyper.com) and geometries were optimized using the semi-empirical AM1 method, employing a gradient limit of 0.001 kcal^{-1} as stopping condition for optimization process. All calculations were carried out at the restricted Hartee-Fack level with no configuration interaction. The optimized geometries were transferred to the dragon software (http://www.disat.unimib.it/vhml) to calculate a total of 1481 descriptors in 18 different classes for each compound available in the data set.

3.3. Descriptor screening

To select the significant descriptors from pool of descriptors, first all descriptors with the same values for about 90% of molecules were removed. Then the remaining descriptors and experimental data were analyzed by the stepwise regression SPSS (version 17) (http://www.spss.com). Only 11 descriptors were selected as the effective ones. These 11 descriptors and the anti-HIV activities were reassembled into a data matrix and used in MLR and ANN modeling. Out of these selected descriptors, a number of 10 descriptors were used as the most feasible ones in the ANN modeling. As it can be seen in the correlation matrix (Table 2), there is no considerable correlation between the selected descriptors.

3.4. ANN generation

A fully connected, three layer, feed-forward ANN with back propagation error was used in this study, whose algorithm was written in MTLAB7.7 (Math Work, Inc., Natick, MA, USA) using the corresponding toolbox in our laboratory, and was run on a personal 3.2 GHz computer. The descriptors appearing in MLR were used as inputs of the networks, and the signal for the output layer demonstrates the anti-HIV activity (pIC50) of the compounds under study.

Table 1. Structural features and pIC_{50} activity for PETT analogs. R_1 and R_2 are defined in Figure 1.

No.	R1	R2	pIC50
1a	phenyl	2-thiazolyl	6.04
2b	2-flourophenyl	2-thiazolyl	7.22
3a	3- flourophenyl	2-thiazolyl	6.82
4a	4- flourophenyl	2-thiazolyl	6.00
5a	2-methoxyphenyl	2-thiazolyl	7.40
6a	3-methoxyphenyl	2-thiazolyl	6.82
7b	4-methoxyphenyl	2-thiazolyl	6.46
8a	2-methylphenyl	2-thiazolyl	7.10
9a	2-azidophenyl	2-thiazolyl	7.52
10a	2-nitrophenyl	2-thiazolyl	6.82
11a	2-hydroxyphenyl	2-thiazolyl	5.96
12b	2-chlorophenyl	2-thiazolyl	6.22
13a	3-ethoxyphenyl	2-thiazolyl	7.22
14a	3-propoxyphenyl	2-thiazolyl	6.70
15a	3-isopropoxyphenyl	2-thiazolyl	6.40
16a	3-phenoxyphenyl	2-thiazolyl	5.96
17a	2,6-dimethoxyphenyl	2-thiazolyl	7.05
18b	2,5-dimethoxyphenyl	2-thiazolyl	6.70
19a	3-bromo -6-methoxyphenyl	2-thiazolyl	7.52
20a	2-fluoro -6-methoxyphenyl	2-thiazolyl	8.00
21a	2-ethoxy -6-flourophenyl	2-thiazolyl	8.00
22b	2,6-diflourophenyl	2-thiazolyl	8.00
23a	2-chloro -6-flourophenyl	2-thiazolyl	8.22
24a	2-pyridyl	2-thiazolyl	6.70
25a	1-methylpyrrol-2-yl	2-thiazolyl	5.72
26a	2-furyl	2-thiazolyl	6.19
27b	phenyl	4-methylthiazol-2-yl	7.00
28a	phenyl	4-ethylthiazol-2-yl	6.22
29a	phenyl	4-propylthiazol-2-yl	6.46
30a	phenyl	4-isopropylthiazol-2-yl	6.70
31a	phenyl	4-butylthiazol-2-yl	5.60
32a	phenyl	4-cyanothiazol-2-yl	6.70
33b	phenyl	4-(triflouromethyl)thiazol-2-yl	6.26
34a	phenyl	4-(ethoxycarbonyl)thiazol-2-yl	6.70
35a	phenyl	5-chlorolthiazol-2-yl	5.62
36b	phenyl	1,3,4- thiadiazol-2-yl	5.72
37a	phenyl	2-pyrazinyl	5.41
38a	phenyl	2-pyridyl	7.70
39a	phenyl	5-bromo-2- pyridyl	7.82
40a	phenyl	5-methyl-2- pyridyl	7.52
41a	phenyl	2-benzothiazolyl	6.70
42b	2,6-difluorophenyl	4-ethylthiazol-2-yl	8.26
43a	2,6-difluorophenyl	4-cyanothiazol-2-yl	8.22
44a	2,6-difluorophenyl	5-bromo-2- pyridyl	9.00
45a	2,6-difluorophenyl	5-methyl-2- pyridyl	8.52
46a	2-ethoxy-6-fluorophenyl	5-methyl-2- pyridyl	8.35
47a	2-ethoxy-6-fluorophenyl	5-bromo-2- pyridyl	8.22
48a	2-pyridyl	5-methyl-2- pyridyl	7.30
49b	2-pyridyl	5-bromo-2- pyridyl	7.82

*** a training set and b test set;**

Table 2. Correlation matrix for selected descriptors.

	nPhX	BEHp8	DECC	RDF105m	Mor26p	CIC2	R5u_A	MATS5p	Mor15p	Mor28u
nPhX	1									
BEHp8	-0.036	1								
DECC	-0.126	0.633	1							
RDF105m	0.154	0.432	0.432	1						
Mor26p	-0.359	-0.301	-0.348	-0.440	1					
CIC2	-0.389	0.556	0.506	0.355	-0.212	1				
R5u_A	-0.104	-0.191	-0.495	-0.167	0.149	-0.106	1			
MATS5p	-0.067	-0.426	0.022	-0.248	0.040	-0.140	0.003	1		
Mor15p	0.245	0.511	0.047	0.222	-0.161	0.259	-0.006	-0.237	1	
Mor28u	0.328	-0.560	-0.362	-0.272	0.064	-0.652	-0.113	0.402	-0.334	1

4. Results and discussion

4.1. Optimization of ANN parameters

For successful training of back propagation ANN and selection of the best ANN model, the effective parameters in ANN performance such as the number of hidden layers, number of neurons in hidden layers, number of input variables, type of training function, transfer function, number of iteration, and momentum values had to be optimized. In the optimization procedure, the data set was randomly partitioned in the training set and test set including 39 and 10 chemicals, respectively. Training set was used for training and optimization weights and biases by the leave-one-out cross validation technique. In this procedure, one compound was removed from the training set. The network was trained using the remaining 38 compounds and then used for prediction of the removed compounds. The process was reiterated for each compound in the training set. Minimization mean square error (MSE) of the training set was selected as criteria in the optimization of parameters. Meanwhile, 10 compounds in the test set were not used during the modeling process, and were kept for evaluation of the constructed ANN model. The architecture and specification of the optimized ANN can be observed in Table 3.

Table 3. Architecture and specification of optimized ANN.

parameter	value
No. of neurons in input layer	10
No. of neurons in hiden layer	4
No. of neurons in output layer	1
Transfer function in hidden layer	tansig
Transfer function in output layer	pureline
Train function	BR
momentum	0.0485
No. of iterations (epoch)	18

4.2. Brief descripton of selected descriptors

The nPhX parameter is a functional group desccriptor which represents the number of halogen atoms bonded to carbon atoms in aromatic ring. This descriptor has a positive coefficient in the model, which indicates that an increase in the number of halogen substituents in aromatic ring causes increase in inhibition of the reverse transcriptase enzyme (e.g. compound numbers 2-4 to 22, 44).

BCUT is a class of molecular descriptors defined as eigenvalues of the modified connectivity matrix, which is also called the Burden matrix B. These descriptors have been demonstrated to reflect relevant aspects of molecular structure, and are therefore useful in similarity searching and comparison. Among BCUT descriptors, BEHp8 (the higest eigenvalue No. 8 of the Burden matrix weighted by atomic polarizabilities) has the highest rank. It was shown that the highest eigenvalues contain contributions from all atoms, and thus reflect the topology of the whole molecule[29, 30] The coefficient for the BEHp8 descriptor has a positive sign in the model, and thus higher values of BEHp8 would be beneficial for activity.

The descriptor DECC [31] belongs to the family of topological descriptors, which is defined as average of absolute sum of the difference between the eccentricity (ηi) and average atom eccentricity ($\bar{\eta}$), where ηi is the maximum distance from the ith vertex to any other vertices, and $\bar{\eta}$ is the average sum of ηi. This distance-based index takes into consideration the distribution of the topological distances in the molecular structure, and hence reflects the topological shape of the compound, describeing the degree of ramification, centricity, and cyclicity.

The 3D-MoRSE descriptors were derived from an equation used in electron diffraction studies. Electron diffraction does not yield atomic coordinates directly but provides diffraction patterns from which the atomic coordinates are derived by mathematical transformations [11, 32]. Some of the 3D-MoRSE (Mor26p, Mor15p, and Mor28u) descriptors appearing in the model are important because they take into account the 3D arrangement of the atoms without depending on the molecular size, and thus are applicable to a large number of molecules with great structural variance.

Radial distribution function (RDF) of an ensemble of N atoms can be interpreted as the probability distribution of finding an atom in spherical volume of certain radius; Eq. 4 represents the RDF code.

$$g(r) = f\sum_{i=1}^{N-1}\sum_{j>i}^{N} A_i A_j e^{B(r-r_{ij})^2} \quad (4)$$

where f is a scaling factor, N is the number of atoms, Ai and Aj are atomic properties of atoms i and j, rij represents the interatomic distance, and B is a smoothing parameter which defines the probability distribution of the individual distances [33]. In the BR-ANN model, RDF105m takes into account the atoms inside virtual spheres 10.5 Å of diameter. These types of descriptors invariance against translation and rotation of the entire molecule, and provide valuable information about interatomic distance in the entire molecule, bond distance, and ring type. The coefficient for RDF105m has a negative sign in the model, which indicates that a lower descriptor value is favorable for the reverese transcriptase enzyme inhibition.

The descriptor CIC2, defined as the complementary information content (neighborhood of order 2) and obtained on the basis of Shannon information theory, takes in to account all atoms in the constitutional formula (hydrogen also being included), and is a topological descriptor. Complementary information content of a system at different levels can be calculated as follows:

$$CIC_k = \log_2 n - \sum_i \frac{n_i}{n}\log_2\frac{n_i}{n} \quad (5)$$

where ni is the number of atoms in the ith class, and n is the total number of atoms in molecule. Division of atoms into different classes depends on the coordination sphere taken in to account. This leads to the different order k. In the two levels, the atom set is decomposed into equivalence classes using their chemical nature and bonding pattern up to the second-order bonded neighbors [34]. This descriptor can express the internal flexibility of the molecule.

Geometry topology and atomic weight assembly (GETAWAY) descriptors are based on the information contained within the molecular influence matrix. They combine the geometrical information in the influence matrix and topological information in the molecular graph weighted by various atomic properties. GETAWAY descriptors contain two sets of molecular descriptors, namely H-GETAWAY and R-GETAWAY descriptors. The influence/distance matrix, where the elements of this matrix are combined with those of the geometry matrix, calculates the R-GETAWAY descriptors [11, 35]. R5u (R-GETAWAY descriptor) is calculated using the leverages between two atoms with the topological distances equal to 5, without any weighing. These descriptors appear as important variables because of the fact that they are highly sensitive to the 3D molecular structure, and are used to compare molecules or even conformers, taking into account their molecular shape, size, and symmetry, and atom distribution. The negative sign for this descriptor indicates that an increase in R5u leads to decrease in the enzyme-inhibitor interaction.

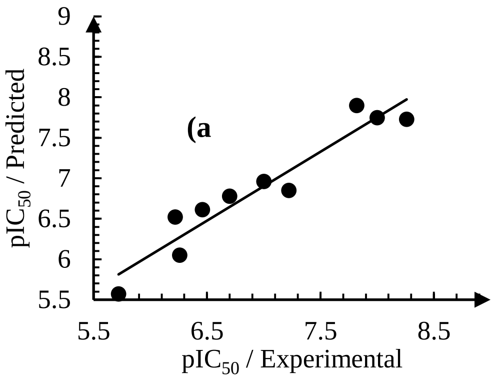
The 2DAUTO class of descriptors [36] also represents the topological structure of the compounds. The 2DAUTO descriptor considered in this study has its origin in autocorrelation of topological structure of Moran (MATS5p). The computation of this descriptor involves the summations of different autocorrelation functions corresponding to the considered length (lag5). At the same time, these descriptors indicate the role of polarizability property of the compounds in deciding the activity.

4.3. Validation of the proposed ANN model

The prediction ability of ANN is its ability to give an acceptable output for a molecule that is not incorporated in the training examples. The prediction ability of the constructed BR-ANN model was evaluated by means of the test set and leave-one-out procedure. In using the test set, the optimized ANN was trained using 39 compounds, and then used for prediction of the pIC_{50} activities for 10 chemicals that were not used in the training procedure. The results illustrated in Table 4 and Figure 2a demonstrates reliability of the model. In the leave-one-out technique, one compound was left out, and the network was trained with the remaining 48 compounds in the data set and then used for prediction of the activity of the discarded compound. The process was repeated until each compound in the data set was removed once. The results (Table 5) show the suitable prediction of the offered model. Figure 3a shows plot of the ANN predicted pIC_{50} against the experimental values. The residual of the ANN predicted values of pIC_{50} versus the experimental values were plotted in Figure 4. The random distribution of the residual around the zero line demonstrates that no systematic error exists in the developed model.

Table 4. Prediction results for the proposed model using the test set.

No.	pIC50				
	experimental	predicted		%E	
		ANN	MLR	ANN	MLR
2	7.22	6.85	6.76	-5.12	-6.37
7	6.46	6.61	6.86	2.32	6.19
12	6.22	6.52	6.78	4.82	9.00
18	6.70	6.78	6.65	1.19	-0.75
22	8.00	7.75	7.50	-3.12	-6.25
27	7.00	6.96	6.59	-0.57	-5.86
33	6.26	6.05	6.08	-3.35	-2.88
36	5.72	5.57	5.61	-2.62	-1.92
42	8.26	7.73	7.53	-6.42	-8.84
49	7.82	7.90	7.97	1.02	1.92



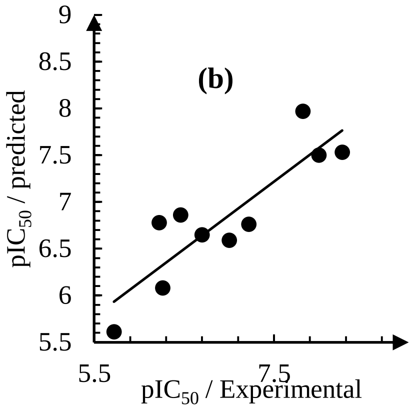


Figure 2. Plot of predicted versus experimental pIC50 values for test set with (a) ANN model (b) MLR model.

For the purpose of comparison, Some MLR models with different numbers of selected descriptors were constructed using training by means of the cross-validation by the leave-one-out method. Molecules in the training set were the same as those in the ANN analysis. The best multiple linear regression model was the one that had the least number of descriptors and high R2adj. The results obtained (Figure 5) showed that the R2adj values increased with increase in the number of descriptors up to 6, and further addition of descriptors into the model did not have any considerable effect on R2adj. Therefore, 6 descriptors were selected as the most feasible ones, which had the following linear equation:

Table 5. Prediction results for the proposed model using leave-one-out procedure.

No.	pIC50				
	experimental	predicted		%E	
		ANN	MLR	ANN	MLR
1	6.04	5.85	5.99	-3.15	-0.83
2	7.22	6.85	6.79	-5.12	-5.96
3	6.82	6.91	6.68	1.32	-2.05
4	6.00	6.24	6.39	4.00	6.50
5	7.40	7.21	7.47	-2.57	0.95
6	6.82	7.09	7.05	3.96	3.37
7	6.46	6.62	6.84	2.48	5.88
8	7.10	7.22	7.05	1.69	-0.70
9	7.52	7.26	7.52	-3.46	0.00
10	6.82	6.70	6.73	-1.76	-1.32
11	5.96	6.07	6.42	1.85	7.72
12	6.22	6.57	6.92	5.63	11.25
13	7.22	6.86	6.84	-4.99	-5.26
14	6.70	6.64	6.65	-0.90	-0.75
15	6.40	6.51	6.35	1.72	-0.78
16	5.96	6.00	6.23	0.67	4.53
17	7.05	6.80	7.03	-3.55	-0.28

18	6.70	6.98	6.72	4.18	0.30
19	7.52	7.52	7.89	0.00	4.92
20	8.00	8.05	7.98	0.63	-0.25
21	8.00	8.33	8.25	4.13	3.13
22	8.00	7.94	7.58	-0.75	-5.25
23	8.22	8.20	8.26	-0.24	0.49
24	6.70	6.56	6.47	-2.09	-3.43
25	5.72	5.60	5.89	-2.10	2.97
26	6.19	6.26	6.03	1.13	-2.58
27	7.00	7.25	6.58	3.57	-6.00
28	6.22	6.47	6.60	4.02	6.11
29	6.46	6.21	6.20	-3.87	-4.02
30	6.70	6.78	6.92	1.19	3.28
31	5.60	5.73	6.01	2.32	7.32
32	6.70	6.77	6. 63	1.04	-1.04
33	6.26	6.02	6.10	-3.83	-2.56
34	6.70	6.78	6.53	1.19	-2.54
35	5.62	5.39	5.31	-4.09	-5.52
36	5.72	5.36	5.68	-6.29	-0.70
37	5.41	5.90	6.26	9.06	15.71
38	7.70	7.16	6.96	-7.01	-9.61
39	7.82	7.83	7.52	0.13	-3.84
40	7.52	7.91	7.28	5.19	-3.19
41	6.70	6.61	6.43	-1.34	-4.03
42	8.26	7.54	7.60	-8.72	-7.99
43	8.22	8.66	8.21	5.35	-0.12
44	9.00	8.55	9.03	-5.00	0.33
45	8.52	8.78	8.78	3.05	3.05
46	8.35	8.02	8.43	-3.95	0.96
47	8.22	8.36	8.45	1.70	2.80
48	7.30	7.16	6.94	-1.92	-4.93
49	7.82	7.78	7.99	-0.51	2.17

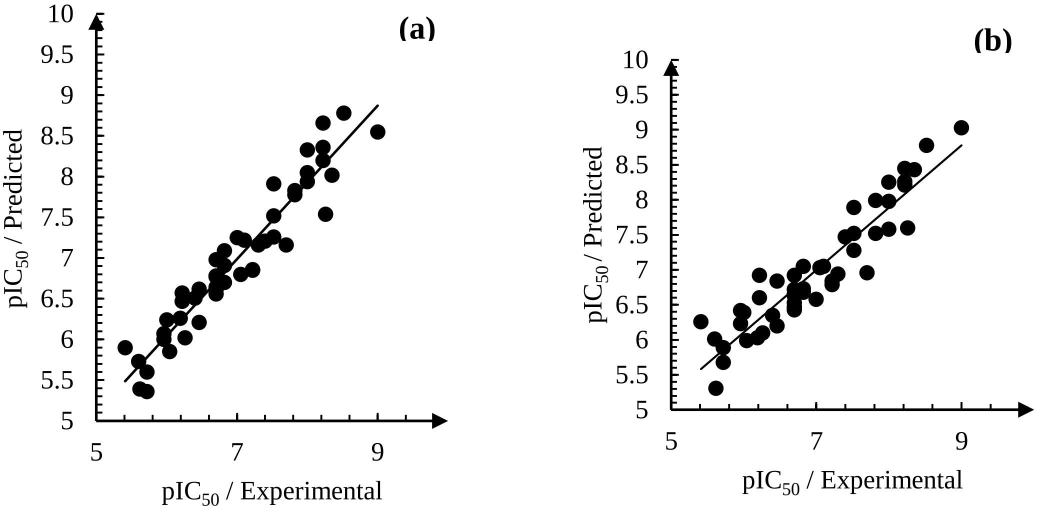


Figure 3. Plot of predicted against experimental pIC50 values for all data sets predicted by LOO cross-validated method with (a) ANN model (b) MLR model.

pIC50 = 4.417 + 0.534nPhX + 3.009 BEHp8 - 2.53DECC - 0.209RDF105m - 3.956 Mor26p - 1.075 CIC2 (6)

To investigate the predictive power of the generated linear model, the test set and leave one-out technique were applied, whose results are demonstrated in Tables 4 and 5, respectively. Figures (2b) and (3b) show plot of the MLR predicted pIC50 against the experimental values for test data and LOO, respectively.

4.4. Y-Randomization or chance correlations

In order to assess robustness of the ANN model and check for the possibility of chance correlations, the Y-randomization test was applied. In this contribution, the dependent variable vector (pIC_{50}) was randomly scrambled, and a new QSAR model was developed by means of the original independent variable matrix (descriptors) and random values of dependent variables. If the original model has no chance correlation, there is a considerable difference in the R^2 values of the original model and new QSAR model constructed using random responses. Several random shuffles of the dependent vectors were carried out, and the results were shown in Table 6. The small values for R2 demonstrate that well results of BR-ANN are not owing to a chance correlation or structural dependency of the training set.

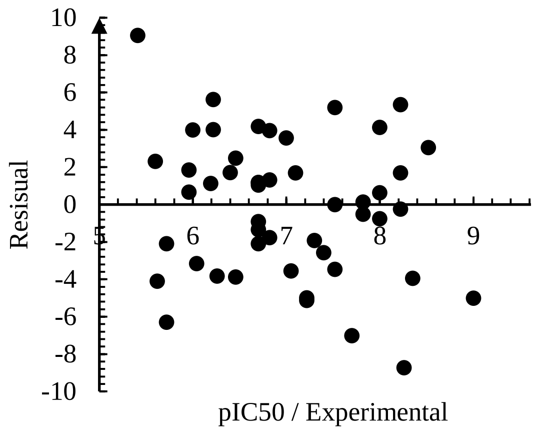


Figure 4. Plot of residual against experimental pIC50 values for all data sets predicted by LOO cross-validated method using ANN model

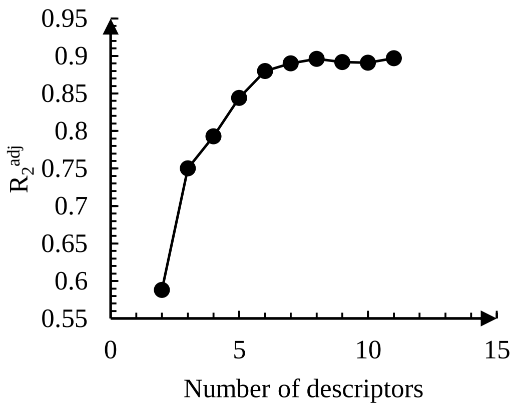


Figure 5. Variation of R^2_{adj} values as a function of number of descriptors

Four statistical parameters were selected to evaluate the prediction power of the constructed model. They were mean square error (MSE), mean absolute error (MAE), squared correlation coefficient (R^2), and mean relative error (MRE). These parameters are calculated as follows:

$$R^2_{adj} = 1-(1-R^2)\frac{n-1}{n-m-1} \tag{7}$$

$$MAE = \frac{\sum_{i=1}^{N}\left|(y_i-\hat{y}_i)\right|}{N} \tag{8}$$

$$MSE = \frac{\sum_{i=1}^{N}(y_i-\hat{y}_i)^2}{N} \tag{9}$$

$$R^2 = 1-\frac{\sum_{i=1}^{N}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{N}(y_i-\bar{y})^2} \tag{10}$$

$$MRE = \frac{\sum_{i=1}^{N}\left|\frac{(y_i-\hat{y}_i)}{y_i}\right|}{N}\times 100 \tag{11}$$

where n is the number of compounds in the training set, m is the number of descriptors involved in the model, y_i is the pIC50 experimental value for compound i, $\hat{y}_i$ is the predicted value for compound i, $\bar{y}$ is the mean of pIC50 experimental values, and N is the number of compounds in the chosen set. The values for these parameters are given in Table 7. Based on these parameters, a comparison between the MLR and ANN models confirm that the ANN model has substantially better and more accurate prediction with respect to the MLR model.

Table 6. R^2test and R^2L.O.O values after several y-randomization tests**.**

iteration	R^2test	R^2L.O.O
1	0.0015	0.00654
2	0.0740	0.0822
3	0.0098	0. 1356
4	0.1599	0.0079
5	0.3606	0.2145
6	0.1059	0.0456
7	0.0624	0.3820
8	0.0034	0.1231
9	0.0237	0.1678
10	0.0055	0.0098

Table 7. Statistical parameters.

parameter	**proposed ANN model**		**MLR model**	
	Test set	**L.O.O**	**Test set**	**L.O.O**
MSE	0.0674	0.068	0.171	0.101
MAE	0.213	0.211	0.355	0.246
R^2	0.913	0.914	0.766	0.870
MRE	3.057	3.029	4.997	3.659

5. Conclusions

In this work, multiple linear regression (MLR) and artificial neural network (ANN) were employed for modeling and predicting the anti-HIV activity of PETT analogs. The ANN approach appeared to be a better model in contrast with the MLR model. The superiority of ANN technique indicates that contribution of some of the descriptors to reverse transcriptase inhibition may be non-linear. This result comes up from the fact that the same parameters were applied for generation of the MLR model and ANN.

Acknowledgment

The authors are thankful to the Shahrood University of Technology Research Council for the support of this work**.**

References

[1] M. Arab Chamjangali, M.Beglari and G.Bagherian, J. Mol. Graphics Model, **26**(2007**)** 360.

[2] K. Zarei and M.Atabati, J. Chin. Chem. Soc, **56***(*2009) 206.

[3] M.Jalali-Heravi and A.Mani-Varnosfaderani, QSAR Comb. Sci, **28** (2009) 946.

[4] H.Bazoui, M. Zahouily, S.Sebti, S.Boulajaaj and D.Zakarya, J. Mol. Model,**8** *(*2002) 1.

[5] D.Weekes and G. B.Fogel, Biosystems, **72** (2003) 149.

[6] M. Jalali-Heravi and F.Parastar, J. Chem. Inf. Model., **40** (2000) 147.

[7] M.Zahouily, J.Rakik, M. Lazar, M. A. Banlaoui, A. Rayadh and N. Komiha, ARKIVOC , **14**(2007) 245.

[8] E.Cichero, S. Cesarini, A. Spallarossa, L. Mosti and P. Fossa, J. Mol. Model, **15** (2009) 871.

[9] B.Hemmateenejad, S. M. H. Tabaei and F.Namvaran, *J. Mol. Struct.: THEOCHEM*, **732***(*2005) 39.

[10] A.Yasri and D.Hartsough, J. Chem. Inf. Model, **41**(2001) 1218.

[11] R.Guha, Methods to improve the reliability, validity and interpretability of QSAR models. PhD thesis, Pennsylvania state university, 2005

[12] A.Habibi-Yangjeh and M.Danandeh-Jenagharad, Indian J. Chem. Sect B, **46B**(2007) 478.

[13] J. T. Leonard and K. Roy, Biorg. Med. Chem, **14**(2006) 1039.

[14] R.Miri, K. Javidnia, B. Hemmateenejad, A. Azarpira and Z. Amirghofran, *Biorg. Med. Chem*, **12***(* 2004) 2529.

[15] K.Roy and J. T.Leonardo, J. Chem. Inf. Model, **45***(*2005) 1352.

[16] R.Vanyúr, K.Héberger and J. Jakus, J. Chem. Inf. Comput. Sci, **43**(2003) 1829.

[17] M.Arab Chamjangali, *Chem. Biol. Drug Des,* **73***(*2009) 456.

[18] M. H.Fatemi and S.Gharaghani, Biorg. Med. Chem, **15**(2007) 7746.

[19] E.Pourbasheer, S.Riah,; M. R.Ganjali and P.Norouzi, Eur. J. Med. Chem, **45***(*2010) 1087.

[20] T.Vasiljević, A. Onjia, Đ.Čokeša and M. Laušević, Talanta, **64**(2004) 785.

[21] S. Y.Tham and S.Agatonovic-Kustrin, *J. Pharm. Biomed. Anal,* **28***(*2002) 581.

[22] J.Caballero, F. M. Zampini, S. Collina and M. Fernández, Chem. Biol. Drug Des, **69**(2007) 48.

[23] A. M.Almerico, M.Tutone and A.Lauria, ARKIVOC *(*2009) 85.

[24] M.Fernández and J.Caballero, Biorg. Med. Chem, **14**(2006) 280.

[25] D. J. C.MacKay , Neural Comput**, 4***(* 1992) 448.

[26] M.Fernández and J. Caballero, J. Mol. Graphics Model. **25**(2006) 410.

[27] M.Fernández, J.Caballero and A.Tundidor-Camba, Biorg. Med. Chem, **14**(2006) 4137.

[28] F. W.Bell, A. S.Cantrell, M.Hoegberg, S. R. Jaskunas, N. G.Johansson, C. L.Jordan, M. D.Kinnick, P.Lind and J. M.Morin, J. Med. Chem, **38**(1995) 4929.

[29] M.Marjanović, M.Kralj, F.Supek, L.Frkanec, I. Piantanida, T.Šmuc and L.Tušek-Božić, J. Med. Chem**,** **50**(2007) 1007.

[30] R.Todeschini and V.Consonni, Handbook of molecular descriptors, *Wiley-VCH Weinheim*: Germany, (2000)

[31] M.Sun, J.Chen, H.Wei, S.Yin, Y.Yang and M.Ji, Chem. Biol. Drug Des, **73** (2009) 644.

[32] J. H.Schuur, P.Selzer and J. Gasteiger, J. Chem. Inf. Comput. Sci, **36***(* 1996) 334.

[33] M. C.Hemmer, V.Steinhauer and J.Gasteiger, Internet J. Vib. Spectro, **19**(1999) 151.

[34] X.Yao, H.Liu, R.Zhang, M.Liu ,Z. Hu , A.Panaye and J. P.Doucet, *Mol.* Pharmaceutics, **2**(2005) 348.

[35] V.Consonni, R.Todeschini , M.Pavan, and P.Gramatica, J. Chem. Inf. Comp. Sci, **42**(2002) 693.

[36] P.Broto, G.Moreau and C.Vandycke, Eur. J. Med. Chem, **19** (1984) 66.