Arabian Journal of Chemistry (2017) **10**, 33–40

King Saud University

Arabian Journal of Chemistry

www.ksu.edu.sa
www.sciencedirect.com

ORIGINAL ARTICLE

Quantitative structure activity relationship study of p38α MAP kinase inhibitors

Eslam Pourbasheer [a], Sajjad Ahmadpour [b], Rohollah Zare-Dorabei [c], Mehdi Nekoei [d,*]

[a] *Department of Chemistry, Payame Noor University, Tehran, Iran*
[b] *Department of Inorganic Chemistry, Faculty of Chemistry, University of Tabriz, Tabriz, Iran*
[c] *Department of Chemistry, Iran University of Science and Technology, Tehran 16846-13114, Iran*
[d] *Department of Chemistry, Shahrood Branch, Islamic Azad University, Shahrood, Iran*

Received 19 October 2012; accepted 14 May 2013
Available online 23 May 2013

KEYWORDS

QSAR;
Genetic algorithm;
Stepwise;
Multiple linear regression;
Pyrazole derivative

Abstract The quantitative structure activity relationship (QSAR) of the novel pyrazole derivatives as inhibitors of p38α mitogen activated protein (MAP) kinase was studied. The suitable set of the molecular descriptors was calculated and the important descriptors using the variable selections of the stepwise (SW) and the genetic algorithm (GA) were selected. The predictive quality of the QSAR models was tested for an external set of nine compounds, randomly chosen out of 44 compounds. A comparison between the attained results indicated the superiority of the genetic algorithm over the stepwise method in the feature selection. The genetic algorithm-multiple linear regression (GA-MLR) model with six selected descriptors was obtained. The accuracy of the proposed model is illustrated using the following evaluation techniques: cross-validation, validation through an external test set, applicability domain, and *Y*-randomization. The analyses may be used to design more potent pyrazole derivatives and predict their activity prior to synthesis.

1. Introduction

The p38α mitogen-activated protein (MAP) kinase is a member of the intracellular family of MAP kinases implicated in the phosphorylation cascade leading to the release of TNFα and other cytokines including interleukin-1beta (IL-1β), interleukin-6 (IL-6) and interleukin-8 (IL-8). The p38 kinases are activated by a variety of stress stimuli including osmotic shock, ionizing radiation, mechanical wear, and cytokine stimulation (Margutti and Laufer, 2007). Activation results in the release of TNFα among other cytokines and the migration of white blood cells to the site of inflammation. The p38α isoform is believed to be the most clinically relevant for the treatment of rheumatoid arthritis (RA) (O'Keefe et al., 2007) hence, p38α has emerged as an attractive target for small molecule drug discovery to blockade the action of TNFα. (Pettus and Wurz, 2008; Wagner and Laufer, 2006; Westra and Limburg, 2006).

* Corresponding author. Tel.: +98 273 2238545.
E-mail address: nekoei_m1@yahoo.com (M. Nekoei).
Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.arabjc.2013.05.009

Novel medicines and experimental measurement of inhibition activity of chemicals are typically developed using a trial and error approach which is time-consuming and costly, thus a great deal of effort has been put into attempting the estimation of activity through statistical modeling. The application of quantitative- structure activity relationship (QSAR) methodologies to this problem has the potential to decrease substantially the time and effort required to discover new medicines or improve current ones in terms of their efficacy. QSARs establish mathematical relationships between physical, chemical, biological, or environmental activities of interest and measurable or computable parameters such as topological, physicochemical, stereo chemical or electronic indices. (Bhatia et al., 2010; Habibi-Yangjeh et al., 2008a, b, 2009; Melagraki et al., 2006; Shahlaei et al., 2011; Hemmateenejad et al., 2011; Yousefinejad et al., 2012). A successful QSAR model is not only constructed to correctly estimate the numerical value of the property or biological activity, but also to give a deeper understanding of what structural features are important for the observed activity. The application of QSAR technique usually requires variable selection for building well-fitted models. In this work, we employed the stepwise (SW) and the genetic algorithm (GA) methods for the variable selection in the multiple linear regression (MLR) method. A limitation of the SW regression search approach is that it presumes that there is a single 'best' subset of X variables and seeks to identify it. There is often no unique 'best' subset. All the possible regression models with a similar number of X variables to the SW regression solution should be subsequently fitted to study whether some other X variables subsets might be better. Nowadays, GA is well-known as an interesting and more widely used variable selection method (Alsberg et al., 2000; Depczynski et al., 2000; Jouanrimbaud et al., 1995). Genetic algorithm is a stochastic method to solve the optimization problems defined by fitness criteria, applying the evolution hypothesis of Darwin and different genetic functions, i.e., crossover and mutation. The aim of this work is to search for an efficient method to build an accurate quantitative relationship between the molecular structure and the p38α MAP Kinase activity of pyrazole derivatives by SW-MLR and GA-MLR methods. The proposed methodology was validated using several strategies: cross-validation, Y-randomization, and external validation using division of the entire data set into training and test sets.

2. Data set and methods

The data set of 44 pyrazole derivatives used for the QSAR analyses was selected from the literature (Wurz et al., 2009, 2010). The data used in this QSAR study consisted of inhibition activity data (IC_{50}), the minimal concentration of compound which affected one inhibitory parameter in 50% of cells. The inhibition activity data [IC_{50} (nM)] for pyrazole derivatives were converted to the logarithmic scale pIC_{50} [$-\log IC_{50}$ (M)] and then used for subsequent QSAR analyses as the response variables. The chemical structures and corresponding pIC_{50} values for studied compounds are presented in Table 1.

2.1. Softwares

A Pentium IV personal computer with the Windows XP operating system was used. Geometry optimization was performed with HYPERCHEM 7.0. DRAGON 2.1 software was utilized to calculate the molecular descriptors. The SPSS software was employed for the simple multiple linear regression model (MLR) analysis. The genetic algorithm (GA)-MLR regression and the other calculations were written in the MATLAB 7.0.

2.2. Descriptor calculation and selection

The main step in every QSAR study is calculating and choosing the structural descriptors as numerical encoded parameters representing the chemical structures. In the present work the molecular descriptors were generated using Dragon software, web version 2.1. Dragon software has been widely used for calculating chemical descriptors in many QSAR studies. It is noticeable, that calculation of these descriptors is easy and fast. An average computing time of 1 min could be considered per structure. A total of 1481 descriptors were calculated for each molecule using this software. Descriptors with constant or almost constant values for all molecules were eliminated. Also, pairs of variables with a correlation coefficient greater than 0.90 were classified as intercorrelated, and only one of them with high correlation with activity data was considered in developing the model. Then, the remaining descriptors were collected in an $n \times m$ data matrix, where $n = 44$ and $m = 574$ are the numbers of compounds and descriptors, respectively. Among the descriptors mentioned above, the most significant molecular descriptors were identified using the genetic algorithm method.

2.3. Genetic algorithm

Nowadays, GA is well-known as an interesting and the most widely employed variable selection method that is used to solve the optimization problems defined by fitness criteria, applying the evolution hypothesis of Darwin and different genetic functions, i.e. cross-over and mutation. To select the most relevant descriptors, the evolution of the population was simulated.(Ahmad and Gromiha, 2003; Hunger and Huttner, 1999; Waller and Bradley, 1999) The population of the first generation was selected randomly. Each individual member in the population, defined by a chromosome of binary values, represented a subset of descriptors. Number of the genes at each chromosome was equal to the number of the descriptors. A gene was given the value of 1, if its corresponding descriptor was included in the subset; otherwise, it was given the value of zero. (Aires-de-Sousa et al., 2002) The number of the genes with the value of 1 was kept relatively low to have a small subset of descriptors. As a result, the probability of generating 0 for a gene was set greater (at least 60%) than the value of 1. The operators used here were cross-over and mutation. The application probability of these operators was varied linearly with a generation renewal (0–0.1% for mutation and 60–90% for cross-over). The population size was varied between 50 and 250 for different GA runs. For a typical run, the evolution of the generation was stopped when 90% of the generations took the same fitness.

3. Results and discussion

3.1. Regression models

For the selection of the most important descriptors, both GA and the SW multiple regression techniques were used. Firstly,

Table 1 Chemical structure and the corresponding experimental and predicted pIC_{50} values by SW-MLR and GA-MLR methods.

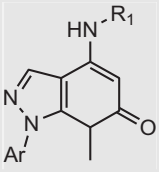
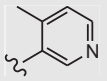
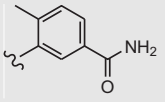
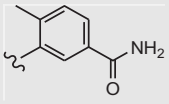
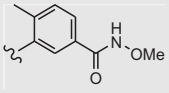
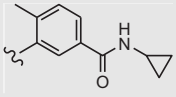
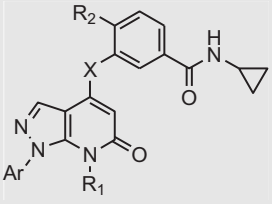
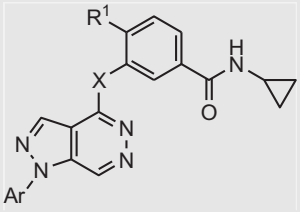
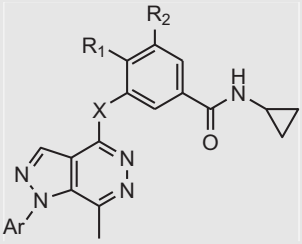
No.	Ar	R_1	R_2	X	Exp.	SW-MLR	GA-MLR
							
1	2,5-Di-F-Ph		–	–	6.35	6.35	6.31
2	2,4-Di-F-Ph		–	–	8.59	8.49	8.49
3	2,5-Di-F-Ph		–	–	8.43	8.45	8.62
4	2,4-Di-F-Ph		–	–	8.96	8.95	8.82
5	2,4-Di-F-Ph		–	–	8.49	8.68	8.50
							
6	3-F-Ph	Me	Me	NH	8.48	8.53	8.56
7	3-F-Ph	Me	Me	O	8.07	8.07	8.06
8[a]	4-F-Ph	Me	Cl	NH	8.82	8.98	8.79
9	4-F-Ph	Et	Me	NH	8.4	8.37	8.43
10	4-F-Ph	Et	Cl	NH	8.92	8.89	8.86
11	4-F-Ph	Et	Me	O	8.4	8.23	8.31
12	2,4-Di-F-Ph	Me	Me	NH	8.49	8.56	8.48
13	2,4-Di-F-Ph	Me	Cl	NH	9	9.01	8.86
14	2,4-Di-F-Ph	Et	Cl	NH	8.96	8.98	8.96
15	2,4-Di-F-Ph	Me	F	NH	7.89	7.80	8.09
16	2,4-Di-F-Ph	Me	Me	O	8.11	8.32	8.09
17	2,5-Di-F-Ph	Me	Me	NH	8.54	8.53	8.51
18	2,5-Di-F-Ph	Me	Cl	NH	8.96	8.96	8.91
19	2,5-Di-F-Ph	Me	Me	O	8.21	8.27	8.22
20[a]	2,5-Di-F-Ph	Me	Cl	O	8.48	8.86	8.16
21	2,6-Di-F-Ph	Me	Me	NH	8.72	8.57	8.86
22[a]	2,6-Di-F-Ph	Me	Cl	NH	8.96	8.89	9.20
23[a]	2,6-Di-F-Ph	Me	F	NH	8.72	7.67	8.42

Table 1 (*Continued*)

No.	Ar	R_1	R_2	X	Exp.	SW-MLR	GA-MLR
							
24	2-Cl-Ph	Cl	–	NH	8.8	8.89	8.87
25[a]	3-F-Ph	Me	–	NH	8.34	8.09	8.26
26	3-F-Ph	Cl	–	NH	8.3	8.24	8.42
27	4-F-Ph	Cl	–	NH	8.07	8.42	8.43
28	2,4-Di-F-Ph	Me	–	NH	7.92	8.10	7.97
29	2,4-Di-F-Ph	Me	–	NH	8.52	8.48	8.50
30	2,4-Di-F-Ph	F	–	NH	7.37	7.45	7.47
31	2,4-Di-F-Ph	Me	–	O	8	8.11	8.06
32[a]	2,4-Di-F-Ph	Cl	–	O	7.44	8.70	7.85
33[a]	2,5-Di-F-Ph	Cl	–	NH	8.68	8.87	8.42
34	2,6-Di-F-Ph	Me	–	NH	8.77	8.80	8.75
35	2,6-Di-F-Ph	Cl	–	NH	8.85	8.90	9.02
36	2,6-Di-F-Ph	Me	–	O	8.52	8.46	8.36
							
37	4-F-Ph	Me	H	NH	8.39	8.25	8.26
38	4-F-Ph	Cl	H	NH	8.64	8.42	8.43
39[a]	4-F-Ph	Me	H	O	8.11	8.09	8.06
40	2,4-Di-F-Ph	Me	H	NH	8.57	8.47	8.50
41[a]	2,4-Di-F-Ph	Me	F	NH	8	8.96	8.33
42	2,4-Di-F-Ph	Cl	H	NH	8.8	8.65	8.72
43	2,4-Di-F-Ph	Me	H	O	8.2	8.12	8.06
44	4-F-Ph	Me	H	NH	8.77	8.70	8.71

[a] Used as test set.

the MLR analysis with a stepwise selection and the variables elimination was employed to model the quantitative structure–activity relationships with a different set of descriptors. In order to build and test model, a data set of 44 compounds was randomly separated into a training set of 35 compounds (80%), which was used to build model and a prediction set of 9 compounds (20%), which was applied to test the built model. The selection of the test set molecules was with respect to distribution in the range of the biological data for the whole set, and their structure diversity. The SW-MLR analysis led to the derivation of one model, with six variables (the closest to the ratio of five training molecules for each descriptor (Hansch et al., 1990)) with low generalization and prediction ability for the test set. It is described by the following equation:

$$\begin{aligned} \mathrm{pIC}_{50} = {} & -25.704(\pm 5.288) + 5.715(\pm 0.775)\mathrm{MWC09} \\ & + 0.480(\pm 0.142)\mathrm{GGI2} \\ & - 13.916(\pm 1.124)\mathrm{GATS5p} + 1.500(\pm 0.644)\mathrm{H5v} \\ & - 3.575(\pm 0.546)\mathrm{R6u} + 12.632(\pm 1.529)\mathrm{Ui} \end{aligned} \quad (1)$$

$$\begin{aligned} N_{\mathrm{train}} = 35,\ & R^2_{\mathrm{train}} = 0.948,\ \mathrm{RMSE}_{\mathrm{train}} = 0.116,\ Q^2_{\mathrm{LOO}} \\ & = 0.476,\ Q^2_{\mathrm{LGO}} = 0.064,\ Q^2_{\mathrm{BOOT}} = 0.844,\ F = 84.851,\ N_{\mathrm{test}} \\ & = 9,\ R^2_{\mathrm{test}} = 0.001,\ \mathrm{RMSE}_{\mathrm{test}} = 0.656 \end{aligned}$$

In this equation, N is the number of compounds, R^2 is the squared correlation coefficient, RMSE is the root mean square error, Q^2_{LOO}, Q^2_{LGO} and Q^2_{BOOT} are the squared cross-validation

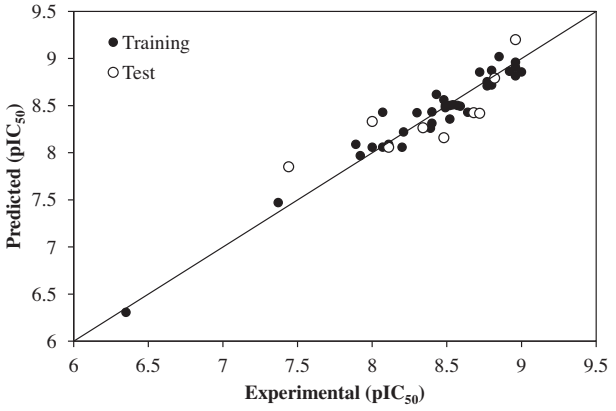


Figure 1 The predicted versus the experimental pIC_{50} values by the GA-MLR modeling.

coefficients for leave one out, leave group out and bootstrapping respectively, and F is the Fisher F statistic. However, this procedure produced acceptable results for the training set, but it did not produce good results for the test set ($R^2_{test} = 0.001$). Therefore the genetic algorithm was used to select the best set of variables. The best model has six parameters because the increase in the number of molecular descriptors has no significant effect on the accuracy of the best model. After the selection of the most important descriptors by genetic algorithm, MLR was performed to build the linear model. This equation and its statistical parameters are presented as:

$$\begin{aligned} pIC_{50} = {} & 10.452(\pm 3.759) + 0.458(\pm 0.042)\text{X2sol} \\ & - 6.257(\pm 1.658)\text{BEHv8} \\ & + 241.370(\pm 24.793)\text{JGI9} \\ & + 9.659(\pm 0.969)\text{GATS4p} \\ & - 5.189(\pm 0.586)\text{HATS8u} \\ & - 6.182(\pm 2.227)\text{R4m}^{+} \end{aligned} \quad (2)$$

$$\begin{aligned} N_{train} = 35, R^2_{train} = 0.946, \text{RMSE}^2_{train} = 0.118, Q^2_{LOO} \\ = 0.916, Q^2_{LGO} = 0.861, Q^2_{BOOT} = 0.894, F = 81.104, N_{test} \\ = 9, R^2_{test} = 0.673, \text{RMSE}_{test} = 0.259 \end{aligned}$$

With the test set, the prediction results were obtained. The experimental and predicted values based on the GA-MLR model are shown in Table 1. Also, Fig. 1 shows the predicted versus experimental pIC_{50} for all of the 44 compounds studied, the training set and the test set. As can be seen, the predicted values for the pIC_{50} are in good agreement with those of the experimental values.

As can be seen from Eqs. (1) and (2), the R^2 and RMSE values in test set improved from 0.001 and 0.656 by SW-MLR model to 0.673 and 0.259 by GA-MLR model respectively. The results illustrated once more that the linear MLR technique combined with a successful variable selection procedure is adequate to generate an efficient QSAR model for predicting the pIC_{50} of compounds.

3.2. Evaluation of the GA-MLR model

The quality of the QSAR model was characterized by the number of compounds used in the study (N), coefficient of determination (R^2), root mean square error (RMSE), and variance ratio (F). For a more exhaustive testing of the predictive power of the model, validation of the model was also carried out using the leave one out (LOO) and the leave group out (LGO) cross-validation techniques on the training set of compounds. For LOO cross-validation, a data point is removed from the set, and the model is recalculated. The predicted pIC_{50} for that point is then compared with its actual value. This is repeated until each data point has been omitted once. For LGO, 20% of the data points are removed from the dataset and the model was refitted; the predicted values for those points were then compared with the experimental values. Again, this is repeated until each data point has been omitted once. The robustness of the proposed models and their predictive ability were also guaranteed by the high Q^2_{BOOT} based on bootstrapping repeated 5000 times (Wehrens et al., 2002). The results produced by the LOO ($Q^2 = 0.915$) and the LGO ($Q^2_{LGO} = 0.860$) cross-validation tests and bootstrapping ($Q^2_{BOOT} = 0.894$) illustrated the quality of the obtained model. Because all of the validation techniques show the obtained GA-MLR model is a valid model so, it can be used to predict the inhibition activity of the components.

The Williams plot, the plot of the standardized residuals versus the leverage, was exploited to visualize the applicability domain (AD) (Netzeva et al., 2005). Leverage indicates a compound's distance from the centroid of X. The leverage of a compound in the original variable space is defined as:

$$h_i = x_i^T (X^T X)^{-1} x_i \quad (3)$$

where x_i is the descriptor vector of the considered compound and X is the descriptor matrix derived from the training set descriptor values. The warning leverage (h^*) is defined as:

$$h^* = 3(p+1)/n \quad (4)$$

Where n is the number of training compounds, p is the number of predictor variables. From the Williams plot (Fig. 2), it is obvious that all compounds in the test set fall inside the domain of the GA-MLR model (the warning leverage limit is 0.60). There are only two chemicals (No. **7** and No. **24** in the training set) which have the leverage higher than the warning h^* value, thus they can be regarded as structural outliers. Fortunately, in this case the data predicted by the model are good for compound numbers **7** and **24**, thus they are "good

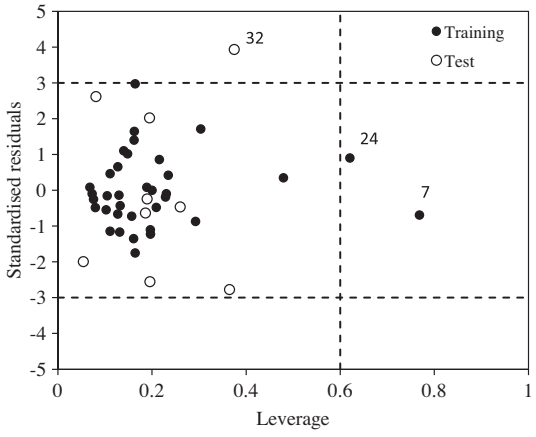


Figure 2 William plot of GA-MLR model.

Table 2 The R^2_{train} and Q^2_{LOO} values after several Y- randomization tests.

Iteration	R^2_{train}	Q^2_{LOO}
1	0.099	0.062
2	0.023	0.141
3	0.053	0.322
4	0.004	0.183
5	0.043	0.149
6	0.069	0.197
7	0.021	0.164
8	0.014	0.269
9	0.002	0.135
10	0.000	0.165

leverage" chemicals. For all the compounds in the training and test sets, their standardized residuals are smaller than three standard deviation units (3δ) except compound number **32**. Thus compound **32** can be as outlier. Because this compound is one of the test set compounds, there is no need to remove this compound from the data set.

The GA-MLR model was further validated by applying *Y*-randomization. Several random shuffles of the *Y* vector (pIC_{50}) were performed and the low R^2 and Q^2 values that were obtained showing that the good results in the original model is not due to a chance correlation or structural dependency of the training set. The results of the *Y*-randomization test are presented in Table 2. The brief description of the selected descriptors by GA-MLR model is summarized in Table 3. The correlation matrix of the six selected descriptors is included in Table 4. From Table 4, it can be seen that the linear correlation coefficient value of each of the two descriptors is <0.692, which means the descriptors are independent in the analysis.

The multi-collinearity between the above six descriptors was detected by calculating their variation inflation factors (VIF), which can be calculated as follows:

$$VIF = 1/1 - r^2 \quad (5)$$

where *r* is the correlation coefficient of the multiple regression between the variables in the model. If VIF equals to 1, then no inter-correlation exists for each variable; if VIF falls into the range of 1–5, the related model is acceptable; and if VIF is larger than 10, the related model is unstable and a recheck is necessary. The corresponding VIF values of the six descriptors are shown in Table 3. As can be seen from this table, most of the variables have VIF values of less than 5, indicating that the obtained model has statistic significance.

3.2.1. Interpretation of the descriptors

The best six-parameter equation for prediction of pIC_{50} for an unknown compound included X2sol, BEHv8, JGI9, GATS4p, HATS8u and R4m+ descriptors. To examine the relative importance as well as the contribution of each descriptor in the model, the value of the mean effect (MF) was calculated for each descriptor (Massart et al., 1997). The MF value indicates the relative importance of a descriptor, compared with the other descriptors in the model. The mean effect values are shown in Table 3. As can be seen the BEHv8, GATS4p and X2sol descriptors have great mean effect values than the other descriptors which means that these descriptors have a large effect on the pIC_{50} of the studied compounds.

The first descriptor is X2sol (solvation connectivity index chi-2), which represents the linear fragment of one carbon atom that is defined in order to model solvation entropy and to describe dispersion interaction in solution. The descriptor

Table 3 The linear model based on the six parameters selected by the GA-MLR method.

Descriptor	Chemical meaning	MF[a]	VIF[b]
Constant	Intercept	–	–
X2sol	Solvation connectivity index chi-2	−2.929	1.990
BEHv8	Highest eigenvalue n. 8 of Burden matrix/weighted by atomic van der Waals volumes	8.294	2.495
JGI9	Mean topological charge index of order9	−0.997	1.111
GATS4p	Geary autocorrelation – lag 4/weighted by atomic polarizabilities	−4.438	1.302
HATS8u	Leverage-weighted autocorrelation of lag 8/unweighted	0.939	1.401
R4m+	R maximal autocorrelation of lag 4/weighted by atomic masses	0.131	1.312

[a] Mean effect.
[b] Variation inflation factors.

Table 4 Correlation coefficient matrix of the selected descriptors by GA-MLR.

	X2sol	BEHv8	JGI9	GATS4p	HATS8u	R4m+
X2sol	1					
BEHv8	0.692	1				
JGI9	0.194	0.144	1			
GATS4p	0.052	0.319	0.099	1		
HATS8u	0.093	0.345	0.157	0.073	1	
R4m+	−0.172	−0.298	−0.026	0.206	−0.503	1

X2sol, solvation connectivity index chi-2, has a positive regression coefficient; hence it has shown positive influence on the activity. Thus, suggesting that a higher value of 2nd order solvation connectivity index would be favorable to the activity.

The second descriptor is BEHv8 (highest eigenvalue number 8 of Burden matrix/weighted by atomic van der Waals volumes), which is one of the BCUT descriptors. The BCUT (Burden, CAS, University of Texas) descriptors are the eigenvalues of a modified connectivity matrix known as the Burden matrix (Burden, 1989). Comparison of the mean effects of the descriptors appearing in the GA-MLR model shows that the BEHv8 of the molecules has the largest effect on the pIC_{50} of the studied compounds. The BEHv8 has a highest mean effect value with negative correlation coefficient in Eq. (2). It can be concluded that BEHv8 displays a great effect in the model and the atomic van der Waals volumes of a molecule are inversely related to pIC_{50} value.

AJGI9 is the mean topological charge index of order 9 which belongs to the Galvez topological charge indices (Todeschini and Consonni, 2000). These indices describe charge transfer between pairs of atoms and therefore global charge transfer in a molecule. This descriptor has a positive sign which indicates that the pIC_{50} is directly related to this descriptor.

Another descriptor in the GA-MLR model is GATS4p (Geary autocorrelation – lag 4/weighted by atomic polarizabilities). The GATS4p belongs to 2D-autocorrelation descriptors (2D) (Todeschini and Consonni, 2000). This set consists of 96 descriptors calculated from the molecular graph by summing the products of atom weights of the terminal atoms of all the paths of the considered path length (the lag).The molecule atoms represent the set of discrete points in space and the atomic property the function evaluated at those points. The physico-chemical property in this case is atomic polarizabilities. GATS4p displays a positive sign, which indicates that the pIC_{50} value is directly related to this descriptor. Hence, it was concluded that by increasing the atomic polarizabilities of a molecule, the value of this descriptor increased, causing an increasing in its pIC_{50} value.

The HATS8u descriptor is one of the GETAWAY descriptors. The GETAWAY (GEometry, Topology, and Atom-Weights AssemblY) descriptors have been recently proposed as chemical structure descriptors derived from a new representation of molecular structure, the molecular influence matrix (MIM) (Consonni et al., 2002). HATS8u is the leverage-weighted, autocorrelation of lag 8/unweighted. The negative sign of the corresponding regression coefficient between pIC_{50} and this descriptor indicates that the pIC_{50} increase with the decrease of these descriptor value.

The final descriptor of the GA-MLR model was the R maximal autocorrelation of lag 4/weighted by atomic masses (R4m+) which is one of the GETAWAY descriptors (Todeschini and Consonni, 2000). This descriptor is related to the mass of the atoms in the molecule. The R4m+ descriptor displays a negative sign, which indicates that the pIC_{50} is inversely related to this descriptor.

From the above discussion we concluded that the solvation connectivity index, the atomic van der Waals volumes, the global charge transfer in a molecule, the atomic polarizabilities and the atomic masses in a molecule play a main role in the p38α MAP kinas inhibition activity of compounds.

4. Conclusion

In the present study, two variable selection methods of stepwise and genetic algorithm were used to construct a quantitative relation between the p38α MAP kinase inhibition activity of pyrazole derivatives and their calculated descriptors. Both methods resulted in a training set with good statistical significance. GA-MLR was superior to SW-MLR at external predictions. Also the solvation connectivity index, atomic van der Waals volumes, global charge transfer in a molecule, atomic polarizabilities and atomic masses proved to be important factors controlling the inhibitory activity of p38α MAP inhibitors. Additionally, the proposed method could also identify and provide some insight into what structural features are related to the inhibitory activity of compounds.

References

Ahmad, S., Gromiha, M.M., 2003. Design and training of a neural network for predicting the solvent accessibility of proteins. J. Comput. Chem. 24, 1313.

Aires-De-Sousa, J., Hemmer, M.C., Gasteiger, J., 2002. Prediction of h-1 nmr chemical shifts using neural networks. Anal. Chem. 74, 80.

Alsberg, B.K., Marchand-Geneste, N., King, R.D., 2000. A new 3d molecular structure representation using quantum topology with application to structure-property relationships. Chemom. Intell. Lab. Syst. 54, 75.

Bhatia, N.M., Mahadik, K.R., Bhatia, M.S., 2010. Exploring the influence of steric, electronic and lipophilic descriptors of 1,3-diarly propenones on their anti-inflammatory activity. DARU 18, 230.

Burden, F.R., 1989. Molecular-identification number for substructure searches. J. Chem. Inf. Comput. Sci. 29, 225.

Consonni, V., Todeschini, R., Pavan, M., 2002. Structure/response correlations and similarity/diversity analysis by getaway descriptors. 1. Theory of the novel 3d molecular descriptors. J. Chem. Inf. Comput. Sci. 42, 682.

Depczynski, U., Frost, V.J., Molt, K., 2000. Genetic algorithms applied to the selection of factors in principal component regression. Anal. Chim. Acta 420, 217.

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

Habibi-Yangjeh, A., Pourbasheer, E., Danandeh-Jenagharad, M., 2008b. Prediction of melting point for drug-like compounds using principal component-genetic algorithm-artificial neural network. Bull. Korean Chem. Soc. 29, 833.

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

Hansch, C., Taylor, J., Sammes, P., 1990. Comprehensive Medicinal Chemistry: The Rational Design, Mechanistic Study & Therapeutic Application of Chemical Compounds. Pergamon, New York.

Hemmateenejad, B., Yousefinejad, S., Mehdipour, A.R., 2011. Novel amino acids indices based on quantum topological molecular similarity and their application to QSAR study of peptides. Amino acids 40, 1169.

Hunger, J., Huttner, G., 1999. Optimization and analysis of force field parameters by combination of genetic algorithms and neural networks. J. Comput. Chem. 20, 455.

Jouanrimbaud, D., Massart, D.L., Leardi, R., Denoord, O.E., 1995. Genetic algorithms as a tool for wavelength selection in multivariate calibration. Anal. Chem. 67, 4295.

Margutti, S., Laufer, S.A., 2007. Are map kinases drug targets? Yes, but difficult ones. ChemMedChem 2, 1116.

Massart, D.L., Vandeginste, B.G.M., Buydens, L.M.C., Jong, S.D.E., Leui, P.J., Smeyers-Verbeke, J., 1997. Hand Book of Chemometrics and Qualimetrics: Part A. Elsevier, The Netherlands.

Melagraki, G., Afantitis, A., Makridima, K., Sarimveis, H., Igglessi-Markopoulou, O., 2006. Prediction of toxicity using a novel rbf neural network training methodology. J. Mol. Model. 12, 297.

Netzeva, T.I., Worth, A.P., Aldenberg, T., Benigni, R., Cronin, M.T.D., Gramatica, P., Jaworska, J.S., Kahn, S., Klopman, G., Marchant, C.A., Myatt, G., Nikolova-Jeliazkova, N., Patlewicz, G.Y., Perkins, R., Roberts, D.W., Schultz, T.W., Stanton, D.T., Van De Sandt, J.J.M., Tong, W., Veith, G., Yang, C., 2005. Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. ATLA – Altern. Lab. Anim. 33, 155.

O'keefe, S.J., Mudgett, J.S., Cupo, S., Parsons, J.N., Chartrain, N.A., Fitzgerald, C., Chen, S.L., Lowitz, K., Rasa, C., Visco, D., Luell, S., Carballo-Jane, E., Owens, K., Zaller, D.M., 2007. Chemical genetics define the roles of p38 alpha and p38 beta in acute and chronic inflammation. J. Biol. Chem. 282, 34663.

Pettus, L.H., Wurz, R.P., 2008. Small molecule p38 map kinase inhibitors for the treatment of inflammatory diseases: novel structures and developments during 2006–2008. Curr. Top. Med. Chem. 8, 1452.

Shahlaei, M., Fassihi, A., Saghaie, L., Arkan, E., Pourhossein, A., 2011. A qsar study of some cyclobutenediones as ccr1 antagonists by artificial neural networks based on principal component analysis. DARU 19, 376.

Todeschini, R., Consonni, V., 2000. Handbook of Molecular Descriptors. Wiley-VCH, Weinheim.

Wagner, G., Laufer, S., 2006. Small molecular anti-cytokine agents. Med. Res. Rev. 26, 1.

Waller, C.L., Bradley, M.P., 1999. Development and validation of a novel variable selection technique with application to multidimensional quantitative structure-activity relationship studies. J. Chem. Inf. Comput. Sci. 39, 345.

Wehrens, R., Putter, H., Buydens, L.M.C., 2002. The bootstrap: a tutorial. Chemom. Intell. Lab. Systems 54, 35.

Westra, J., Limburg, P.C., 2006. P38 mitogen-activated protein kinase (mapk) in rheumatoid arthritis. Mini-Rev. Med. Chem. 6, 867.

Wurz, R.P., Pettus, L.H., Xu, S.M., Henkle, B., Sherman, L., Plant, M., Miner, K., Mcbride, H., Wong, L.M., Saris, C.J.M., Lee, M.R., Chmait, S., Mohr, C., Hsieh, F., Tasker, A.S., 2009. Part 1: Structure–activity relationship (sar) investigations of fused pyrazoles as potent, selective and orally available inhibitors of p38 alpha mitogen-activated protein kinase. Bioorg. Med. Chem. Lett. 19, 4724.

Wurz, R.P., Pettus, L.H., Henkle, B., Sherman, L., Plant, M., Miner, K., Mcbride, H.J., Wong, L.M., Saris, C.J.M., Lee, M.R., Chmait, S., Mohr, C., Hsieh, F., Tasker, A.S., 2010. Part 2: Structure–activity relationship (sar) investigations of fused pyrazoles as potent, selective and orally available inhibitors of p38 alpha mitogen-activated protein kinase. Bioorg. Med. Chem. Lett. 20, 1680.

Yousefinejad, S., Hemmateenejad, B., Mehdipour, A.R., 2012. New autocorrelation QTMS-based descriptors for use in QSAM of peptides. J. Iran. Chem. Soc. 9, 569.