Regular Article

ORGANIC CHEMISTRY RESEARCH
Published by the
Iranian Chemical Society
www.orgchemres.org

Org. Chem. Res., Vol. 7, No. 2, 190-199, July 2021.
DOI: 10.22036/org.chem.2023.327725.1274

Application of Monte Carlo and QSAR Techniques of Several Methotrexate Derivatives as Anticancer Drugs

R. Sayyadikordabadi[a], O. Alizadeh[a], Gh. Ghasemi[a], B. Motahary[b], R. Rajabei Nezhad[c] and K. Akhavan[a]

[a]*Department of Chemistry and Chemical Engineering, Rasht Branch, Islamic Azad University, Rasht, Iran*
[b]*Department of Computer Engineering, Rasht Branch, Islamic Azad University, Rasht, Iran*
[c]*Department of Fisheries, Bandar Anzali Branch, Islamic Azad University, Bandar Anzali, Iran*

(Received 2 February 2022, Accepted 18 August 2023)

The present study demonstrates a quantitative structure-activity relationship (QSAR) between the half-maximal inhibitory concentration (IC_{50}) values of 31 different Methotrexate derivatives, using multiple linear regression (MLR), artificial neural network (ANN), simulated annealing algorithm (SA), and genetic algorithm (GA). Furthermore, CORAL software was used for multiple probability simulation of the studied derivatives. The results obtained from the MLR-MLR, MLR-SA, SA-ANN, MLR-GA, and GA-ANN approaches were compared; GA-ANN combination showed the best performance according to the correlation coefficient (R^2) and root-mean-square error (RMSE). From Monte Carlo simulations, it was found that the presence of double bonds, the presence of nitrogen and oxygen, the absence of sulfur and phosphorus, and connected sp^2 carbon to the ring are the most important molecular features that affect the biological activity of the drug. It was concluded that the simultaneous application of GA-ANN and Monte Carlo methods can provide a more comprehensive understanding of the relationship between a drug's physicochemical, structural, or theoretical molecular descriptors and its biological activity, leading to accelerating the development of new drugs.

Keywords: QSAR, Methotrexate derivatives, Monte Carlo method, Genetic algorithm

INTRODUCTION

Methotrexate, a folate antagonist, has been utilized as a treatment for autoimmune and inflammatory diseases. Purine and thymidylate synthesis require the active form of folic acid that is reduced to tetrahydrofolate by dihydrofolate reductase (DHFR) [1,2]. To regenerate tetrahydrofolate from dihydrofolate [3], this enzyme is essential for intracellular folate metabolism.

Quantitative structure-activity relationship (QSAR) methods use mathematical equations to establish a relationship between chemical structures and biological activities [4]. Several QSAR techniques include multiple linear regression (MLR), simulated annealing algorithm (SA) [5,6], genetic algorithm (GA) [7], and partial least squares (PLS) to be applied in the development of a quantitative relationship between the structural descriptors and the physical or chemical properties [8,9].

CORAL has recently been suggested as a competent software for the expert QSAR studies. By employing the Monte Carlo method, the most significant and simplified molecular input-line entry system (SMILES)-based descriptors are identified and their correlation weights are calculated to predict an endpoint (*e.g.*, -log(IC_{50})). The molecular structure is represented by SMILES, which are lines of symbols [10,11].

In this work, multiple linear regression (MLR) and artificial neural network (ANN) as modeling tools and

*Corresponding author. E-mail: sayyadi_04@yahoo.com simulated annealing (SA) and genetic algorithm (GA) as optimization techniques besides the CORAL software were applied to investigate the QSAR of Methotrexate derivatives. Various QSAR models were utilized to find the best descriptor in inhibitory activity of methotrexate derivatives, and the obtained results were compared.

THEORY AND COMPUTATIONAL METHODS

Linear and Non-linear Methods

The geometry optimization of methotrexate derivatives was carried out by Gaussian 03W [12] at B3LYP/6-31g level of theory. 3226 molecular descriptors, including topological, geometrical, MoRSE [13,14], RDF [14,15], GETAWAY [16], auto-correlations [4], and WHIM [9,17,18] groups, were calculated for each of the 25 compounds using the Dragon program. Subsequently, the SPSS [19] program was used to reduce the number of descriptors by selecting a three-stage objective characteristic. These steps involve i) removing descriptors that have the same value for at least 70% of compounds; ii) descriptors with correlation coefficient less than 0.25 with a logarithm half-maximal inhibitory concentration (-logIC$_{50}$) as a dependent variable were removed [20]; iii) by carrying out these two steps, the number of descriptors were reduced to 858 and then a stepwise MLR procedure was employed to select the appropriate descriptors of these 858 descriptors. Low standard deviation, least numbers of independent variables, high ability of prediction, high F statistic value [21], high correlation coefficient (R^2), and the lowest RMSE are characteristics of an ideal model. The definition of the RMSE is as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - y_o)^2}{n}} \qquad (1)$$

In the Eq. (1), y_i is the desired result, y_o is the predicted value of the model, and n is the number of molecules in the dataset. The method employed is shown in Fig. 1. In these approaches, including GA-ANN, SA-ANN, MLR-SA, MLR-GA, 858 descriptors were considered as possible inputs of the ANN and fed into the input layer of the ANN. All the artificial networks employed were three-layer and the Levenberg-Marquart algorithm [22] was applied for the training of the networks. Modelling and optimization calculations were conducted using MATLAB 2014a. The objective of these networks was to identify the non-linear relationship between the structural descriptors and the inhibitory activity of methotrexate derivatives.

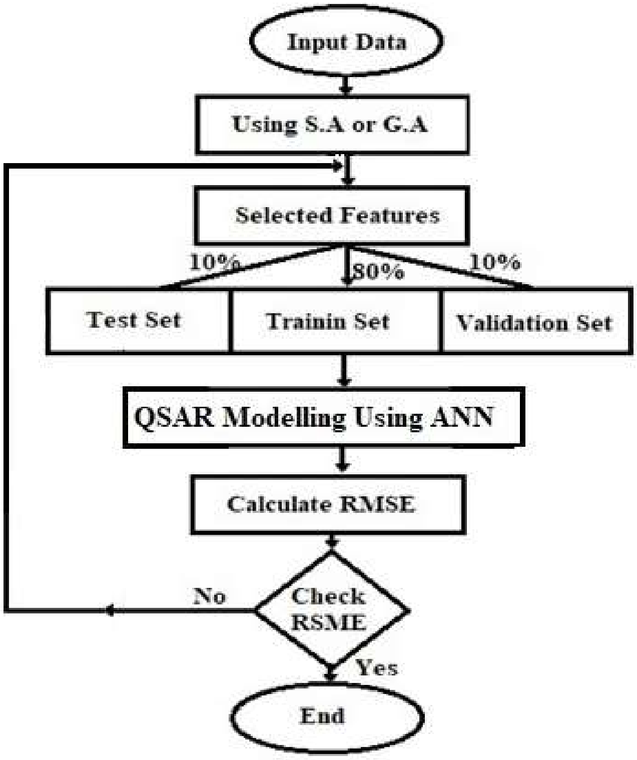


Fig. 1. The employed method for finding optimum descriptors of the ANN methods.

Monte Carlo Method

CORAL [23] software was used for the calculation of descriptor correlation weight (DCW) of the 31 methotrexate derivatives with a hybrid optimization scheme including hydrogen-suppressed molecular graph (HSG) and SMILES representation of molecular structures. An overall number of 300 runs were performed by modelling using CORAL software for thresholds of 1 up to 3 and 100 epochs. Each epoch [24] is a sequence of computations used to find a new set of modified correlation weights of the model. The SMILES-based and Graph-based optimal descriptors are achieved using the following relations:

$$\mathrm{DCW(T,\ Nepoch)}^{\mathrm{SMILES}} = \alpha\sum \mathrm{CW(Sk)} + \beta\sum \mathrm{CW(SSk)} + \gamma\sum \mathrm{CW(SSSk)} + \mathrm{x.CW(NOSP)} + \mathrm{y.CW(HALO)}\ \mathrm{z.CW(BOND)} \qquad (2)$$

$$\mathrm{DCW(T,\ Nepoch)}^{\mathrm{Graph}} = \sum \mathrm{CWAk} + \alpha\sum \mathrm{CW(^{0}ECk)} + \beta\sum \mathrm{CW(^{1}ECk)} + \gamma\sum \mathrm{CW(^{2}ECk)} + \delta\sum \mathrm{CW(^{3}ECk)} \qquad (3)$$

where, S_k, SS_k, and SSS_k are the names for one, two, and three component SMILES attributes. The presence or absence of chemical elements can be demonstrated by NOSP (nitrogen, oxygen, sulfur, and phosphorus) and HALO (fluorine, chlorine, and bromine). In addition, 'BOND' signifies chemical bonds that are either double (=), triple (#), or stereo (@ or @@). A_k in Eq. (3) indicates the occurrence of the C, N, and O atoms in the HSG and HFG molecular graphs. The α, β, γ, and δ coefficients and combinations of their values are used to define various versions of the graph-based optimal descriptor and can be 1 or 0. The hybrid objective function for finding the optimal descriptors is defined as [25]:

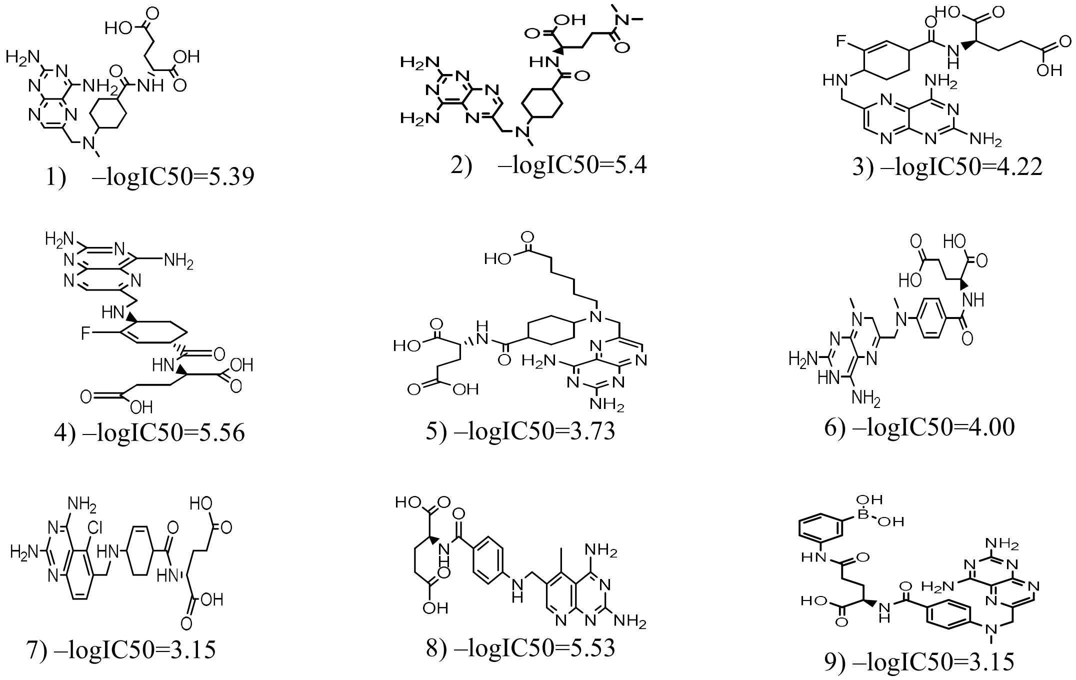
$$\mathrm{DCW(T,\ Nepoch)}^{\mathrm{Hybrid}} = \mathrm{DCW(T,\ Nepoch)}^{\mathrm{SMILES}} + \mathrm{DCW(T,\ Nepoch)}^{\mathrm{Graph}} \qquad (4)$$

RESULTS AND DISCUSSION

Linear Methods

The schematic structures of methotrexate compounds can be found in Fig. 2.

The optimized parameters have been reported in Table 1S in supplementary file. SPSS and Unscrambler programs were employed for linear calculation, including MLR-MLR, MLR-PCR, and MLR-PLS1 methods. The RMSE and the correlation coefficient (R^2) for biological activity in MLR–PCR, MLR–PLS1, and MLR–MLR for the predicted activity were found to be [0.5565 1.1065], [0.5597 1.1025], and [0.7080 0.8989], respectively. Furthermore, the calculated parameters indicate that MLR–MLR is the most effective linear method.



1) –logIC50=5.39

2) –logIC50=5.4

3) –logIC50=4.22

4) –logIC50=5.56

5) –logIC50=3.73

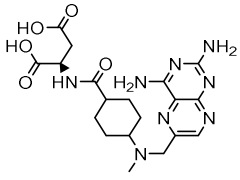
6) –logIC50=4.00

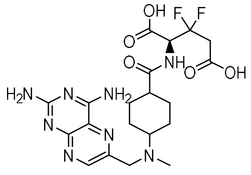
7) –logIC50=3.15

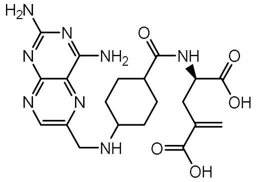
8) –logIC50=5.53

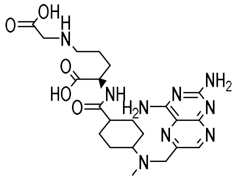
9) –logIC50=3.15

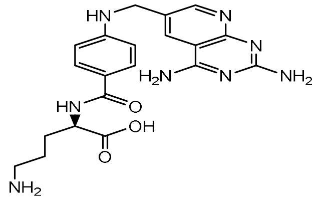
Fig. 2. The schematic structures of methotrexate compounds used to construct QSAR models.

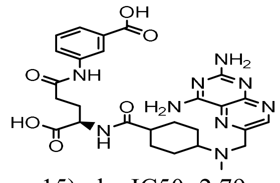

10) –logIC50=4.49

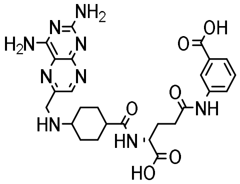

11) –logIC50=2.87

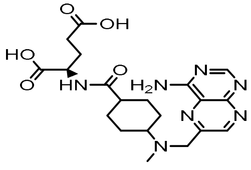

12) –logIC50=0.05

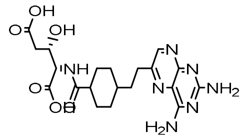

13) –logIC50=1.3

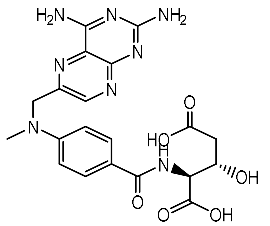

14) –logIC50=1.06

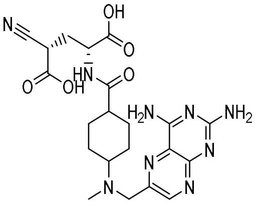

15) –logIC50=2.79

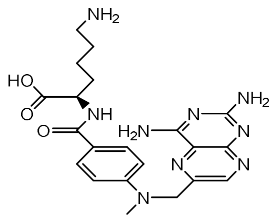

16) –logIC50=2.21

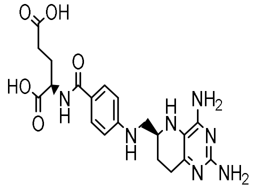

17) –logIC50=0.6

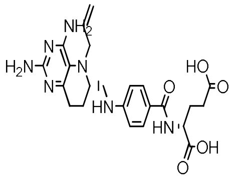

18) –logIC50=2.22

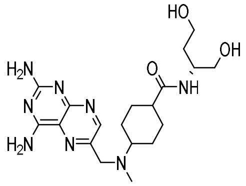

19) –logIC50=2.27

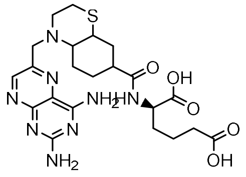

20) –logIC50=1.74

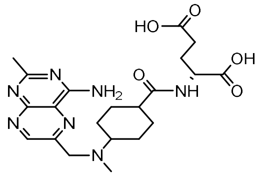

21) –logIC50=1.5

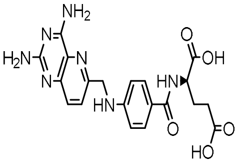

22) –logIC50=0.61

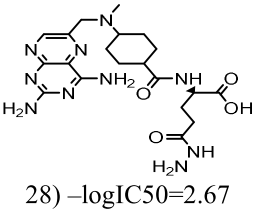

23) –logIC50=0.06

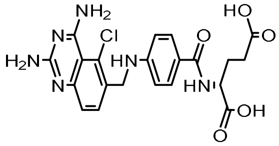

24) –logIC50=0.3


25) –logIC50=1.49


26) –logIC50=0.65


27) –logIC50=2.52


28) –logIC50=2.67


29) –logIC50=2.72

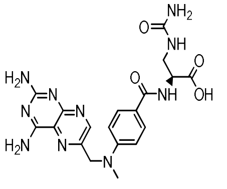

30) –logIC50=2.02

Fig. 2. Continued.

Non-linear Methods

As described in theory and computational method section, in order to establish the target models, the 858 descriptors in the gas phase were fed into the NN to explore the best descriptor. Table 1 displays the statistical parameters of all non-linear QSAR models. 80%, 10%, and 10% of data sets were randomly selected as training, validation, and test sets in non-linear methods.

Based on Table 1, GA-ANN (with a RMSE value of 0.1494 and R^2 of 0.9472) is the best approach among all the studied non-linear methods. Definitions of the selected descriptors using GA-ANN are given in Table 2.

In Table 2, BEHv1 is the Burden eigenvalue descriptors that the B matrix defines as the number of atoms, bond order between two atoms, and the electronegativity of the atoms [14]. RDF140u and RDF090e are independent of the atom number, *i.e.*, the size of a molecule; in addition, these descriptors provide further valuable information, for instance, about bond distances, ring types in planar and non- planar systems, and atom types [14]; ESPm10x indicates edge adjacency indices. The edge adjacency relationships in molecular graphs have been used to define a new topographic index [14]. H1m and R5v+ (Table 2) are GETAWAY (Geometry, Topology, and Atom-Weights Assembly) descriptors encode the geometrical information obtained from the molecular matrix, the topological information obtained from the molecular graph, and the information obtained from atomic weights, which are specially designed with the aim of matching the 3D-molecular geometry [14]. DP02 designates Randic molecular profile descriptors, derived from the distance distribution moments of the geometric matrix G as the average row sum of its entries raised to the k^{th} power and normalized by the factor k! [14]. Since RMSE is highly dependent on the range of the dependent variable, high values of RMSE are due to the possible errors in the experimental data employed. However, the -logIC_{50} values -in our data set were in the range of 0.05 to 5.53 and the RMSE of the predicted set in the GA-ANN model was 0.1494 in the gas phase, which are acceptable in comparison with previous works [26,2].

Table 1. Statistical Parameters of Different Non-linear QSAR Models

Predicted		Train	
R^2	RMSE	R^2	RMSE
0.8855	0.3330	0.9031	0.2994
0.9206	0.2218	0.9244	0.2049
0.9145	0.2472	0.8988	0.2372
0.9472	0.1494	0.9463	0.1216

The graphs of the most effective descriptors in the gas phase (DP02, RDF140u, ESPm10x, BEHv1, RDF090e, H1m and R5v+) versus the empirical negative logarithm half-maximal inhibitory concentration (-logIC50) are plotted in Fig. 3.

Descriptors were selected using the GA–ANN model, as shown in Table 2, employed to build the final model. The figure demonstrates that the changes in the selected

Table 2. Definition of the Selected Descriptors Using the GA-ANN Method

Descriptor	Definition	Type
DP02	Molecular profile no. 2	Randic molecular profiles
RDF140u	Radial Distribution Function-14.0/unweighted	RDF descriptor
ESpm10x	Spectra moment 10 from edge adj. matrix/weighted by edge degrees	Edge adjacency indices
BEHv1	Highest eigenvalue n. 1 of Burden matrix/weighted by atomic van der Waals volumes	Burden eigenvalues
RDF090e	Radial Distribution Function-9.0/Weighted by atomic Sanderson electronegativities	RDF descriptor
H1m	H autocorrelation of lag 1/weighted by atomic masses	GETAWAY descriptors
R5v+	R maximal of lag 5/weighted by atomic masses van der Waals volumes	GETAWAY descriptors

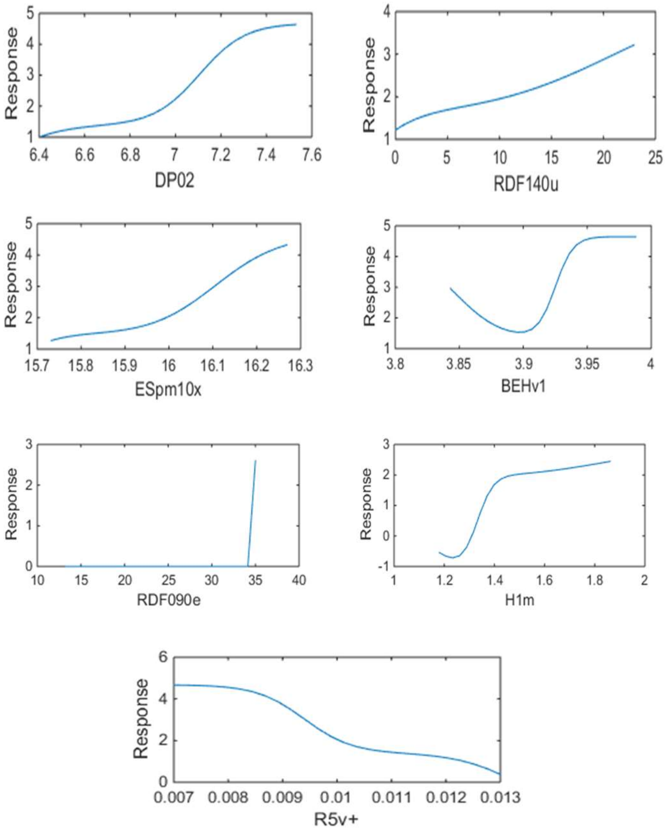


Fig. 3. Plot between experimental -logIC50 (respunse) versus DP02, RDF140u, ESPm10x, BEHv1, RDF090e, H1m, and R5v+ descriptors in the gas phase.

191 tors may have significant effects, so that changing the descriptor values from minimum to maximum can lead to three to five time changes in the value of the response (-logIC50). According to these results, except RV5+, which should be minimized, the other descriptors should be kept on their maximum values. Among the studied derivatives (Fig. 2), compound No. 8 has the high empirical negative logarithm of the half-maximal inhibitory concentration and thus a low empirical half-maximal inhibitory concentration (IC50), making it the best drug among them. To assess the interpreted results, the status of each descriptor for compound No. 8 is plotted in Fig. 4. One can see that the descriptor states are in good agreement with the interpreted results. Note that the outcome of descriptor effects determines the value of the response (-logIC50).

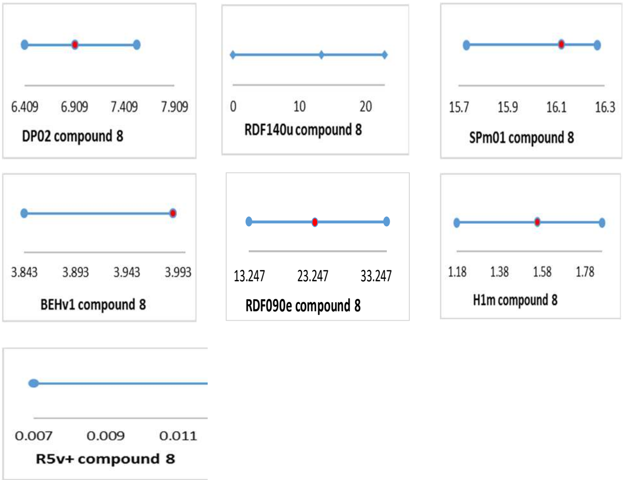


Fig. 4. DP02, RDF140u, ESPm10x, BEHv1, RDF090e, H1m, and R5v descriptors in compounds No. 4 and No. 8 obtained from GA-ANN method.

Results of the Monte Carlo Method

The statistical parameters of the models obtained using molecular graphs (HSG) and SMILES are shown in Table 3. Th performances of the models are compared with each other by the criterion of the predictability in test set (Rm^2) which should be larger than 0.5 [23], correlation coefficient (R^2) in each set, cross-validated correlation coefficient (Q^2), and standard error of estimation (s). The difference between R^2m and R'^2m values (ΔRmTEST) was used as another criterion in this issue. The depicted results in Table 3 reveal that a split 1, a threshold of 3, and a probe of 2 give the best results.

The variation of correlation coefficient (test set) with respect to the threshold and the number of epochs is plotted in Fig. 5. This figure confirms that 3 and 80 are the most appropriate values for threshold and number of epochs, respectively.

The plot showing the variation of observed versus predicted -logIC_{50} values are illustrated in Fig. 6. An acceptable correlation between the calculated and empirical

Table 3. The Split Models in Monte Carlo Method

Split 1: (T = 3, probe = 2)

-logIC50 = -22.1252009 (± 0.6242217) + 0.3652079 (± 0.0092170) * DCW(3,100)
n = 19, R^2 = 0.8879, Q^2 = 0.8549, s = 0.595 (training set)
n = 6, R^2 = 0.9995, Q^2 = 0.9984, s = 1.41 (calibration set)
n = 6, R^2 = 0.9719, Q^2 = 0.9515, s = 0.849 (test set), R^2m TEST = 0.6320

Spit 2: (T = 1, probe = 3)

-logIC50 = -30.3749259 (± 0.0259107) + 0.4347987 (± 0.0003267) * DCW(1,100)
n = 19, R^2 = 0.9992, Q^2 = 0.9991, s = 0.049 (training set)
n = 7, R^2 = 0.9999, Q^2 = 0.9997, s = 1.56 (calibration set)
n = 5, R^2 = 0.9258, Q^2 = 0.8448, s = 1.92, R^2m TEST = 0.4312

Spit 3: (T = 3, probe = 1)

-logIC50 = -27.7454941 (± 0.4084093) + 0.4070119 (± 0.0054781) * DCW(3,100)
n = 17, R^2 = 0.9682, Q^2 = 0.9586, s = 0.296 (training set)
n = 8, R^2 = 0.9995, Q^2 = 0.9991, s = 1.98(calibration set)
n = 6, R^2 = 0.5230, Q^2 = 0.4501, s = 1.86 (test set), R^2m TEST = 0.3425

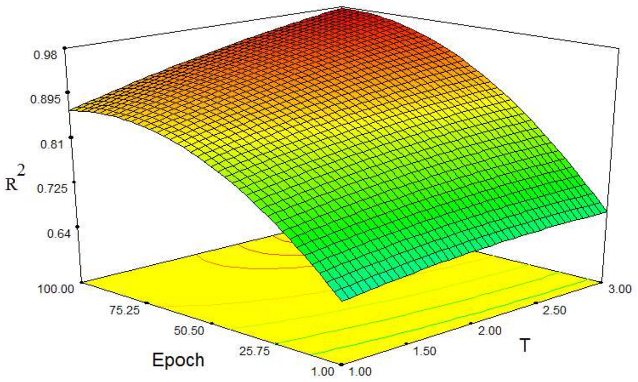


Fig. 5. The variation of correlation coefficient for test set of threshold and number of epochs. 3-D surface plot of R^2 according to the threshold and the number of epochs.

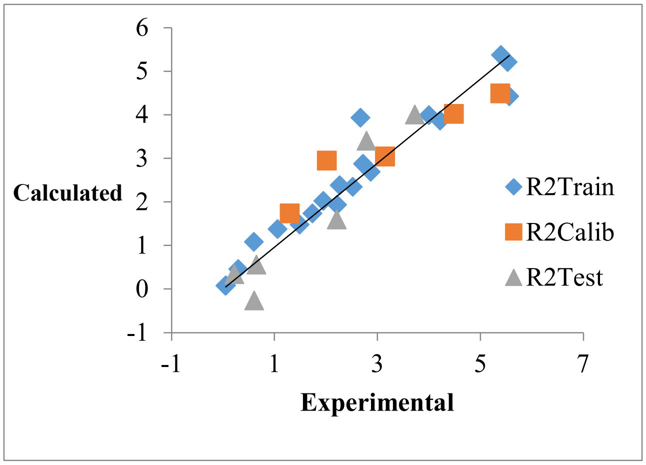


Fig. 6. The correlation between experimental and predicted -logIC50 calculated using Eq. (3).

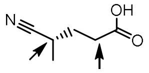
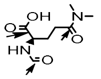
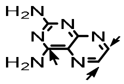
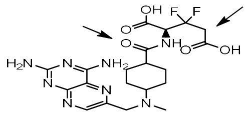
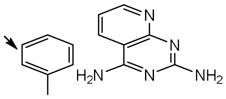
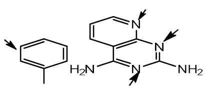
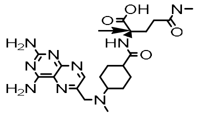
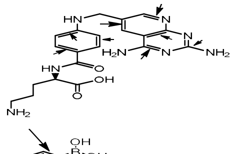
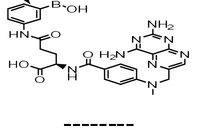
values of -logIC50 can be observed in this figure that approves the appropriateness of the developed model.

Molecular features are sorted according to their correlation weights and are given in Table 4. Molecular feature with negative correlation weights were omitted due to their inverse effect on the -logIC50 value. The higher correlation weigh of a molecular feature, the lower value of IC50; therefore, the features are more significant. Definitions of the molecular features are given by Kumar and Chauhan (Table 5) [28].

Table 4. SMILES Attributes with Positive Correlation Weights for Split 1

SMILES attributes	CWs	SMILES attributes	CWs
1...........	6.60296	EC0-O...1...:	4.13383
2...........	6.40964	=...(.......	3.92591
3...........:	6.16322	C...1.......:	3.90834
C...2.......:	5.24903	EC0-C...1...:	3.86631
C...3.......:	4.67554	NOSP11000000	3.39175

Table 5. Definition of the Molecular Features

Molecular features	Definition	Scheme
HALO00000000	Absence of F, Cl, Br	--------
C...C.......	Presence of carbon-carbon bonds (sp^3)	R_3C-CR_3
C…(…C…	sp^3 carbon atoms with branching	
++++O---B2==	Presence of oxygen and double bonds	
C…=…….	sp^2 carbon atom	
(...........	Branching in molecular skeleton	
O...........	Presence of oxygen	--------
1...........	Presence of rings	
++++N---B2==	Presence of nitrogen and double bond	
=	Double bond	
@	Stereo specific bond	
#	Triplet bond	
=...(.......	Presence of double bond in combination with rings	
C...1......	Presence of cyclic ring with branching	
C...2......, C...3......	Presence of three and two of sp^2 carbon	--------
NOSP11000000	Presence of nitrogen and oxygen but absence of sulfur and phosphorus	--------

According to Table 4 and Table 5, the presence of cyclic ring (1...........), double bond in combination with rings (=...(.......), cyclic ring with branching (C...1.......), three and two of sp^2 carbon in molecule (C...2......., C...3.......), and nitrogen and oxygen but absence of sulfur and phosphorus (NOSP11000000) are the most important molecular features that might be considered in designing new drugs.

CONCLUSION

The obtained results from QSAR models showed that GA-ANN combination was better than the other models used and also proved that DP02, RDF140u, ESPm10x, BEHv1, RDF090e, H1m, and R5v+ descriptors were more significant than the other descriptors as well as predicting biological activity of methotrexate substitution patterns. Thus, this work predicts a new design for this class of drugs and the DP02, RDF140u, ESPm10x, BEHv1, RDF090e, H1m, and R5v+ descriptors values are the maximum. Therefore, the aforementioned descriptors are the best descriptors in the gas phase, and physicochemical descriptors including van der Waals volumes, weighted by atomic masses, and weighted by atomic Sanderson electronegativities should be maximized in designing new drugs. The obtained results can be applied to design new anti-cancer drugs. Additionally, Monte Carlo method was utilized to find out the quality of the effects of structural descriptors on the biological activity of the studied drugs. Structural descriptor including independent of the number of atoms, the size of a molecule, ring types of planar and non-plane systems, and atom types should be maximized. It can be concluded that simultaneous use of Monte Carlo, GA-ANN, and ICA-MLR methods gives deeper and more comprehensive knowledge of the effect of molecular and structural descriptors on the activity of drugs and provides better insights for designing new drugs.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the support provided by the Islamic Azad University of Rasht.

REFERENCES

[1] B.N. Cronstein, Pharmacol. Rev. 57 (2005) 163.

[2] R.J. Lippens, Am. J. Pediatr. Hematol. Oncol 6 (1984) 379.

[3] J.L. Grem, S.A. King, J.M. Sorensen, M.C. Christian, Invest. New. Drugs. 9 (1991) 281.

[4] V. Consonni, R. Todeschini, M. Pavan, M. J. Chem. Inf. Comput. Sci. 42 (2002) 693.

[5] S. Kirkpatrick, Jr.C.D. Gelatt, M.P. Vecchi, Science 220 (1983) 671.

[6] V.O. Černý, J. Optim. Theory. Appl. 45 (1985) 41.

[7] L.M. Schmitt, Theor. Comput. Sci. 259 (2001) 1.

[8] D. Bertsimas, J. Tsitsiklis. Stat. Sci. 8 (1983) 10.

[9] P. Gramatica, V. Consonni, R. Todeschini, Chemosphere 38 (1999) 371.

[10] A.P. Toropova, A.A. Toropov, E. Benfenati, G. Gini, D. Leszczynska, J. Leszczynski, J. Comput. Chem. 32 (2011) 2727.

[11] A.A. Toropov, A.P. Toropova, S.E. Martyanov, E. Benfenati, G. Gini, D. Leszczynska, J. Leszczynski, J. Chemom. Intell. Lab. 109 (2011) 94.

[12] E.B. DeMelo, M.M. Ferreira, Eur. J. Med. Chem. 44 (2009) 3577.

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

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

[15] R. SayyadikordAbadi, O. Alizadehd. Rev. Roum. Chim. 63 (2018) 171.

[16] V. Consonni, R. Todeschini, M. Pavan, P. Gramatica, J. Chem. Inf. Comput. Sci. 42 (2002) 693.

[17] R. SayyadikordAbadi, O. Alizadeh, S. Tajadodi Paskiabei, J. Korean Chem. Soc. 60 (2016) 225.

[18] R. Todechine, *QSAR Group*, http://www.disat.unimib.it/chem.

[19] SPSS, Version 19 (2010); available at http://www.spssscience.com.

[20] M. Jalali-Heravi, M.F. Parastar, J. Chem. Inf. Comput. Sci. 40 (200) 147.

[21] M.H. Fatemi, S. Gharaghani, Bioorg. Med. Chem. 15 (2007) 7746.

[22] K. Levenberg, Quart. J. Mech. Appl. Math. 2 (1944) 164.

[23] J. Veselinović, A. Veselinović, A.A Toropov, A.P. Toropova, I. Damnjanović, G. Nikolić, Scientific Journal of the Faculty of Medicine in Niš, 31 (2014) 95.

[24] A.M. Veselinović, J.B. Milosavljević, A.A. Toropov, G.M. Nikolić, Eur. J. Pharm. Sci. 48 (2013) 532.

[25] (http://www.insilico.eu/coral)

[26] R. SayyadikordAbadi, A. Alizadehdakhel, Rev. Roum. Chim. 63 (2018) 931.

[27] R. Sayyadi Kord Abadi, A. Alizadehdakhel, Russ. J. Phys. Chem. B. 11 (2017) 307.

[28] A. Kumar, S. Chauhan, SAR. QSAR. Environ. Res. 28 (2017) 179.