A 2D/3D-QSAR STUDY ON BIOLOGICAL ACTIVITIES OF 1,2-ETHYLENDIAMINE DERIVATIVES AS ANTI-TUBERCULOSIS DRUGS

GHASEM GHASEMI, REIHANEH MOHAMADZADE*

Department of Chemistry, Rasht Branch, Islamic Azad University, Rasht, Iran

ABSTRACT

In this work quantitative structure-activity relationship (QSAR) study has been done on 1,2-ethylenediamine derivatives as anti-tuberculosis drugs. Genetic algorithm (GA), artificial neural network (ANN), multiple linear regressions (stepwise-MLR) and Imperialist Competitive Algorithm (ICA), were used to create the nonlinear and linear QSAR models. The root-mean square errors of the training set and the test set for GA–ANN models using the jack-knife method, were 0.1402, 0.1304 and $Q^2 = 0.94$. Also, the R and R^2 values 0.85, 0.73 in the gas phase were obtained from a GA-stepwise-MLR model. Q2 of training set for PLS was 0.52. The results obtained from this work indicate that ANN and ICA models are more effective than other statistical methods and exhibit reasonable prediction capabilities. The best descriptors are G3u, HATS2e, F02(C-N), GGI10, RDF040m, Mor22p, Mor05p, TIC4, H4e, H-052, G2m and G1e.

KEYWORDS: Tuberculosis, quantitative structure-activity relationship, 1,2-ethylenediamine derivatives, Genetic Algorithm and Imperialist Competitive Algorithm, Artificial Neural Network

INTRODUCTION

Tuberculosis in humans is mainly caused by Mycobacterium tuberculosis[1]. The infection is transmitted by respirable droplets generated during forceful expiratory activity such as coughing.

Tuberculosis infection can be either active or latent. The World Health Organization (WHO) estimates that within the next 20 years about 30 million people will be infected with the bacillus. The clinical management of TB has relied heavily on a limited number of drugs such asisonicotinic acid, hydrazide, rifampicin, ethambutal, streptomycin, ethionamide, pyrazinamide, fluroquinolones etc[3-4].

In imperialistic competition algorithm(ICA), all the empires try to take possession of colonies of other empires and control them. This competition gradually brings about a decrease in the power of weaker empires and an increase in the power of more powerful ones. This competition is modelled by just picking some (usually one) of the weakest colonies of the weakest empires and making a competition among all empires to possess the colony, or colonies [5-7].

The search for quantitative relations between chemical structure and biological activity is the subject of quantitative structure-activity relationships, the purpose of which is to explain why a given drug produces its particular effect, and ultimately to predict the effect of newly synthesized chemical compounds.

One of the important features of mathematical model is its ability to predict the activity of molecules not yet synthesized or those with limited *in vivo* and *in vitro* experimental information, for economic or ethical reasons.

Partial least squares (PLS), is highly sensitive to extreme values of variables, which do not contribute to a predictive model[8]. The situation becomes worse when more variables are introduced to the models. Thus, the larger the number of variables, the less the predicted value of the developed model[9]. Therefore, a variable selection step is necessary prior to building PLS models. To solve this problem, GA and PLS have been combined in a variable selection of QSAR and QSPR modelling. Genetic algorithms are best known for their ability to efficiently search large spaces and have been widely applied in different fields[10-11]. Thus, GA can be used for improving the prediction of QSAR modelling.

The more commonly used techniques in construction of QSAR models are PLS, principle component regression (PCR) and ANN[12-14]. PLS is insensitive to co-linearity among the predictor variables and allows one to handle data sets where the number of variables is larger than the number of observations. Thus, for large data sets PLS is preferable. In addition, PLS analysis provides equations describing the relationship between one or more dependent variables and a group of explanatory variables.

The RMSE can be calculated for prediction or validation samples (RMSEP) and for calibration samples (RMSEC). RMSEC (all validation methods) is calculated as:

$$\mathrm{RMSEC} = \frac{1}{\mathbf{yWeight}}\sqrt{\mathbf{ResYCalVar}} \qquad (1)$$

Res Ycal Var is Residual calibration variance

DATA SET

The data set consists of the experimental bioavailabilities of 21 structurally-diverse chemicals as reported by Protopopova[15]. Twenty-seven compounds with MICs of 15.6 mM were tested on Vero cells to determine in vitro cytotoxicity (IC_{50}) and to establish a selectivity index (SI). Five of the most potent compounds were tested for in vivo efficacy in a murine model of chronic tuberculosis infection.

MATERIALS AND METHOD

Actual half-maximal inhibitory concentration (IC_{50}) values of all compounds were selected from literature. This set contained the effective concentration activities of 21diketo analogues. A set of seven compounds was randomly removed from the dataset to be used as the prediction set (PSET). The log ($1/IC_{50}$) of this set spanned the entire dataset. The remaining compounds were utilized as the training set (TSET). In the simplest form of bootstrapping, instead of repeatedly analyzing subsets of the data, sub samples of the data are repeatedly analyzed. Each sub sample is a random sample with replacement from the full sample.

The structure and biological data of 21 molecules were obtained from literature[15]. The 3D structures of the molecules were generated using the built optimum option of Chemoffice. The structures were then fully optimized based on the ab initio method using the DFT level of theory (Figure 1). Dragon (version 5.5) was employed to calculate the molecular descriptors. All calculations were performed using the Gaussian 09W program series. Geometric optimization of compounds was carried out using the B3LYP method employing a 6–31G (2d) basis set.

The independent variables were molecular descriptors and the dependent variables were the actual half-maximal inhibitory concentration (IC_{50}) values. More than 3226 theoretical descriptors were selected and calculated. Unscrambler (version 9.7) was used for analysis of data and statistical methods. For each compound in the training sets, a correlation equation was derived using the same descriptors. The equation was then used to predict log ($1/IC_{50}$) values for the compounds from the corresponding test sets.

Stepwise MLR, GA, ANN and ICA were used to select the most appropriate descriptor from all descriptors.

A genetic algorithm (Figure 2, Table3), ICA (Figure 2,3), ANN, MLR (Table 2), PLS (Table 1), Principal component regression(Table1) and least absolute shrinkage and selection operator (Table 2) were used to create the QSAR models.

First, descriptors that had the same values for at least 75% of compounds within correlation coefficients less than 0.4 with the dependent variable were regarded redundant and removed. Finally, since highly correlated descriptors provide approximately identical information, a pairwise correlation was performed. When their correlations coefficient exceeded 0.95, one of two descriptors was randomly removed. These descriptors would be used as inputs of the ANN. GA was utilized as the mean for non-linear feature selection.

e-mail: ghassemi47@gmail.com

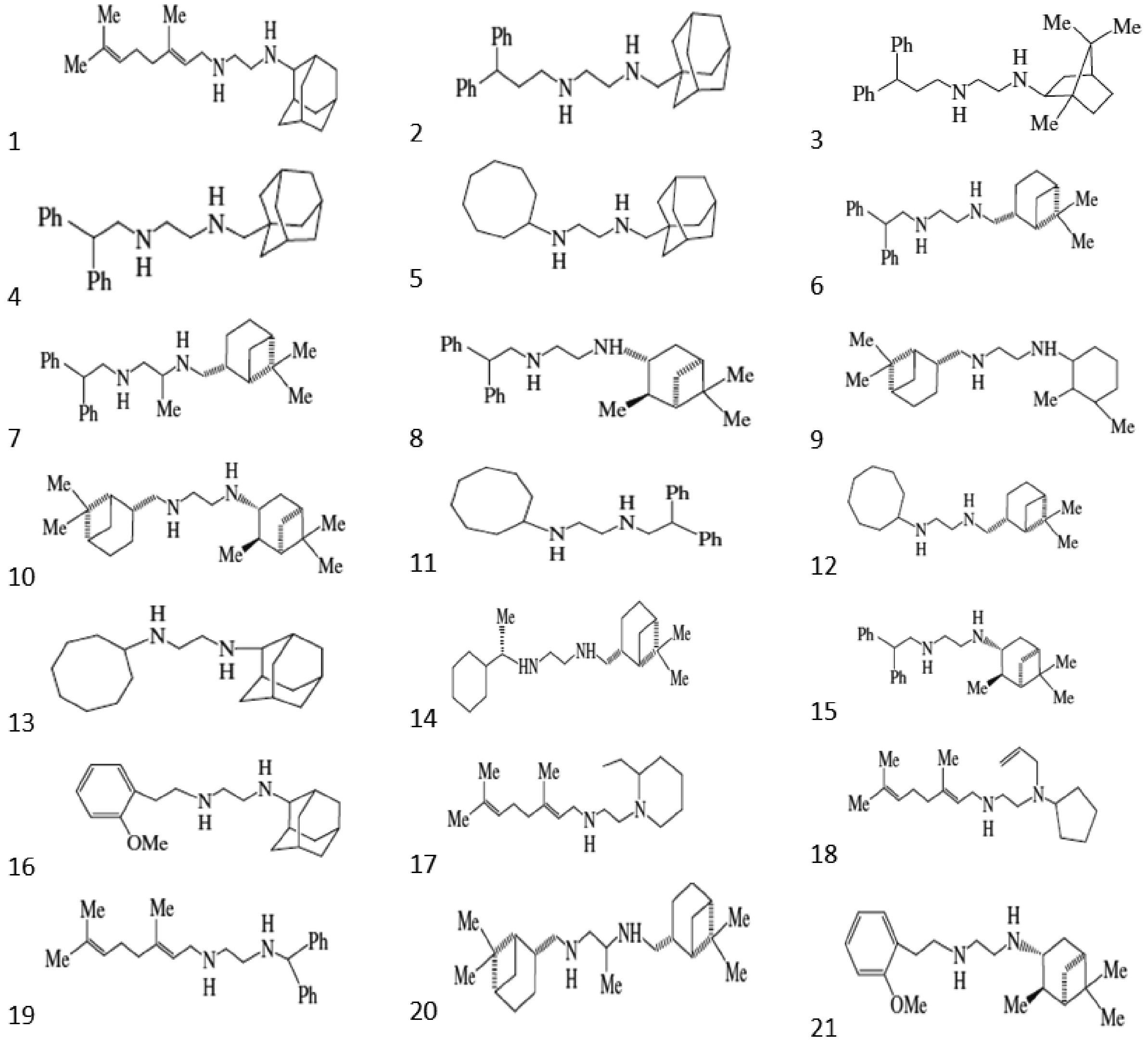

Fig. 1 The molecular structure of ethylenediamine analogues.

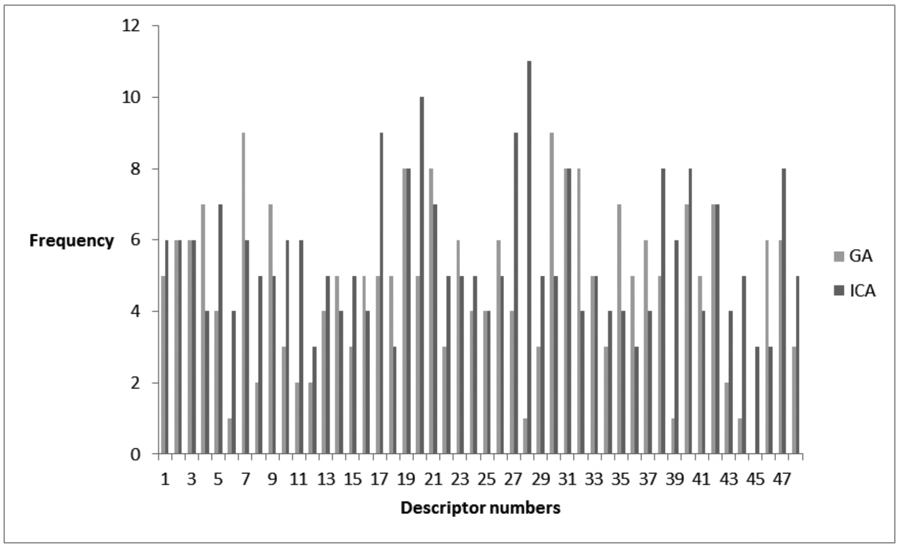

Fig. 2. The results of GA and ICA.

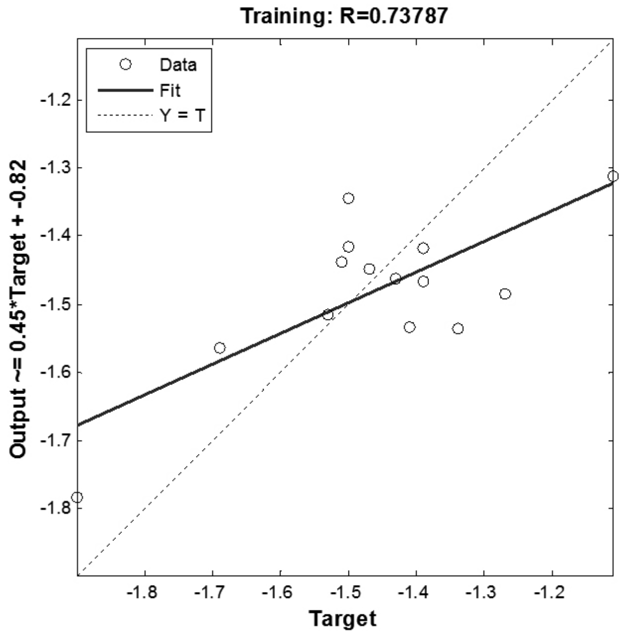

Fig. 3. Plot of output versus target data using ICA method.

RESULTS AND DISCUSSION

In this work, QSAR between oral bioavailabilities of some drugs and their molecular structural descriptors were investigated by using linear and nonlinear techniques. After calculations of descriptors,the different methods were performed on the remaining descriptors to select the most important of them.

Statistical parameters of the different QSAR models are shown in Table 2. It can be seen that the RMSE values for the ANN and ICA method are better than those for the other methods. The descriptors selected using the methods described above were used to construct linear and nonlinear models using GA, ICA, ANN, MLR, PLS and PCR.

The efficiency of the QSAR model for predicting log (IC_{50}) was estimated using the internal cross-validation method which resulted a prediction for log (1/IC50) using MLR, PLS and PCR (Table 1). Considering experimental error, the overall prediction for log (1/IC50) was satisfactory.

Table1 shows the linear variable selection methodsused to select the most significant descriptors.The most significant descriptors selected are G3u, HATS2e, F02(C-N), GGI10, RDF040m, Mor22p, Mor05p, TIC4, H4e, H-052, G2m and G1e (Table 4).

Atomic masses, symmetry directional WHIM index, electronegativities, frequency of C-N at topological distance, radial distribution function, atomic van der Waals volumes, atomic polarizabilities and symmetry were important descriptors in this study.

Atom polarizabilities are linearly correlated with their hardness. The atomic properties considered are partial charges, electron densities and polarizabilities, calculated by computational chemistrymethods; moreover, bond properties have been proposed as the difference between the property values of the atoms forming the bond. GG_k is The topological charge index .

The radial distribution function (RDF) descriptors are based on the distance distribution in the molecule. The radial distribution function of an ensemble of *n* atoms can be interpreted as the probability distribution of finding an atom in a spherical volume of radius *R*.

3D MoRSE descriptors (3D Molecule Representation of Structures based on Electron diffraction) are derived from Infrared spectra simulation using a generalized scattering function.

WHIM descriptors are based on the statistical indices calculated on the projections of atoms along principal axes .They are built in such a way as to capture relevant molecular 3D information regarding the molecular size, shape, symmetry and atom distribution with respect to invariant reference frames. The algorithm consists of performing a Principal Components Analysis on the centered Cartesian coordinates of a molecule by using a weighted covariance matrix obtained from different weighing schemes for the atoms.

HATS descriptors are computed on a Hydrogen-filled molecule. We construct the Molecular Influence matrix *H* as follows. Let *M* be the geometric distance matrix having *n* rows and 3 columns, where we have one row for each of the *n* atoms present in the molecule and one column for each of the x-, y-, z-coordinates of the atoms in the molecule. The atomic coordinates are assumed to be calculated with respect to the geometric center of the molecule.

Bond character is closely related to the capacity of bonded atoms to exchange electrons, and such capacity is commonly well represented by the electronegativity x of the bonded atoms.

The molecular weight can be viewed as a simple linear atom contribution model, where the group contributions are atomic masses. In the first case, large training sets are used to obtain reliable estimates of the group contributions. Usually a battery of group contributions (a field of scalar parameters) is defined taking into account several structural characteristics of the molecules, also sometimes adding extra terms (correction factors) referring to special substructures.

R^2 = 0.73 for the stepwise-MLR model. The result of the ICA was Output = 0.28 Target + 1.1 with a training R equal to 0.74(Figure3).

For MLR, we found that: Out = 0.841-4.807 G3u -8.655 G1e

For MLR-ICA, we had: Out: = 0.386+0.39Le3-0.496 Mor32m

For MLR-GA, we found that: Out= -1.115-10.104G1e+7.301G2v

As can be seen in this table, there is correlation between selected molecular descriptors [Table5]. We have correlation:

HATS2e with Mor05p
Mor32m with Mor05p
G1e with Mor32m
F02(C-N) with H052
RDF040m with TIC4
F02(C-N) with Mor05p

In studies with different methods and different goals, compounds 1, 4, 11, 15 and 17 among 21 studied compounds have the lowest deviation and are suggested as the best compounds to make anti-TB drugs.

Table 1.Experimental and predicted values of log ($1/IC_{50}$) using Jack – Knife, PLS PCR model.

Sample	Observed Log($1/IC_{50}$)	Calculated (Jack-knife)	Calculated GA-PCR	Calculated GA-PLS
1	-1.4100	-1.3947	-1.592	-1.570
2	-1.1100	-1.5221	-1.349	-1.316
3	-1.4700	-1.4271	-1.412	-1.419
4	-1.5000	-1.4162	-1.392	-1.347
5	-1.5300	-1.4322	-1.512	-1.507
6	-1.5000	-1.3818	-1.372	-1.353
7	-1.5100	-1.4358	-1.312	-1.333
8	-1.3900	-1.4375	-1.276	-1.277
9	-1.2700	-1.4458	-1.473	-1.497
10	-1.3900	-1.4400	-1.458	-1.462
11	-1.4300	-1.4475	-1.540	-1.540
12	-1.3400	-1.4657	-1.541	-1.550
13	-1.9000	-1.5433	-1.680	-1.710
14	-1.6900	-1.3898	-1.512	-1.524
15	-1.4400	-1.4657	-1.480	-1.484
16	-1.9000	-1.5183	-1.569	-1.572
17	-1.6500	-1.4101	-1.615	-1.637
18	-1.7000	-1.4023	-1.599	-1.610
19	-1.3200	-1.4303	-1.527	-1.479
20	-1.3400	-1.4413	-1.403	-1.416
21	-1.3200	-1.4308	-1.495	-1.505

Table 2. The statistical parameters of different constructed QSAR models.

Methods	RMSE $_1$	RMSE $_2$	R^2	P value	Fratio	Standard error of predicted value	Q^2	R^{2}pre
Jack-knife	0.1267	0.1395	-	-	-	-	0.94	
ICA	-	0.1728	-	-	-	-		
LASSO	-	-	0.93	-	-	0.027		
GA-MLR	-	-	0.73	-	-	-		
GA-PLS	0.0018	0.0235	0.81	-	-	-	0.52	0.68
GA	0.1350	0.1748	-	-	-	-		
RS	-	-	-	0.04	1.7	-		

Table 3. Descriptors values for GA.

Molecule	TIC4	Mor22p	H4e	H-052	G2m	G1e	F02[C-N]
1	294.896	0.435	2.800	2	0.165	0.151	5
2	311.908	0.229	2.753	2	0.186	0.154	4
3	332.163	-0.185	2.907	4	0.155	0.149	5
4	291.415	0.662	2.671	1	0.156	0.142	4
5	268.226	0.471	3.289	4	0.152	0.159	5
6	360.725	-0.009	3.118	2	0.152	0.139	4
7	384.897	0.045	3.198	5	0.147	0.144	5
8	363.970	-0.101	2.903	4	0.152	0.146	5
9	358.928	0.106	3.247	4	0.154	0.154	5
10	312.226	0.226	3.114	4	0.181	0.163	5
11	272.904	-0.057	2.411	5	0.190	0.153	5
12	287.394	0.141	2.920	5	0.160	0.160	5
13	252.253	0.557	2.613	6	0.154	0.169	6
14	298.149	0.143	2.506	5	0.178	0.167	5
15	284.265	-0.028	2.819	5	0.177	0.159	5
16	270.457	0.434	2.387	4	0.147	0.163	5
17	281.455	0.291	2.444	6	0.174	0.159	6
18	303.510	0.218	2.494	4	0.148	0.166	6
19	311.735	-0.070	2.181	0	0.150	0.143	5
20	302.654	0.288	3.316	5	0.167	0.155	5
21	300.743	-0.020	2.383	5	0.177	0.146	5

Table 4. The mean of selected Descriptors.

Descriptor Symbol	Descriptor group	Meaning	Method
MATS6m	2D autocorrelations	Morgan autocorrelation-lag6/weighted by atomic masses	LASSO
G3u	WHIM	3st component symmetry directional WHIM index	LASSO
HATS2e	GETAWAY	Leverage-weighted autocorrelation of lag 2/weighted by atomic Sanderson electronegativities	LASSO
F02(C-N)	2D frequency finger prints	frequency of C-N at topological distance 02	LASSO
GGI10	Topological	Topological charge index of order 10	ICA
RDF040m	RDF	Radial Distribution function -4/Weighted by atomic masses	ICA
Mor22p	3D-MoRSE	3D-MoRSE-signal22/ Weighted by atomic masses	ICA,GA-ANN
Mor05p	3D-MoRSE	3D-MoRSE-signal05/ Weighted by atomic polarisabilities	ICA
TIC4	Information indices(2D)	Total information content index (neighborhood symmetry of 4-order)	GA-ANN
H4e	GETAWAY	H autocorrelation of lag 4 / weighted by atomic Sanderson electronegativities	GA-ANN
H-052	Atom-centered fragments (1D)	H attached to C0(sp3) with 1X attached to next C	GA-ANN
G2m	WHIM	2st component symmetry directional WHIM index / weighted by atomic masses	GA-ANN
G1e	WHIM	1st component symmetry directional WHIM index/weighted by atomic Sanderson electronegativities	GA-MLR
G2v	WHIM	2st component symmetry directional WHIM index/weighted by atomic van der waals valumes	GA-MLR
G3u	WHIM	3st component symmetry directional WHIM index/Unweighted	MLR
Le3	WHIM	3rd component symmetry size directional WHIM index/ weighted by atomic Sanderson electronegativities	MLR-ICA
Mor32m	3D-MoRSE	3D-MoRSE signal 32/ weighted by atomic masses	MLR-ICA

Table 5. Correlation matrix between selected descriptors.

	TIC 4	MATS 6m	GG I10	RDF 040m	Mor 32m	Mor 05p	Mor 22p	G 3u	G 2m	G 2v	Le 3	G1e	H 4e	HATS 2e	H0 52	F02 (C_N)
TIC4	1															
MATS6m	0.19	1														
GGI10	0.33	0.07	1													
RDF040m	0.65	-0.01	0.30	1												
Mor32m	-0.47	-0. 9	-0.2	-0.37	1											
Mor05p	-0.65	-0.22	-0.37	-0.65	0.56	1										
Mor22p	-0.53	-0.35	-0.26	-0.11	0.25	0.09	1									
G3u	-0.25	-0.19	-0.49	-0.33	0.20	0.21	0.45	1								
G2m	-0.32	0.18	0.10	-0.08	0.09	0.16	-0.11	-0.31	1							
G2v	0.01	0.05	0.25	0.14	-0.18	-0.05	-0.4	-0.37	0.63	1						
Le3	0.21	0.08	0.06	0.06	-0.67	-0.13	-0.29	-0.04	0.03	0.1	1					
G1e	-0.52	-0.25	-0.07	-0.51	0.67	0.59	0.36	0.25	0.24	0.07	-0.48	1				
H4e	0.35	-0.08	0.20	0.32	0.06	-0.37	0.06	0.1	-0.06	0.06	0.23	0.04	1			
HATS2e	-0.83	-0.42	-0.37	-0.57	0.25	0.69	0.50	0.28	0.24	-0.18	-0.11	0.5	-0.3	1		
H052	-0.38	0.06	-0.06	-0.29	0.15	0.49	0.03	-0.10	0.29	0.27	-0.09	0.44	-0.17	0.37	1	
F02(C_N)	-0.37	-0.11	-0.21	-0.27	0.61	0.63	0.12	0.13	-0.02	0.12	-0.43	0.61	-0.16	0.19	0.61	1

CONCLUSION

In the present study, MLR, PLS ,GA, ANN, ICA and ANN were used as linear and nonlinear models to their calculated molecular descriptors. The calculated statistical parameters of these models revealed that ANN was better than others which means that there are some linear and nonlinear relations between selected molecular descriptors and their structures. ANN and ICA were successfully used to develop a QSAR model for ethylene diamine derivatives that provided the best results in comparison with other methods. This attempt to correlate log (1/IC$_{50}$) with theoretically calculated molecular descriptors led to a relatively successful QSAR model that relates these derivatives.

CONFILICT OF INTEREST: The authors have no conflict of interest.

REFERENCES

[1] F. J. DuMelle, P. C. Hopewell, *TB Notes News let*. 1 , 23–27. (2000)

[2] J. B . Jr.Bass , L. S . Farer , P. C. Hopewell , *Am. J. Respir. Crit Care Med* . 149, 1359–1374. (1994)

[3] C. R. Jr.Horsburgh , S. Feldman, R. Ridzon, *Clin. Infect Dis* . 31, 633–639. (2000)

[4] P. A. Gross, T. L. Barrett , E. P. Dellinger, P. J. Krause , W. J. Martone, J. E. Jr. McGowan , R. L. Sweet, R. P. Wenzel, *Clin. Infect Dis* . 18, 421. (1994)

[5] R. Hosseini, H. Salehipoor, *Int. J. Struct. Stab. Dy* . 12, No. 3, 1250019 . (2012)

[6] E. Atashpaz-Gargari , C. Lucas. Imperialist competitive algorithm: an algorithm for optimization inspired by imperialistic competition. Evolutionary Computation, CEC 2007. IEEE Congress on , 2007; pp. 4661-4667.

[7] E. Atashpaz-Gargari , C. Lucas. Designing an optimal PID controller using Colonial Competitive Algorithm. First Iranian Joint Congress on Intelligent and Fuzzy Systems, 2007.

[8] C. Sarbu , C. Onisor, M. Poša, *Talanta* .75, 651–657. (2008)

[9] M. P. Freitas, J. A. Martins , *Talanta.* 67, 182–186. (2005)

[10] K. Valko, *J. Chromatog*. A. 1037, 299–310. (2004)

[11] K. Tang, T. Li, *Anal. Chem. Acta*. 476 , 75–92. (2003)

[12] H. Gonzalez-Diaz, I. Bonet, C. E. Teran, *Eur. J. Med.Chem*. 42, 580–585. (2007)

[13] S. Vilar, L. Santana , E. Uriarte, *J. Med. Chem*. 49, 1118–1124. (2006)

[14] C. Tang, P. A. Almeida Fishwick Times series forecasting using neural networks vs. Box–Jenkins methodologySimulations, Simulations Councils, 1991; pp. 303–310.

[15] Marina. Protopopova, Colleen. Hanrahan, Boris Nikonenko, *J. Antimicrob. Chemother* . 56, 968–974. (2005)

[16] R. Todeschini, V. Consonni , Handbook of Molecular Descriptors; WILEY-VCH, Verlag GmbH. Vol. 11 , 2000; p 516.