Available online at www.sciencedirect.com
ScienceDirect
Journal of Molecular Graphics and Modelling 26 (2007) 360–367
Journal of Molecular Graphics and Modelling
www.elsevier.com/locate/JMGM

Prediction of cytotoxicity data (CC_{50}) of anti-HIV 5-pheny-l-phenylamino-1*H*-imidazole derivatives by artificial neural network trained with Levenberg–Marquardt algorithm

M. Arab Chamjangali [a,*], M. Beglari [b], G. Bagherian [a]

[a] College of Chemistry, Shahrood University of Technology, Shahrood, P.O. Box 36155-316, Iran
[b] Shahrood Girls Technical and Professional Institute, Haft-e-Tir SQ, Shahrood, P.O. Box 36155-555, Iran

Received 29 October 2006; received in revised form 9 January 2007; accepted 12 January 2007
Available online 18 January 2007

Abstract

A Levenberg–Marquardt algorithm trained feed-forward artificial neural network in quantitative structure–activity relationship (QSAR) was developed for modeling of cytotoxicity data for anti-HIV 5-pheny-l-phenylamino-1*H*-imidazole derivatives. A large number of descriptors were calculated with Dragon software and a subset of calculated descriptors was selected with a stepwise regression as a feature selection technique. The 28 molecular descriptors selected by stepwise regression, as the most feasible descriptors, were used as inputs for feed-forward neural network. The neural network architecture and its parameters were optimized. The data were randomly divided into 31 training and 11 validation sets. The prediction ability of the model was evaluated using validation data set and ''one-leave-out'' cross validation method. The root mean square errors (RMSE) and mean absolute errors for the validation data set were 0.042 and 0.024, respectively. The prediction ability of ANN model was also statistically compared with results of linear free energy related model. The obtained results show the validity of proposed model in the prediction of cytotoxity data of corresponding anti-HIV drugs.

Keywords: ANN; Levenberg–Marquardt algorithm; QSAR; Anti-HIV; Imidazole

1. Introduction

Acquired immunodeficiency syndrome (AIDS), which is caused by the human immunodeficiency virus type 1 (HIV-1), has become a major worldwide pandemic [1,2]. Three million people had died from AIDS and 40 million were living with HIV-1 or AIDS at the end of 2003 [3]. From the beginning of anti-HIV-1 chemotherapy development, HIV-1 reverse transcriptase (RT) has been one of the main targets. Anti-AIDS drugs fall into three categories, the nucleoside reverse transcriptase inhibitors (NRTIs) that act as chain terminators to block the elongation of the HIV-1 viral DNA strand, the non-nucleoside reverse transcriptase inhibitors (NNRTIs) that directly inhibit RT enzyme by binding to the allosteric site near the polymerase active site and the protease inhibitors (PIs) [4–7]. Highly active antiretroviral therapy (HAART) regimens, which are based on triple or quadruple combinations of NRTIs, NNRTIs and PIs, reduce HIV to very low levels, but are unable to extricate the infection and long period therapies lead to the emergence of drug resistant mutant strains [8]. Thus, it is strongly desired to develop new anti-HIV-1 agents with superior efficacy and safety profiles. The activity data for drug like chemicals can be conveniently assayed using cell culture. Once a well-designed subset of chemicals is tested, one can develop quantitative structure–activity relationship (QSAR) models to understand the structural basis of biological activity and the potential activity of untested chemical of the same class.

QSAR approach has become very useful in the prediction of physical and chemical activities and properties [9]. This approach is based on the assumption that variation of the behavior of the compounds, as expressed by any measured physical or chemical activities (properties) can be correlated with changes in molecular features of the compounds termed descriptors [10]. The main steps involved in QSAR include: data collection, molecular geometry optimization, molecular

* Corresponding author. Tel.: +98 273 3335441.
E-mail address: marab@shahroodut.ac.ir (M. Arab Chamjangali).

1093-3263/$ – see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2007.01.005

descriptor generation, descriptor selection, model development and finally model performance evaluation [11]. QSAR models can be formulated based on experimentally derived descriptors or parameters, which can be computed from molecular structure without the input of experimental data. Whereas the experimentally based QSARs work well with narrow classes of chemicals, such parameters are not available for diverse groups of chemicals. Another advantage of theoretically calculated descriptors is that they are available for any molecule, real or hypothetical. Therefore, the latter group of descriptors can be used in the evaluation of compounds not yet synthesized.

One of the important steps involved in QSARs studies is model building. There are several major approaches in QSAR modeling. One is the use of multivariate mathematical–statistical methods such as multiple linear regressions (MLR) [12–15] and partial least squares (PLS) projection of latent structures [16–19]. These methods are linear modeling approaches and have been developed to extract the maximum information from complex data matrices based on their linear behaviors. The other approach is the use of artificial neural networks (ANNs), which offer attractive possibilities for non-linear modeling and optimization when underlying mechanisms are very complex. ANNs are computational simulations of biological networks. An ANN consists of many pathways and nodes organized into a sequence of layers. The first layer is an input layer with one node for each variable or feature of the data. The last layer is an output layer consisting of one node for each variable to be investigated. In between, there is a series of one or more hidden layer(s) consisting of a number of nodes, which are responsible for learning. Nodes of one layer are connected to the nodes of the succeeding layer. Each connection is represented by a number called weight. Initially, a learning phase is defined in which each of the input parameter is applied to a processing element. The weights between these parameters are adjusted until the output is correct. The system can then be applied to unknowns.

ANNs have been widely applied to QSAR studies as a powerful non-linear modeling technique. Some applications of ANNs to the QSAR studies of anti-HIV activity of novel compounds are as follows: study inhibition of HIV replication (IC_{90}) for 55 cyclic urea derivatives [20]; predicting anti-HIV activity for a set of 1-[(2-hydroxyethoxy)methyl]-6-(phenylthio)thymine (HEPT) derivatives [21]; QSAR analysis for a set of 4,5,6,7-tetrahydro-5-methylimidazo[4,5,1-jk][1,4]benzodiazepin-2(1H)-ones (TIBO) derivatives [22]; prediction of anti-HIV activity for a set of 107 inhibitors of the HIV-1 reverse transcriptase derivatives of 1-[(2-hydroxyethoxy)methyl]-6-(phenylthio)thymine [23–26] and anti-HIV-1 activities prediction of 20 tetrapyrrole derivatives [27]. Some evidence show that ANNs modeling give better statistical results both in fitting and prediction, in comparison with linear modeling approaches in QSAR studies [22–25].

As far as we are concerned there are no reports on the use ANNs in the QSAR studies for the 5-pheny-l-phenylamino-1*H*-imidazole derivatives, thus the aim of the current work is to provide an application of ANN to the structure–anti-HIV-1 activity relationship of 5-pheny-l-phenylamino-1*H*-imidazole derivatives. The results obtained by ANN will be statistically compared with those given by multiple linear regressions (MLR).

2. Data and methodology

2.1. Data set

The data used in this QSAR study consisted of cytotoxicity data (CC_{50}), the 50% cytotoxic concentration to reduce MT-4 cell viability, for 42 derivatives of 5-pheny-l-phenylamino-1*H*-imidazole that have been reported by Lagoja et al. [28]. The activity data [CC_{50} (μM)] for 5-phenyl-1-phenylamino-1*H*-imidazole derivatives (Fig. 1 and Table 1) were converted to the logarithmic scale [−log CC (μM)] and then used for subsequent QSAR analyses as the response variables.

2.2. Calculations

The two-dimensional structures of the molecules were drawn using HyperChem 7 software [29]. The final geometries were obtained with the semi-empirical AM1 method in Hyperchem program. All calculations were carried out at the restricted Hartree-Fock level with no configuration interaction. The molecular structures were optimized using the Polak–Ribiere algorithm until the root mean square gradient was 0.001 Kcal mol^{-1}. The resulted geometry was transferred into the Dragon program package, which was developed by Milano Chemometrics and QSPR Group [30], to calculate 1481 descriptors in 18 different classes. Multiple linear regression analysis was carried out for selection of molecular descriptors, using the stepwise strategy in SPSS [31] (for Windows, 13.0) software.

2.3. Descriptor selection

The selection of significant descriptors, which relate the cytotoxicity data to the molecular structure, is an important step in QSAR modeling. To select the significant structural descriptors among 1481, the following steps were taken:

(i) All descriptors with same values for all molecules were omitted.
(ii) The input variables in MLR must not be highly correlated. Therefore, one of the two descriptors that has the pair wise

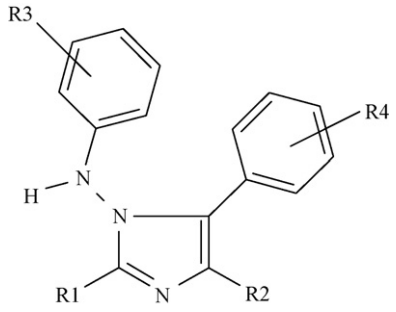

Fig. 1. Structure of 5-pheny-l-phenylamino-1*H*-imidazole compounds.

Table 1
Structural features and cytotoxicity data of 5-pheny-l-phenylamino-1*H*-imidazole compounds. R_1, R_2, R_3 and R_4 are defined in Fig. 1

No.[a]	R_1	R_2	R_3	R_4	Cytotoxicity data
1a	CH_3	SH	3-Cl	H	1.176
2a	CH_3	SH	4-F	H	0.657
3a	CH_3	SH	4-Cl	H	0.872
4a	CH_3	SH	3-Cl	4-Br	1.478
5a	CH_3	SH	3-Cl	3-Cl	1.509
6a	CH_3	SH	3-Cl	4-Cl	1.389
7a	CH_3	SH	3-Cl	4-OCH_3	1.412
8a	CH_3	SH	H	H	0.718
9a	CH_3	SH	3-Br	H	1.354
10a	CH_3	SH	3-NO_2	H	1.316
11a	CH_3	SH	3-F	H	0.966
12a	CH_3	SH	3-CH_3	H	1.054
13a	$(CH_3)_2CH$	SH	3-CH_3	H	1.37
14a	C_2H_5	SH	3-CH_3	H	1.271
15a	C_6H_5	SH	3-Cl	H	1.44
16a	CH_3	SH	3-OCH_3	H	1.42
17a	CH_3	SH	3-Cl	3-CN	1.003
18a	CH_3	SH	3-Cl	3-$OCOCH_3$	1.172
19a	CH_3	SH	3-Cl	3-COOH	0.921
20a	CH_3	SH	3-CH_3	3-COOH	0.728
21a	CH_3	SH	4-C_2H_5	H	1.463
22a	CH_3	SH	4-CH_3S	H	1.275
23a	CH_3	H	3-Cl	H	1.757
24a	CH_3	H	3-Cl	3-Cl	1.923
25a	CH_3	H	3-CH_3	H	1.568
26a	C_2H_5	H	3-CH_3	H	1.555
27a	CH_3	H	3-OCH_3	H	1.463
28a	CH_3	H	3-Cl	3-CN	1.434
29a	CH_3	H	3-Cl	3-$CONH_2$	0.584
30a	CH_3	H	3-CH_3	3-$CONH_2$	0.79
31a	CH_3	H	3-Cl	3-COOH	0.794
32b	CH_3	SH	2-Cl	H	0.791
33b	CH_3	SH	3-Cl	3-Br	1.357
34b	CH_3	SH	2,5-Cl	H	0.832
35b	C_2H_5	SH	3-Cl	H	1.409
36b	CH_3	SH	3-CH_3	3-CN	1.341
37b	CH_3	H	3-Cl	3-Br	1.785
38b	CH_3	H	H	H	1.282
39b	CH_3	H	4-F	H	1.511
40b	CH_3	H	3,5-CH_3	H	1.690
41b	CH_3	H	3-CH_3	3-CN	1.350
42b	CH_3	H	3-Cl	3-$OCOCH_3$	1.480

[a] a: training set and b: prediction set.

correlation coefficient above 0.9 ($R > 0.9$) and has a large correlation coefficient with the other descriptors in each class was eliminated.

(iii) The selected descriptors from each class and the experimental cytotoxicity data were analyzed by the stepwise regression SPSS (Version 13.0) software.

2.4. Artificial neural network

All feed-forward ANNs used in this paper are three-layer networks with 28 units in the input layer, 4 neurons in the hidden layer and 1 unit in the output layer. Each neuron in any layer is fully connected with the neurons of a succeeding layer. Fig. 2 shows an example of the architecture of such ANN. The Levenberg–Marquardt back propagation algorithm (TRAILM) was used for ANN training and the linear functions were used as the transformation functions in hidden and output layers. The ANN algorithms were written in MATLAB 6.1 [32], using the corresponding toolbox. All programs were run on a personal computer (Pentium 266) with Windows XP-2000 operational system.

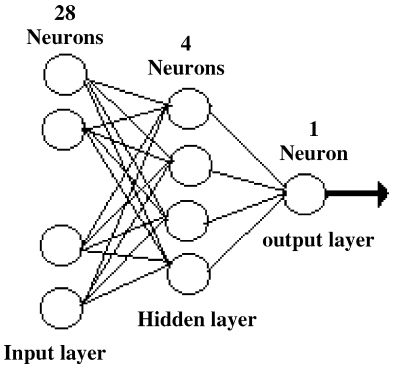

Fig. 2. Used three layer ANN.

3. Results and discussion

3.1. Selected descriptors

Descriptor selection was carried out according to the steps described in Section 2.3. Applying stepwise regression showed that only 44 descriptors of the total calculated ones have significant relationships with cytotoxicity data of anti-HIV derivatives for 5-pheny-l-phenylamino-1*H*-imidazole. From these selected descriptors, a number of 28 descriptors were used as the most feasible descriptors in ANN modeling. A full list of 28 selected descriptors and their chemical meaning are given in Table 2. Some of more effective descriptors and their relevance to the anti-HIV activity of chemicals are describe as follows.

Number of hydrogen attached to heteroatom (H-050) and number of oxygen atoms (nO) were taken in modeling because of the presence of hydrogen bond donor group at meta position of the phenyl ring present at 5 position of the imidazole nucleus reduces cytotoxicity. Mean size is a simple and very significant property of a molecule [33]. Easily obtained from light scattering experiments, a common measure of mean size such as the radius of gyration (RGyr) provides valuable information on interactions of molecule with its surrounding medium or its target. The 3D-MoRSE (3D-molecular representation of structure based on electron diffraction), descriptors are based on the idea of obtaining information from the 3D atomic coordinates by the transform used in electron diffraction studies for preparing theoretical scattering curves [34]. Some of 3D-MoRSE (Mor22m, Mor13m, Mor04e, Mor09p, Mor30p, Mor04p, Mor16v, Mor10u and Mor16u) descriptors appearing in the model are important because they take into account the 3D arrangement of the atoms without ambiguities (in contrast with those coming from chemical graphs), and also because they do not depend on the molecular size, thus being applicable to a large number of molecules with great structural variance and being a characteristic common to all of them. Weighted

Table 2
Descriptors and their meaning

No.	Symbol	Class	Meaning
1	H-050	Atom-centred fragments	Hydrogen attached to heteroatom
2	RGyr	Geometrical	Radius of gyration (mass weighted)
3	Mor22m	3D-MoRSE	3D-MoRSE-signal 22
4	GATS7v	2D-Autocorrelation	Geary autocorrelation-lag 7
5	RDF035m	RDF	Radial distribution function-3.5
6	Mor13m	3D-MoRSE	3D-MoRSE-signal 13
7	E3m	WHIM	3rd component accessibility directional WHIM index
8	Mor04e	3D-MoRSE	3D-MoRSE-signal 04
9	MATS4e	2D-Autocorrelation	Moran autocorrelation-lag 4
10	Mor09p	3D-MoRSE	3D-MoRSE-signal 13
11	H3p	GETAWAY	H autocorrelation of lag 3
12	MATS1v	2D-Autocorrelation	Moran autocorrelation-lag 1
13	Mor30p	3D-MoRSE	3D-MoRSE-signal 30
14	RDF110p	RDF	Radial distribution function-11.0
15	G2u	WHIM	2nt component symmetry directional WHIM index
16	RDF085m	RDF	Radial distribution function-8.5
17	Mor04p	3D-MoRSE	3D-MoRSE-signal 04
18	Mor16v	3D-MoRSE	3D-MoRSE-signal 16
19	BEHe7	BCUT	Highest eigenvalue n.7 of Burden index
20	JGI3	Galvez Topological Charge Indices	Mean topological charge index of order 3
21	RDF040u	RDF	Radial distribution function-4.0
22	nO	Constitutional	Number of oxygen atoms
23	qneg	Charge	Maximum negative charge
24	RDF100e	RDF	Radial distribution function-10.0
25	Mor10u	3D-MoRSE	3D-MoRSE-signal 10
26	RDF060m	RDF	Radial distribution function-6.0
27	Mor16u	3D-MoRSE	3D-MoRSE-signal 16
28	RDF065e	RDF	Radial distribution function-6.5

The letters in the end of descriptor symbols mean: m: weighted by atomic masses; v: weighted by atomic Vander Waals volumes; e: weighted by Sanderson electronegativities; p: weighted by atomic polarizibilities; and u: unweighted.

holistic invariant molecular (WHIM) descriptors [35] are based on the projections of the atoms through principal axes or principal components, with the main purpose of capturing relevant three-dimensional information, in present case the atomic distribution with respect to the invariant reference. These numerical variables (G2u and E3m) point to the influence of the molecular conformation of chemical (drug) during its interaction with virus. In the Galvez Charge Indices terms, the presence of heteroatoms is taken into account by introducing their electronegativity values (according to Pauling's scale taking chlorine as standard value = 2) in the corresponding entry of the main diagonal of the adjacency matrix. These descriptors contain important information on the relationship between the compound structures and its activities by describing the molecular topology and the charge transfer through the molecule [36,37] mainly relative positions of electron-withdrawing substituents change around the two-phenyl ring skeleton. This kind of descriptor (JGI3) was appeared as an important variable due to this fact that the presence of electron-withdrawing substituents at the para position of the phenyl ring of 1-phenylamino fragment is not favorable for the cytotoxicity. The radial distribution function (RDF) descriptors [38] can be interpreted as the probability distribution of finding an atom in a spherical volume of certain radius. The RDF descriptors (RDF035m, RDF110p, RDF085m, RDF040u, RDF100e, RDF060m and RDF065e) are important due to this fact that the absence of any substituents at 2 and 3 positions of the phenyl ring of 1-phenylamino fragment reduces the cytotoxicity. GETAWAY descriptors (H3p) derived from the molecular influence matrix (H) whose elements are obtained through the atomic Cartesian coordinates values [39,40]. H3p deals with atoms at three bonds of distance and the sum is weighted by atomic polarizibilities. This type of elaborated three-dimensional descriptors is able to determine the entire shape and size of the inhibitor. The 2DAUTO class descriptors also represent the topological structure of the compounds. The 2DAUTO descriptors considered in the study have their origin in autocorrelation of topological structure of Moran (MATS4e and MATS1v) and of Geary (GATS7v). The computation of these descriptors involve the summations of different auto-correlation functions corresponding to the different fragment lengths and lead to different autocorrelation vectors corresponding to the lengths of the structural fragments [41]. At the same time, these descriptors indicate the role of physicochemical properties such as mass, volume, and/or polarizability of the compounds in deciding the activity and indicate that the length, width and overall size of meta substituents on the phenyl ring of 1-phenylamino fragment are conductive factors for the cytotoxicity. According to classical chemical theory, all chemical interactions are by nature either electrostatic (polar) or orbital (covalent). Electrical charges in the molecule are obviously the driving force of electrostatic interactions. Thus,

charge-based descriptor [42] was employed as chemical reactivity indices or as measures of weak intermolecular interactions. 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 [43,44].

3.2. ANN optimization

Many factors affect successful training of the back propagation neural networks including the number of hidden layers, kind of training algorithm, transformation (activation) functions, initial weights, the number of neurons in input and hidden units, learning rate and momentum rate. Thus, these factors must be optimized for obtaining a predictable ANN model. In the ANN optimization procedure, the data set for 42 derivatives of 5-pheny-l-phenylamino-1*H*-imidazole was randomly divided into two subsets: training (31) and prediction (11). The minimization of the root mean squares errors (RMSE) of prediction set was used as criteria in ANN optimization.

Unfortunately, there are neither theoretical result available nor satisfactory empirical rules that would enable us to determine the number of hidden layers and number of neurons contained in these layers. However, for most of the applications of ANN to chemistry, one hidden layer seems to be sufficient [45]. Thus, a three layer ANN consisting of input, output and one hidden layer was selected in this study.

After constructing the network, its training is necessary for it to approach an accurate prediction. One of the most extended training algorithms for ANNs structures is back propagation [46]. This method basically consists of a gradient descent technique. This simple gradient descent suffers from several convergence problems. One is vanishing gradient at the solution, meaning that the algorithm takes small steps toward a solution. The other is that the algorithm takes large steps when the gradient is large, which may cause overshooting the local minima. These convergence problems make the training with back propagation neural network slow and hard to learn. The added momentum term in the back propagation might help some convergence problems. Alternatively, various second order learning methods had been proposed. Among these second order methods, Levenberg–Marquardt (LM) algorithm is one of the popular and is very well suited to neural network training where the performance index is the mean squared error [47–49]. The Levenberg–Marquardt algorithm is a variation of Newton's method that was designed for minimizing functions that are sums of squares of other non-linear functions. This algorithm provides a nice compromise between the speed of Newton's method and the guaranteed convergence of steepest descent. Therefore, the Levenberg–Marquardt back propagation algorithm (TRAILM function in Matlab toolbox) was used for ANN training.

The transfer function of the output nodes depends upon the required output of the networks. If this output is qualitative, a sigmoid transfer function is usually used. If the output is quantitative, a linear function may be used. Further studies were carried out to select a proper transfer function in the hidden layer. The sigmoid and linear functions were tested and the results obtained showed that linear function as hidden layer transfer function gives a better prediction ability. Thus, the linear functions were used as the transformation functions in both hidden and output layers.

In order to find the most feasible descriptors, the number of descriptors (number of inputs) used for the training set was optimized. ANNs with different number of hidden nodes and input numbers up to 44 were trained. The results obtained showed that ANN with 28 input nodes had the lowest RMSE error and that such a number of descriptors was chosen for further optimization.

Preliminary studies showed that change of learning rate in the range of 0.001–1.0 has no considerable effect on the RMSE of prediction set in the ANNs with various numbers of hidden layer neurons. To select the best learning times (training epochs) and number of neurons in hidden layer, RMSE error values of the prediction set were calculated for all combinations of number of neurons in the hidden layer (ranged from 2 to 10) and the learning epochs (ranged from 10 to 400 with a step of 10) with learning rate of 0.01. The network training was stopped when one of the default performance goals was reached. The results showed that RMSE varies from minimum value of 0.046 to maximum value of 0.1 for all combinations. According to the results, numbers of hidden layer neurons and epochs were selected 4 and 100, respectively. With this ANN optimized architecture, the required time for training and prediction steps is about 30–70 s.

3.3. ANN validation and statistical parameters

The predictive ability of an ANN is its ability to give a satisfactory output for a molecule that is not included in the ANN learning examples. To determine the predictive aspect, prediction set and leave-one-out procedure were used. In using prediction set, the optimized ANN was learned with 31 training set and then was used for prediction of cytotoxicity data for 11 chemicals in the prediction set. The results are shown in Table 3. In the leave-one-out procedure one compound was

Table 3
Prediction results of the proposed model using prediction set

No.	Cytotoxicity data				
	Actual	Predicted		%E	
		ANN	MLR [50]	ANN	MLR
32	0.791	0.789	0.796	0.2	−0.63
33	1.357	1.36	1.518	−0.22	−11.9
34	0.832	0.928	0.935	−11.5	−12.4
35	1.409	1.436	1.301	−1.9	7.7
36	1.341	1.311	1.147	2.3	14.5
37	1.785	1.777	1.817	0.45	−1.8
38	1.282	1.265	1.096	1.3	14.5
39	1.511	1.6	1.361	−5.6	9.9
40	1.690	1.69	1.719	0	−1.7
41	1.350	1.355	1.447	−0.37	−7.2
42	1.480	1.45	1.440	2	2.7

Table 4
Prediction results of proposed model using leave-one-out procedure

No.	Cytotoxicity data				
	Actual	Predicted		%E	
		ANN	MLR [50]	ANN	MLR
1	1.176	1.158	1.301	1.5	−10.6
2	0.657	0.647	1.061	1.5	−61.5
3	0.872	0.883	0.774	−1.3	11.2
4	1.478	1.492	1.301	−0.95	12.0
5	1.509	1.557	1.480	−3.2	1.9
6	1.389	1.470	1.301	−5.8	6.3
7	1.412	1.372	1.301	2.8	7.9
8	0.718	0.718	0.796	0.0	−10.9
9	1.354	1.338	1.381	1.2	−2.0
10	1.316	1.280	1.338	2.7	−1.8
11	0.966	0.958	1.159	0.83	−20.0
12	1.054	1.088	1.291	−3.2	−22.5
13	1.37	1.355	1.291	1.1	5.8
14	1.271	1.275	1.291	−0.31	−1.6
15	1.44	1.393	1.301	3.3	9.7
16	1.42	1.447	1.353	−1.9	4.7
17	1.003	1.032	1.158	−2.9	−15.4
18	1.172	1.176	1.140	−0.34	2.7
19	0.921	0.890	0.706	3.4	23.3
20	0.728	0.753	0.695	−3.4	4.5
21	1.463	1.456	1.415	0.48	3.3
22	1.275	1.247	1.162	2.2	8.9
23	1.757	1.727	1.601	1.7	8.9
24	1.923	1.988	1.780	−3.4	7.4
25	1.568	1.531	1.590	2.4	−1.4
26	1.555	1.520	1.590	2.2	−2.2
27	1.463	1.482	1.652	−1.3	−12.9
28	1.434	1.452	1.457	−1.3	−1.6
29	0.584	0.540	0.711	7.5	−21.8
30	0.79	0.857	0.700	−8.5	11.4
31	0.794	0.840	1.005	−5.8	−26.6
32	0.791	0.772	0.796	2.4	−0.63
33	1.357	1.369	1.518	−0.88	−11.9
34	0.832	0.900	0.935	−8.2	−12.4
35	1.409	1.437	1.301	−2.0	7.7
36	1.341	1.282	1.147	4.4	14.5
37	1.785	1.777	1.817	0.45	−1.8
38	1.282	1.230	1.096	4.1	14.5
39	1.511	1.588	1.361	−5.1	9.9
40	1.690	1.686	1.719	0.24	−1.7
41	1.350	1.363	1.447	−0.96	−7.2
42	1.480	1.450	1.440	2.027	2.7

removed from the data set, the network was trained with the remaining compounds and used to predict the discarded compound. The process was repeated for each compound in the data set. The results obtained by applying leave-one-out procedure are summarized in Table 4. These results are satisfying and show that the ANN gives correct predictions. Four general statistical parameters were selected to evaluate the prediction ability of the constructed model. These parameters are root mean square error of prediction, relative error of prediction (REP), mean absolute error (MAE) and square of correlation coefficient (R^2). These parameters are calculated as follows:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \quad (1)$$

$$\mathrm{REP}\,(\%) = \frac{100}{\bar{y}} \times \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \quad (2)$$

$$\mathrm{MAE}\,(\%) = \frac{100}{n} \times \sqrt{\sum_{i=1}^{n}|(y_i - \hat{y}_i|} \quad (3)$$

$$R^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (4)$$

where y_i is the actual cytotoxicity data for compound i, $\hat{y}_i$ the predicted cytotoxicity data for compound i, $\bar{y}$ the mean of actual value of cytotoxicity data in the prediction set and n is the number of prediction samples. The values for RMSE, REP, MAE and R^2 are given in Table 5. A simple comparison between ANN prediction results and reported best MLR results [50] shows that ANN has a better and more accurate prediction compared with MLR method.

4. Conclusion

A Levenberg–Marquardt algorithm trained neural network was applied to analyze the QSAR of 5-pheny-l-phenylamino-1*H*-imidazole compounds. Based on our knowledge, this work is the first report on the use of ANN combined with Levenberg–Marquardt algorithm in QSAR studies. The results obtained show that this ANN modeling was able to establish a satisfactory relationship between the molecular descriptors and the anti-HIV activity. ANN approach would seem to have a great potential for determining quantitative structure-anti-HIV-1 activity relationships in comparison with reported MLR method [50].

Acknowledgement

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

Table 5
Statistical parameters

Parameter	Proposed ANN method		MLR method [50]	
	Prediction set ($n = 11$)	Leave-one-out set ($n = 42$)	Prediction set ($n = 11$)	Total data set ($n = 42$)
MAE	0.028	0.026	0.077	0.120
REP	3.60	3.27	10.25	12.07
RMSE	0.0418	0.037	0.119	0.135
R^2	0.9808	0.9877	0.8500	0.8313

References

[1] M.S. Gottlieb, R. Schroff, H.M. Schanker, J.D. Weisman, P.T. Fan, R.A. Wolf, A.N. Saxon, Pneumocystis carinii pneumonia and mucosal candidiasis in previously healthy homosexual men: evidence of a new acquired cellular immunodeficiency, Engl. J. Med. 305 (1981) 1425–1431.

[2] F. Barre-Sinoussi, J.C. Chermann, F. Rey, M.T. Nugeyre, S. Chamaret, J. Gruest, C. Dauguet, C. Axler-Blin, F. Vezinet-Brun, C. Rouzioux, W. Rozenbaum, L. Montagnier, Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS), Science 220 (1983) 868–871.

[3] UNAIDS/WHO AIDS Epidemic Update, December 2003, UNAIDS/WHO, Geneva, Switzerland, 2003.

[4] E. De Clercq, Toward improved anti-HIV chemotherapy: therapeutic strategies for intervention with HIV infections, J. Med. Chem. 38 (1995) 2491–2517.

[5] M. Artico, Non-nucleoside anti-HIV-1 reverse transcriptase inhibitors (NNRTIs): a chemical survey from lead compounds to selected drugs for clinical trials, Farmaco 51 (1996) 305–331.

[6] E. De Clercq, Highlights in the development of new antiviral agents, Mini-Rev. Med. Chem. 2 (2002) 163–175.

[7] R.F. Schinazi, J.R. Mead, P.M. Feorino, Insights into HIV chemotherapy, AIDS Res. Hum. Retroviruses 8 (1992) 963–990.

[8] A.M. Vandamme, K. Van Vaerenbergh, E. De Clercq, Anti-human immunodeficiency virus drug combination strategies, Antivir. Chem. Chemother. 9 (1998) 187–203.

[9] X.J. Yao, Y.W. Wang, X.Y. Zhang, R.S. Zhang, M.C. Liu, Z.D. Hu, B.T. Fan, Radial basis function neural network-based QSPR for the prediction of critical temperature, Chemom. Intell. Lab. Syst. 62 (2002) 217–225.

[10] X.J. Yao, M.C. Liu, X.Y. Zhang, Z.D. Hu, B.T. Fan, Radial basis function network-based quantitative structure–property relationship for the prediction of Henry's law constant, Anal. Chim. Acta 462 (2002) 101–117.

[11] A. Yasri, D. Hartsough, Toward an optimal procedure for variable selection and QSAR model building, J. Chem. Inf. Comput. Sci. 41 (2001) 1218–1227.

[12] M. Kukla, H. Breslin, R. Pauwels, C. Fedde, M. Miranda, M. Scott, R. Sherrill, A. Raeymaekers, J. Van Gelder, K. Andries, M.A.C. Janssen, E. De Clercq, P. Janssen, Synthesis and anti-HIV-1 activity of 4,5,6,7-tetrahydro-5-methylimidazo [4,5,1-jk] [1,4] benzodiazepin-2 (lH)-one (TIBO) derivatives, J. Med. Chem. 34 (1991) 746–751.

[13] R. Miri, K. Javidnia, B. Hemmateenejad, A. Azarpira, Z. Amirghofran, Synthesis, cytotoxicity, QSAR, and intercalation study of new diindenopyridine derivatives, Bioorg. Med. Chem. 12 (2004) 2529–2536.

[14] S. Bajaj, S.S. Sambi, A.K. Madan, Topochemical model for prediction of anti-HIV activity of HEPT analogs, Bioorg. Med. Chem. Lett. 15 (2005) 467–469.

[15] A.D. Pillai, S. Rani, P.D. Rathod, F.P. Xavier, K.K. Vasu, H. Padh, V. Sudarsanam, QSAR studies on some thiophene analogs as anti-inflammatory agents: enhancement of activity by electronic parameters and its utilization for chemical lead optimization, Bioorg. Med. Chem. 13 (2005) 1275–1283.

[16] J.T. Leonard, K. Roy, QSAR by LFER model of HIV protease inhibitor mannitol derivatives using FA-MLR, PCRA, and PLS techniques, Bioorg. Med. Chem. 14 (2006) 1039–1046.

[17] L. Lin, W.Q. Lin, J.H. Jiang, G.L. Shen, R.Q. Yu, QSAR analysis of substituted bis[(acridine-4-carboxamide)propyl]methylamines using optimized block-wise variable combination by particle swarm optimization for partial least squares modeling, Eur. J. Pharmaceut. Sci. 25 (2005) 245–254.

[18] F.A. De Lima Ribeiro, M.M. Castro Ferreira, QSAR model of the phototoxicity of polycyclic aromatic hydrocarbons, J. Mol. Struct. THEOCHEM 719 (2005) 191–200.

[19] C.H.T. De Paula da Silva, S.M. Sanches, C.A. Taft, A molecular modeling and QSAR study of suppressors of the growth of trypanosoma cruzi epimastigotes, J. Mol. Graph. Model. 23 (2004) 89.

[20] M. Fernández, J. Caballero, Modeling of activity of cyclic urea HIV-1 protease inhibitors using regularized-artificial neural networks, Bioorg. Med. Chem. 14 (2006) 280–294.

[21] D. Weekes, G.B. Fogel, Evolutionary optimization, backpropagation, and data preparation issues in QSAR modeling of HIV inhibition by HEPT derivatives, Biosystems 72 (2003) 149–158.

[22] L. Douali, D. Villemin, D. Cherqaoui, Exploring QSAR of non-nucleoside reverse transcriptase inhibitors by neural networks: TIBO derivatives, Int. J. Mol. Sci. 5 (2004) 48–55.

[23] M. Jalali-Heravi, F. Parastar, Use of artificial neural networks in a QSAR study of anti-HIV activity for a large group of HEPT derivatives, J. Chem. Inf. Comput. Sci. 40 (2000) 147–154.

[24] L. Douali, D. Villemin, D. Cherqaoui, Comparative QSAR based on neural networks for the anti-HIV activity of HEPT derivatives, Curr. Pharm. Des. 9 (2003) 1817–1826.

[25] L. Douali, D. Villemin, A. Zyad, D. Cherqaoui, Artificial neural networks: non-linear QSAR studies of HEPT derivatives as HIV-1 reverse transcriptase inhibitors, Mol. Divers. 8 (2004) 1–8.

[26] H. Bazoui, M. Zahouily, S. Boulaajaj, S. Sebti, D. Zakarya, QSAR for anti-HIV activity of HEPT derivatives, SAR QSAR Environ. Res. 13 (2002) 567–577.

[27] R. Vanyur, K. Heberger, J. Jakus, Prediction of anti-HIV-1 activity of a series of tetrapyrrole molecules, J. Chem. Inf. Comput. Sci. 43 (2003) 1829–1836.

[28] I.M. Lagoja, C. Pannecouque, A. Van Aerschot, M. Witvrouw, Z. Debyser, J. Balzarini, P. Hardewijn, E. De Clercq, *N*-aminoimidazole derivatives inhibiting retroviral replication via a yet unidentified mode of action, J. Med. Chem. 46 (2003) 1546.

[29] HyperChem Release 7, HyperCube, Inc., http://www.hyper.com.

[30] R. Todeschini, Milano Chemometrics and QSPR Group, http://www.disat.unimib.it/vhml.

[31] SPSS for Windows, Statistical Package for IBM PC, SPSS Inc., http://www.spss.com.

[32] MATLAB 6.1, The Math Works Inc., Natick, MA.

[33] G.A. Arteca, Analysis of shape transitions using molecular size descriptors associated with inner and outer regions of a polymer chain, J. Mol. Struct. THEOCHEM 630 (2003) 113–123.

[34] J.H. Schuur, P. Selzer, J. Gasteiger, The coding of the three dimensional structure of molecules by molecular transforms and its application to structure-spectra correlations and studies of biological activity, J. Chem. Inf. Comput. Sci. 36 (1996) 334–344.

[35] R. Todeschini, M. Lasagni, E. Marengo, New molecular descriptors for 2D and 3D structures, Theory, J. Chemom. 8 (1994) 263–272.

[36] J. Galvez, R. Garcia, M.T. Salabert, R. Soler, Charge indexes. New topological descriptors, J. Chem. Inf. Comput. Sci. 34 (1994) 520–525.

[37] J. Galvez, R. Garcia-Domenech, J.V. de Julihn-Ortiz, R. Soler, Topological approach to drug design, J. Chem. Inf. Comput. Sci. 35 (1995) 272–284.

[38] M.C. Hemmer, V. Steinhauer, J. Gasteiger, The prediction of the 3D structure of organic molecules from their infrared spectra, J. Vib. Spectrosc. 19 (1999) 151–164.

[39] V. Consonni, R. Todeschini, M. Pavan, Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 1. Theory of the novel 3D molecular descriptors, J. Chem. Inf. Comp. Sci. 42 (2002) 682–692.

[40] V. Consonni, R. Todeschini, M. Pavan, P. Gramatica, Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 2. Application of the novel 3D molecular descriptors to QSAR/QSPR studies, J. Chem. Inf. Comp. Sci. 42 (2002) 693–705.

[41] P. Broto, G. Moreau, C. Vandycke, Molecular structures: Perception, autocorrelation descriptor and SAR studies, Eur. J. Med. Chem. 19 (1984) 66–70.

[42] M. Karelson, V.S. Lobanov, A.R. Katritzky, Quantum-chemical descriptors in QSAR/QSPR studies, Chem. Rev. 96 (1996) 1027–1043.

[43] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, Wiley-VCH, Weinheim, Germany, 2000.

[44] R.S. Pearlman, K.M. Smith, Metric validation and the receptor-relevant subspace concept, J. Chem. Inf. Comp. Sci. 39 (1999) 28–35.

[45] J.J. Zupan, J. Gasteiger, Neural Networks for Chemists. An Introduction, VCH Publishers, Weinheim (Germany), 1993.

[46] D.E. Rumelhart, G.E. Hinton, J.L. McClelland, in: D.E. Rumelhart, J.L. McClelland (Eds.), A General Framework for Parallel distributed Processing, vol. 1, Foundations, MIT Press, Cambridge, MA, 1986.
[47] M.T. Hagan, M.B. Menhaj, Training feedforward networks with the marquardt algorithm, IEEE Trans. Neural Netw. 5 (1994) 989–993.
[48] A.J. Adeloye, A. De Munari, Training feed forward networks with the marquardt algorithm 36-artificial neural network based generalized storage–yield–reliability models using the Levenberg–Marquardt algorithm, J. Hydro. 326 (2006) 215–230.
[49] E. Tourwe, R. Pintelon, A. Hubin, Extraction of a quantitative reaction mechanism from linear sweep voltammograms obtained on a rotating disk electrode. Part I: theory and validation, J. Electroanal. Chem. 594 (2006) 50–58.
[50] K. Roy, J.T. Leonard, 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, Bioorg. Med. Chem. 13 (2005) 2967–2973.