ICC

Iranian Chemical Communication

Original Research Article

Payame Noor University http://icc.journals.pnu.ac.ir

Prediction of IC50 of 2,5-diaminobenzophenone organic derivatives antimalarial compounds using informatics-aided genetic algorithm

Rashid Heidarimoghadam[a], Seyede Shima Mortazavi[b], Abbas Farmany[c,*]

[a]Department of Ergonomics, Health Sciences Research Center, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran

[b]Young Researchers & Elite Club, Hamedan Branch, Islamic Azad University, Hamedan, Iran

[c]Dental Research Center, School of Dentistry, Hamadan University of Medical Sciences, Hamadan, Iran

Received: 2 July 2018, Accepted: 20 September 2018, Published: 1 October 2018

Abstract

In the present paper, informatics-aided quantitative structure activity relationship (QSAR) models using genetic algorithm-partial least square (GA-PLS), genetic algorithm-Kernel partial least square (KPLS), and Levenberg-Marquardt artificial neural network (LM ANN) approach were constructed to access the antimalarial activity (pIC_{50}) of 2,5-diaminobenzophenone derivatives. Comparison of errors and correlation coefficients was obtained by the models as it illustrated that the LM ANN approach works with a high correlation coefficient and low prediction error. This model was applied to the prediction of pIC_{50} values of 2,5-diaminobenzophenone derivatives.

Keywords: *P. falciparum* malaria; antimalarial compounds; 2,5-diaminobenzophenones; QSAR.

Introduction

As a parasitic (sporozite) disease, malaria with mortality rate of 25% is a major worldwide health problem [1]. In 2011, 216 million cases of this disease were reported [2,3]. A parasite which is passed through a human to another via mosquitoes *Anopheles* infection causes malaria. Also, mosquitoes in climates with specific temperatures can carry malaria [2]. Recently, the antimalarial activity of 2,5-diaminobenzophenone based compounds with the farnesyl transferase inhibition effects was reported [4]. Due to their activity against multi-drug resistant, the experimental and theoretical aspects of this particular compounds need to be investigated. So, developing QSAR models with informatics-aided drug-design has a key role in the understanding of the effectiveness and mechanism evaluation of new drug compounds [5-8]. This paper aimed to develop a QSAR model for 2,5-diaminobenzophenone derivative. Following the variables selection, the linear and nonlinear regressions (e.g. PLS and KPLS) and a neural network

*Corresponding author: Abbas Farmany
Tel: +98 (918) 1432750, Fax: +98 (81) 38381939
E-mail: a.farmany@ut.ac.ir

(L-M ANN) were employed to construct the QSAR models.

Experimental

Materials and Methods

Data set

Typically, IC_{50} is used to calculate the antagonist drug efficiency. It exhibits upon what amount of a careful substance/molecule is alluring on restrain 50% biological progression. Also, its quantitative measure show what amount of a specific material (drug *etc*...) is required for a biological process hindering. In this study, to indicate a greater potency, IC_{50} was converted to pIC_{50} scale (-log IC_{50}).

Descriptor generation

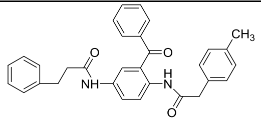
The data set used in this work is pIC_{50} of 92 derivatives of 2,5-diaminobenzophenones molecules obtained from the literature [9] which is shown in Table 1.

HyperChem (Version 7.0 Hypercube, Inc) was used to draw the molecular structures. Streamline AM1 method was used to optimize the structures. To calculate the molecular descriptor the Dragon 2.1 software was used.

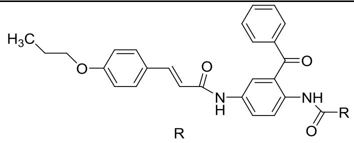
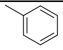
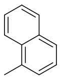
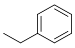
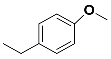
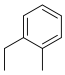
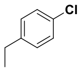
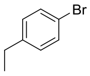
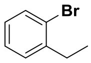
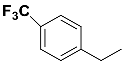
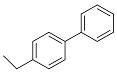
Data pretreatment

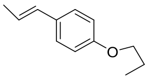
All of the constant variables were removed by analyzing the calculated descriptors. To choose the collinear descriptors (r > 0.9), the existence of redundancy in the data matrix was checked. A set of collinear descriptors with a highest correlation was retained and the other descriptors were deleted. The descriptors were set in an n × m data matrix (D), where n = 92 and m=1042. Note that n and m are the number of the compounds and the descriptors, respectively.

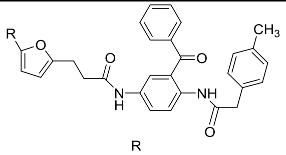
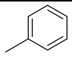
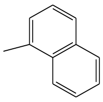
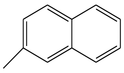
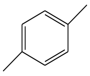
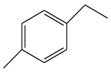
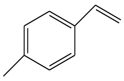
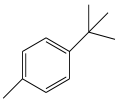
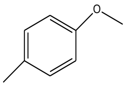
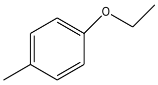
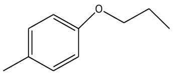
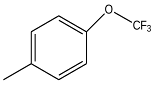
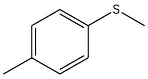
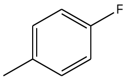
Table 1. The data set, structure and the corresponding observed pIC_{50} values

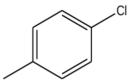
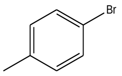
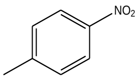
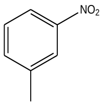
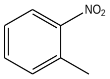
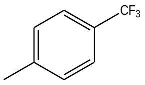
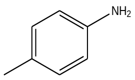
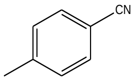
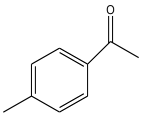
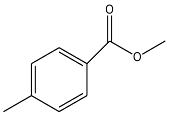
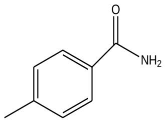
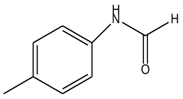
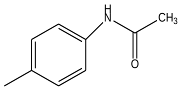
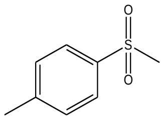
Compound	**Structure**	**PIC_{50}**
Compound		PIC_{50}
1	R	5.57
2	-H	5.24
3p	$-NO_2$	5.19
4v	-CHO	4.4
5	$-COOCH_3$	6.00
6	$-CF_3$	5.24
7p	-Cl	5.26
8v	-Br	5.49
9	$-NH_2$	5.26

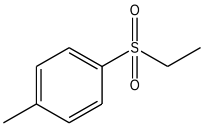
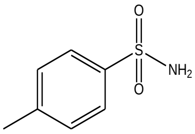
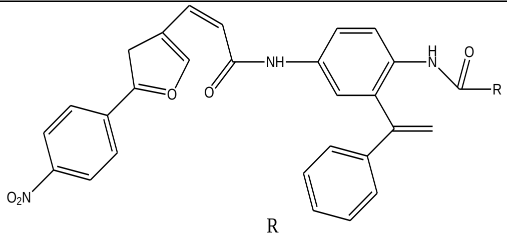
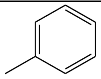
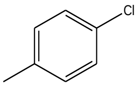
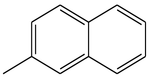
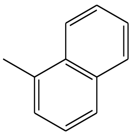
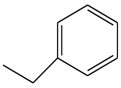
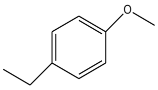
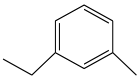
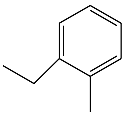
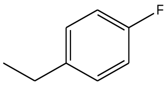
10	-CH$=$C(CN)$_2$	4.37
11	-CH$_3$	5.85
12	-O-CH$_3$	5.89
13p	-CH$_2$-CH$_3$	5.92
14v	-CH(CH$_3$)$_2$	5.92
15	-C(CH$_3$)$_3$	5.52
16	-O-CH$_2$-CH$_3$	6.07
17	-O-(CH$_2$)$_3$-CH$_3$	5.96
18p	-O-(CH$_2$)$_2$-CH$_3$	6.47

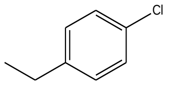
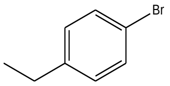
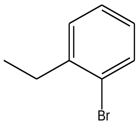
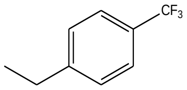
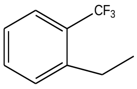
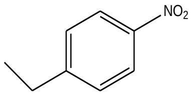
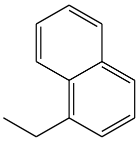
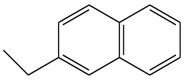
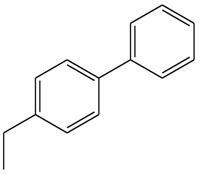
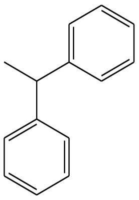
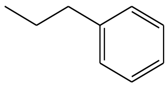
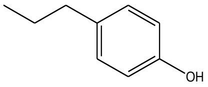
Compound	 R	PIC$_{50}$
19v		5.05
20		5.47
21		5.6
22		5.89
23v		5.62
24p		6.46
25		6.51
26		6.00
27		6.92
28v		4.62

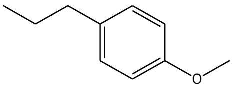
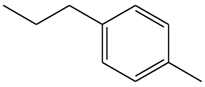
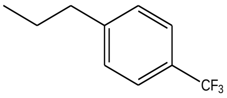
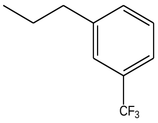
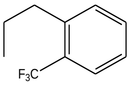
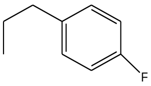
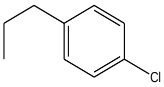
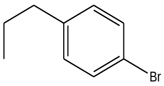
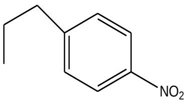
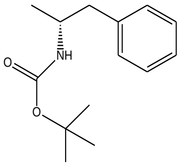
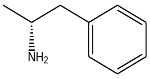
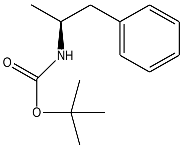
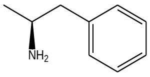
29p		4.64

Compound		PIC$_{50}$
30		6.38
31		6.00
32v		6.70
33		6.92
34		7.06
35		7.07
36p		6.52
37		6.89
38v		6.52
39		6.12
40		6.68
41p		7.08
42		6.49

43		6.84
44v		6.9
45		7.12
46p		6.23
47		6.17
48		7.11
49v		6.59
50		6.66
51p		7.17
52		6.77
53v		6.25
54		6.55
55		6.25
56p		7.43

57		7.22
58v		6.7
Compound	 R	PIC_{50}
59		5.52
60p		5.6
61		5.6
62		6.11
63v		6.57
64		6.49
65p		6.82
66		6.19
67v		6.64

68		7.19
69p		7.15
70v		6.00
71		7.33
72		6.00
73v		5.85
74p		6.60
75		6.68
76		6.00
77		5.26
78v		6.51
79		5.48

80p		5.89
81		6.36
82		7.21
83v		6.2
84		5.96
85		6.36
86p		6.89
87v		6.77
88		5.85
89		6.05
90p		5.92
91v		6.38
92		6.24

V: Validation set
P: Prediction set

Descriptor selection by genetic algorithm

Generally, in a chromosome, the absence or presence of a descriptor is coded as 0 or 1. Each string contains 561 genes which represent the status of descriptors (presence or absence). For each GA runs, the population size was changed. In a typical GA run, the generation evaluating was stopped, when more than 80% of generations had the same fitness [10, 11]. The population size of this study was 30 chromosomes.

Nonlinear model

Artificial neural network

To investigate the feature sets, a three-layer artificial neural network ANN with a back propagation sigmoid transfer function was employed. The model generation was made using the training set descriptors. In this sense, the validation sets were used for the network overtraining cut-off. The model predictivity was verified using validation set descriptors [12-14].

Result and Discussion

Linear model

Results of GA-PLS model

Based on the highest square correlation coefficient (R^2), the least root mean squares error (RMSE) and relative error (RE), the best mode is obtained. The GA-PLS model which was constructed in this study is based on the 21 descriptors in 9 latent variables. In this model, for training and validation descriptors R^2, RE and RMSE were (0.837, 0.741), (5.28, 7.39) and (0.041, 0.096), respectively. The predicted pIC_{50} values for training and set descriptors are presented in Figure 1a. According to Figure 1a, in the PLS model, the number of latent variables is not more than the independent variables. This allows the model to extract more structural information from descriptors which minimize the prediction error.

Nonlinear model

Results of GA-KPLS model

In this study a radial basis kernel function was used for nonlinear model construction,

$$K(x, y) = \exp(\| x - y \|^2 / c)$$

(1)

where $c = rm\sigma^2$ [15]. Using GA-KPLS selection method, 13 descriptors in 5 latent variables space were chosen. As a result, (0.872, 0.785), (4.79, 6.52) and (0.038, 0.084) were obtained as R^2, RE and RMSE for training and test sets, respectively. The related GA-KPLS of the predicted and experimental values of pIC_{50} were shown in Figure 1b. The results of GA-KPLS model are superior to GA-PLS. It is interesting that higher R^2 and lower RMSE and RE were obtained instead of linear model.

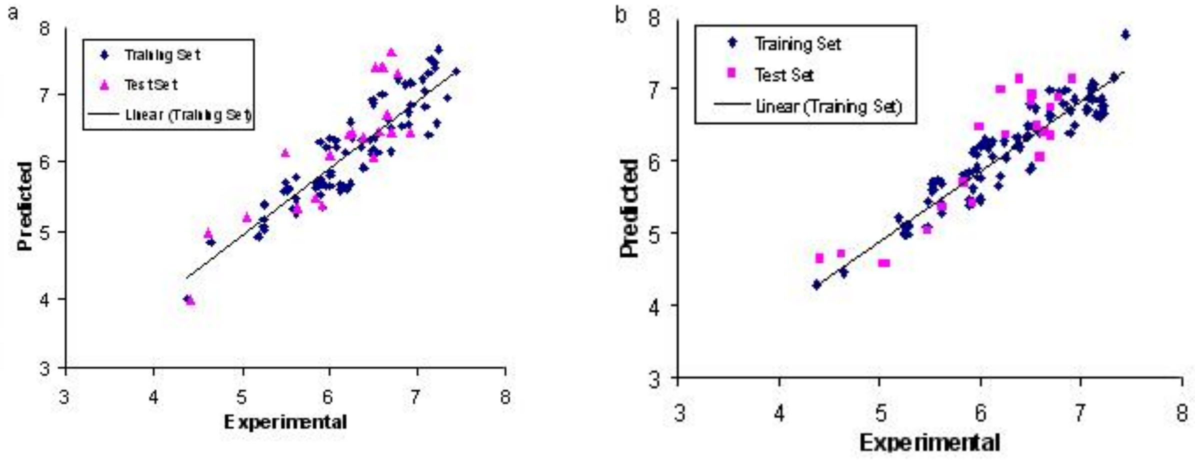


Figure 1. Plots of predicted pIC_{50} against the experimental values by (a) GA-PLS model and (b) GA-KPLS model

Results of LM ANN model

The ANN model was generated using three groups of descriptors: calibration, validation and prediction. The number of neurons in the hidden layer, learning rate and momentum were optimized. The retention relationship was obtained using a back-propagation feed-forward neural network [16]. In this algorithm, to obtain the minimum error function, training process diminishes the network outputs and the expected values difference [17]. In this work, we used a network with nine input layer, four hidden layer and one output layer. A bias unit with constant activation was connected to units in the hidden and output layers. The calibration RMSE was used to evaluate the performance of ANN algorithm. Optimum numbers were the number of neurons in the hidden layer with the minimum RMSE. Similar way was used to optimize the learning rate and momentum. R^2 and RE values of calibration, prediction and validation were (0.952, 0.930, 0.894) and (3.61, 4.27, 5.69), respectively. RMSE of calibration, prediction and test sets were obtained as (0.029, 0.050, 0.065), respectively. In this study, as compared to the counterparts in other models, higher R^2 and lower RMSE and RE were obtained for validation set (Figure 2a,b). Figures 3a,b shows the residuals ($pIC_{50}^{predicted} – pIC_{50}^{experimental}$) instead of $pIC_{50}^{experimental}$ which is obtained by L-M ANN algorithm. Distribution of residual on the both side of zero line shows that the neural network algorithm works without a systematic error. Comparison between R^2, RE and RMSE values of the developed algorithms shows the superiority of the L-M ANN model. Unlike the regression analysis, the neural network with a key strength has the potency of flexible mapping by manipulating implicitly. The results of this study shows the reproducibility of L-M ANN to pIC_{50} prediction of 2,5-diaminobenzophenones derivatives.

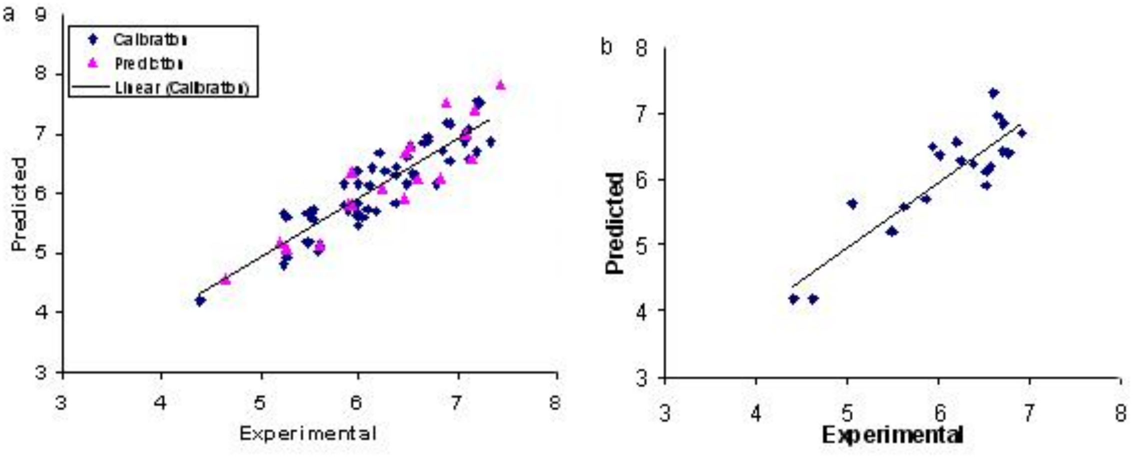


Figure 2. Plot of predicted pIC_{50} obtained by L-M ANN against the experimental values (a) calibration and prediction sets of molecules and (b) for test (validation) set

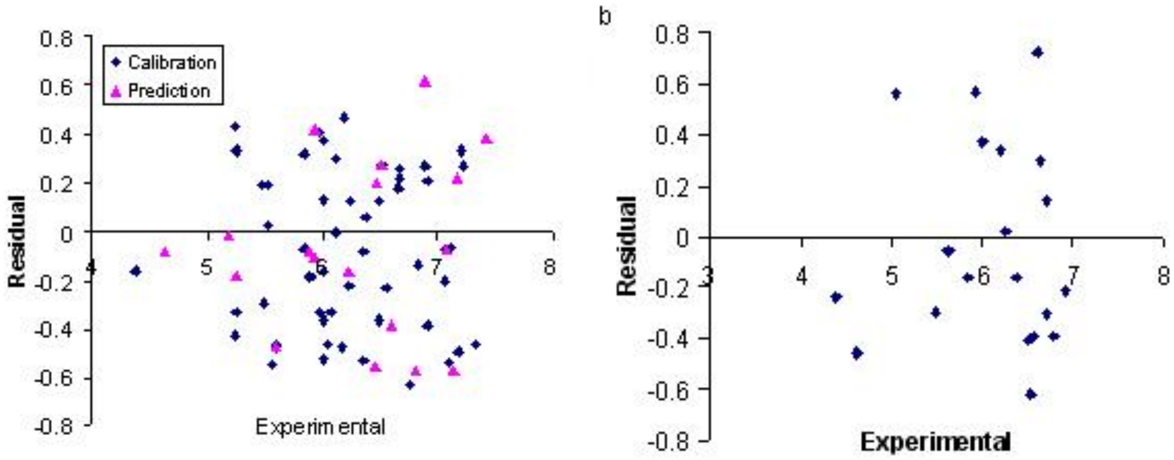


Figure 3. Plot of residuals obtained by L-M ANN against the experimental pIC_{50} values (a) training set of molecules and (b) for test set

Model validation and statistical parameters

To evaluate the predictivity of the developed algorithms, both the internal (LGO-CV)) and external (validation set) were used. For LGO-CV, a compound was removed and after training with the remained compounds, the discarded compound was predicted. This process was repeated for all compounds. The data set descriptors were distributed/divided into three sub-data sets as calibration, prediction and test sets. The calibration and prediction sets were used for model generation and overfitting the network and validation set was used to evaluate the predictive power of external set [18-20].

In this work we used 54 components in calibration set; 18 components in prediction set and 20 components in validation sets. So the predictive power of this study was measured as its ability to predict the partition of unknown derivatives of 2,5-diaminobenzophenones. To evaluate the model predictive ability, each predicted pIC_{50} value was compared with the experimental acidity constant [21-26].

Conclusion

As a conclusion, the GA-PLS, GA-KPLS and L-M ANN algorithms were used to predict the antimalarial activities of 2,5-diaminobenzophenones derivatives. Low errors and high correlation coefficients indicated proper predictability of L-M ANN model. Applying the extended model to a dataset of 20 compounds demonstrates the reliability and accuracy of the model. Comparing three models revealed the superiority of the L-M ANN to predict the pIC_{50} of 2,5-diaminobenzophenones derivatives.

Acknowledgments

We would like to thank Ehsan Hashemi for discussions.

References

[1] V. Moorthy, Z. Reed, P. G. Smith, *Vaccine,* **2007***, 25,* 5115-5123.

[2] S. Mharakurwa, Ph.E. Thuma, D.E. Norris, M. Mulenga, V. Chalwe, J. Chipeta, Sh. Munyati, S. Mutambu, P.R. Mason, *Acta Trop*., **2012**, *13*, 531-547.

[3] R.I. Chima, C.A. Goodman, A. Mills, *Health Pol*., **2003**, *63*, 17-36.

[4] L. Hviid, *Microbes and Infec.*, **2007**, *9*, 772-776.

[5] M. Nekoei, *Iran. Chem. Commun.*, **2017**, *5*, 79-98.

[6] M. Bouachrin, Y. Bouakarai, F. Khalil, *Chem. Method.*, **2017**, *1*, 173-193.

[7] (a) M. Shahpar, S. Esmaeilpoor, *Asian J. Green Chem.*, **2017**, *1*, 116-129. DOI: 10.22631/ajgc.**2017**.94413.1010 (b) M. Shahpar, S. Esmaeilpoor, *Chem. Method*., **2017**, 1(2), 98-120. DOI: 10.22631/chemm.**2017**.96397.1008.

[8] H. Noorizadeh, S. Esmaeilpoor, Z. Moghadam, S. Nosratolahy, *Iran. Chem. Commun.*, **2014**, *2,* 283-299.

[9] R.A. Cormanich, M.P. Freitas, R. Rittner, *J. Braz. Chem. Soc.*, **2011**, *22*, 37-642.

[10] H. Noorizadeh, A. Farmany, M. Noorizadeh, *Quim. Nova*, **2011**, *34*, 1398-1404.

[11] E. Pourbasheer, S. Riahi, MR. Ganjali, P. Norouzi, *Eur. J. Med. Chem*., **2009**, *44*, 5023–5028.

[12] N. Singh, A. Basant, S. Malik, G.J. Sinha, *Intell. Lab. Syst.*, **2009**, *99*, 150-160.

[13] B. Jančić-Stojanović, D. Ivanović, A. Malenović, M. Medenica, *Talanta*, **2009**, *78*, 107–112.

[14] M. Jalali-Heravi, M. Asadollahi-Baboli, P. Shahbazikhah, *Eur. J. Med. Chem.*, **2008**, *43*, 548-556.

[15] H. Noorizadeh, A. Farmany, *Drug. Test. Anal*., **2012**, *4*, 151-157.

[16] A.A. D'Archivio, M.A. Maggi, P. Mazzeo, F. Ruggieri, *Anal. Chim. Acta.*, **2008**, *628*, 162 – 172.

[17] B. Jančić, M. Medenica, D. Ivanović, S. Janković, A. Malenović, *J Chromatogr A*, **2008**, *1189*, 366-373.

[18] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, **2000**. Wiley/VCH, Weinheim.

[19] (a) O. Deeb, *Chemom. Intell. Lab. Syst.*, **2010**, 104, 181-194. (b) S. Sajjadifar, *Chem. Method.*, 1(1), 1-11. DOI: 10.22034/chemm.**2017**.49740.

[20] D. Pran Kishore, C. Balakumar, A. Raghuram Rao, P. P. Roy, K. Roy, *Bioorg. Med. Chem. Lett.*, **2011**, *21*, 818-823.

[21] M. Arab Chamjangali, M. Beglari, G. Bagherian, *J. Mol. Graphics. Modell.*, **2007**, *26*, 360-367.

[22] H. Noorizadeh, *A.* Farmany, *J. Chinese Chem. Soc.*, **2010**, *57*, 1268-1277.

[23] H. Noorizadeh, A. Farmany, A. Khosravi, *J. Chin. Chem. Soc*, **2010**, *57*, 982-991.

[24] A.K. Zhokhov, A.Y. Loskutov, I.V. Rybal’chenko, *J. Anal. Chem.*, **2018**, *73*, 207-220.

[25] N. Fan, S. Zhang, T. Sheng, L. Zhao, Z. Liu, J. Liu, X. Wang, _*Chem. Bio. Drug Des*., **2018**, 398-407.

[26] U. Judycka, K. Jagiello, M. Gromelski, L. Bober, J. Błażejowski, T. Puzyn, _*J. Chromatogr. B*, **2018**, *1095*, 8-14.

How to cite this manuscript: Rashid Heidarimoghadam, Seyede Shima Mortazavi, Abbas Farmany. "Prediction of IC50 of 2,5-diaminobenzophenone organic derivatives antimalarial compounds using informatics-aided genetic algorithm". *Iranian Chemical Communication*, 2019, 7 (1), 39-51.