ICC

Iranian Chemical Communication

Original Research Article

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

Quantitative structure-activity relationship (QSAR) study of CCR2b receptor inhibitors using SW-MLR and GA-MLR approaches

Mehdi Nekoei

Department of Chemistry, Faculty of Basic Sciences, Shahrood Branch, Islamic Azad University, Shahrood, Iran

Received: 6 February 2016, Accepted: 2 May 2016, Published: 2 May 2016

Abstract

In this paper, the quantitative structure activity-relationship (QSAR) of the CCR2b receptor inhibitors was scrutinized. Firstly, the molecular descriptors were calculated using the Dragon package. Then, the stepwise multiple linear regressions (SW-MLR) and the genetic algorithm multiple linear regressions (GA-MLR) variable selection methods were subsequently employed to select and implement the prominent descriptors having the most significant contributions to the activities of the molecules. A combined data set including numerical values of inhibition activity data (IC_{50}) of 103 CCR2b receptor derivatives was adopted for our simulations. This study revealed that both SW-MLR and GA-MLR methods consisted of six molecular descriptors. The adopted descriptors belong to topological, charge, RDF and atom-centered fragments classes. A comparison of results by the two methodologies indicated the superiority of GA-MLR over the SW-MLR method. The authenticity of the proposed model (GA-MLR) was further confirmed using the cross-validation, validation through an external test set and *Y*-randomization.

Keywords: Quantitative structure-activity relationship (QSAR); CCR2b receptor inhibitors; genetic algorithm (GA); stepwise (SW); multiple linear regression (MLR); molecular descriptor.

Introduction

Chemokines are a family of low molecular weight secreted proteins acting as leukocyte specific chemoattractants. C-C chemokine receptors type 2 (CCR2) belonging to the G protein coupled receptors (GPCRs) family are expressed on monocytes, macrophages, basophiles, mast cells and T lymphocytes. There are two alternatively spliced forms of CCR2 receptors, namely CCR2a and CCR2b which differ only in their carboxyl-terminal tails. These receptors play important roles in the recruitment of monocytes/macrophages, T cells, and are directly related to many diseases such as inflammation, HIV and pulmonary fibrosis. CCR2 receptors are implicated in a diversity of inflammatory responses by interaction with chemokine receptors situated in the cell surface of leukocytes followed

*Corresponding author: Mehdi Nekoei

Tel: +98 (23) 32394530, Fax: +98 (23) 32394537

E-mail: m_nekoei1356@yahoo.com

by chemotaxis and infiltration into the adjacent tissue [1].

Monocyte chemotactic protein-1 (MCP-1) is a member of the CC-chemokines family and has been implicated in the acute diseases such as atherosclerosis[2], rheumatoid arthritis [3], glomerulonephritis [4] and multiple sclerosis [5]. Using a trial and error approach to evaluate the activity and property of chemical compounds and medicines is a time-consuming and costly process. The quantitative-structure activity relationship (QSAR) methodologies are powerful ways to overcome this restriction. These techniques are usually based on statistically determined linear or non-linear models that correlate the chemical behavior of compounds and their descriptors. The main interest in development of predictive QSAR models is owing to their high capability of predicting activities and/or properties of compounds, particularly for those which are experimentally immeasurable for many reasons, including their instability, toxicity or cost. In fact, QSARs create mathematical relationships between chemical, physical, biological or environmental activities of measurable and computable parameters such as physicochemical, stereochemical, and topological descriptors [6].

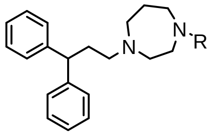
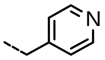
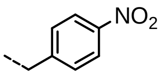
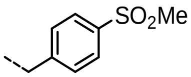
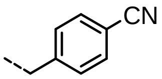
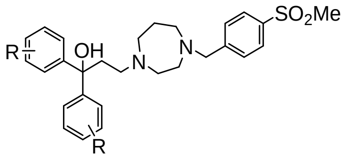
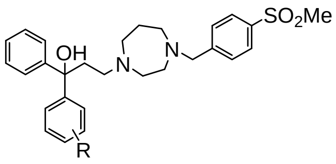
Nowadays, the use of statistical approaches has attracted an increased interest [7]. There are various tools to control, optimize and predict diverse physical and chemical properties of broad sets of organic compounds. Of these, one is the genetic algorithms (GA) [8-10] and other ones are the stepwise (SW) [11-13] and particle swarm optimization (PSO) [14,15] feature selection methods. The application of QSAR technique usually requires a reliable variable selection method for building well-fitted models. In this work, we used the genetic algorithm (GA) and stepwise (SW) methods for the variable selection in combination with multiple linear regression (MLR) strategy. According to the obtained results, these approaches enable us to get satisfactory statistical parameters engaged with the simulation of inhibition activities (IC_{50}) of CCR2b receptors. Moreover, to the best of our knowledge, this is the first report concerning predictions of activities of CCR2b receptors using SW-MLR and GA-MLR methods.

Computational section

Data set and methods

An integrated data set consisting of inhibition activities (IC_{50}) of 103 homopiperazine and diamine derivatives selected from three recently-published papers was used for QSAR analyses [16-18]. The inhibition activity data [IC_{50} (nM)] for the CCR2b receptor derivatives were converted to a logarithmic scale pIC_{50} [-log IC_{50} (M)] and subsequently used for exploring QSARs as the response variables. The chemical structures of the studied compounds and their corresponding pIC_{50} values are displayed in Table1.

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

No.	**R_1**	**R_2**	**Exp.**	**SW-MLR**	**GA-MLR**
					
1		-	4.96	4.58	4.61
2		-	4.72	4.80	4.88
3		-	4.89	4.51	4.72
4[a]		-	4.37	4.64	4.64
					
5	H	-	5.39	5.16	5.43
6	4-NMe_2	-	4.35	4.58	4.91
7[a]	4-OH	-	4.68	5.62	5.17
8	3-OH	-	5.82	5.63	5.60
9	4-F	-	5.15	5.57	5.24
10	3-F	-	4.47	5.49	5.32
11	4-Cl	-	4.96	5.65	5.44
					

No.	**R$_1$**	**R$_2$**	**Exp.**	**SW-MLR**	**GA-MLR**
12	3-OH	-	6.15	5.46	5.67
13	3-CH$_2$OH	-	5.40	5.34	5.47
14[a]	3-NH$_2$	-	5.38	5.57	5.84
15	3-NHMe	-	5.30	5.46	5.47
16[a]	3-OMe	-	5.10	5.35	5.42
17	3-F	-	4.92	5.29	5.25
18[a]	3-Me	-	4.80	5.37	5.74

Table 1. Continued

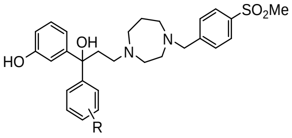
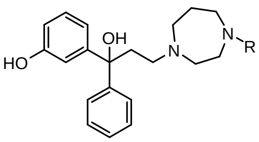
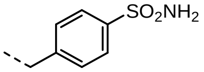
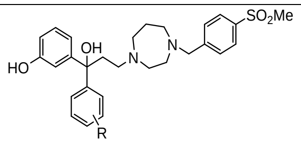
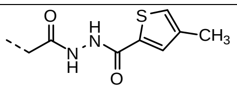
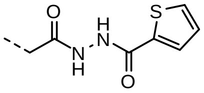
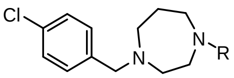
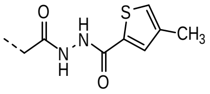
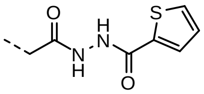
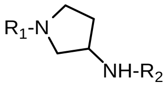
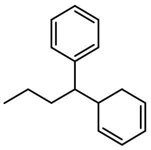
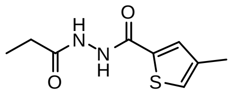
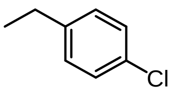
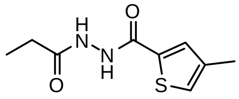
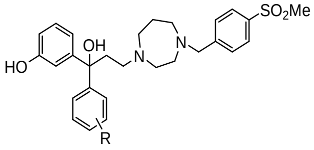
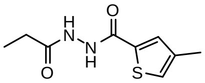
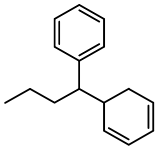
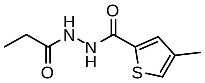
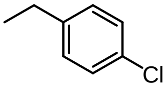
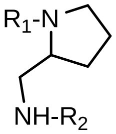
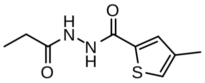
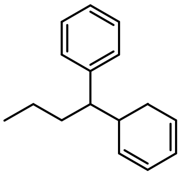
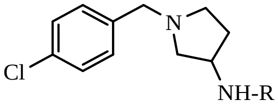
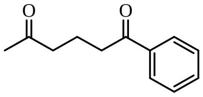
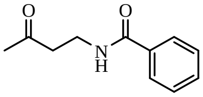
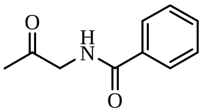
No.	**R$_1$**	**R$_2$**	**Exp.**	**SW-MLR**	**GA-MLR**
					
19	3-F	-	5.62	5.53	5.70
20	3-Cl	-	5.35	5.58	5.23
21	4-F	-	5.82	5.49	4.85
22	4-Cl	-	5.82	5.67	5.59
23	3,5-DiF	-	5.03	5.64	5.18
					
24		-	5.82	5.45	5.68

Table 1. Continued

No.	R_1	R_2	Exp.	SW-MLR	GA-MLR
					
25[a]		-	5.19	4.05	4.23
26		-	4.55	3.99	4.17
					
27		-	5.13	4.94	4.95
28		-	4.52	4.09	4.13
					
29[a]			4.48	4.45	4.34
30			4.39	4.70	4.93

No.	R$_1$	R$_2$	Exp.	SW-MLR	GA-MLR
					
31			4.77	5.17	4.99
32			5.11	5.05	5.32
					
33			4.59	4.78	5.13
					
34[a]		-	4.96	5.11	4.82
35		-	4.66	4.85	4.63
36		-	4.38	4.87	4.96

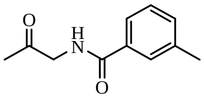
37		-	5.16	5.30	5.37

Table 1. Continued

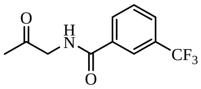
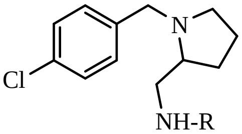
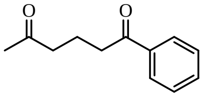
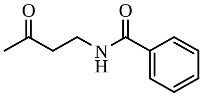
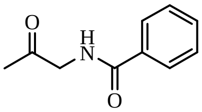
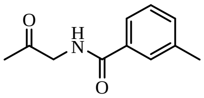
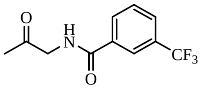
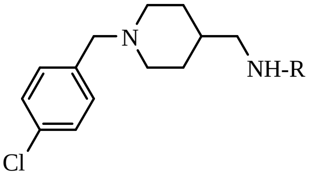
No.	R$_1$	R$_2$	Exp.	SW-MLR	GA-MLR
38		-	6.43	5.86	5.92
					
39		-	4.72	4.20	4.42
40		-	4.40	4.16	4.37
41		-	4.18	4.31	4.23
42		-	4.85	4.73	4.78
43		-	5.64	5.63	6.02
					

Table 1. Continued

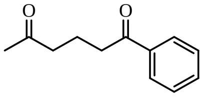
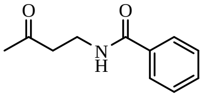
No.	R_1	R_2	Exp.	SW-MLR	GA-MLR
44		-	4.30	4.73	4.47
45		-	4.19	4.76	4.35

Table 1. Continued

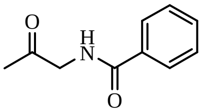
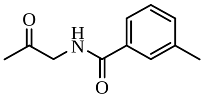
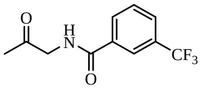
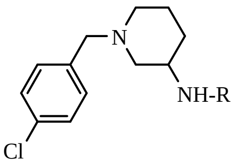
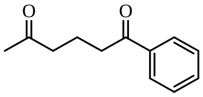
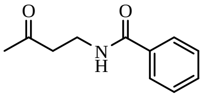
No.	R_1	R_2	Exp.	SW-MLR	GA-MLR
46		-	4.96	4.85	4.74
47		-	5.26	5.32	5.42
48[a]		-	6.15	6.07	5.91
					
49		-	4.06	4.03	4.18
50[a]		-	4.13	4.32	4.41

Table 1. Continued

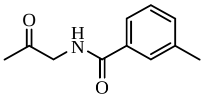
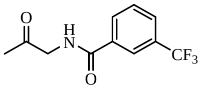
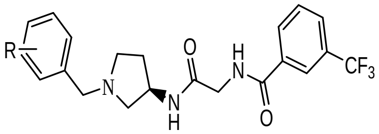
No.	**R$_1$**	**R$_2$**	**Exp.**	**SW-MLR**	**GA-MLR**
51		-	4.77	4.78	4.72
52		-	6.18	5.60	5.84
					
53	H	-	6.16	6.01	6.11
54	2-Cl	-	6.20	6.22	6.24
55	2-CH$_3$	-	6.02	6.46	6.68
56	2-OCH$_3$	-	5.87	6.48	6.48
57	3-CH$_3$	-	5.51	6.33	6.51
58	3-OCH$_3$	-	5.50	6.35	6.29
59	4-Cl	-	6.74	6.13	6.23
60	4-CH$_3$	-	6.94	6.71	6.70
61	4-OCH$_3$	-	6.94	6.69	6.63
62	4-Et	-	7.23	6.98	7.02
63[a]	4-Br	-	6.78	6.56	6.76
64	4-Vinyl	-	6.92	6.59	6.72
65	4-CH$_3$S	-	6.69	6.65	6.53
66[a]	4-OH	-	6.65	6.59	6.64
67	4-NHAc	-	6.52	6.96	6.05

Table 1. Continued

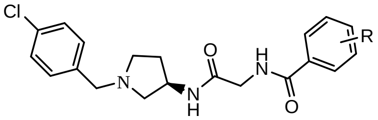
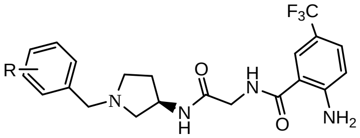
No.	R_1	R_2	Exp.	SW-MLR	GA-MLR
68	4-OCF_3	-	6.21	7.07	6.47
69[a]	4-F	-	6.02	6.77	6.40
70[a]	4-NO_2	-	5.81	6.97	6.76
71[a]	4-CN	-	5.58	6.29	6.12
72	2,4-$(CH_3)_2$	-	7.27	6.71	6.96
73	2,4-Cl_2	-	6.52	6.41	6.42
74	4-OH, 3-OCH_3	-	6.82	6.83	6.78
75	2-Naphthyl	-	6.12	5.91	5.74
					
76	3-CH_3	-	5.62	5.33	5.50
77	3-Cl	-	5.62	5.32	5.36
78	4-CH_3	-	5.00	5.64	5.66
79	3-F	-	5.36	5.31	5.45
80	3-Br	-	6.11	5.94	5.51
81	3-OCF_3	-	6.31	6.20	6.15
82	3-NO_2	-	6.08	5.98	5.93
83	2-NH_2, 5-NO_2	-	6.68	6.54	6.62
84	2-NH_2, 5-Cl	-	6.14	6.02	6.03
85	2-NH_2, 5-Br	-	6.19	6.69	6.47
86	2-NH_2, 5-I	-	6.51	6.82	6.25
87	2-NH_2, 5-OCF_3	-	7.06	7.17	6.63

Table 1. Continued

No.	R_1	R_2	Exp.	SW-MLR	GA-MLR
88[a]	2-NH_2, 5-CF	-	7.59	7.23	7.20
					
89[a]	4-Cl	-	7.59	7.47	7.53
90[a]	4-Br [a]	-	6.94	7.58	7.28
91	4-CH_3	-	7.70	7.25	7.51
92[a]	4-Et	-	7.96	7.44	7.53
93	4-Vinyl	-	7.72	7.54	7.53
94	4-OCH_3[a]	-	7.70	7.53	7.48
95	4-OH	-	7.38	7.39	7.44
96[a]	4-Cl, 3-NH_2	-	8.39	7.82	8.06
97	4-CH_3, 3-NH_2	-	8.21	8.05	8.24
98	4-OCH_3, 3-NH_2	-	8.28	8.24	8.43
99	4-OH, 3-NH_2	-	7.85	7.91	8.03
100[a]	4-OCH_3, 3-OH	-	7.28	7.85	7.84
101	4-OH, 3-OCH_3	-	7.41	7.57	7.40
102	2,4-$(CH_3)_2$	-	8.49	7.91	8.22
103	2,4-Cl_2	-	7.02	7.43	7.62

[a]Used as test set

Software

A Pentium IV personal computer with the Windows XP operating system was used. The geometry optimization of compounds using the (MM+) and (AM1) methods was performed using the HYPERCHEM 7.0 package. The GA-MLR and the other advanced calculations were performed in the MATLAB 7.0 environment.

Descriptors calculation and selection

The calculation and selection of the descriptors as reliable parameters representing the chemical structures from diverse points of view are of prime importance in QSAR-based simulations. Dragon software was used to calculate chemical descriptors in a broad spectrum of such studies. It should be pointed out that calculation of these descriptors is easy and fast. Using Dragon software, 1481 descriptors for each molecule were calculated. In a preliminary step, constant and near constant variables were eliminated because they do not interpret meaningful concepts related to the structure of compounds in the data set. Variables with correlation coefficients higher than 0.9 were selected for developing suitable models. Finally, the remaining descriptors were collected in an n × m data matrix, where n = 103 and m = 356 refer to the numbers of compounds and descriptors, respectively. Among the descriptors mentioned earlier, the most significant molecular descriptors were identified using the genetic algorithm (GA) and stepwise (SW) methods.

Genetic algorithm

The genetic algorithms (GAs) take inspiration from natural selection, Darwin’s evolutionary theory and other genetic functions, e.g. cross-over and mutation. GAs have a great potential for solving certain types of difficult problems in fast, suitable and credible ways [19]. Nowadays, GA is one of the most widely used variable selection methods. It was developed by John Holland in 1975 and in recent years it has been utilized to resolve and optimize a variety of problems [20-24]. In Figure 1, a flowchart on genetic algorithm process has been demonstrated to give a deeper insight to the process operation.

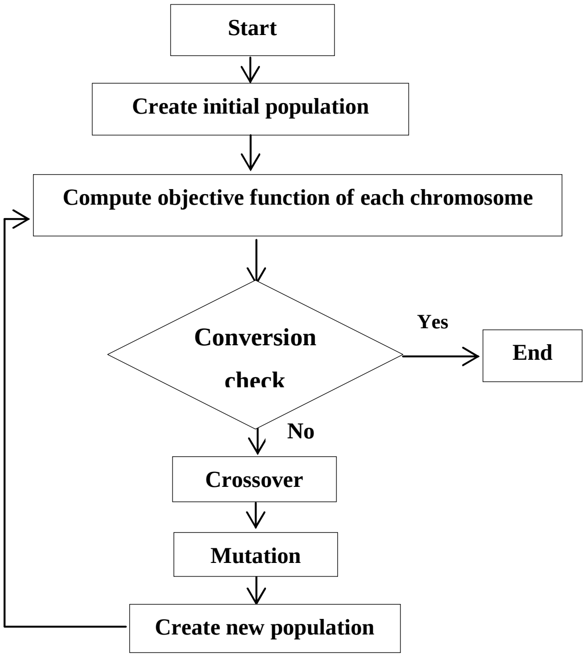

Start
Create initial population
Compute objective function of each chromosome
Conversion check
Yes
End
No
Crossover
Mutation
Create new population

Figure 1. General flowchart of the genetic algorithm approach

To select the most related descriptors, the evolution of the population was simulated [25-27]. The first step in GA performance is the random selection of individuals for generation of the first population. Each individual in the population, was defined by a chromosome of binary (0 & 1) values. Number of genes equals selected descriptors. For the genes encoded with binary system, if gene was given the value of 1, its corresponding descriptor was included in the subset; otherwise, it was given the value of zero [28]. The number of the genes with the value=1 was kept relatively low to have a small subset of descriptors. The operators used here were cross-over and mutation. The population size was varied over the range 50-250 for different GA runs. For a typical run, the evolution of the generation was stopped when 90% of the generations achieved the same fitness.

Results and discussion

In this attempt, we employed both the GA-MLR and SW-MLR techniques for the selection of the most significant descriptors. In the first step, we employed the variable elimination step prior to the MLR analysis followed by the use of SW selection to model the structure-activity relationship with a different set of descriptors. Training and test sets including 82 (80%) and 21 (20%) compounds, respectively, were randomly selected from a date set of 103 compounds. The SW-MLR analysis led to the derivation of a model including six variables. The linear model is described by the following equation:

pIC_{50}= -19.296 (±4.523) + 25.306 (±6.985) X0A + 0.119 (±0.016) PCWTe + 3.388 (±0.836) Jhetp +0.036 (±0.008) RDF065m - 0.315 (±0.064) RDF145v - 0.087 (±0.030) H-052 (eq. 1)

N_{train}= 82, R^2_{train}=0.887, $RMSE_{train}$= 0.372, Q^2_{LOO}=0.858, Q^2_{LGO}=0.838, Q^2_{BOOT}=0.850, F=97.855, N_{test}=21, R^2_{test}=0.807, $RMSE_{test}$= 0.573

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 coefficients for “leave one out,” “leave group out” and “bootstrapping,” respectively, and *F* is the Fisher F statistic.

In another part of our study, the genetic algorithm was employed as the variable selection procedure to select the best variables, while the MLR was carried out to build the model. This equation and its statistical parameters are:

pIC_{50}= -16.25 (±5.382) + 44.168 (±6.261) X0A - 112.295 (±17.837) X5A - 0.095 (±0.015) MDDD + 0.092 (±0.014) PCWTe - 0.241 (±0.049) RDF145m + 0.062 (±0.014) RDF065p (eq. 2)

N_{train}= 82, R^2_{train}=0.895, $RMSE_{train}$= 0.358, Q^2_{LOO}=0.871, Q^2_{LGO}=0.833, Q^2_{BOOT}=0.867, F=106.387, N_{test}=21, R^2_{test}=0.862, $RMSE_{test}$= 0.477

Subsequently, the built model was used to compute the pIC_{50} values of the compounds present in the test set. In accordance with the GA-MLR simulation, the R^2 values for training and test sets were found to be 0.895 and 0.862, respectively. Therefore, when using GA-MLR the coefficient of determination (R^2) for the training and test sets are more than those obtained by SW-MLR.

Of the six descriptors used by SW-MLR, two (X0A and Jhetp), two (RDF065m and RDF145v), one (PCWTe) and one (H-052) respectively belong to topological class, RDF, charge and atom-centred fragments descriptors. On the other hand, in the model developed using GA-MLR, half of descriptors are topological descriptors (X0A, X5A and MDDD), while two (RDF145m and RDF065p) and one (PCWTe) are related to RDF and charge descriptors. Another important point is implementation of six molecular descriptors, of which two (X0A and PCWTe) are common in both models.

The experimental and predicted values based on GA-MLR and SW-MLR models are shown in Table 1. In addition, the general statistical parameters of the two models are summarized in Table 2. A simple comparison of the proposed linear models (SW-MLR and GA-MLR) models shows that the RSME of GA-MLR method for both training and test sets were lower than the RMSE of SW-MLR method, whereas the R^2 and Q^2_{LOO} of GA-MLR were higher than those of SW-MLR (Table 2). From Table 2, it can be seen that GA-MLR model also gives higher *F* values; therefore, one can conclude that the GA-MLR is more robust and better than the SW-MLR.

Table 2. Statistical results of SW-MLR and GA-MLR models

	Training set			Test set		
	R^2	RMSE	F	R^2	RMSE	F
SW-MLR	0.887	0.372	97.855	0.807	0.573	10.130
GA-MLR	0.895	0.358	106.387	0.862	0.477	14.083

As can be seen from equations 1 and 2 and their results, the value of R^2 is enhanced in test set over the range 0.807-0.862 by SW-MLR and GA-MLR models, respectively. From the results, it is clear that the MLR technique combined with SW and GA variable selection procedures generated productive QSAR models for predicting the pIC_{50} of compounds. However, in view of the superiority of GA-MLR, we focused on it in further studies.

Figure 2 shows the predicted pIC_{50} values versus the experimental ones using the GA-MLR modeling approach. Accordingly, a convergency of points towards the most probable line (R^2_{train}=0.895; R^2_{test}=0.862) is seen for both training and test sets.

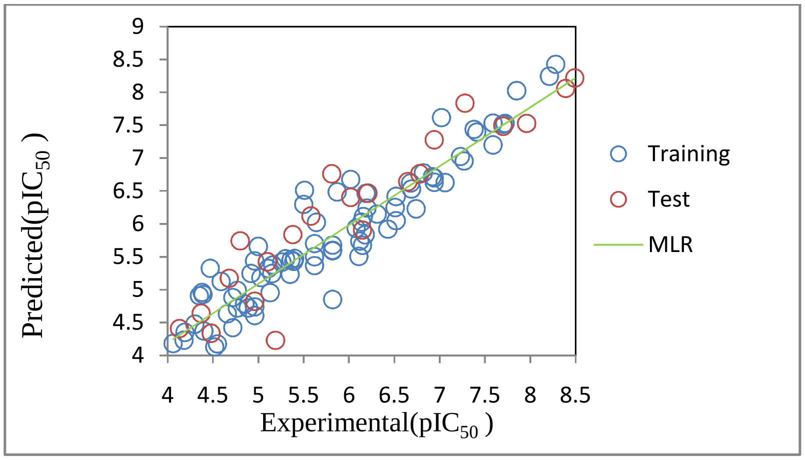


Figure 2. The predicted versus the experimental pIC_{50} values by the GA-MLR modeling

An important step in QSAR study is evaluating predictive capability of models. In this study, we employed cross-validation utilizing different variables, such as “leave one out” (LOO) and “leave group out” (LGO), the number of compounds (N), the coefficient of determination (R^2), the root mean square error (RMSE) and the variance ratio (F). The Q^2_{LOO} was calculated by the following equation:

$$Q^2_{LOO} = 1 - \frac{PRESS}{SSR} = 1 - \frac{\sum_{i=1}^{n}(y_{\exp} - y_{pred})^2}{\sum_{i=1}^{n}(y_{pred} - \overline{y}_{pred})^2}$$

(eq. 3)

The y_{exp} and y_{pred} values are the experimental and predicted values for training set, and $\overline{y}_{pred}$ is the mean experimental value of the samples in the training set.

The sturdiness of the proposed models and its predictive abilities were assured through the high Q^2_{BOOT} approach[29]. The results of the LOO (Q^2_{LOO} = 0.871) and the LGO (Q^2_{LGO} = 0.833) cross-validation tests and bootstrapping (Q^2_{Boot}= 0.867) reveal that the proposed model is of satisfactory quality. Therefore, since all of the validation techniques confirm the validity of the GA-MLR model, it can be used to predict the inhibition activity of the components.

To evaluate the applicability domain (AD) of a model, application of a plot of standardized cross-validated residuals against leverage values, namely the William plot, has been proposed by Gramatica [30]. This plot is primarily employed to recognize the response outliers and structurally influential compounds in the model. The leverage indicates the distance of a compound from the centroid of X and for a compound in the original variable space is defined as[31]:

$$h_i = x_i^T \left(X^T X\right)^{-1} x_i \qquad \text{(eq. 4)}$$

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 [32]

h^* = 3 (p+1)/ n (eq. 5)

where n and p are the number of training compounds and number of predictor variables, respectively. The presence of outliers and compounds structurally influential in determining model parameters (i.e. compounds with high leverage value (h) greater then warning leverage (h*)), was verified by the William plot acquired by plotting hat values versus standardized residuals. From Figure 3, it is immediately evident that two compounds (No. 5 and No. 21 in the training set) have leverage values higher than the warning leverage h* value, thus they can be regarded as structural outliers. Fortunately, in two aforementioned cases, the respective data predicted by the model are good; thus they are considered as ‘‘good leverage’’ compounds. In summary, the prediction plot (Figure 2) and applicability domain (AD) plot (Figure 3), confirm the suitability of the built model and appropriate divisions of the whole data set into training and test sets.

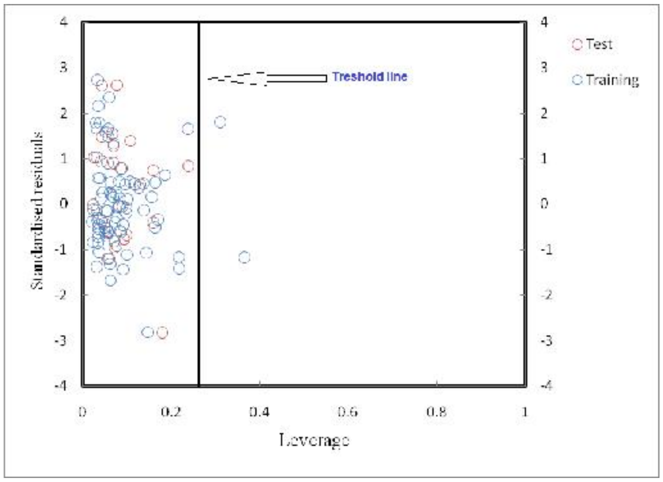


Figure 3. William plot of GA-MLR model

A brief description of the selected descriptors by GA-MLR model is summarized in Table 3. As this Table shows, six utilized descriptors for prediction of pIC_{50} of CCR2b receptors include X0A, X5A, MDDD, RDF145m and RDF065p. The first, second and third implemented molecular descriptors in the developed GA-MLR model, namely X0A (Average connectivity index chi-0) and X5A (Average connectivity index chi-5), MDDD (mean distance degree deviation) belong to the topological class. The positive and negative signs attributed to these variables denote, respectively, their direct and inverse relationships with the pIC_{50} value. Subsequently, the increase in these average connectivity indices as well as mean distance degree deviation of the molecules results in an increase and decrease in their pIC_{50}, respectively. Topological descriptors include valence and non-valence molecular connectivity indices calculated from the hydrogen-suppressed formula of the molecule, encoding information about the size, composition and the degree of branching of a molecule. In fact, these descriptors are based on the graphical representations of the molecules and are considered as numerical quantifiers of molecular topology calculated through the application of algebraic operators to matrices representing molecular graphs whose values are independent of vertex numbering or labeling. Topological descriptors are generally sensitive to one or more structural features of the molecule involving size, shape, symmetry, branching and cyclicity, encoding chemical information concerning atom type and bond multiplicity.

The fourth descriptor is PCWTe. It occurs among the charge descriptors and implies a partial charge weighted topological electronic descriptor. The positive sign of PCWTe shows its reinforcement impact on pIC_{50}.

RDF145m and RDF065p are the fifth and sixth descriptors appearing in the model. RDF145m represents the Radial distribution function – 14.5 / weighted by atomic masses whereas RDF065p is the Radial distribution function – 6.5 / weighted by atomic polarizabilities. RDF145m and RDF065p are two members of the radial distribution function (RDF) descriptors. The RDF descriptors are based on the geometrical interatomic distance and constitute a radial distribution function code[33].

RDF145m descriptors possess a negative sign, indicating that pIC_{50} is inversely related to this descriptor; therefore, increasing the RDF145m of molecules leads to an appreciable decrease in their respective pIC_{50} values. On the other hand, the other radial distribution function (RDF) descriptor namely RDF065p has a positive sign. Therefore, it exerts a constructive influence on pIC_{50} value.

Table 3. Concise description of the six parameters selected by the proposed GA-MLR method

Descriptor	**Chemical meaning**	**MF_j [a]**	**VIF [b]**
Constant	Intercept	-	-
X0A	Average connectivity index chi-0	1.433	2.304
X5A	Average connectivity index chi-5	-0.420	2.818
MDDD	Mean distance degree deviation	-0.156	3.669
PCWTe	Partial charge weighted topological electronic descriptor	0.098	1.991
RDF145m	Radial distribution function – 14.5 / weighted by atomic masses	-0.005	1.261
RDF065p	Radial distribution function – 6.5 / weighted by atomic polarizabilities	0.050	3.090

[a]Mean effect
[b]Variation inflation factors

In order to appraise the robustness of the model, the *Y*-randomization test was employed [34]. Accordingly, the dependent variable vector (pIC_{50}) is randomly shuffled and a new QSAR model is built using the variable matrix. The newly obtained models are anticipated to have low R^2 and Q^2 values showing that the good results in the original model used were not due to a chance correlation or structural dependency of the training set. The results of the *Y*-randomization test are presented in Table 4. The correlation matrix of the six selected descriptors is included in Table 5 and indicates the linear correlation coefficient value of each pair of descriptors. As seen from numerical values of R^2 from bivariate analysis, selected descriptors behave independently in the proposed model.

Table 4. The R^2_{train} and Q^2_{LOO} values after several *Y*-randomization tests

Iteration	R^2_{train}	Q^2_{LOO}
1	0.099	0.000
2	0.027	0.082
3	0.068	0.000
4	0.088	0.000
5	0.081	0.000
6	0.077	0.000
7	0.037	0.065
8	0.055	0.013
9	0.061	0.005
10	0.125	0.015

Table 5. Correlation coefficient matrix of the selected descriptors by GA-MLR

	X0A	X5A	MDDD	PCWTe	RDF145m	RDF065p
X0A	1					
X5A	-0.424	1				
MDDD	0.093	-0.733	1			
PCWTe	0.640	-0.145	-0.056	1		
RDF145m	-0.107	-0.201	0.284	-0.347	1	
RDF065p	-0.011	-0.624	0.801	-0.001	0.131	1

To verify the inter correlation of descriptors, variance inflation factor (VIF) analysis was performed, which can be calculated as follows:

$$VIF = \frac{1}{1-r^2}, \qquad \text{(eq. 6)}$$

where r is the multiple correlation coefficient. The corresponding *VIF* values of the six descriptors are shown in Table 3. If a value of *VIF* falls within the range 1-5, the related model is acceptable. As can be seen in Table 3, all of the variables have *VIF* values of less than 5, indicating that the obtained model is significant.

To examine the relative importance as well as the contribution of each descriptor in the model, the value of the mean effect (MF_j) was calculated for each descriptor [35], which is defined as:

$$MF_j = \frac{j\sum_{i=1}^{n} d_{ij}}{\sum_{i}^{m} j\sum_{i}^{n} d_{ij}} \qquad \text{(eq. 7)}$$

The *MF_j* value indicates the relative importance of a descriptor, compared with the other descriptors in the model. The mean effect values of each molecular descriptor are shown in Table 3. The results explicitly argue that the X0A topological descriptor with the highest MF_j value has the most effect in the linear constructed model.

Conclusion

Quantitative relationships between molecular structures and CCR2b inhibitory activities of 103 amine derivatives were discovered by two linear regression methods (SW-MLR and GA-MLR). Both methods resulted in training sets with good statistical significance. Results show that GA-MLR has a superior power in modeling this relationship. The exactness and prediction capability of the proposed models is illustrated using various criteria such as cross-validation and *Y*-randomization. The prediction results and the experimental values are in good agreement. In addition, the average connectivity index, mean distance degree deviation, partial charge weighted topological electronic descriptor and radial distribution function are seen to be important factors controlling the inhibitory activity of CCR2b inhibitors. The proposed method also shows that structural features are related to the inhibitory activities of compounds.

Acknowledgements

The authors express their special thanks to the office for research affair of Tehran University and Islamic Azad University, Shahrood Branch for financial supports of the research proposal and the time dedicated for performing the calculations of the study.

References

[1] J. Saunders, C.M. Tarby, *Drug Discov. Today*, **1999**, *4*, 80-92.

[2] M. Navab, S.Y. Hama, T.B. Nguyen, A.M. Fogelman, *Coron. Artery Dis.*, **1994**, *5*, 198-204.

[3] J.H. Gong, L.G. Ratkay, J.D. Waterfield, I. Clarc-Lewis, *J. Exp. Med.*, **1977**, *186*, 131-137.

[4] C.M. Lloyd, A.W. Minto, M.E. Dorf, A. Proudfoot, T.N.C. Wells, D.J. Salant, J.C. Gutierrz-Ramos, *J. Exp. Med.*, **1977**, *185*, 1371-1380.

[5] K.J. Kennedy, R.M. Strieter, S.L. Kunkel, N.W. Lukacs, W.J. Karpus, *J. Neuroimmunol.*, **1998**, *92*, 98-108.

[6] V. Consonni, R. Todeschini, M. Pavan, *J. Chem. Inf. Comput. Sci.*, **2002**, *42*, 682-692.

[7] K. Benyounis, A. Olabi, *Adv. Eng. Softw.*, **2008**, *39*, 483-496.

[8] M. Mohammadhosseini, *Asian J. Chem.*, **2012**, *24*, 3814-3820.

[9] M. Mohammadhosseini, O. Deeb, A. Alavi-Gharabagh, M. Nekoei, *Anal. Chem. Lett.*, **2012**, *2*, 80-102.

[10] M. Mohammadhosseini, M. Nekoei, *Asian J. Chem.*, **2013**, *25*, 349-352.

[11] M. Mohammadhosseini, H.A. Zamani, H. Akhlaghi, M. Nekoei, *J. Essent. Oil-Bear. Plants*, **2011**, *14*, 559-573.

[12] E.P. Box, N.R. Draper, Empirical Model Building and Response Surfaces, Wiley and Sons, EUA, 1987.

[13] M. Nekoei, M. Mohammadhosseini, A. Alavi-Gharahbagh, *Anal. Bioanal. Electrochem.*, **2009**, *1*, 159-168.

[14] M. Mohammadhosseini, *Anal. Chem. Lett.*, **2013**, *3*, 226-248.

[15] M. Mohammadhosseini, *J. Chem. Health Risks*, **2014**, *4*, 75-95.

[16] M. Imai, T. Shiota, K.-i. Kataoka, C.M. Tarby, W.J. Moree, T. Tsutsumi, M. Sudo, M.M. Ramirez-Weinhouse, D. Comer, C.-M. Sun, *Bioorg. Med. Chem. Lett.*, **2004**, *14*, 5407-5411.

[17] W.J. Moree, K.-i. Kataoka, M.M. Ramirez-Weinhouse, T. Shiota, M.

Imai, M. Sudo, T. Tsutsumi, N. Endo, Y. Muroga, T. Hada, *Bioorg. Med. Chem. Lett.*, **2004**, *14*, 5413-5416.

[18] W.J. Moree, K.-i. Kataoka, M.M. Ramirez-Weinhouse, T. Shiota, M. Imai, T. Tsutsumi, M. Sudo, N. Endo, Y. Muroga, T. Hada, *Bioorg. Med. Chem. Lett.*, **2008**, *18*, 1869-1873.

[19] D. Goldberg, Genetic Algorithms & Engineering Optimization, Wiley, New York, 1989.

[20] M. Nekoei, M. Salimi, M. Dolatabadi, M. Mohammadhosseini, *Monatsh. Chem.*, **2011**, *142*, 943-948.

[21] M. Nekoei, N. Goudarzi, S. Nekoei, M. Mohammadhosseini, *Anal. Chem. Lett.*, **2014**, *4*, 14-28.

[22] M. Nekoei, M. Mohammadhosseini, E. Pourbasheer, *Med. Chem. Res.*, **2015**, *24*, 3037-3046.

[23] M. Nekoei, M. Salimi, M. Dolatabadi, M. Mohammadhosseini, *J. Serb. Chem. Soc.*, **2011**, *76*, 1117-1127.

[24] G. Mitsuo, C. Runwei, Genetic Algorithms & Engineering Optimization, John Wiley & Sons, 2002.

[25] S. Ahmad, M.M. Gromiha, *J. Comput. Chem.*, **2003**, *24*, 1313-1320.

[26] J. Hunger, G. Huttner, *J. Comput. Chem.*, **1999**, *20*, 455-471.

[27] C.L. Waller, M.P. Bradley, *J. Chem. Inf. Comput. Sci.*, **1999**, *39*, 345-355.

[28] M. Shahlaei, A. Fassihi, L. Saghaie, E. Arkan, A. Pourhossein, *Daru*, **2011**, *19*, 376.

[29] R. Wehrens, H. Putter, L.M. Buydens, *Chemometr. Intell. Lab. Syst.*, **2000**, *54*, 35-52.

[30] P. Gramatica, *QSAR Comb Sci.*, **2007**, *26*, 694.

[31] S. Riahi, M.R. Ganjali, E. Pourbasheer, P. Norouzi, *Chromatographia*, **2008**, *67*, 917-922.

[32] S. Riahi, E. Pourbasheer, M.R. Ganjali, P. Norouzi, *Chem. Biol. Drug Des.*, **2009**, *73*, 558-571.

[33] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, John Wiley & Sons2008.

[34] D. Bonchev, Information-Theoretic Indices for Characterization of Chemical Structures, RSP/Willey,, Chichester, UK, 1983.

[35] D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S.D.E. Jong, Leui, P. J., J. Smeyers-Verbeke, Hand Book of Chemometrics and Qualimetrics: Part A., Elsevier1977.