Research Article

Chemical Review and Letters
journal homepage: www.chemrevlett.com
ISSN (online): 2645-4947 (print) 2676-7279

QSAR Study of Antiproliferative Drug Against A549 by GA-MLR and SW-MLR Methods

Somayeh Alimohammadi[a], Aliasghar Hamidi[b], Parinaz Pargolghasemi[c], Nasim Nourani[b,d], Mir Saleh Hoseininezhad-Namin[b,d*]

[a].*Faculty of Medicine, Shahid Beheshti University of Medical Sciences, Tehran, Iran,*
[b]. *Biotechnology Research Center, Tabriz University of Medical Sciences, Tabriz, Iran*
[c].*Department of Chemistry, Payame Noor University (PNU), P. O. Box, 19395-3697 Tehran, Iran*
[d]. *Students Research Committee, Tabriz University of Medical Sciences, Tabriz, Iran*

ARTICLE INFO

Article history:
Received
Received in revised form
Accepted
Available online

Keywords:
lung Cancer
Quinoline derivative
QSAR
Multiple linear regressions
Genetic algorithm

ABSTRACT

Quantitative structure-activity relationship (QSAR) is the most extensively used computational methodology for analogue-based design. In this research, QSAR model was used to predict antiproliferative properties of 4-(2-fluorophenoxy) quinoline derivatives against A549(human lung adenocarcinoma). For this purpose, we used the multiple linear regressions (MLR) for the construction of a model to predict drug activity and Stepwise (SW) and genetic algorithm (GA) methods used to build the model. The data were selected from 31 molecules with specific pharmacological activity. They were divided into two sets train and test data. The resulting model was tested using statistical methods such as external test set and cross-validation to ensure its authenticity. The results showed that GA-MLR approach had good predictive power and higher data rates compared with SW-MLR (Q^2_{LOO} = 0.877, R^2_{Train} =0.933). The results obtained in this study can be used to design drugs with higher performance and pharmacological activity to treat lung cancer.

1. Introduction

Cancer is characterized by the virulent tumors and virulent neoplasm which may be defined as abnormal, excessive, uncoordinated, and autonomous proliferation of cells. [1]. Cancer has been seriously threatening the health and life of humans for a long period and has become the leading disease-related cause of deaths of human population [2]. Lung cancer is one of the most common types of cancer that occurs in women and men [3]. Radiation therapy and surgery for treatment of cancer are only successful when the cancer is found at early localized stage. However, chemotherapy in contrast is the mainstay in treatment of malignancies because of its ability to cure widespread or metastatic cancers [4].

4-(2-fluorophenoxy) quinoline derivatives bearing an imidazolonehas anti-cancer properties and act as antiproliferative against A549(human lung adenocarcinoma) drugs are essential to fight lung cancer [5]. Given the importance of anti-cancer drugs, it seems necessary design and predictions new drugs with more activities and spend less time and cost to do the synthesis [6]. Computational methodologies have emerged as an imperative tool for any drug discovery program, playing key role from hit identification to lead optimization [7]. In this research, the famous method, which is called quantitative structure–activity relationship (QSAR) has been developed, and used for predicting the biological behavior of compounds by utilizing molecular structures and experimental data. Through this method biological properties can be obtained easily without any experimental efforts for synthesis and testing the novel compounds [8]. These characteristics made this method to be expanded and used in several fields, and recently have been employed for screening the biological activities of drugs in drug design [9]. QSAR model can be generated by collecting the experimental data, and then calculating the theoretical parameters for new designed compounds. The experimental data are related to the biological

* Corresponding author. e-mail: *hoseininezhad.s@tbzmed.ac.ir*

properties that considered as dependent variables such as toxicity, bioavailability and activity for creating the model [10]. Among these descriptors only relevant variables should be selected which are in correlation with biological activities. Therefore, one of the essential steps in the QSAR method is employing a technique to select the respective variables [11]. Innovation in the variable selection tools effected in developing of important methods such as stepwise [12], simulated annealing and genetic algorithms (GAs) [13]. After the respective descriptors have been obtained, the model is built by using various modeling methods such as multiple linear regression (MLR) [14], support vector machine (SVM) [15] and partial least squares (PLS) [16]. The main goal of this paper is development of an effective model using SW-MLR and GA-MLR approaches, and considering the most important descriptors which affect the activities of molecules.

2. Results and Discussion

2.1. Stepwise-multiple linear regression method

First, the data set of 31 derivatives was partitioned into a training set of 25 compounds and a test set of 6 compounds (80 and 20%, respectively, of the total number of compounds) (Table 1). By using Stepwise-multiple linear regression method (SW-MLR) seven descriptors was selected with most associated to the pIC_{50} which includes: RDF020m, Jhetp, R6v+, RDF100m, E1v, HATS4v, Mor32e.

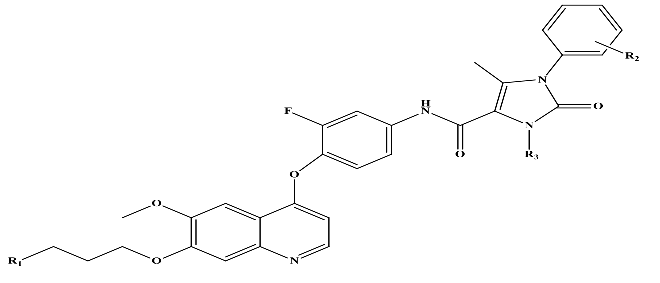
The following formula (1) was obtained:

pIC_{50}= 12.188 (±1.148) + 14.692 (±5.485) HATS4v + 0.5358 (±0.0869) Mor32e - 0.1417 (±0.0316) RDF020m - 12.485 (±1.5746) Jhetp + 3.952 (±0.8528) E1v + 23.889 (±5.1379) R6v+ + 0.0175 (±0.0047) RDF100m

Predicted pIC_{50} values for each molecule by using the SW model have been reported in Table1. Figure 1 displays the predicted error values with little error (-0.1 to 0.1). Statistical analysis values of SW method are shown in Table 2. Also contribution of each descriptor obtained in SW model shown in Figure 2. Statistical data of model showed very good R^2_{Train} (0.948) value and the value for Test (0.88) also showed relatively acceptable.

In the next stage, the genetic algorithms method used to provide better prediction models of Test and compare the resulting of the models.

Table 1. Chemical structures and the corresponding experimental and predicted pIC50 values by SW-MLR and GA–MLR methods



Num	R1	R2	R3	pIC_{50}	SW-MLR	GA-MLR
1	morpholinyl	H	CH_3	6.00	6.01	6.08
2	pyrrolidinyl	H	CH_3	5.83	5.82	5.85
3	piperidinyl	H	CH_3	6.10	6.05	6.07
4[a]	4-methylpiperidinyl	H	CH_3	6.18	6.19	6.11
5	4-methylpiperazinyl	H	CH_3	6.22	6.27	6.21
6	4-methylpiperidinyl	4-CH_3	CH_3	5.95	5.97	5.97
7	4-methylpiperazinyl	4-CH_3	CH_3	6.08	6.12	6.09
8	4-methylpiperidinyl	4-OCH_3	CH_3	5.98	6.00	5.93
9[a]	4-methylpiperazinyl	4-OCH_3	CH_3	6.02	6.06	5.92
10	4-methylpiperidinyl	4-F	CH_3	6.28	6.27	6.23
11	4-methylpiperazinyl	4-F	CH_3	6.37	6.33	6.29
12	4-methylpiperidinyl	4-Cl	CH_3	6.38	6.41	6.43
13	4-methylpiperazinyl	4-Cl	CH_3	6.44	6.46	6.40
14	4-methylpiperidinyl	4-Br	CH_3	6.18	6.23	6.3
15	4-methylpiperazinyl	4-Br	CH_3	6.32	6.30	6.30
16	morpholinyl	3,4-2F	CH_3	6.21	6.28	6.21
17[a]	4-methylpiperidinyl	3,4-2F	CH_3	6.27	6.35	6.23
18	4-methylpiperazinyl	3,4-2F	CH_3	6.40	6.39	6.35
19	4-methylpiperidinyl	3-Cl-4-F	CH_3	6.37	6.42	6.39
20[a]	4-methylpiperazinyl	3-Cl-4-F	CH_3	6.42	6.50	6.30
21	4-methylpiperidinyl	4-OCF_3	CH_3	6.21	6.27	6.22
22	4-methylpiperazinyl	4-OCF_3	CH_3	6.36	6.33	6.38
23	4-methylpiperidinyl	4-CF_3	CH_3	6.27	6.25	6.31

24[a]	4-methylpiperazinyl	4-CF_3	CH_3	6.30	6.22	6.35
25	4-methylpiperidinyl	3-CF_3	CH_3	6.06	6.09	6.11
26	4-methylpiperazinyl	3-CF_3	CH_3	6.11	6.1	6.08
27	4-methylpiperidinyl	2-CF_3	CH_3	6.35	6.31	6.41
28	4-methylpiperazinyl	2-CF_3	CH_3	6.43	6.42	6.38
29	4-methylpiperidinyl	2-Cl	CH_3	6.49	6.5	6.48
30[a]	4-methylpiperazinyl	2-Cl	CH_3	6.51	6.49	6.51
31	4-methylpiperazinyl	2-Cl	H	6.60	6.57	6.60

[a] Test set

Table 2. Statistical results of different QSAR models

	Training set				Test set		
	R^2	RMSE	F	Q^2_{LOO}	R^2	RMSE	F
SW-MLR	0.963	0.043	63.158	0.751	0.881	0.587	-2.84
GA-MLR	0.933	0.048	34.008	0.877	0.916	0.075	-1.86

2.2. Genetic algorithm-multiple linear regression method

In this step for choose the best descriptors with highest associated to pIC_{50} used genetic algorithm as a subset based on MLR method. Descriptors with the highest correlation were selected which including MATS7m, Mor28u, R5v +, MATS8v, Mor21u, Mor32v, RDF040v descriptors. Formula (2) was obtained from Train data in qsar model. The accuracy of this formula was evaluated by Data Test.

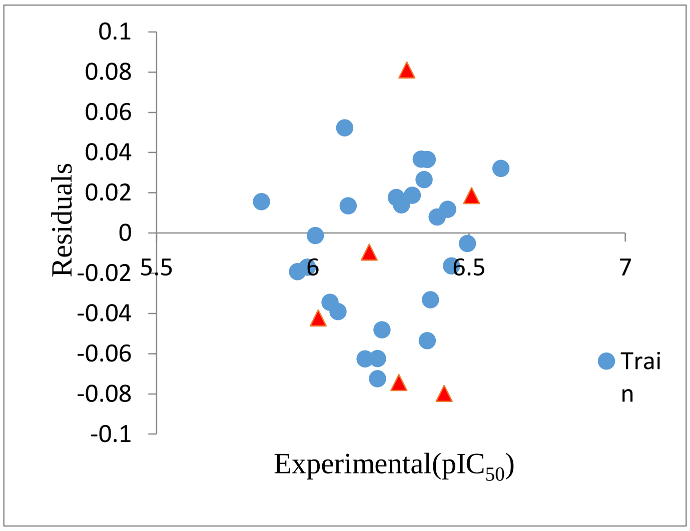


Figure 1. The residuals to the experimental pIC_{50} values by SW-MLR model for train and test set

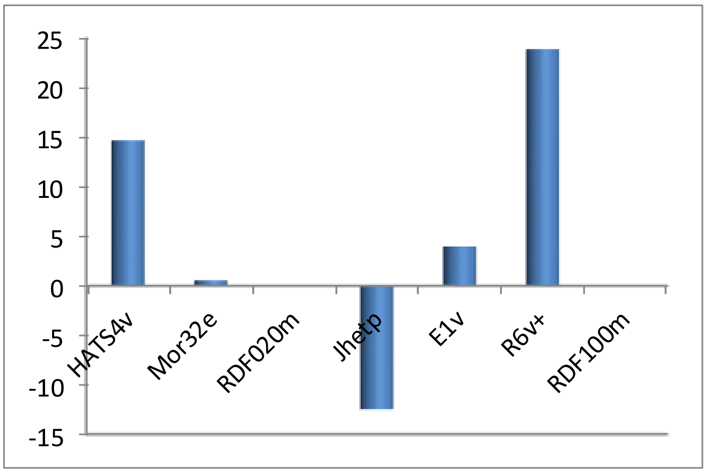


Figure 2. Contribution of each descriptor obtained in SW model

Formula 2:

pIC_{50}=19.61 (±4.975) + 24.97 (±4.782) MATS7m + 0.3525 (±0.0723) Mor28u + 12.22 (±7.411) R5v + - 0.4855 (±0.9412) MATS8v - 0.1345 (±0.0578) Mor21u + 0.5993 (±0.2384) Mor32v + 0.0449 (±0.0189) RDF040v

N_{Train} = 25, R^2_{Train} = 0.933, RMSE $_{Train}$ = 0.048, F_{Train} = 34.008

N_{Test} = 6, R^2_{Test} = 0.916, $RMSE_{Test}$ = 0.075, F_{Test} = -1.861

R^2_{adj} = 0.906, Q^2_{LOO} = 0.877, Q^2_{LGO} = 0.803

N represents the number of molecules, R2 represents squared correlation coefficient, RMSE represents the root mean square error, F represents the Fisher F statistic, R^2_{adj} represents the adjusted R^2, Q^2_{LOO} and Q^2_{LGO} represents coefficients for leave-one-out and leave group out respectively.

The data obtained show that the predictive potency of this model is very good. Also the R^2 data shows that if this model evaluated by test data, the potency of prediction could be very well (0.916), which confirm the validity of the results. Train RMSE values and the Test values are 0.048 and 0.075 respectively, this data shows that the error of method is very small and negligible. The result of our calculations for Q^2_{LOO} and Q^2_{LGO} are 0.877 and 0.803 respectively. High levels of Q^2_{LOO} and Q^2_{LGO} values prove the ability of the presented model in support of the internal validation. Experimental values and predicted values shown in Table1. These values indicate that the predicted values are very close to the experimental values. Figure (3) show the predicted values of train and test against the experimental pIC_{50} values. Contribution of each descriptor obtained in GA-MLR model displayed in Figure (4). As shown in Figure 4, MATS7m and R5v+ descriptors compare to other descriptors have more effective and contribution in prediction of pharmaceutical activity. Error for pIC_{50} of drugs prediction to both train and test were between -0.13 and +0.13, but this value for most molecules are between +0.07 and -0.07. These error amounts show that there aren't significant errors for each of molecules (Fig.5). As seen in Figure 5 dispensing errors are completely random and there is no systematic error, so this method is statistically acceptable. The correlation coefficient matrix for selected descriptor in GA-MLR method is reported in the Table 3. As you can see, there is no dependency between selected descriptors. The definition and class of best descriptors were indicated in Table 4.

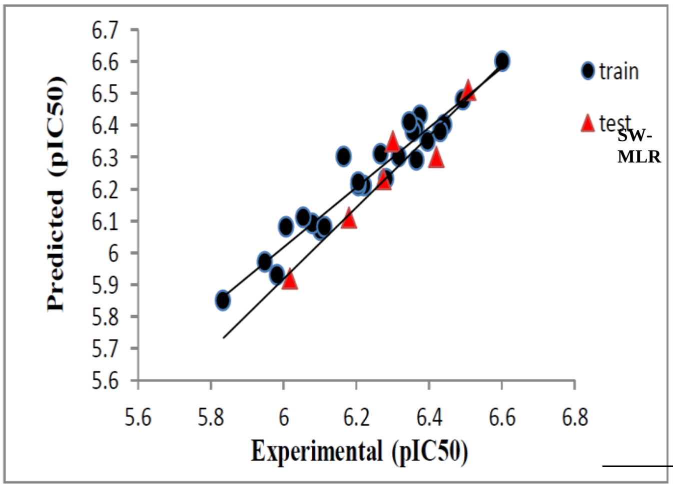


Figure 3. The predicted versus the experimental pIC_{50} by GA-MLR

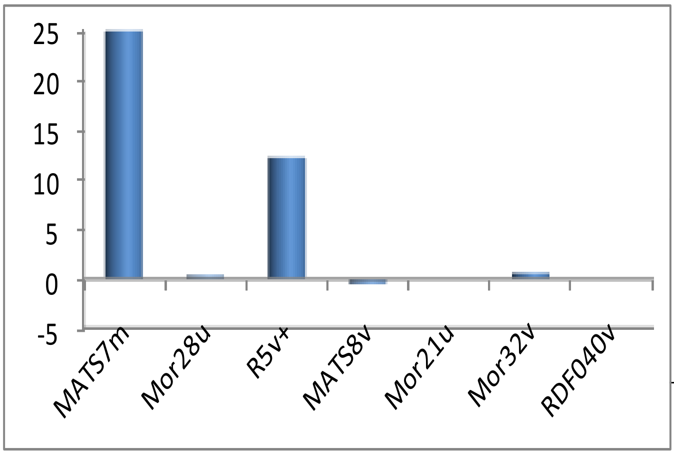


Figure 4. Contribution of each descriptor obtained in GA-MLR model

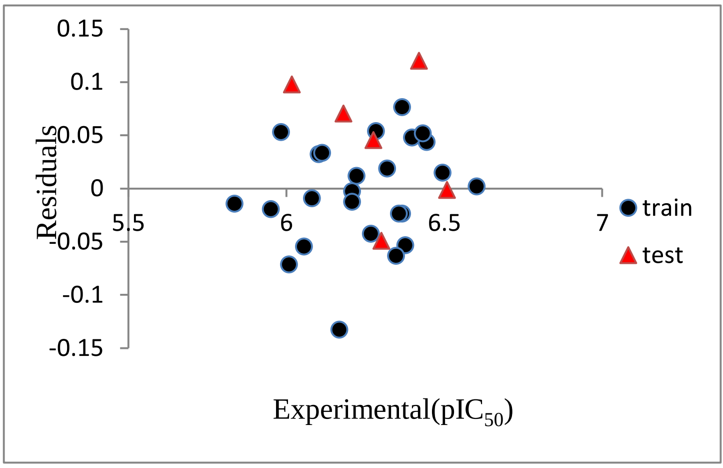


Figure 5. The residuals to the experimental pIC50 values by GA- MLR model for train and test set

Table 3. The correlation coefficient of selected descriptors based on GA-MLR and SW-MLR

	HATS4v	Mor32e	RDF020m	Jhetp	E1v	R6v+	RDF100m
HATS4v	1	0	0	0	0	0	0
Mor32e	0.425	1	0	0	0	0	0
RDF020m	0.379	0.006	1	0	0	0	0
Jhetp	0.118	-0.043	0.036	1	0	0	0
E1v	0.752	0.305	0.427	0.222	1	0	0
R6v+	0.414	0.072	-0.062	0.497	0.287	1	0
RDF100m	0.474	0.006	0.096	0.006	0.312	-0.007	1

	MATS7m	Mor28u	R5v+	MATS8v	Mor21u	Mor32v	RDF040v
MATS7m	1	0	0	0	0	0	0
Mor28u	-0.128	1	0	0	0	0	0
R5v+	-0.070	0.452	1	0	0	0	0
MATS8v	0.134	0.158	0.585	1	0	0	0
Mor21u	-0.056	-0.034	0.170	-0.059	1	0	0
Mor32v	0.465	0.242	0.041	-0.022	-0.305	1	0
RDF040v	-0.111	0.298	0.505	0.207	-0.473	0.161	1

Table 4. Definition and class of best descriptors

No.	Symbol	Class	Definition
1	MATS7m	2D autocorrelations	Moran autocorrelation - lag 7 / weighted by atomic masses
2	Mor28u	3D-MoRSE descriptors	3D-MoRSE - signal 28 / unweighted
3	R5v+	GETAWAY descriptors	R maximal autocorrelation of lag 5 / weighted by atomic van der Waals volumes
4	MATS8v	2D autocorrelations	Moran autocorrelation - lag 8 / weighted by atomic van der Waals volumes
5	Mor21u	3D-MoRSE descriptors	3D-MoRSE - signal 21 / unweighted
6	Mor32v	3D-MoRSE descriptors	3D-MoRSE - signal 32 / weighted by atomic van der Waals volumes
7	RDF040v	RDF descriptors	Radial Distribution Function - 4.0 / weighted by atomic van der Waals volumes

2.3. Comparison of SW-MLR and GA-MLR

Both SW-MLR and GA-MLR methods provided good result, but when created models in SW, GA method were assayed by the test set, genetic algorithm method had higher predictive power. Statistical values in the Table 2 indicated that in the genetic algorithm increase in R^2 and F, also decrease in RMSE be seen. As shown in Table 3 correlation coefficient between each descriptor in SW-MLR is less than 0.75 and in GA-MLR is less than 0.58. Therefore, selected descriptors in GA-MLR method have low correlation and compare to SW-MLR method are more independent. All of the result shows higher predictive power of genetic algorithm.

3. Materials and methods

3.1. Data set

In this paper, the activities of 31anti-cancer drugs were derived from literature [5]. All of the IC_{50} values were inverted to –log (IC_{50}) and indicated as pIC_{50} during this research (Table1).

31molecules of dataset were divided into training and test sets (25 molecules training set and 6 molecules test set).

3.2. Molecular descriptors

The chemical structures of molecules were drawn in Hyperchem software (Version 7.0) [17]. Then optimization procedure was carried out to achieve stable conformation of molecules. To do this first using the method of molecular mechanics force field (MM+) and then optimization was performed using optimization methods semi empirical (AM1) with the root mean square gradient of 0.01 kcal mol^{-1}. Using Hyperchem was obtained descriptors diverse as the highest filled orbital (HOMO), lowest vacant orbital (LUMO), Lipophilicity (LogP), etc. Moreover, calculated 1497 descriptor by DRAGON software [18]. Removal operation done to eliminate problems such as the random chance of descriptors and correlations between choice descriptors. In the first stage, removed descriptors with the almost constant value, or 90% of fixed amounts. In the second stage descriptors were selected which independent variables more than 9.0 correlation with other independent variables, And Removed variable is the lowest correlation with the pIC_{50}. After omissions remained 386 descriptors for operations final analysis.

3.3. Variable selection

In this step the best descriptors should be chosen among remaining descriptors. The biggest challenge of qsar studies is providing a good model with selecting the smallest number of descriptors among the remaining descriptors [19]. There are many ways to predict the dependence between descriptors and dependent variable. In this study we used stepwise multiple linear regressions (SW-MLR) and genetic algorithm multiple linear regressions (GA-MLR) methods. Calculations were performed by using spss software (version 18) and Matlab 6.5 software [20, 21] for done SW-MLR and GA-MLR methods respectively. In both methods validity of model evaluated by statistical methods.

4. Conclusion

In this research, qsar study was used to find reliable model for antiproliferative properties of 4-(2-fluorophenoxy) quinoline derivatives against A549. For the selection of significant descriptors and construction of the model, SW-MLR and GA-MLR were utilized as powerful and suitable techniques. In each of the methods seven descriptor with the most correlation selected. Models were assayed by test set to evaluate the accuracy. When the results of SW-MLR and GA-MLR methods were compared, one can deduce that GA-MLR had the best outputs. According to the proposed models, the new drugs 4-(2-fluorophenoxy) quinoline derivatives can be synthesized with higher pharmacological effect. So GA-MLR method can be used for the calculation of the activities of novel compounds.

References

[1] J.B. Vieira, F.S. Braga, C.C. Lobato, C.F. Santos, J.S. Costa, J.A.H. Bittencourt, D.S. Brasil, J.O. Silva, L.I. Hage-Melim, W.J.C. Macêdo, A QSAR, Pharmacokinetic and Toxicological Study of New Artemisinin Compounds with Anticancer Activity, *Molecules.*, 19 (2014) 10670-10697.

[2] M.H. Bohari, H.K. Srivastava, G.N. Sastry, Analogue-based approaches in anti-cancer compound modelling: the relevance of QSAR models, *Org. Med. Chem. Lett.*, 1 (2011) 1-12.

[3] J. Yan, Y. Pang, J. Sheng, Y. Wang, J. Chen, J. Hu, L. Huang, X. Li, A novel synthetic compound exerts effective anti-tumour activity in vivo via the inhibition of tubulin polymerisation in A549 cells, *Biochem. Pharmacol.*, 97 (2015) 51-61.

[4] G.M. Cragg, P.G. Grothaus, D.J. Newman, Impact of Natural Products on Developing New Anti-Cancer Agents†, *Chem. Revi.*, 109 (2009) 3012-3043.

[5] W. Liao, G. Hu, Z. Guo, D. Sun, L. Zhang, Y. Bu, Y. Li, Y. Liu, P. Gong, Design and biological evaluation of novel 4-(2-fluorophenoxy) quinoline derivatives bearing an imidazolone moiety as c-Met kinase inhibitors, *Bioorg. Med. Chem.*, 23 (2015) 4410-4422.

[6] T. Fujita, QSAR and drug design, Elsevier, 1995.

[7] J.C. Madden, M.T. Cronin, Structure-based methods for the prediction of drug metabolism, (2006).

[8] E. Pourbasheer, R. Aalizadeh, M.R. Ganjali, P. Norouzi, J. Shadmanesh, QSAR study of ACK1 inhibitors by genetic algorithm–multiple linear regression (GA–MLR), *J. Saudi. Chem. Soc.*, 18 (2014) 681-688.

[9] W. Li, Y. Tang, Y.-L. Zheng, Z.-B. Qiu, Molecular modeling and 3D-QSAR studies of indolomorphinan derivatives as kappa opioid antagonists, *Bioorg. Med. Chem.*, 14 (2006) 601-610.

[10] A. Habibi-Yangjeh, E. Pourbasheer, M. Danandeh-Jenagharad, Prediction of basicity constants of various pyridines in aqueous solution using a principal component-genetic algorithm-artificial neural network, *Monatsh. Chem.*, 139 (2008) 1423-1431.

[11] A. Habibi-Yangjeh, E. Pourbasheer, M. Danandeh-Jenagharad, Application of principal component-genetic algorithm-artificial neural network for prediction acidity constant of various nitrogen-containing compounds in water, *Monatsh. Chem.*, 140 (2009) 15-27.

[12] R.R. Hocking, A Biometrics invited paper. The analysis and selection of variables in linear regression, *Biometrics.*, 32 (1976) 1-49.

[13] Q. Shen, Q.-Z. Lü, J.-H. Jiang, G.-L. Shen, R.-Q. Yu, Quantitative structure–activity relationships (QSAR): studies of inhibitors of tyrosine kinase, European journal of pharmaceutical sciences, 20 (2003) 63-71.

[14] E. Pourbasheer, S. Riahi, M.R. Ganjali, P. Norouzi, QSAR study on melanocortin-4 receptors by support vector machine, *Eur. J. Med. Chem.*, 45 (2010) 1087-1093.

[15] V. Vapnik, Statistical learning theory. 1998, in, Wiley, New York, 1998.

[16] H. Khajehsharifi, M. Sadeghi, E. Pourbasheer, Spectrophotometric simultaneous determination of ceratine, creatinine, and uric acid in real samples by orthogonal signal correction–partial least squares regression, *Monatsh. Chem.*, 140 (2009) 685-691.

[17] P. Gramatica, P. Pilutti, E. Papa, Validated QSAR prediction of OH tropospheric degradation of VOCs: splitting into training-test sets and consensus modeling, *J. Chem. Inf. Comput. Sci.*, 44 (2004) 1794-1802.

[18] R. Todeschini, V. Consonni, A. Mauri, M. Pavan, DRAGON-Software for the calculation of molecular descriptors, Web version, 3 (2003).

[19] R. Leardi, R. Boggia, M. Terrile, Genetic algorithms as a strategy for feature selection, *J. Chemom.*, 6 (1992) 267-281.

[20] I. MathWorks, Genetic Algorithm and Direct Search Toolbox for Use with MATLAB: User's Guide, in, MathWorks, 2005.

[21] P. Pargolghasemi, M.S. Hoseininezhad-Namin, A. Parchehbaf Jadid, Prediction of Activities of BRAF (V600E) Inhibitors by SW-MLR and GA-MLR Methods, *Curr. Comput. Aided. Drug. Des.*, 13 (2017) 249-261.

How to Cite This Article

Somayeh Alimohammadi; Aliasghar Hamidi; Parinaz Pargolghasemi; Nasim Nourani; Mir Saleh Hoseininezhad-Namin. "QSAR study of antiproliferative drug against A549 by GA-MLR and SW-MLR methods". Chemical Review and Letters, 2, 4, 2019, 193-198. doi: 10.22034/crl.2020.220465.1037