

Quantitative structure activities relationships of some 2-mercaptoimidazoles as CCR2 inhibitors using genetic algorithm-artificial neural networks

L. Saghaie¹, M. Shahlai^{2,*} and A. Fassihi¹

¹Department of Medicinal Chemistry and Isfahan Pharmaceutical Research Center, School of Pharmacy and Pharmaceutical Sciences, Isfahan University of Medical Sciences, Isfahan, I.R. Iran.

²Department of Medicinal Chemistry and Nano Drug Delivery Research Center, Faculty of Pharmacy, Kermanshah University of Medical Sciences, Kermanshah, I.R. Iran.

Abstract

Quantitative relationships between structures of twenty six of 2-mercaptoimidazoles as C-C chemokine receptor type 2 (CCR2) inhibitors were assessed. Modeling of the biological activities of compounds of interest as a function of molecular structures was established by means of genetic algorithm multivariate linear regression (GA-MLR) and genetic algorithm (GA-ANN). The results showed that, the pIC₅₀ values calculated by GA-ANN are in good agreement with the experimental data, and the performance of the artificial neural networks regression model is superior to the multivariate linear regression-based (MLR) model. With respect to the obtained results, it can be deduced that there is a non-linear relationship between the pIC₅₀s and the calculated structural descriptors of the 2-mercaptoimidazoles. The obtained models were able to describe about 78% and 93% of the variance in the experimental activity of molecules in training set, respectively. The study provided a novel and effective approach for predicting biological activities of 2-mercaptoimidazole derivatives as CCR2 inhibitors and disclosed that combined genetic algorithm and GA-ANN can be used as a powerful chemometric tools for quantitative structure activity relationship (QSAR) studies.

Keywords: Quantitative structure activity relationship; Multivariate linear regression; Artificial neural networks; CCR2 inhibitors; 2-mercaptoimidazoles

INTRODUCTION

Chemokines or chemotactic cytokines are a large group of small (~ 8-15 kDa) proteins that relate to each other structurally and functionally and insert a significant function in leukocyte migration and activation (1-5). Chemokines mediate their influences through activation of particular proteins in surface of the cells belonging to well-known seven transmembrane spanning G-protein coupled receptors family (GPCR). Monocyte chemoattractant protein (MCP-1/CCL2) is a part of the CC chemokine subgroup which is attached to the CC chemokine receptor 2 (CCR2) expressed on the greater number of blood born monocytes (6). Disturbance of the MCP-1/CCR2 route in rodent models of inflammatory and autoimmune diseases by genetic deletion of either MCP-1(7)

or CCR2 (8-10) and use of peptidyl CCR2 antagonists (11) or anti-MCP-1 antibodies (12) propose that inhibition of CCR2 may supply possible therapies for a variety of sicknesses including rheumatoid arthritis (11,12) multiple sclerosis (13-15) and atherosclerosis (10, 16-18). These outlooks have stimulated the search for small molecule MCP-1/CCR2 antagonists in a large number of research laboratories (19). Lately, quantitative structure activity relationships (QSARs) have been employed widely to generate models in order to calculate and predict biological or toxicological values of drug candidate compounds using computational descriptors solely extracted from molecular structure.

For the first time, McCulloch and Pitts (20) proposed artificial neural networks (ANN) as a technique of data mining employing a neural

*Corresponding author: M. Shahlai, this paper is extracted from the Ph.D thesis NO. 389440
Tel. 0098 831 4276489, Fax. 0098 831 4276493
Email: mshahlai@kums.ac.ir

network's information processing units (neurons) as centers of data analyzing that are organized in layers. An ANN is a tool for processing the input information. ANN is built based on the structure and function of the human brain as a template. Central nervous system of human is consisted of a series of neurons interconnected to each other by synapses. Information transfer between these neurons via a series of action potentials has been proved by scientists (21). Various ANN algorithms have advantages such as adaptive learning behavior, capability of parallel distributed processing and good generalization property for unseen data.

The ANN method has several benefits over traditional regression methods, since they need known input data set without any suppositions (22). The ANN generates a mapping of the input and output variables, which can afterwards be employed to predict wanted output as a function of appropriate inputs (23). A multi layer neural network can estimate any smooth and measurable relationship between input and output vectors by optimizing a fitted set of connecting weights and transfer functions (22). ANN models could describe any nonlinear relationship between calculated descriptors of drug like compounds and their bioactivities (24,25). Therefore; it is more desirable to

apply a non parametric method such as feed forward back propagation neural network QSAR modeling to characterize such a nonlinear relationship (26).

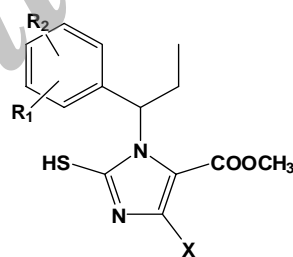
Here we describe multiple linear regressions (MLR) as a linear method and back propagation ANN as a nonlinear technique for investigating of the relationship between the structure and the CCR2 antagonist activity of some 2-mercaptoimidazoles compounds. We further make a comparison between the two different methods to verify their efficacy in modeling in the inhibitory activity of the studied compounds.

MATERIALS AND METHODS

Computer hardware, software and preparation of data set

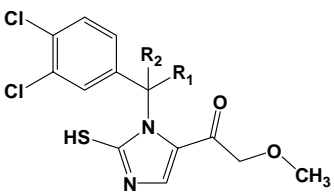
All calculations were run on a desktop computer with Windows XP operating system. Bioactivities of 26 C-C chemokine receptor type 2 (CCR2) antagonists were taken from the literature (27), and are presented in Table 1. These values were converted from IC_{50} to pIC_{50} (-logarithm of IC_{50}). The two-dimensional structures of molecules were drawn by Hyperchem 7.0 software. The ultimate conformations were calculated with the semi-empirical AM1 technique.

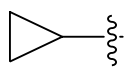
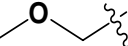
Table 1. Structures of some 2-mercaptoimidazoles as CCR2 Inhibitors used in this study

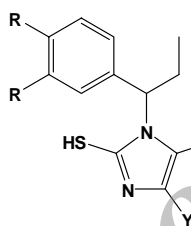


Compound	R ₁	R ₂	X	pIC ₅₀
1	3-Cl	4-Cl	H	6.7
2	3-F	4-F	H	5.9
3	3-Br	4-Br	H	6.6
4	3-F	4-CF ₃	H	5.7
5	3-CF ₃	4-F	H	6
6	3-Cl	4-Cl	COOCH ₃	8
7	3-F	4-F	COOCH ₃	8
8	3-F	4-CF ₃	COOCH ₃	7.7
9	3-CF ₃	4-F	COOCH ₃	7
10	3-Cl	4-Cl	CH ₃	6.6

Table 1. (Continued)



Compound	R ₁	R ₂	pIC ₅₀
11	H	Methyl	6.6
12	H	Propyl	6.8
13	H		5.7
14	H		6.5
15	H	Cyclohexyl	5



Compound	R	X	Y	pIC ₅₀
16	Cl	CONH ₂	H	7.3
17	Cl	CONHCH ₃	H	6.5
18	Cl	CN	H	6.2
19	F	COOEt	H	6.5
20	Cl	COOi-Prop	H	6.5
21	F	CONH ₂	H	6.3
22	F	CONHCH ₃	H	5.7
23	Cl	CONHCH ₃	CONHCH ₃	6.4
24	Cl	CONH ₂	CONH ₂	7.2
25	Cl	CN	CONH ₂	6.9
26	Cl	CN	COOCH ₃	7.4

The molecular structures were optimized using the Polak-Ribiere algorithm until the root mean square gradient was 0.01 kcal mol⁻¹. The z-matrix of structures was provided by the Hyperchem and transferred to the Gaussian 98 program (28). Whole conformation optimization was carried out taking the most extended conformation as starting geometries. Semi empirical molecular orbital calculation (AM1) of the structures was performed again to avoid trapping in local minimal using Gaussian 98 program. The obtained conformation was

relocated to Dragon program package, which was developed by Milano Chemometrics and QSAR Group (29). Dragon software was employed to calculate a large number of descriptors including geometrical, topological, functional group and constitutional. The name and number of calculated descriptors are listed in Table 2. After calculation of descriptors, in the preprocessing step, the estimated descriptors were investigated for descriptors that have constant values for all studied molecules and those discerned were deleted from data matrix.

Table 2. Short description of some descriptors used in this study including their name and number.

Type of descriptor	Molecular description	Number of calculated descriptor
Constitutional	Mean atomic van der Waals volume (Mv) (scaled on Carbon atom), no. of heteroatoms, no. of multiple bonds (nBM), no. of rings, no. of circuits, no of H-bond donors, no of H-bond acceptors, no. of Nitrogen atoms (nN), chemical composition, sum of Kier-Hall	32
Topological	electrotopological states (Ss), mean atomic polarizability (Mp), number of rotatable bonds (RBN) mean atomic Sanderson electronegativity (Me), etc. Narumi harmonic topological index (HNar), total structure connectivity index (Xt), information content index (IC), mean information content on the distance degree equality (IDDE), total walkcount, path/walk-Randic shape indices (PW3, PW4, PW5, Zagreb indices, Schultz indices, Balaban J index (such as MSD) Wiener indices, Information content index (neighborhood	231
Geometrical	symmetry of 2-order (IC2), ratio of multiple path count to path counts (PCR), Lovasz-Pelikan index (leading eigenvalue) (LP1), total information content index (neighborhood symmetry of 1-order) (TIC1), reciprocal hyper-detour index (Rww), Average connectivity index chi-5 (X5A), piID (conventional bond-order ID number), etc. 3D Petijean shape index (PJI3), asphericity (ASP), gravitational index, Balaban index, Wiener index, length-to-breadth ratio by WHIM (L/Bw), etc.	39
Functional group	Number of total secondary C(sp ³) (nC _s), number of total tertiary carbons (nC _t), number of H bond acceptor atoms (nHAcc), number of secondary amides (aliphatic) (nCONHR), number of unsubstituted aromatic C (nCaH), number of ethers (aromatic) (nRORPh), number of ketones (aliphatic) (nCO), number of tertiary amines (aliphatic) (nNR ₂), number of phenols (nOHPh), number of total primary C(sp ³) (nC _p), etc.	17

To reduce the redundancy existed in the calculated descriptors, the correlation among descriptors and with the bioactivity of the molecules was checked and collinear descriptors (i.e. $R^2 > 0.90$) were detected. Among the collinear descriptors, one with the highest correlation with bioactivity kept for model building phase and the others were removed.

MATLAB software (version 7.1 Math Work Inc.) was used for developing some scripts to perform ANN regression modeling and model validation.

The data set was split into a training set and a prediction set using Kenard and Stones algorithm (30). According to Tropsha the best models would be built when this algorithm was used (31). The training set of 21 molecules was employed to adjust the parameters of the developed QSAR models, and the test set of 5 compounds was employed to assess its prediction capability.

Feature selection using genetic algorithm

Where the number of independent variables is more than investigated molecules, feature selection is necessary for avoiding chance correlation and selecting the most informative descriptors. However, selecting the sufficient and informative descriptors for biological activity in QSAR studies is not easy because there are no universal rules that manage this selection. Genetic algorithm (GA) is one of the best methods to feature selection in model building. The GA used here was demonstrated in other literature (33) and does not present for brevity.

Multiple linear regression

The general purpose of multivariate linear regressions (MLR) is to quantify the relationship between several independent or predictor variables and a dependent variable. Independent or predictor variables could be various physicochemical descriptors of

molecules, their principle components or latent variables. After building the model, the activity value of each ligand would then be calculated using the developed model. A multi linear model can be represented as:

$$y = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + \dots + a_nx_n + \beta \quad (1)$$

where n is the number of independent variables, a_1, \dots, a_n denote the regression coefficients, β is the error and y is the dependent variable. Regression coefficients signify the independent contributions of each molecular descriptor.

Artificial neural networks

An itemized explanation of the theory behind ANN has been sufficiently explained elsewhere (34-36). The pertinent rule of supervised learning in an ANN is that it obtains numerical types inputs (the training data) and conveys them into preferred outputs. The input and output neurons may be joined to the 'external world' and to other neurons inside the artificial network. The method in which each neuron conveys its input is dependent on the so called weights and bias of the neurons, which are adjustable. The output values of each neuron rely on both the weight strengths and bias values. Also, the outputs depend on the weighted sum of all its inputs which are usually conveyed using a nonlinear weighting function. For the at hand goals, the big strength of ANNs systems arise from this fact that it is conceivable to train this systems. Training is carried out through successively introducing the networks with certain inputs and outputs and adjusting the connection weights and biases between the individual neurons. This procedure is corroborated until the output neurons of the network match the preferred outputs to a desired degree of accuracy. Though, training can be carried out by using the back propagation algorithm. In order to train the network using this algorithm, the differences between the ANNs output and its preferred output are estimated after each epoch. The changes in the values of the weights can be calculated by using following equation:

$$\Delta w_{ij}(n) = \eta \delta_i O_j + \alpha \Delta w_{ij}(n-1) \quad (2)$$

where, Δw is the change in the values of weights for each network neuron, δ_i is the actual error of neuron i , and O_j is the output of neuron j . The coefficients η and α are the learning rate and the momentum factor, respectively. These coefficients manage the velocity and the efficacy of the learning course. These parameters would be optimized before training the network. Equation like Equation (2) can be employed for the bias settings.

The ANN can apply qualitative as well as quantitative inputs, and it does not need an unambiguous relationship connecting the inputs and the outputs. Though in statistics the analysis is limited to a known number of possible interactions, more expressions can be checked for interactions by the ANNs. In addition, by permitting more information to be analyzed at the same time, more complicated and delicate interactions can be investigated using this method.

Validation of QSAR models

Some of common parameters used for checking predictability of proposed models are root mean square error (*RMSE*), square of the correlation coefficient (R^2), an predictive residual error sum of squares (*PRESS*). These parameters were calculated for each model as follows:

$$RMSE = [1/n \sum_{i=1}^n (\hat{y}_i - y_i)^2]^{1/2} \quad (3)$$

$$R^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 / \sum_{i=1}^n (y_i - \bar{y})^2 \quad (4)$$

$$PRESS = \sum_{i=1}^n (\hat{y}_i - y_i)^2 \quad (5)$$

where, y_i is the true bioactivity of the investigated compound i , \hat{y}_i represents the calculated bioactivity of the compound i , \bar{x} the mean of true activity in the studied set, and n the total number of molecules used in the studied sets.

The value of R^2 can be usually raised by adding the additional independent variables to the generated model, even if the added independent variable does not cause to the decrease of the unexplained variance of the dependent variable. Consequently, the use of

needs particular consideration. Hence, it is better to employ another statistics, known as *adjusted R²* (R_{adj}^2) in association with R^2 . R_{adj}^2 can be calculated using following equation:

$$R_{adj}^2 = 1 - (1 - R^2) \left(\frac{n-1}{n-p-1} \right) \quad (6)$$

where, n is the number of molecules in studied data set and p is the number of independent variables in generated model.

The actual efficacy of generated QSAR models is not just their capability to reproduce known data that is confirmed by their fitting power (R^2), but is chiefly their feasibility of predictive application. Hence, the QSAR model estimations were carried out maximizing the explained variance in prediction, assigned by the leave-one-out cross-validated correlation coefficient, Q^2 .

Also, the predictive ability of the regression model generated on the training set molecules is estimated on the predictions of test set compounds, by the external R_p^2 defined as follows (37):

$$R_p^2 = 1 - \frac{\sum_{i=1}^{test} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{test} (y_i - \bar{y}_{tr})^2} \quad (7)$$

Where, \bar{y}_t the averaged value of the bioactivity for the training set; the summations cover all the molecules in the test set.

An accepted technique employed by researchers to defend their generated models against the danger of chance correlation between dependent and independent variables has been y-randomization that is., fortuitous correlation without any predictability for developed model. Y randomization is a method that is said to be "probably the most powerful validation procedure" (39).

Applicability domain of the model

The presence of response outliers (i.e. molecules with standardized residuals greater than two standard deviation units) in the investigated data set and compounds very effective in determining figures of merit and statistical parameters of developed model [i.e.

molecules with high leverage value (h) greater than $3k'/n$ where k' is the number of model variables plus one, and n the number of the molecules applied in model development] were confirmed by the Williams plot (38).

RESULTS

The structures of 26 molecules were built and optimized and a large number of descriptors (columns of X block) were estimated for each molecule using its molecular structure. In order to obtain the relationship between the biological activities as dependent and molecular structures as independent variables, logarithms of the inverse of biological activity ($\log 1/IC_{50}$) of 26 molecules were used. After dividing the molecules into calibration and validation sets, based on Kennard and Stones algorithm, different models using training set were built. Developed models were used to predict the activity of molecules in test set to evaluate the performance of models.

To determine the degree of homogeneities in the original data set and recognize potential clusters in the studied molecules, principle component analysis (PCA) was performed within the calculated pixels space for all of the molecules. PCA is a valuable multivariate statistical approach in which new orthogonal variables called principal components or PCs are derived as linear combinations of the original variables. These new generated variables are sorted on the basis of information content (i.e. explained variance of the original dataset). Priority of PCs demonstrates their higher quota in the explained variance, so most of the information is retained in the early few PCs. A main characteristic in PCA is that the generated PCs are uncorrelated. PCs can be used to obtain scores which present most of the original variations in the original data set in a smaller number of dimensions.

Here, using three more significant PCs (eigenvalues > 1), which explain 77.57 % of the variation in the data (56.74 %, 12.74 % and 8.09%, respectively) distribution of molecules over the three first principal components is shown in Fig.1. As can be seen in this figure, no cluster exists in dataset.

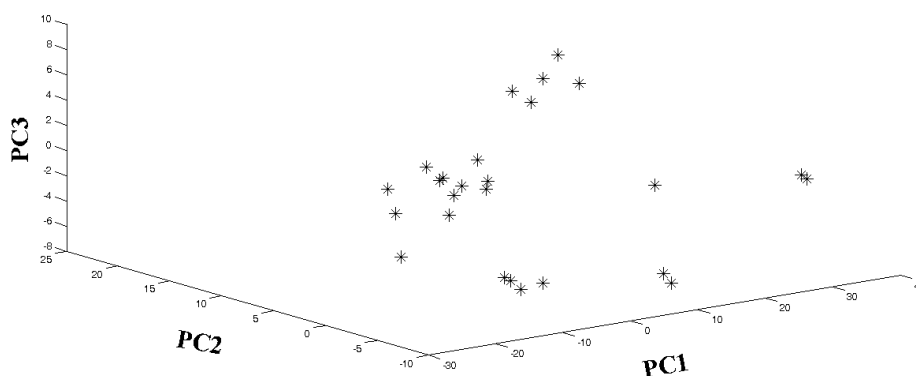


Fig. 1. Principal components analysis of the calculated descriptors of all molecules in the data set.

Table 3. The result of MLR analysis with different type of descriptors for training set molecules

Number	Descriptor class	MLR equation	R^2	Adjusted R^2	S.E.	RMSE _{CV}	Q^2	F
1	Constitutional	$pIC_{50} = -6.791 (\pm 3.533) + 23.638 (\pm 3.926) \times RBF + 12.061 (\pm 4.651) \times Mv$	0.681	0.644	0.476	0.507	0.886	18.187
2	Topological	$pIC_{50} = 3.384 (\pm 2.205) + 9.310 (\pm 2.869) \times JhteZ + 0.046 (\pm 0.015) \times T(O..O) + 10.349 (\pm 4.205) \times J$	0.761	0.717	0.425	0.457	0.887	17.014
3	Geometrical	$pIC_{50} = 7.941 (\pm 0.718) + 0.119 (\pm 0.023) \times G(O..O) + 0.003 (\pm 0.001) \times DDI$	0.613	0.567	0.525	0.574	0.906	13.463
4	Functional group	$pIC_{50} = 10.117 (\pm 0.859) - 0.943 (\pm 0.239) \times nCaH - 0.276 (\pm 0.102) \times NCRH2$	0.640	0.598	0.506	0.530	0.830	15.103

After determination of homogeneity in dataset, models were built using training set. Before model building step, the pretreatment phase was carried out on pool of calculated descriptors. This pretreatment was begun with the deletion of constant descriptor for all molecules. Also for reduction of redundancy among retained descriptors, if two or more descriptors were highly correlated, only one descriptor with the highest correlation and dependent variable was picked and others were deleted. This pretreatment phase helps to accelerate the descriptor selection and decreases the probability of including unrelated descriptors in final model.

Developed models were used to predict the activity of molecules in test set to evaluate the performance of the developed models.

Dragon software was used for calculating four different classes of descriptors including constitutional, geometrical, topological, and functional group descriptors. The following

procedure was employed to choose the most informative descriptors using the training set in each class. A certain MLR model was built with calculated descriptor of each class. The method for the selection of descriptor in developed model was a stepwise feature selection. The most significant molecular descriptors among the pool of calculated descriptors were identified using multiple linear regression analysis with a stepwise selection method. The developed MLR model for each class and its statistical parameters were reported in Table 3.

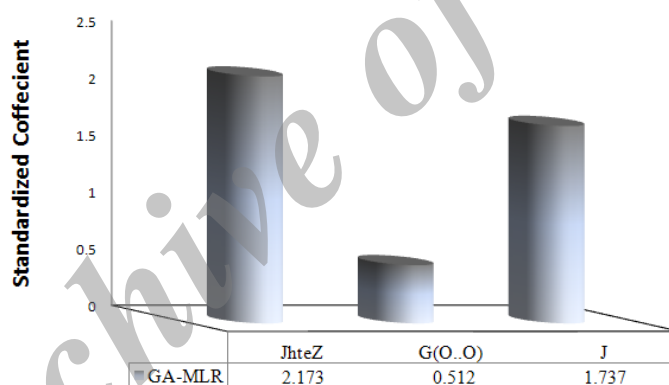
As can be seen in this Table, it was recognized that only 9 descriptors are enough to relate the bioactivity of investigated molecules to their structures. Table 4 shows the selected descriptors, their definitions and classes. A number of the calculated descriptors estimated for each molecule encoded similar information about the molecule of interest. Hence, it was desirable to examine the pool of

Table 4. List of selected descriptors for each class, their definitions by genetic algorithm.

No.	Descriptor name	Definition	Descriptor class
1	RBF	Ratable Band Fraction	Constitutional
2	Mv	Mean Atomic van der Waals volume (scaled on Carbon atom)	Constitutional
3	JhteZ	Balaban type index from Z weighted distance matrix (Barysz matrix)	Topological
4	T (O..O)	Sum of topological distances between O..O	Topological
5	J	Balaban J index	Topological
6	G (O..O)	Sum of geometrical distances between O..O	Geometrical
7	DDI	D/D index	Geometrical
8	nCaH	Number of unsubstituted aromatic C (sp ²)	Functional group
9	NCrH2	Number of ring secondary C (SP ³)	Functional group

Table 5. Correlation coefficient (R^2) matrix for the descriptors selected by MLR in various classes.

Descriptor name	RBF	Mv	JhteZ	T (O..O)	J	DDI	G (O..O)	nCaH	nCrH2
RBF	1.00								
Mv	-0.38	1.00							
JhteZ	0.88	-0.13	1.00						
T (O..O)	0.69	-0.31	0.55	1.00					
J	0.90	-0.30	0.90	0.59	1.00				
DDI	0.38	-0.53	0.14	0.59	0.26	1.00			
G (O..O)	0.70	-0.32	0.57	0.90	0.60	0.62	1.00		
nCaH	-0.67	0.07	-0.66	-0.63	-0.65	-0.53	-0.68	1.00	
nCrH2	-0.72	0.05	-0.80	-0.19	-0.74	0.32	-0.19	0.25	1.00

**Fig. 2.** Standardized coefficient of descriptors appeared in GA-MLR.

calculated descriptor and eliminate those which show high correlation with each other. Correlation coefficient (R^2) descriptors matrix for the descriptors selected in various MLR equations is shown in Table 5. As you can see any descriptors correlated ($R^2 > 0.92$) was assigned as criterion for correlated descriptor.

The most significant molecular descriptors among the selected descriptors were identified using a genetic algorithm (GA) selection method. Then these descriptors selected by GA were used as input of multiple linear regression analysis. The best equation obtained

for the pIC_{50} of the 2-mercaptoimidazoles derivatives was:

$$pIC_{50} = 3.847 (\pm 2.165) + 9.562 (\pm 2.773) \times JhteZ + 0.062 (\pm 0.021) \times G(O..O) + 10.894 (\pm 4.079) \times J$$

$$n = 21, \quad R^2 = 0.778, \quad F = 18.741 \quad (8)$$

For evaluation of the predictability of the generated GA-MLR model, the optimized model was applied for prediction of pIC_{50} values of all compounds in the calibration and prediction set. The calculated pIC_{50} for each molecule is summarized in Table 6.

Table 6. The experimental pIC_{50} and the predicted values of the studied molecules^a.

No.	pIC_{50} observed	pIC_{50} calculated (GA-MLR)	RE (GA-MLR)	pIC_{50} calculated (GA-ANN)	RE (GA-ANN)
1*	6.7	6.49	-0.03	6.59	-0.02
2	5.9	6.15	0.04	6.01	0.02
3	6.6	6.70	0.01	6.42	-0.03
4	5.7	6.16	0.08	6.01	0.05
5	6	6.08	0.01	6.17	0.03
6	8	7.87	-0.02	7.80	-0.03
7	8	7.59	-0.05	7.45	-0.02
8	7.7	7.51	-0.03	7.58	-0.02
9	7	7.46	0.06	7.07	0.01
10	6.6	6.56	-0.01	6.44	-0.03
11*	6.6	6.72	0.02	6.47	-0.02
12*	6.8	6.16	-0.10	6.92	0.02
13	5.7	5.76	0.01	5.60	-0.02
14	6.5	6.91	0.06	7.00	0.03
15	5	5.04	0.01	5.09	0.02
16	7.3	6.56	-0.11	7.15	-0.02
17*	6.5	6.32	-0.03	6.64	0.02
18	6.2	6.85	0.09	6.15	-0.01
19	6.5	5.97	-0.09	6.52	0.00
20	6.5	6.12	-0.06	6.49	0.00
21	6.3	6.18	-0.02	6.47	0.03
22	5.7	5.98	0.05	5.80	0.02
23*	6.4	6.86	0.07	6.52	0.02
24	7.2	7.10	-0.01	6.94	-0.04
25	6.9	7.02	-0.05	6.75	-0.02
26	7.4	6.49	-0.03	6.59	-0.03

*Molecules assigned as test set by Kennard and Stones algorithm

It must be noted that positive values in the regression coefficients indicate that the given descriptor contributes positively to the value of pIC_{50} , whereas negative values indicate that the increase in the value of the descriptor lead to a decrease in the value of pIC_{50} . Said another way, increasing J_{HteZ} , $G(O..O)$ and J will increase pIC_{50} of the investigated 2-mercaptoimidazoles derivatives. The standardized regression coefficient reveals the significance of an individual descriptor presented in the regression model. The increase in the absolute value of a coefficient, leads to an increase in the weight of the variable in the model. The effects of various descriptors on the biological activity are shown in Fig.2. As can be seen the effects of J_{HteZ} and J as two topological indexes are more significant than the other appeared descriptor in final GA-MLR. Experimental

versus predicted values of pIC_{50} for molecules, obtained by the GA-MLR modeling, is shown graphically in Fig. 3A.

The statistical parameters calculated for the developed MLR model are presented in Table 7. The correlation coefficient R^2 , Q^2 , and $RMSE$ for the prediction set are 0.78, 0.89 and 0.36, respectively. The chemical applicability domain of the developed GA-MLR model and the trustworthiness of the predictions are also confirmed by the leverage method. Values of leverage could be calculated for both training and test compounds. Calculating leverage for training set is useful for determining the compounds which influence the model in a way that they result in an unstable model. On the other hand calculating leverage for objects that were not used in model building (such as test set) is useful for assigning the applicability domain of the model.

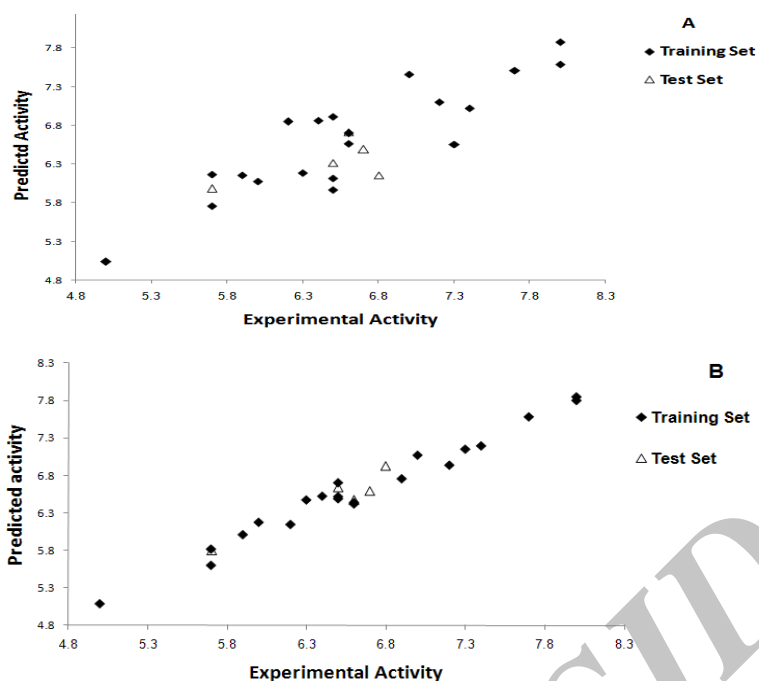


Fig. 3. Plots of predicted activities versus experimental activities for (A) GA-MLR, and (B) GA-ANN.

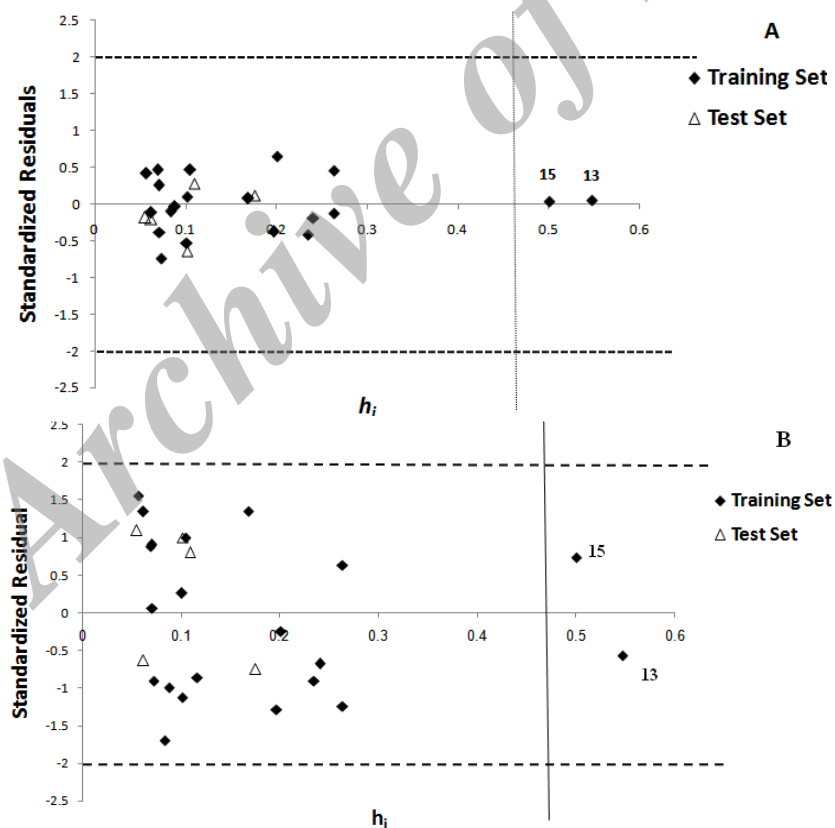


Fig. 4. Plot of standardized residuals versus leverage values (Williams plot) for (A) GA-MLR and, (B) GA-ANN. The compounds included in the training and test sets, are denoted differently; the response outliers and structurally influential compounds, explained in the text, are denoted using numbers. The horizontal lines are the 2.0σ limit and the vertical one is the warning value of leverage ($h^* = 0.470$).

In the Williams plot chemicals influential on the structural domain of the model, described by a hat value exceeding the threshold one (vertical line in Fig. 4A and 4B), can be demonstrated as compounds with unusual structural characteristics badly depicted in the training set, which could influence the calculated descriptors, selection for a better modeling of those chemicals. Outliers are compounds which their standardized residual values pass the threshold value (here, $\pm 2\sigma$, horizontal dashed line in Fig. 4A and 4B) could be correlated with errors in the measured values of bioactivities. Consideration of Williams plot (Fig 5A) implies that there is no response outlier in investigated data set. Two molecules namely 13 and 15 have a leverage value higher than warning leverage limit (0.476) but they have standard residual values between ± 2.0 standard deviation units. Hence these molecules can be considered as influent in fitting performance of model but there are no strong reasons to consider them as outliers to delete from studied data set. Williams plot showed further the trustworthiness of the predictions from another side.

In this nonlinear model, a network including a fully connected three layer, feed forward ANN model trained with a back propagation learning algorithm was used. GA-ANN had an input layer including neurons with the number of descriptors selected as the input of model (3 neurons), a hidden layer of neurons in which the number of neurons must be optimized and also a transfer function, and a single neuron output layer corresponding to the activity vector that its elements are calculated bioactivities of studied molecules by network. There are no exact theoretical principles for choosing the appropriate network topology, so before the training of network, the adjustable parameters such as number of nodes in the hidden layer, transfer function, learning rate and etc. were optimized. In order to evaluate the ANN, root mean square error of cross validation (RMSECV) was used. The values resulting from hidden layer are transferred to the last layer, which contains a single neuron representing the predicted activity. For output

layer a linear transfer function was chosen. Also for hidden layer, a sigmoid transfer function, as a more flexible transfer function, was selected.

To optimize the value of network parameters on performance of developed model, some various configurations of ANN with different values of neuron in hidden layer ($n_H = 2, 3, 4, 5, 6, \text{ and } 7$), learning rate (from 0.1 to 1) and momentum (from 0.1 to 1) were built, and output of each network on the basis of RMSECV was evaluated. A special technique using response mesh plot was employed to optimize number of node in hidden layer, learning rate, and momentum. In Fig. 5, the mesh plots of output of developed model (on the basis of RMSECV) as a linear function of learning rate and momentum in six different numbers of nodes in hidden layer are shown. It must be noted that for inhibition of overfitting in the generated ANN model, the training of the network must be performed when the RMSECV of calculated activity by network is in the minimum value.

The results show that 5 nodes in hidden layer, learning rate of 0.6 and momentum of 0.1 are the optimum parameters of model. After optimization of previous parameters, the number of iterations must be optimized. Fig. 6 shows a plot of the RMSECV for training set versus the number of epochs which represents the estimation of the extent of training period. It can be seen from this figure that while training of network was performed for the training set; RMSECV initially decreases and then begins to increase after approximately 1900 epochs. This position is commencing point of overtraining of network and then 1900 iteration was chosen as the optimum number of epoch. The generated nonlinear model was then trained using the training set for optimization of the weights and biases. For estimate of the predictability of the generated ANN, a trained network was applied for prediction of the pIC_{50} s values in the test set which were not used in the modeling step. The predicted activity of molecules calculated by GA-ANN is plotted against the experimental values in Fig. 3B and is reported in Table 6. As expected, the calculated values are in good agreement with experimental values.

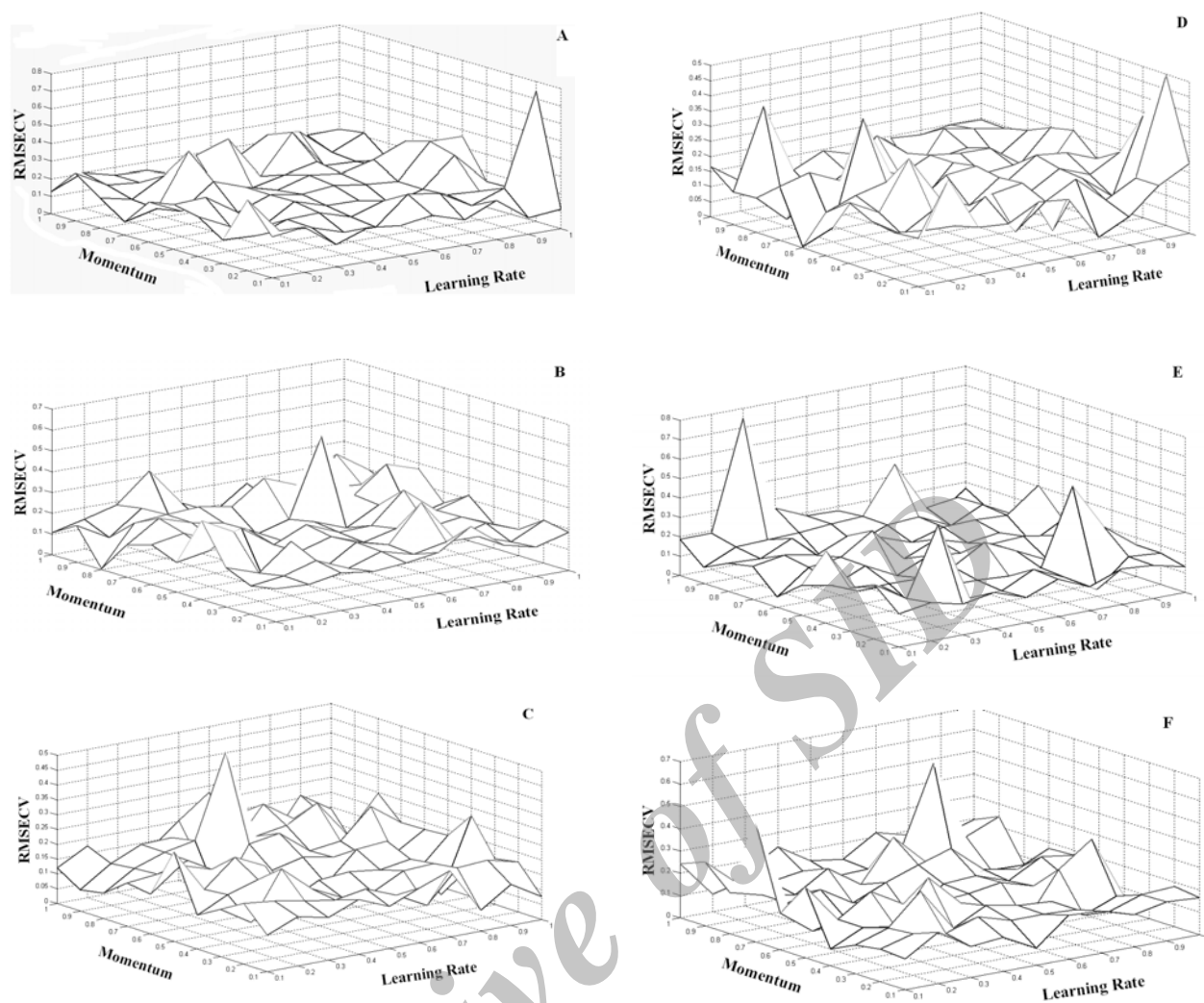


Fig. 5. Mesh counter plots of output of GA-ANN (on the basis of RMSECV) to optimize networks parameters including linear rate, momentum, and number of hidden layer nodes (n_H) (A) $n_H = 2$; (B) $n_H = 3$; (C) $n_H = 4$; (D) $n_H = 5$; (E) $n_H = 6$; and (F) $n_H = 7$.

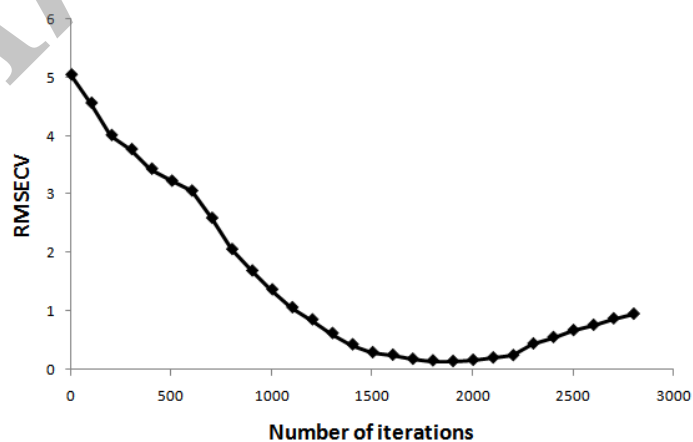


Fig. 6. Plot of RMSECV for training set versus the number of iterations

Table 7. Statistical parameters obtained for the developed model for anti tuberculosis inhibitor activity of investigated compounds.

Parameter	GA-MLR		GA-ANN	
	Training Set	Test Set	Training Set	Test Set
<i>Data set</i>				
<i>N</i>	21	5	21	5
R^2	0.78	0.35	0.93	0.91
<i>RMSE</i>	0.36	0.34	0.22	0.12
<i>PRESS</i>	2.68	0.58	1.01	0.07
Q^2	0.89		0.78	
$RMSE_{CV}$	0.442		0.458	
$PRESS_{CV}$	3.92		4.03	

N: Number of objects in data set, R^2 : Correlation coefficient of experimental and predicted activities, *RMSE*: Root mean square error, *PRESS*: Predicted error sum of square, R^2_{CV} : Correlation coefficient of leave one out cross validation, $RMSE_{CV}$: Root mean square error of cross validation, $PRESS_{CV}$: Predictive residual sum of square of cross validation

Table 8. R^2 and Q^2 obtained in two models by Y randomization.

Iteration	GA-MLR		GA-ANN	
	R^2	Q^2	R^2	Q^2
1	0.12	0.13	0.20	0.07
2	0.11	0.02	0.11	0.12
3	0.05	0.16	0.12	0.16
4	0.07	0.01	0.20	0.13
5	0.11	0.15	0.05	0.12
6	0.21	0.17	0.03	0.20
7	0.15	0.24	0.06	0.05
8	0.01	0.01	0.10	0.05
9	0.09	0.03	0.08	0.09
10	0.08	0.03	0.15	0.07

Table 7 compares the results obtained using the GA-MLR and GA-ANN models. The R^2 , *RMSE* and *PRESS* of the models for training and test sets reveal the potential of the ANN model for prediction of pIC_{50} s values of various 2-mercaptoimidazoles as CCR2 inhibitors.

RMSE and *PRESS* of 0.34 and 0.58 for the test set by the GA-MLR model should be compared with the values of 0.12 and 0.07 by the GA-ANN model. It can be seen from Table 7 that although parameters appearing in the GA-MLR model are used as inputs for the generated GA-ANN model, the statistics have shown a large improvement. These improvements are because pIC_{50} s values of 2-mercaptoimidazoles reveal nonlinear correlations with the selected descriptors by genetic algorithm.

Same with GA-MLR, to better estimate the developed GA-ANN, the Williams plot was constructed to verify the presence of outliers

and/or molecules with high influence on the results (Fig.4B). As discussed for GA-MLR leverage values and standardized residuals in prediction of activity of molecules are reported, respectively, on x and y axes. In this plot, reference lines are also reported both for leverage critical value (0.470) and for standardized residuals critical value ($\pm 2\sigma$) Molecules with leverage greater than the critical value can be considered as objects with too much influence on the regression model.

In the same manner, molecules with a standardized residual greater than the critical value are described by a poor prediction value. By examining the applicability domain of the GA-ANN from the Williams plot (Fig. 4B) it can be seen that neither of the molecules in the set is recognized as a response outlier based on the 2σ criterion for the total molecules. On the other hand, on the basis of leverage approach two compounds from the investigated molecules are recognized as structurally

influential chemicals: molecules 13 and 15. These results are same with GA-MLR model.

In order to avoid chance correlations which are possible because of a large number of generated columns (independent variables), and examine the robustness of developed models, Y-randomization test has been applied to models. The dependent variable vector is randomly permuted and a new QSAR models is constructed using the original independent variable matrix. The new modeling was expected to have low R^2 and Q^2 values. For sureness, a number of iterations were carried out. If the results show high R^2 and Q^2 , it implies that an acceptable QSAR model cannot be obtained. Several random shuffles of the Y vector were performed on the generated models and the results are shown in Table 8. The low R^2 and Q^2 values show that the good results in our original model are not due to a chance correlation or structural dependence of the training set.

DISCUSSION

By interpreting the descriptors included in the model, it is possible to obtain valuable chemical insights into the biological activity. For this reason, a brief explanation of the three descriptors that were employed in the generated GA-MLR model is provided below. JhetZ is Balaban type index from Z weighted distance matrix and J is Balaban J index. Both JhetZ and J are belonging to topological descriptor class. Presence of these descriptors in the final MLR model, basically accounts for size, shape, and branching, thus steric contribution to biological activity. The structures of almost all 2-mercaptoimidazoles included in this QSAR study are very similar to each other. These structures have an imidazole ring in the center of molecule and three substituents in positions 1, 2, and 3 that more or less have similar structure. Therefore, appearing of the topological descriptors such as J and JhetZ in the model is not unusual. These topological descriptors encode the compactness and the degree of branching of a molecule.

$G(O..O)$ is sum of geometrical distance between two oxygen atoms in studied molecules.

Because this descriptor belongs to the geometrical group of descriptors, some geometrical properties including angles between atoms, dihedrals angles, and atomic distances are probably important features in the effectiveness of these compounds as CCR2 inhibitors. The variables appeared in the GA-MLR model encode different aspects of topological and geometrical molecular structure.

One of the most important reasons for this study is comparison of ability of linear QSAR model building methods (such as MLR) and non linear QSAR model building techniques (such as ANN) in predicting the inhibitory activity of some 2-mercaptoimidazoles as CCR2 inhibitors. To obtain robust and accurate models, the ANN models should be trained by subset of descriptors instead of all generated descriptors.

As discussed above, a genetic algorithm technique was applied as a feature selection method to choose the most relevant subset of descriptors. Said another way, to find robust and predictable model, layered feed forward back propagation neural network model was trained with subsets of descriptors instead of all calculated descriptors.

Therefore, an ANN model was developed by using the three descriptors appearing in the MLR model as its inputs. Since these descriptors were selected by GA, QSAR model was called GA-ANN.

As a result, it was discovered that a correctly selected and trained neural network could reasonably represent the dependence of the CCR2 receptor inhibitory activities of 2-mercaptoimidazoles on the descriptors. The optimized neural network could then simulate the complex nonlinear relationship between the pIC_{50s} value and the descriptors.

CONCLUSION

QSAR were built for the CCR2 receptor inhibitory activity of some 2-mercaptoimidazoles by using the GA-MLR and GA-ANN methods. Comparison of the GA-MLR and GA-ANN models reveal superiority of the GA-ANN model over the GA-MLR model. Because the improvement of

results acquired by using the non-linear model is substantial, it can be deduced there is a non-linear relationship between the pIC_{50} s and the calculated structural descriptors of the 2-mercaptoimidazoles. In the final models, importance of topological descriptors is considerable (including JhetZ and J). Presence of these descriptors in the final models, basically accounts for size, shape, and branching, thus steric contribution to biological activity.

REFERENCES

1. Feria M, Diaz-Gonzalez F. The CCR2 receptor as a therapeutic target. *Expert Opin Ther Pat.* 2006;16:49-57.
2. Bachmann MF, Kopf M, Marsland BJ. Chemokines: more than just road signs. *Nat Rev Immunol.* 2006;6:159-164.
3. Carter PH. Chemokine receptor antagonism as an approach to anti-inflammatory therapy: 'just right' or plain wrong? *Curr Opin Chem Biol.* 2002;6:510-525.
4. Gao Z, Metz WA. Unraveling the chemistry of chemokine receptor ligands. *Chem Rev.* 2003;103:3733-3752.
5. Schwarz MK, Wells TN. New therapeutics that modulate chemokine networks. *Nat Rev Drug Disc.* 2002;1:347-358.
6. Katschke KJ Jr, Rottman JB, Ruth JH, Qin S, Wu L, LaRosa G, *et al.* Differential expression of chemokine receptors on peripheral blood, synovial fluid, and synovial tissue monocytes/macrophages in rheumatoid arthritis. *Arthritis Rheum.* 2001;44:1022-1032.
7. Lu B, Rutledge BJ, Gu J, Fiorillo J, Lukacs NW, Kunkel SL, *et al.* Abnormalities in monocyte recruitment and cytokine expression in monocyte chemoattractant protein 1-deficient mice. *J Exp Med.* 1998;187:601-608.
8. Kuziel WA, Morgan SJ, Dawson TC, Griffin S, Smithies O, Ley K, *et al.* Severe reduction in leukocyte adhesion and monocyte extravasation in mice deficient in CC chemokine receptor 2. *Proc Natl Acad Sci USA.* 1997;94:12053-12058.
9. Kurihara T, Warr G, Loy J, Bravo R. Defects in macrophage recruitment and host defense. *J Exp Med.* 1997;186:1757-1762.
10. Boring L, Gosling J, Chensue SW, Kunkel SL, Farese RV, Broxmeyer HE, *et al.* Impaired monocyte migration and reduced type 1 (Th1) cytokine responses in C-C chemokine receptor 2 knockout mice. *J Clin Invest.* 1997;100:2552-2561.
11. Gong JH, Ratkay LG, Waterfield JD, Clark-Lewis I. An antagonist of monocyte chemoattractant protein 1 (MCP-1) inhibits arthritis in the MRL-lpr mouse model. *J Exp Med.* 1997;186:131-137.
12. Ogata H, Takeya M, Yoshimura T, Takagi K, Takahashi K. The role of monocyte chemoattractant protein-1 (MCP-1) in the pathogenesis of collagen-induced arthritis in rats. *J Pathol.* 1997;182:106-114.
13. Huang DR, Wang JT, Kivisakk P, Rollins BJ, Ransohoff MR. Absence of monocyte chemoattractant protein 1 in mice leads to decreased local macrophage recruitment and antigen-specific T helper cell type 1 immune response in experimental autoimmune encephalomyelitis. *J Exp Med.* 2001;193:713-726.
14. Izikson L, Klein RS, Charo IF, Weiner HL, Luster AD. Adjustment for known risk factors for transplant rejection confirmed the univariate findings for possession of the CCR2-64I allele (OR, 0.20; P = 0.032) and homozygosity for the 59029-A allele (OR, 0.26; P = 0.027). *J Exp Med.* 2000;192:1075-1080.
15. Fife BT, Huffnagle GB, Kuziel WA, Karpus WJ. CC chemokine receptor 2 is critical for induction of experimental autoimmune encephalomyelitis. *J Exp Med.* 2000;192:899-905.
16. Dawson TC, Kuziel WA, Osahar TA, Maeda N. Absence of CC chemokine receptor-2 reduces atherosclerosis in apolipoprotein E-deficient mice. *Atherosclerosis.* 1999;143:205-211.
17. Gu L, Okada Y, Clinton SK, Gerard C, Sukhova GK, Libby P, *et al.* Absence of monocyte chemoattractant protein-1 reduces atherosclerosis in low density lipoprotein receptor-deficient mice. *J Mol Cell.* 1999;2:275-281.
18. Daly C, Rollins B. Monocyte chemoattractant protein-1 (CCL2) in inflammatory disease and adaptive immunity: therapeutic opportunities and controversies. *J Microcirculation.* 2003;10:10247-10257.
19. Dawson J, Miltz W, Mir AK, Wiessner C. Targeting monocyte chemoattractant protein-1 signalling in disease. *Expert Opin Ther Targets.* 2003;7:35-48.
20. McCulloch WS, Pitts W. A logical calculus of ideas immanent in nervous activity. *Bull Math Biophys.* 1943;5:115-133.
21. Hodgkin AL, Huxley AF. A quantitative description of membrane current and its application to conduction and excitation in nerve. *J Physiol.* 1952;117:500-544.
22. Gardner MW, Dorling SR. Artificial neural networks (the multilayer perceptron) a review of applications in the atmospheric sciences. *Atmos Environ.* 1998;32:2627-2636.
23. Schalkoff R. *Pattern Recognition: Statistical, Structural and Neural Approaches.* Wiley NY. 1992.
24. Shahlaei M, Fassihi A, Saghale L. Application of PC-ANN and PC-LS-SVM in QSAR of CCR1 antagonist compounds: A comparative study. *Euro J Med Chem.* 2010;45:1572-1582.
25. Arkan E, Shahlaei M, Pourhossein A, Fakhri K, Fassihi A. Validated QSAR analysis of some diaryl substituted pyrazoles as CCR2 inhibitors by various linear and nonlinear multivariate chemometrics methods. *Eur J Med Chem.* 2010;45:3394-3406.

26. van Deventer JSJ, Liebenberg SP, Lorenzen L, Aldrich C. Dynamic modelling of competitive elution of activated carbon in columns using neural networks. *Miner Eng.* 1995;8:1489-1501.
27. Van Lomen G, Doyon J, Coesemans E, Boeckx S, Cools M, Buntinx M, *et al.* 2-Mercaptoimidazoles, a new class of potent CCR2 antagonists. *Bioorg Med Chem Lett.* 2005;15:497-500
28. Frisch MJ, Trucks MJ, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, *et al.* Gaussian 98, Revision A7, Gaussian Inc. Pittsburgh PA. 1998.
29. Todeschini R. Milano Chemometrics and QSPR Group, <http://michem.disat.unimib.it>.
30. Kennard RW, Stone LA. Computer aided design of experiments. *Technometrics* 1969;11:137-149.
31. Tropsha A, Gramatica P, Gombar VK. The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb Sci.* 2003;22:69-77.
32. Ghasemi J, Saaidpour S. Quantitative structure–property relationship study of n-octanol–water partition coefficients of some of diverse drugs using multiple linear regression. *Anal Chem Acta.* 2007;604:99-106.
33. Mohajeri A, Hemmateenejad B, Mehdipour A, Miri R. Modeling calcium channel antagonist activity of dihydropyridine derivatives using quantum topological molecular similarity indices analyzed by GA-PLS and GA-PC-PLS. *J Mol Graphics Modell.* 2008; 26: 1057-1065.
34. Hagan MT, Demuth HB, Beal M. *Neural Network Design.* Boston; PWS: 1996.
35. Haykin S. *Neural Network.* Prentice-Hall, Englewood Cliffs. NJ. 1994
36. Bose PK, Liang P, *Neural Network, Fundamentals.* New York; McGraw-Hill: 1996.
37. Atkinson AC. *Plots, Transformations and Regression,* Clarendon Press: Oxford UK; 1985. p. 282.
38. Golbraikh A, Tropsha A. Beware of $q^2!$, *J Mol Graph Model.* 2002;20:269-276.
39. Naes T, Isaksson T, Fearn T, Davies T. *A User-Friendly Guide to Multivariate Calibration and Classification,* NIR Publications, Chichester, UK, 2004.

Archive of SID