Research in Pharmaceutical Sciences, August 2015; 10(4): 307-325
Received: Jan 2014
Accepted: Mar 2014

School of Pharmacy & Pharmaceutical Sciences
Isfahan University of Medical Sciences

Original Article

Quantitative structure–activity relationship study of $P2X_7$ receptor inhibitors using combination of principal component analysis and artificial intelligence methods

Mehdi Ahmadi[1], and Mohsen Shahlaei[2,*]

[1]*Student Research Committee, Kermanshah University of Medical Sciences, Kermanshah, I.R. Iran.*
[2]*Meical Biology Research Center, Kermanshah University of Medical Sciences, Kermanshah, I.R. Iran.*

Abstract

$P2X_7$ antagonist activity for a set of 49 molecules of the $P2X_7$ receptor antagonists, derivatives of purine, was modeled with the aid of chemometric and artificial intelligence techniques. The activity of these compounds was estimated by means of combination of principal component analysis (PCA), as a well-known data reduction method, genetic algorithm (GA), as a variable selection technique, and artificial neural network (ANN), as a non-linear modeling method. First, a linear regression, combined with PCA, (principal component regression) was operated to model the structure–activity relationships, and afterwards a combination of PCA and ANN algorithm was employed to accurately predict the biological activity of the $P2X_7$ antagonist. PCA preserves as much of the information as possible contained in the original data set. Seven most important PC's to the studied activity were selected as the inputs of ANN box by an efficient variable selection method, GA. The best computational neural network model was a fully-connected, feed-forward model with 7−7−1 architecture. The developed ANN model was fully evaluated by different validation techniques, including internal and external validation, and chemical applicability domain. All validations showed that the constructed quantitative structure–activity relationship model suggested is robust and satisfactory.

Keywords: QSAR; $P2X_7$ receptor antagonists; Artificial neural network (ANN); Principal component analysis (PCA); Genetic algorithm (GA)

INTRODUCTION

The $P2X_7$ receptor is extremely expressed by cells of hemopoietic origin. This receptor is an ATP-gated ion channel belonging to the family of ATP-sensitive ionotropic P2X receptors (1,2). The $P2X_7$ receptor has a 595-amino acid sequence within average of 40% homology to other members of purinergic P2X receptors, while contains a carboxyl-terminal which is considerably longer than that of other P2X receptors (1).

The P2X7 protein possesses features such as low affinity for adenosine 5'-triphosphate (ATP) differentiating it from other P2X receptors (1). These features possibly reflect the preference of this protein for ATP as its endogenous ligand. Neumerous possible biological activities including maturation and release of interleukin-1 beta (IL-1β) and ATP-induced apoptosis for $P2X_7$ receptor have been proposed (1,3). In brief, studies have demonstrated the potential role of $P2X_7$ receptor in modulating IL-1β and possibly glutamate, to decrease nociception in neuropathic pain models (4,5).

In modern medicinal chemistry, it becomes to a greater extent essential to handle huge sets of structural data (6). The analysis of large data sets can be interesting as may consist of over a thousand descriptors calculated from different molecules each sampled from thousands molecular geometries. Even when the raw descriptors are converted into a data matrix, it could be dealt with more than a hundred descriptor vectors for each molecule. Such a large number of variables lead to multicollinearity, and to redundancies among

*Corresponding author: M. Shahlaei
Tel: 0098 831 4276489, Fax: 0098 831 4276493
Email: mshahlaei@kums.ac.ir

the descriptors (6). As a result, it becomes more complicated to disclose patterns present in the original data. Advanced data mining methods dealing with such difficulties become *tremendously* more important (6,7). In this study, techniques are used that permit us to better understand the structure of large sets of structural data.

Data mining can be defined as the procedure of extracting helpful information from large data sets (8). Until now, a number of data mining approaches have been developed, but often a single data mining method is insufficient and, instead, more than a few methods must be used to support a single application (8). However, using different procedures to large databases causes a computational problem. A simple solution would be to reduce the amount of data by taking a subset of representative molecules from a given data set (8). Alternatively, a data compression method such as principal component analysis (PCA) can be used.

PCA has been extensively used in data mining to study data structure (6). In PCA, new orthogonal variables (latent variables or PC's) are calculated by maximizing variances of the data (6). The number of the latent variables (factors) is much lower than the number of original descriptors, so that the data can be visualized in a low-dimensional PC defined space (6,9,10). While PCA really reduces the dimensionality of the space, it does not reduce the number of the original descriptors (the independent variables in a typical quantitative structure–activity relationship (QSAR) study), as it uses all the original descriptors to produce the new latent variables (principal components) (6,9,10). For interpretation purposes and future investigations or model building, it would often be very useful to reduce the number of variables. PC selection can be attained either by choosing informative PC's (PC's with maximum variance) or using stochastic methods such as genetic algorithm. Several approaches exist and most of them carry out feature reduction using stepwise forward and/or backward techniques (6,9,10). Jolliffe (11) compared a number of methods, mainly working on preserving most of the variation of the data. McCabe (12) developed techniques to remain as much information as possible by optimizing four numerical criteria (6). Rannar and coworkers (13) chose variables that span the original space as well as possible by a combination of PCA and partial least squares. In data mining, it is of importance to select a small subset of variables that can reproduce as closely as possible the structure of the complete data (6). Krzanowski (14) developed such a method based on Procrustes analysis. As the technique seeks variables by a stepwise procedure (backward elimination), there is no assurance to find the best global subset. Moreover, with hundreds or thousands of independent variables, as is often the case in data mining, intensive calculation is needed to perform PCA in each elimination step (6). In this study, a method is presented that uses a genetic algorithm (GA) to search for the best subset instead of a classical variable selection such as backward elimination procedure (6).

QSAR models can be generated employing a number of methods, including a variety of statistical methods (e.g., principal component regression (PCR)). For predicting biological activity, PCR has emerged as the statistical method of choice (15,16). Artificial neural network (ANN) as a representative artificial intelligence method stands for a non-linear technique that has emerged as a potential alternative to linear regression techniques such as PCA (6,17-19).

ANN are not constrained by a known mathematical equation between dependent and independent variables, and have the power to model any arbitrarily complicated nonlinear relationship (16). Developers of ANN QSAR models do not require formal training in statistical methodology, and models can be generated by users with a minimum of theoretical and mathematical knowledge (16). There are a large number of researches suggesting that ANN models may offer significantly better predictive performance than traditional statistical approaches such as multiple linear regression (MLR) for certain problems such as QSAR (9,10,20,21). Thus, ANNs may represent an attractive alternative to PCA as a statistical modeling technique under certain circumstances.

In the present study both ANN and PCR techniques were used for modeling of the observed $P2X_7$ antagonist activity of 49 studied molecules. The capability of the developed QSAR model was assessed by means of the prediction of $P2X_7$ receptor antagonist activity of test set for which biological activity data have been reported.

In this study, we applied semi-empirical method to derive structural-chemical descriptors for the QSAR study of the 49 $P2X_7$ antagonist activities of purine analogues. First, a linear regression, combined with PCA was operated to model the structure–activity relationships, and after that a combination of PCA and ANN algorithm was employed to predict the biological activity of the $P2X_7$ antagonist accurately.

In this study, it is shown that ANN was superior to linear PCR in providing a good prediction of $P2X_7$ antagonist activity of purine analogues.

MATERIALS AND METHODS

Data sets

A data set including 49 $P2X_7$ receptor inhibitors was collected from literature (22,23) (Table 1). The majority of the tested inhibitors are efficient $P2X_7$ receptor inhibiting agents showing pIC_{50} values from 5.559 to 7.619. The 2D structure of each molecule was drawn by Chem Draw version 8.0 (2004) and then converted into 3D structure by Chem3D. The resulted structures were optimized using parametric method. The generated 3D structure of each molecule was visually examined to guarantee that the chirality of the chiral agent is correctly prepared and structure of molecules was not duplicated (24). Molecules were further separated into the training (32 compound), validation (7 compound) and test (10 compounds) sets based on Kennard and Stone algorithm (25,26).

The best situation of this stage of model building is dividing data set to guarantee that both training and test sets individually cover the total space occupied by original data set. Then ideal splitting of the data set as each of objects in test set is close to at least one of the objects in the training set. Various methods were used as tools for splitting the whole original data set to the training and test sets.

The Kennard–Stone algorithm (25) selects a set of molecules in studied set of data, which are 'uniformly' distributed over the space defined by the candidates.

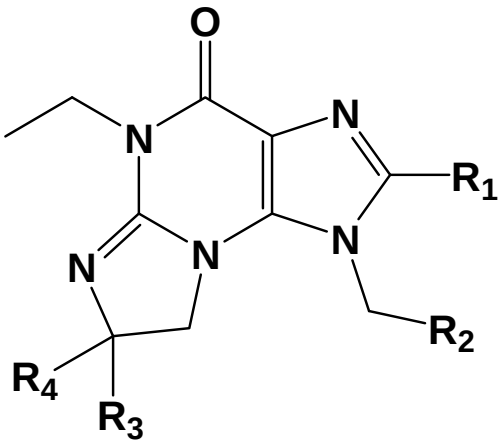
This is a classic technique to extract a representative set of molecules from a given data set. In this technique the molecules are selected consecutively. The first two objects are chosen by selecting the two farthest apart from each other. The third sample chosen is the one farthest from the first two objects, etc. Supposing that *m* objects have already been selected (m<n), the (m + 1)*th* sample in the calibration set is chosen using the following criterion:

$$\max_{m<r\le n}(\min(d_{1r}, d_{2r}, \ldots d_{mr}) \quad (1)$$

Where, *n* stands for the number of samples in the training set, d_{jr}, $j=1,\ldots,m$ are the squared euclidean distances from a candidate sample *r*, not yet included in the representative set, to the *m* samples already included in the representative set. One more benefit of the Kennard–Stone method is that it may be used to any matrix of predictors; there are no restrictions regarding the matrix multicollinearity. The other advantage is that the test molecules all fall inside the measured region and the training set molecules map the measured region of the input variable space completely with respect to the induced metric.

Overfitting problem or poor generalization capability occurs when a typical ANN model overlearns during the training phase. A too well-trained model may not carry out well property prediction on unseen data set due to its lack of generalization capability. A technique to solve the problem is the early stopping method in which the training process is concluded as soon as the overtraining signal appears. This technique needs the data set to be divided into three subsets including training set, test set, and validation sets. The training and the validation sets are the norm in all model formation procedures. The test set is employed to test the trend of the prediction accuracy of the model trained at some point of the training process. At later training steps, the validation error increases.

Table 1. Main structure and details of the compounds used in this study.



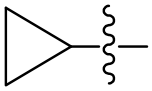
Compd.	R_1	R_2	R_3	R_4
1	4F-Ph	Ph	H	Isopropyl
2	4F-Ph	3F,4F-Ph	H	Isopropyl
3	4F-Ph	2F,4F-Ph	H	Isopropyl
4	4F-Ph	4F-Ph	H	Isopropyl
5	4F-Ph	3F,5F-Ph	H	Isopropyl
6	4F-Ph	3,4,5-TriF-Ph	H	Isopropyl
7	4F-Ph	3F,4FPh	H	tButyl
8	4F-Ph	3F,4F-Ph	H	

Table 1. (continued)

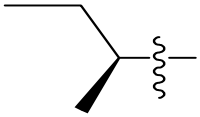
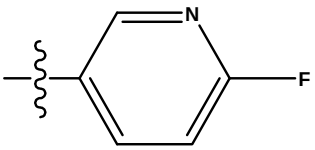
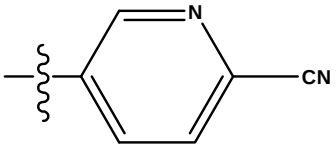
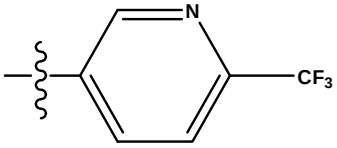
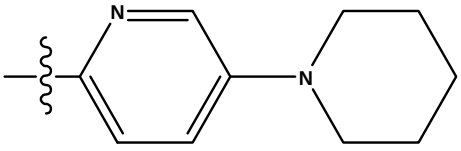
Compd.	R_1	R_2	R_3	R_4
9	4F-Ph	3F,4F-Ph	H	
10	4F-Ph	3F,4F-Ph	Me	Me
11	4-CF3OPh	3F,4F-Ph	H	Isopropyl
12	3-CF_3OPh	3F,4F-Ph	H	Isopropyl
13	4-CN-Ph	3F,4F-Ph	H	Isopropyl
14	4-CN, 3F-Ph	3F,4F-Ph	H	Isopropyl
15		3F,4F-Ph	H	Isopropyl
16		3F,4F-Ph	H	Isopropyl
17		3F,4F-Ph	H	Isopropyl
18		3F,4F-Ph	H	Isopropyl

Table 1. (continued)

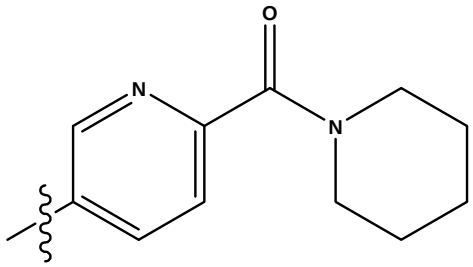
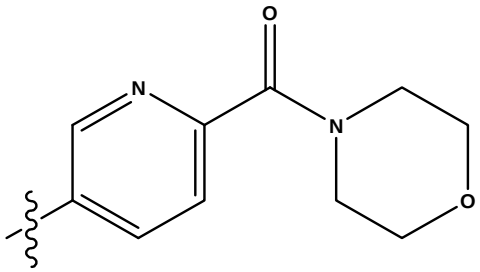
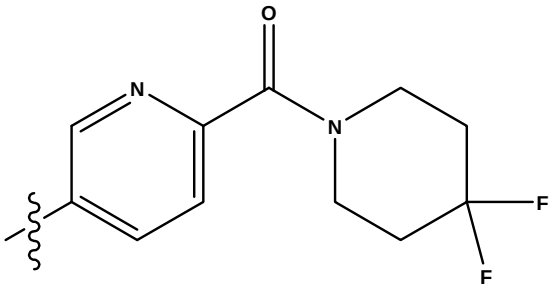
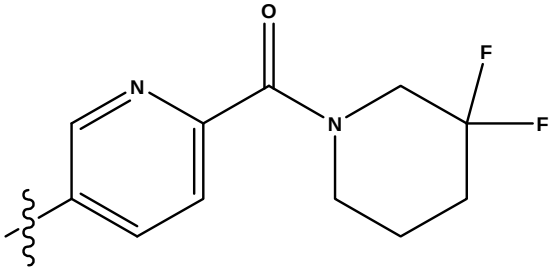
Compd.	R_1	R_2	R_3	R_4
19		3F,4F-Ph	H	Isopropyl
20		3F,4F-Ph	H	Isopropyl
21		3F,4F-Ph	H	Isopropyl
22		3F,4F-Ph	H	Isopropyl
23	Biphenyl	2F-Ph	H	Isopropyl
24	Biphenyl	3F-Ph	H	Isopropyl
25	Biphenyl	4F-Ph	H	Isopropyl
26	Biphenyl	2F,3F-Ph	H	Isopropyl

Table 1. (continued)

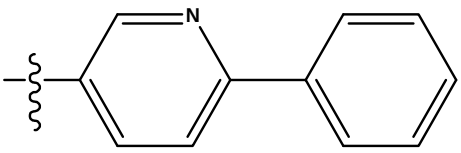
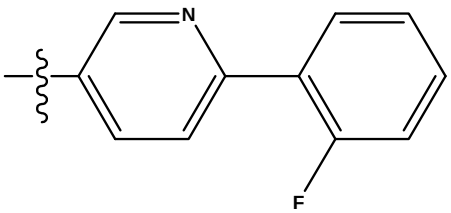
Compd.	R_1	R_2	R_3	R_4
27	Biphenyl	2F,4F-Ph	H	Isopropyl
28	Biphenyl	2F,5F-Ph	H	Isopropyl
29	Biphenyl	3F,5F-Ph	H	Isopropyl
30	Biphenyl	3F,4F-Ph	H	Isopropyl
31	Biphenyl	3F,4F-Ph	Me	Me
32	Biphenyl	3F,4F-Ph	-	Cyclohexyl
33	Biphenyl	4F-Ph	Me	Me
34	Biphenyl	Biphenyl	Me	Me
35	Biphenyl	Biphenyl	Me	Me
36		3F,4F-Ph	Me	Me
37		3F,4F-Ph	Me	Me

Table 1. (continued)

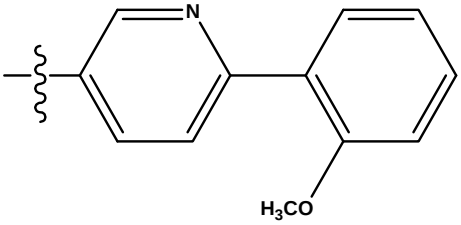
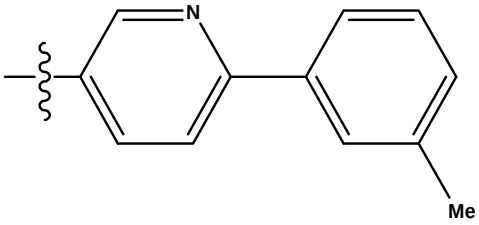
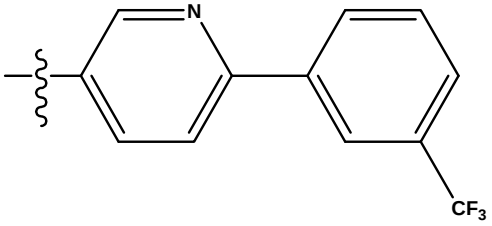
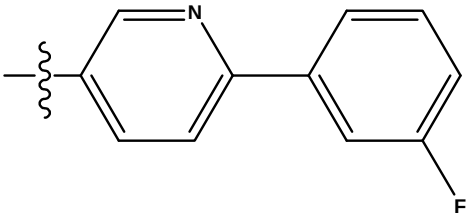
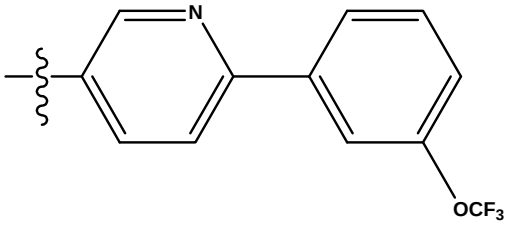
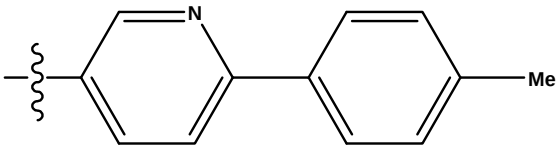
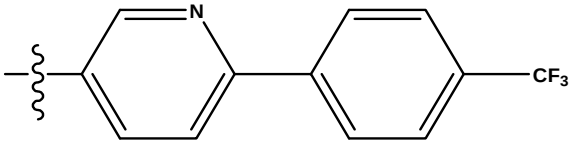
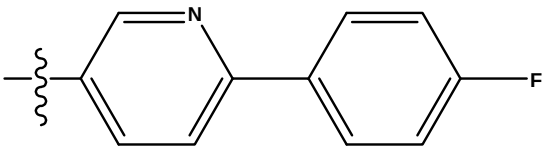
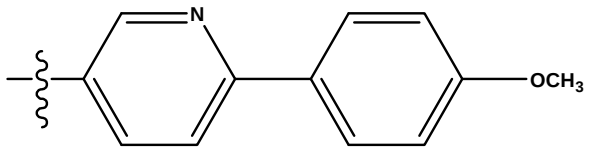
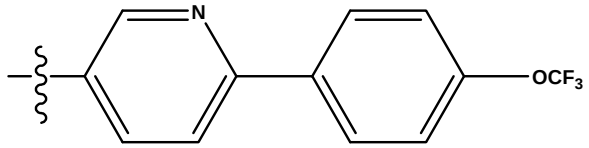
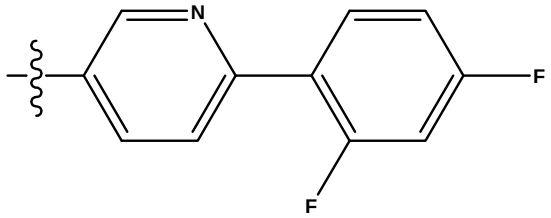
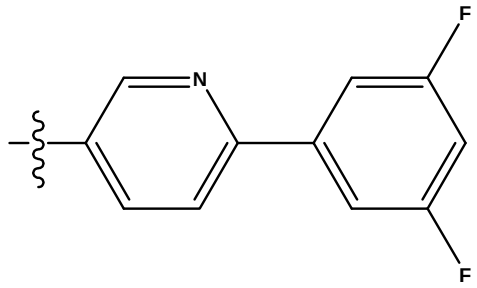
Compd.	R_1	R_2	R_3	R_4
38		3F,4F-Ph	Me	Me
39		3F,4F-Ph	Me	Me
40		3F,4F-Ph	Me	Me
41		3F,4F-Ph	Me	Me
42		3F,4F-Ph	Me	Me
43		3F,4F-Ph	Me	Me
44		3F,4F-Ph	Me	Me
45		3F,4F-Ph	Me	Me

Table 1. (continued)

Compd.	R_1	R_2	R_3	R_4
46		3F,4F-Ph	Me	Me
47		3F,4F-Ph	Me	Me
48		3F,4F-Ph	Me	Me
49		3F,4F-Ph	Me	Me

This is the moment when the model should stop to be trained to solve the overfitting problem. To achieve this task, the extracted PCs matrix was classified as discussed above. Then, the training and validation sets were employed to optimize the ANN performance.

Molecular descriptors

Structural descriptors have been routinely employed for quantitative description of different features of compounds such as topological and geometrical characteristics (24,26). In this study, a total of 285 theoretical structural descriptors were calculated.

This set of descriptors was manually selected from descriptors calculated in Dragon software (version 2.1) by eliminating those descriptors that were obviously redundant or irrelevant to the prediction of pharmaceutical agents (26,27).

The name and number of calculated theoretical descriptors included 12 descriptors in the class of functional group, 218 descriptors in the class of topological descriptors, 28 descriptors in the class of geometrical descriptors, and 27 descriptors in the class of constitutional indexes.

These theoretical descriptors were computed from the optimized 3D structure of each molecule, but not all of the descriptors are essential for the QSAR modeling.

In order to decrease interference of multicollinearity before model building, these molecular descriptors were preprocessed. The procedure includes: (i) exclusion of descriptors which have the identical value for more than 90% of the molecules; (ii) exclusion of descriptors with relative standard deviation less than 0.05; (iii) for each pair of descriptors with Pearson correlation coefficient over 0.9, only one descriptor, which has the higher correlation with the biological activity, remained (28).

Principal component analysis

Calculated descriptors were exported to the MATLAB environment for the purpose of PCA. The complete data set, defined by the descriptors in the columns (in this study, 285 descriptors) and the molecules in the rows, was auto scaled through mean centering by column. PCA models the maximum directions of variation in a data set by projecting the molecules as a swarm of points in a space spanned by PC's. Each PC is a linear function of a number of original descriptors, resulting in a reduction of the original number of descriptors. PC's describe, in decreasing order, the most variation among the molecules, and because they are calculated to be orthogonal to one another, each PC can be interpreted independently. This allows an overview of the data structure by revealing relationships between the molecules as well as the detection of deviating molecules (29). To find these sources of variation, the original data matrix of descriptors, defined by *X(n,m)*, is decomposed into the molecule space, the descriptor space, and the error matrix. The latter represents the variation note explained by the extracted PC's and is dependent on the problem definition. The approach describing this decomposition is presented as:

$$X(n,m) = T(n,k)P(k,m)^T + E(n,m) \quad (2)$$

where, *X* is the independent descriptor matrix, *T* is the scores matrix, *P* is the loadings matrix, *E* is the error matrix, *n* is the number of molecules, *m* is the number of descriptors, and *k* is the number of PC's used (29).

In a PCR analysis, a model formation step was carried out with stepwise selection and elimination of PCs to model the binding pIC_{50} relationships with different vectors of scores. On the other hand, in PCR procedure, all calculated scores were collected in a single data matrix and the best subset of descriptors was obtained by stepwise regression.

Genetic algorithm

The GA is used to select the descriptors that are most significant for the molecular data set. GA is a stochastic optimization technique that has been inspired by evolutionary principles (30,31). One of the most important aspects of GA is that it studies many possible solution simultaneously, each of which explores different regions in space spanned by input variable (32). In this study, GA was tried for selecting significant PC's. In this way of using GA, an individual in the population is represented by the string of bits that encoding the selected descriptor. The first step in a GA is to create a gene pool of *n* individuals. Each individual (chromosome) contains some PC's that in the first generation these PC's are selected randomly from a table including calculated PC's and in a way such that no two individuals can be found that contain exactly the same set of descriptors (33). This individual was used as an input of ANN. The fitness score of each individual in this generation is determined by mean square error (*MSE*) of the network.

In the next step regeneration happen, so that the new offspring contains characteristics from both of its parents. Two individuals are selected probabilistically on the basis of their fitness scores and serve as parents. The selection strategy that used in this program was random selection method. Next step is applying a crossover operator that each parent contributes a random selection of half of its descriptors and an offspring is built by combining these two halves of genetic code. At last, this offspring is subjected to a random mutation in one of its gene, i.e. one variable is replaced by another. This selection–crossover–mutation procedure is repeated until all of the *n* parents in the gene pool are replaced by their offspring. The fitness score (*RMSE* and *RMSECV*) of each member of this new generation is again evaluated by using network, and the reproductive cycle is continued until a desired number of generation or target fitness score is reached. In the routine GA implemented in the MATLAB, some modifications were made (33). Here, for the computation of the fitness score of each chromosome a nonlinear model was built using descriptors consist in each chromosome separately by ANN technique and the values of MSE were calculated by means of this model. This process was used for each chromosome separately (33,34). The basic design of the algorithm combined genetic algorithms and ANN is summarized in the flow diagram shown in Fig. 1.

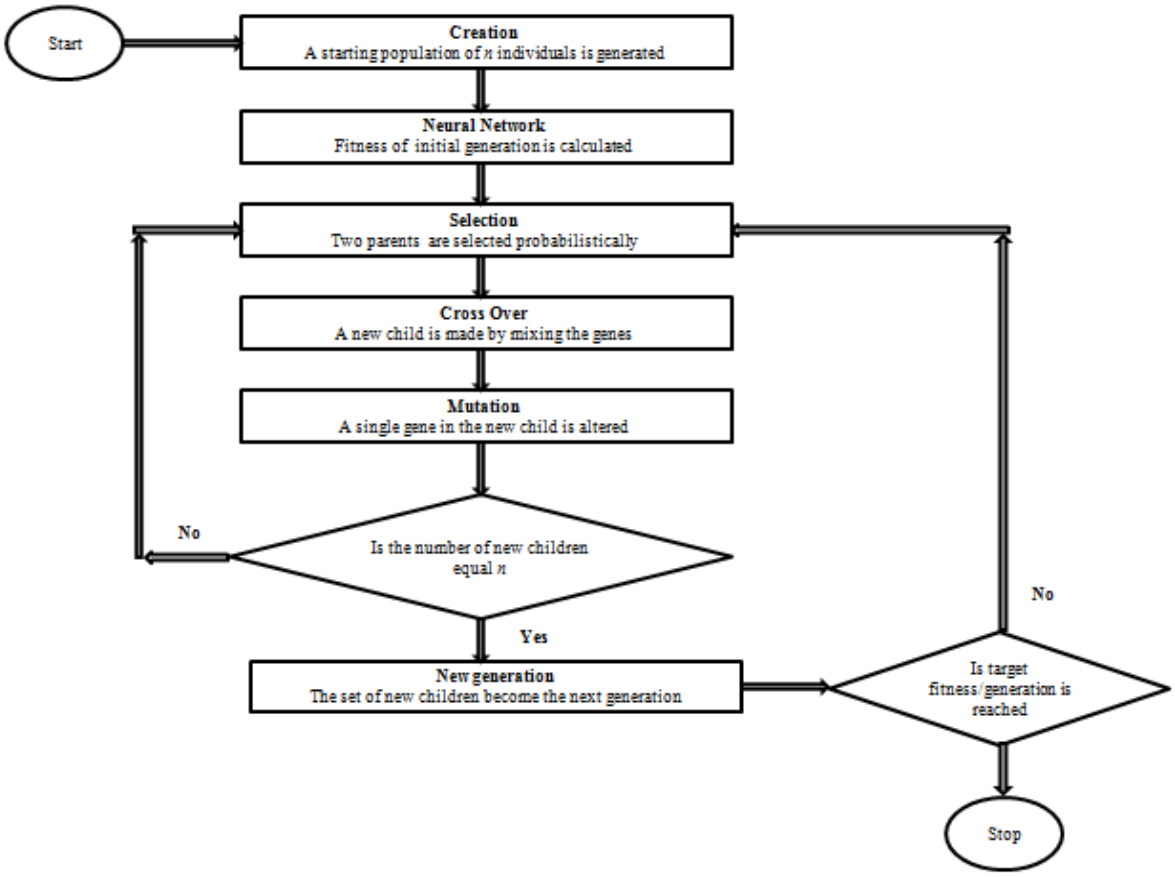


Fig. 1. The basic design of the algorithm combined genetic algorithms and artificial neural network used in this study.

Feature selection with GA–PCA strategy

The GA–PCA based coupling technique used to select the suitable features was performed using our own written routines for GA and PCA in MATALB environment. These m-codes are optimization tools based on the GA strategy in MATALB. Steps were applied in this process can be described as follows: (1) definition and encoding of chromosome; (2) population initialization; (3) evaluation of each chromosome; (4) protection of chromosome; (5) selection of best chromosomes; (6) crossover and mutation operations; (7) stoping if a halt condition is satisfied, otherwise going to step 3 (22). More description about the theory of this GA strategy can be found in the literature (22,35,36).

Artificial neural network

One method for providing a more flexible form of regression is to use a feed-forward neural network with error back-propagation learning algorithm. This is a computational system whose design is based on the architecture of biological neural networks which consists of artificial 'neurons' joined so that signals from one neuron can be passed to many others (23).

Clarification of the theory of the artificial neural networks in details has been adequately described elsewhere (37) but little relevant remarks is presented. ANN are parallel computational tools consisting of computing units named neurons and connections between neurons named synapses that are arranged in a series of layers (25).

Back propagation artificial neural network includes three layers. The first layer namely input layer has n_i neurons whose function is the reception of information (*i.e.* inputs) and their transferring to all neurons in the next layer called the hidden layer whose number is assined by n_h. The neurons in the hidden layer calculate a weighted sum of the inputs that is subsequently transformed by a linear or non-linear function. The last layer, the calculated response vector, is the output layer whose

neurons handle the output from the network (38). The task of synapses is the connection of input layer to hidden layer and hidden layer to output layer. The manner in which each node transforms its input depends on the "weights" and bias of the node, which are modifiable. On the other hand the output value of each node depends on both the weight, and biases values. In addition, depend on the weighted sum of all network inputs which are normally transformed by a nonlinear or linear transform function determine the outputs of the network (39).

The relation between response, Y_O, of the network and a vector input, X_i, can be written as following if the number of neurons in the output layer is equal to 1 (same with our condition in here):

$$Y_O = \sum_{J=1}^{N_H} W_J f\left(\sum_{I=1}^{N_I} W_{JI} X_I + b_I \right) \quad (3)$$

where, b_j is the bias term, W_{JI} is the weight of the connection between the *Ith* neuron of the input layer and the *Jth* neuron of the hidden layer, and f is the transformation function of the hidden layer. In the training process, the weights and bias of the network which are the adjustable parameters of the network are determined from a set of objects, which is known as training set.

Through the training of the network, the connection weights are regulated so that error of calculated responses and observed values were minimized. For this, a nonlinear transfer function makes a connection between the inputs and the outputs. Commonly neural network is adjusted, or trained, so that a particular input leads to a specific target output.

There are numerous algorithms available for training ANN models. We used back propagation algorithm here for training of the network. In this algorithm several steps for minimizing of networks were performed and the update of weight for the *(n + 1)th* pattern is given as:

$$W_{JI,n+1} = W_{JI,n} + \alpha \Delta W_{JI,n} \quad (4)$$

By using following equation, the descent down the error surface is calculated (40):

$$\Delta W_{JI,n} = -\mu \frac{\partial E}{\partial W_{JI,n}} \quad (5)$$

where, α and μ are momentum and learning rate, respectively.

With respect to above description, some adjustable parameters including number of nodes in input and hidden layers, transfer function of the hidden and transfer function output layers, momentum (optimum value in this study was 0.16), number of iteration for training of the network (17000 epoch), and learning rate (0.84) evaluated by obtaining those which result in minimum prediction are present in the ANN.

As described above, in order to avoid overfitting or underfitting, a validation set was used in the ANN modeling. Evaluation of ANN was performed on an external set (validation set) that consisted of molecules belonging to neither the training set nor the test set (41).

RESULTS

A lot of descriptors were calculated for each studied molecule using Dragon. In order to calculate a relationship with independent variable, logarithms of the inverse of biological activity (Log $1/IC_{50}$) data of 49 molecules were used. After dividing the molecules into training, validation, and test sets, building of QSAR models using training and validation sets was carried out (42).

PCA was performed on the training set after deleting constant descriptors. Results of PCA considering 20 first PC's and also their eigenvalues are reported in Table 2. In this Table, the eigenvalues, the percent of variances explained by each eigenvalue and the cumulative percent of variances are reported (34). After acquiring PC's, linear regression with stepwise factor selection was carried out. The cross-validation technique used was eliminating only one molecule at a time and then performing PCR on the remaining of training set (leave one out technique). The activity of the left-out object was predicted by using this developed model. This process was repeated until each molecule in the training set had been gone out once.

Table 2. The result of principal component analysis on the total descriptors.

Component	Eigenvalues	% of variance explained	Cumulative variance
1	177.11	62.15	62.15
2	53.39	18.73	80.88
3	16.90	5.93	86.81
4	9.68	3.39	90.20
5	5.84	2.05	92.26
6	4.93	1.73	93.99
7	3.13	1.10	95.08
8	2.80	0.98	96.06
9	2.29	0.80	96.87
10	1.61	0.56	97.43
11	1.15	0.40	97.84
12	0.97	0.34	98.18
13	0.82	0.29	98.47
14	0.66	0.23	98.70
15	0.62	0.22	98.91
16	0.55	0.19	99.11
17	0.44	0.15	99.26
18	0.37	0.13	99.39
19	0.29	0.10	99.49
20	0.24	0.09	99.58

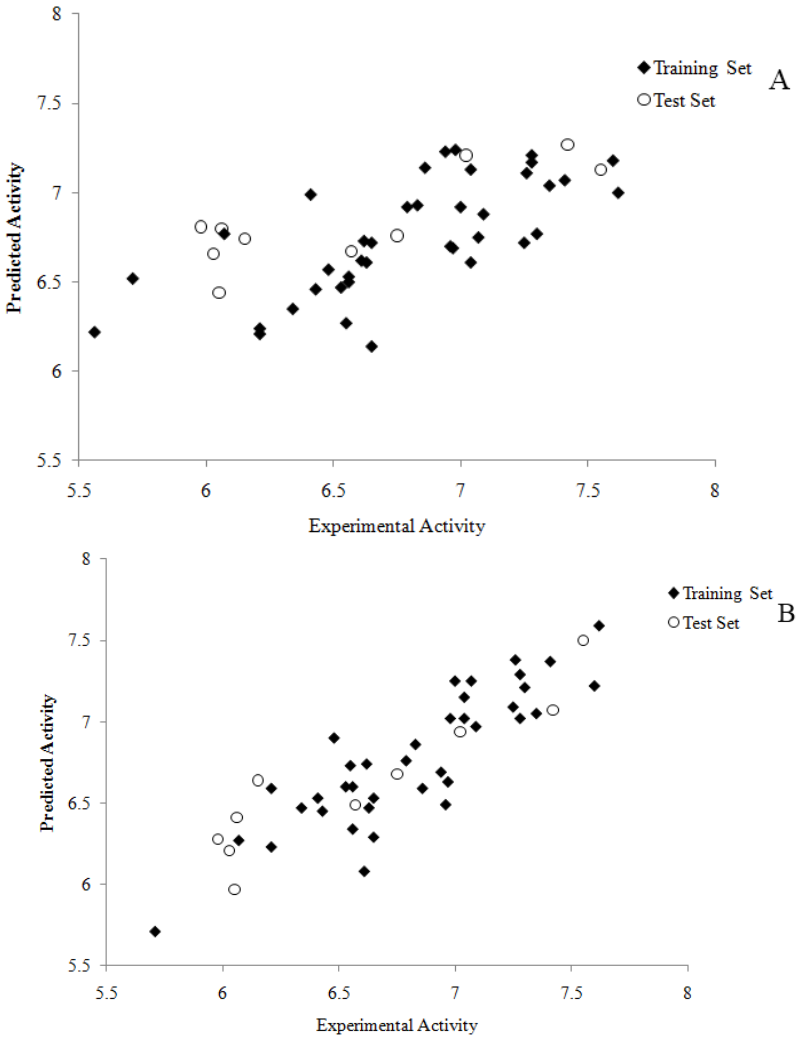


Fig. 2. pIC_{50} estimated by modeling versus experimental values for training and test sets A; PCR, B; GA-PC-ANN.

Table 3. The experimental pIC_{50} and the predicted values of the training set and test set.

Name	Activity	PCR		GA-PC-ANN	
		Predicted activity	Statistical errors	Predicted activity	Statistical errors
1	5.56	6.22	-0.66	5.45	0.11
2	6.53	6.47	0.06	6.60	-0.07
3*	6.05	6.44	-0.39	5.97	0.08
4	6.21	6.21	0.00	6.23	-0.02
5	6.34	6.35	-0.01	6.47	-0.12
6	6.56	6.50	0.06	6.34	0.21
7	6.21	6.24	-0.03	6.59	-0.38
8	6.61	6.62	-0.01	6.08	0.52
9	6.48	6.57	-0.09	6.90	-0.43
10	6.43	6.46	-0.02	6.45	-0.02
11	6.65	6.14	0.51	6.53	0.12
12	6.55	6.27	0.28	6.73	-0.18
13	6.65	6.72	-0.07	6.29	0.36
14	6.79	6.92	-0.13	6.76	0.03
15	6.62	6.73	-0.12	6.74	-0.13
16	6.83	6.93	-0.09	6.86	-0.03
17	7.09	6.88	0.20	6.97	0.12
18	6.94	7.23	-0.29	6.69	0.25
19	7.04	7.13	-0.09	7.02	0.01
20	7.25	6.72	0.53	7.09	0.16
21	7.28	7.21	0.06	7.02	0.26
22*	7.42	7.27	0.15	7.07	0.35
23*	6.03	6.66	-0.63	6.21	-0.18
24*	6.57	6.67	-0.10	6.49	0.08
25	6.63	6.61	0.02	6.47	0.15
26*	6.06	6.80	-0.74	6.41	-0.35
27	6.07	6.77	-0.69	6.27	-0.19
28*	5.98	6.81	-0.83	6.28	-0.29
29	6.97	6.69	0.28	6.63	0.34
30*	6.75	6.76	-0.01	6.68	0.07
31	7.07	6.75	0.32	7.25	-0.18
32	6.41	6.99	-0.58	6.53	-0.12
33	6.56	6.53	0.03	6.60	-0.04
34*	6.15	6.74	-0.58	6.64	-0.49
35	7.04	6.61	0.43	7.15	-0.11
36	7.62	7.00	0.62	7.59	0.03
37	7.28	7.17	0.12	7.29	0.00
38	7.00	6.92	0.07	7.25	-0.26
39	7.60	7.18	0.43	7.22	0.38
40	6.98	7.24	-0.25	7.02	-0.04
41*	7.55	7.13	0.42	7.50	0.05
42	6.96	6.70	0.26	6.49	0.47
43	7.41	7.07	0.34	7.37	0.03
44	6.86	7.14	-0.27	6.59	0.27
45	7.35	7.04	0.31	7.05	0.30
46	7.30	6.77	0.53	7.21	0.09
47	5.71	6.52	-0.81	5.71	0.00
48*	7.02	7.21	-0.19	6.94	0.08
49	7.26	7.11	0.15	7.38	-0.12

*Molecules used as external test set.

For evaluation of the predictive power of the generated PCR, the optimized model was applied for prediction of pIC_{50} values of all molecules in the training and test sets. The calculated pIC_{50} for each molecule and statistical error of prediction by model are summarized in Table 3. Experimental versus predicted values for pIC_{50} values of training and test set, obtained by the PCR modeling is shown graphically in Fig. 2A.

This model was validated by some statistical parameters such as PRESS and RMSE reported in Table 4. It is clear that this model on the basis of statistical parameters is weak. With respect to these results, it is a good idea to try a nonlinear regression method, artificial neural network, to obtain robust and predictive QSAR model able to describe a relationship between the structure and $P2X_7$ antagonist activity of the studied purine analogues.

In the non-linear model used, a network including a fully connected three layer, feed-forward ANN model trained with a back-propagation learning algorithm was used. The input of the network was the GA selected PC's. Seven PC's were selected by GA and were applied as input of the networks. In order to evaluate the ANN, MSE 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. Various ANN architectures were run with the seven selected PC's as input. In each run, the neuron architecture and parameters were optimized to reach the lowest MSE as the performances of the resulted models.

It must be considered that for inhibition of overfitting in the ANN model, the training of the network for the prediction of activity must be stopped when the MSE of the test set commences to increase while MSE of training set continues to decrease. Therefore, training of the network was stopped when overtraining began (43).

As mentioned above, before training the network, the number of nodes in the hidden layer must be optimized. For this purpose, several training of network was performed with different numbers of hidden nodes from 1 to 15. The MSE for training sets was obtained for different numbers of neurons at the hidden layer, and the minimum value of MSE was verified as the optimum value. A typical plot of MSE for training set versus the number of nodes in the hidden layer is shown in Fig. 3. It is clear that 7 nodes in the hidden layer is the optimum value.

The network was trained using training data and it was evaluated by prediction molecules. The predicted activity of the ANN calculated values of pIC_{50} are plotted against the experimental values in Fig. 2B and are reported in Table 3 and as expected, the calculated values are in good agreement with experimental values. The statistical parameters for the nonlinear model are represented in Table 4. As it is observed, the model obtained by the PC-GA-ANN has superior qualities relative to those obtained by PCR. This means that there is nonlinear relationship between the calculated PC's and the activity of the antagonists used in this study.

As a result, it was found that correctly opted and trained neural network could practically represent dependence of the activity of $P2X_7$ receptor antagonist to the extracted PC's from various geometrical, topological, and other calculated descriptors. Then, the optimized neural network could simulate the complicated nonlinear relationship between pIC_{50} value and the PC's.

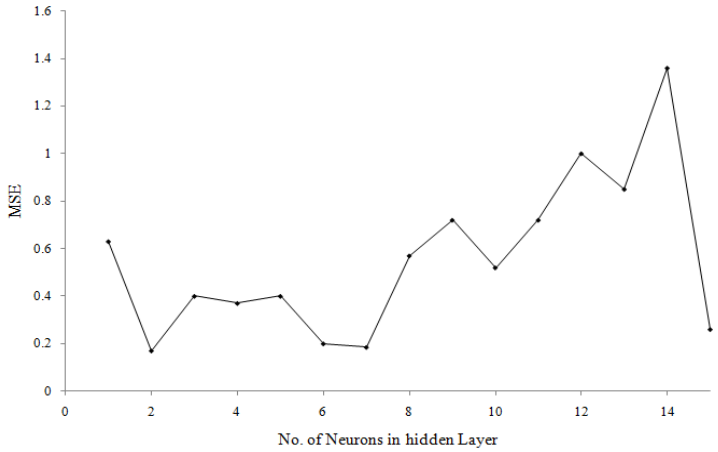


Fig. 3. Plot of MSE for training sets versus the number of nodes in hidden layer.

Table 4. Statistical parameters obtained for the QSAR models.

Statistical parameters	PCR			GA-PC-ANN		
	Training	Test set	Dataset	Training set	Test set	Dataset
N	39	10	49	39	10	49
R^2	0.487	0.668	0.449	0.792	0.844	0.794
RMSE	0.335	0.487	0.371	0.220	0.251	0.227
PRESS	4.389	2.374	6.763	1.893	0.632	2.526
R^2_{cv}	0.349			0.771		
RMSEcv	0.384			0.205		
PRESScv	5.630			1.702		

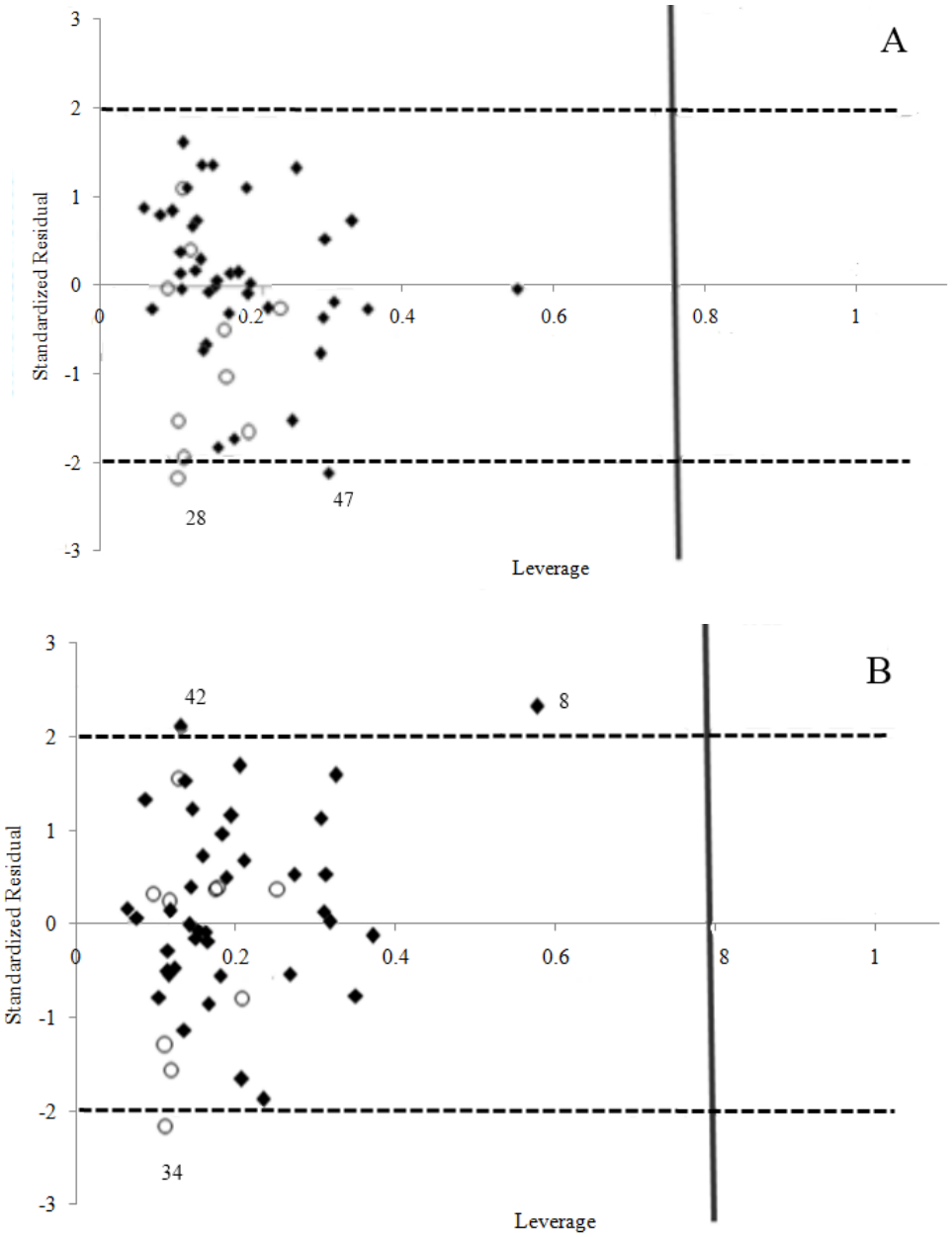


Fig. 4. Applicability domains of developed A; PCR and B; GA-PC-ANN.

DISCUSSION

The statistical parameters obtained by developed ANN model for the training and test sets were reported in Table 4. These results demonstrated some significant differences between PC-GA-ANN and PCR models. It can be seen from this Table that statistical results of the ANN model is better than other method. Also these results reveal that GA is superior method for feature-selection in this QSAR study.

All QSAR models have to constantly be verified for their applicability of domain (AD), in order to generate reliable predicted activities for compounds that are not too structurally dissimilar (44). The Williams plot verified the presence of outliers (i.e. molecules with standardized residuals greater than two standard deviation units) and the compounds that were very influential in determining model parameters i.e. molecules with high leverage value (*h*) (44,45) greater than 3 *(m + 1)/n*, where *m* is the number of the variables used in model formation, and *n* the number of the molecules employed to calculate the model. Also the data predicted by the models were verified for reliability by their leverage; with the intention that only predicted activities for molecules belonging to the chemical domain of the training set would be proposed (44). Actually, the leverage can be employed as a quantitative measure of the model applicability domain appropriate for assessing the degree of extrapolation: it represents a sort of compound "distance" from the model experimental space. The prediction for a compound having a high leverage value (h>h*, the warning limit leverage (WLL) h*= (3m/n) must be considered as unreliable (44). On the other hand, when the molecule has a leverage value lower than the WLL, the chance of accordance between the predicted and the experimental values is as high as that for the training set compounds (46). In this work each developed QSAR model was verified for the AD of the studied compounds to validate the prediction reliability.

Fig. 4A and 4B shows the AD of compounds of PCR and GA-PC-ANN employed in this study. Compounds influencing the structural domain of the model can be defined as molecules characterized by unusual structural features and badly represented in the training set. On the other hand, the outliers, whose standardized residual values exceed the cut off value of 2 standard deviation units, could be associated with the experimental error (4,47).

By examining the AD of the developed models (1) from the Williams plot (Fig. 4), it can be seen that two compounds for PCR model (molecule 28 and 47) and three compounds (molecule 34, 42 and 8) for GA-PC-ANN model are identified as a response outlier based on the 2σ rule for the training set. On the basis of Williams plot leverage method neither of the molecules in the studied set is identified as structurally influential chemicals. According to the results presented in the Williams plot, it is evident that neither of studied compounds is outlier.

CONCLUSION

A novel approach is suggested to select features from high-dimensional data in data mining. A GA procedure is used to search for the best subset in a PCA matrix and results is combined with ANN for $P2X_7$ receptor antagonist activity prediction. The performance of the approach was validated on the $P2X_7$ receptor inhibitory data by comparison of PC regression methods. The obtained results represent that the suggested approach (combination of GA and ANN) leads to a better subset of variables than original linear PCR method. The effectiveness of the GA is explained by the selection of the best set of PC's. The main aspect of developed model based on a neural network is its ability to allow for flexible mapping of the selected PC's by manipulating their functional dependence implicitly, unlike regression analysis. Models based on ANN handle both linear and nonlinear relationships between dependent and independent variables without adding complexity to the model. This capability offset the larger computing time required by a neural network simulation. This study shows that

combination of PCA and artificial intelligence methods are potentially helpful for the prediction of $P2X_7$ receptor inhibitory activity from a set of inhibitors. Also, it indicates that a non-linear method such as ANN is superior to a linear method such as PCR in building prediction models. In addition, the analysis of these GA selected PC's in the developed model can provide helpful clues to the structural and physicochemical characteristics of molecules contributing to the $P2X_7$ receptor inhibitory activity; this may help to provide reference information for ligand-based $P2X_7$ receptor antagonist drug design.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the Vice Chancellor for Research and Technology, Kermanshah University of Medical Sciences for financial support of the study. This article resulted from Pharm. D thesis of Mehdi Ahmadi, major of Pharmacy, Kermanshah University of Medical Sciences, Kermanshah, Iran.

REFERENCES

1. Gartland A, Buckley K, Bowler W, Gallagher J. Blockade of the pore-forming $P2X_7$ receptor inhibits formation of multinucleated human osteoclasts *in vitro*. Calcif Tissue Int. 2003;73:361-369.
2. Rassendren F, Buell GN, Virginio C, Collo G, North RA, Surprenant A. The permeabilizing ATP receptor, $P2X_7$ cloning and expression of a human cDNA. J Biol Chem. 1997;272:5482-5486.
3. Zheng LM, Zychlinsky A, Liu CC, Ojcius DM, Young JD. Extracellular ATP as a trigger for apoptosis or programmed cell death. J Cell Biol. 1991;112:279-288.
4. Kušić H, Rasulev B, Leszczynska D, Leszczynski J, Koprivanac N. Prediction of rate constants for radical degradation of aromatic pollutants in water matrix: a QSAR study. Chemosphere. 2009;75:1128-1134.
5. Carroll WA, Donnelly-Roberts D, Jarvis MF. Selective $P2X_7$ receptor antagonists for chronic inflammation and pain. Purinergic signalling. 2009;5:63-73.
6. Guo Q, Wu W, Questier F, Massart D, Boucon C, De Jong S. Sequential projection pursuit using genetic algorithms for data mining of analytical data. Anal Chem. 2000;72:2846-2855.
7. Czarnik AW. Peer Reviewed: Combinatorial Chemistry. Anal Chem. 1998;70:378A-386A.
8. Daszykowski M, Walczak B, Massart D. Representative subset selection. Anal Chim Acta. 2002;468:91-103.
9. Saghaie L, Shahlaei M, Fassihi A. Quantitative structure activities relationships of some 2-mercaptoimidazoles as CCR2 inhibitors using genetic algorithm-artificial neural networks. Res Pharm Sci. 2013;8:97.
10. Shahlaei M, Fassihi A, Saghaie L. Application of PC-ANN and PC-LS-SVM in QSAR of CCR1 antagonist compounds: a comparative study. Eur J Med Chem. 2010;45:1572-1582.
11. Jolliffe IT. Discarding variables in a principal component analysis. I: Artificial data. Appl Stat. 1972; 21:160-173.
12. McCabe GP. Principal variables. Technometrics. 1984;26:137-144.
13. Rännar S, Wold S, Russell E. Selection of spanning variables in PCA. Many Variables in Multivariate Projection Methods, Ph D Thesis, Department of Organic Chemistry, Umeå University, Sweden. 1996.
14. Krzanowski W. Selection of variables to preserve multivariate data structure, using principal components. Appl Stat. 1987;36:22-33.
15. Hosmer Jr DW, Lemeshow S. Applied logistic regression. 3rd edition. New Jersey: John Wiley & Sons; 2004. p. 153-222.
16. Fatemi M, Jalali-Heravi M, Konuze E. Prediction of bioconcentration factor using genetic algorithm and artificial neural network. Anal Chim Acta. 2003;486:101-108.
17. Guerriere MR, Detsky AS. Neural networks: what are they? Ann Intern Med. 1991;115:906-907.
18. Hinton GE. How neural networks learn from experience. Sci Am. 1992;267:145-151.
19. Shahlaei M, Madadkar-Sobhani A, Fassihi A, Saghaie L, Shamshirian D, Sakhi H. Comparative quantitative structure–activity relationship study of some 1-aminocyclopentyl-3-carboxyamides as CCR2 inhibitors using stepwise MLR, FA-MLR, and GA-PLS. Med Chem Res. 2012;21:100-115.
20. Shahlaei M, Fassihi A, Nezami A. QSAR Study of some 5-methyl/trifluoromethoxy-1H-indole-2, 3-dione-3-thiosemicarbazone derivatives as anti-tubercular agents. Res Pharm Sci. 2009;4:123-131.
21. Shahlaei M, Nazari Z. Computational neural network analysis of the affinity of 2-pyridyl-3, 5-diaryl pyrroles analogs for the human glucagon receptor using density functional theory. Med Chem Res. 2014;23:2046-2061.
22. Matasi JJ, Brumfield S, Tulshian D, Czarnecki M, Greenlee W, Garlisi CG, *et al*. Synthesis and SAR development of novel $P2X_7$ receptor antagonists for the treatment of pain: Part 1. Bioorg Med Chem Lett. 2011;21:3805-3808.
23. Brumfield S, Matasi JJ, Tulshian D, Czarniecki M, Greenlee W, Garlisi C, *et al*. Synthesis and SAR development of novel $P2X_7$ receptor antagonists for the treatment of pain: Part 2. Bioorg Med Chem Lett. 2011;21:7287-7290.

24. Sutter J, Peterson T, Jurs P. Prediction of gas chromatographic retention indices of alkylbenzenes. Anal Chim Acta. 1997;342:113-122.
25. Kennard R, Stone L. Computer aided design of experiments. Technometrics. 1969;11:137-148.
26. Saghaie L, Shahlaei M, Madadkar-Sobhani A, Fassihi A. Application of partial least squares and radial basis function neural networks in multivariate imaging analysis-quantitative structure activity relationship: study of cyclin dependent kinase 4 inhibitors. J Mol Graphics Model. 2010;29:518-528.
27. Mauri A, Consonni V, Pavan M, Todeschini R. Dragon software: An easy approach to molecular descriptor calculations. MATCH Commun Math Comput Chem. 2006;56:237-248.
28. Tan N, Rao H, Li Z, Li X. Prediction of chemical carcinogenicity by machine learning approaches. SAR QSAR Environ Res. 2009;20:27-75.
29. Habibi-Yangjeh A, Pourbasheer E, Danandeh-Jenagharad M. Application of principal component-genetic algorithm-artificial neural network for prediction acidity constant of various nitrogen-containing compounds in water. Monatsh Chem Chem Mon. 2009;140:15-27.
30. Goldberg DE, Holland JH. Genetic algorithms and machine learning. Mach Learn. 1988;3:95-99.
31. Habibi-Yangjeh A. QSAR study of the 5-HT1A receptor affinities of arylpiperazines using a genetic algorithm–artificial neural network model. Monatsh Chem Chem Mon. 2009;140:523-530.
32. Shahlaei M. Descriptor selection methods in quantitative structure–activity relationship studies: a review study. Chem Rev. 2013;113:8093-8103.
33. Habibi-Yangjeh A, Pourbasheer E, Danandeh-Jenagharad M. Prediction of melting point for drug-like compounds using principal component-genetic algorithm-artificial neural network. Bull Korean Chem Soc. 2008;29:833-841.
34. Fatemi M. Prediction of ozone tropospheric degradation rate constant of organic compounds by using artificial neural networks. Anal Chim Acta. 2006;556:355-363.
35. Cong Y, Li B-k, Yang Xg, Xue Y, Chen Yz, Zeng Y. Quantitative structure–activity relationship study of influenza virus neuraminidase A/PR/8/34 (H1N1) inhibitors by genetic algorithm feature selection and support vector regression. Chemometr Intell Lab. 2013;127:35-42.
36. Leardi R, Boggia R, Terrile M. Genetic algorithms as a strategy for feature selection. J Chemometrics. 1992;6:267-281.
37. Leardi R. Application of a genetic algorithm to feature selection under full validation conditions and to outlier detection. J Chemometrics. 1994;8:65-79.
38. Caballero J, Fernández M. Linear and nonlinear modeling of antifungal activity of some heterocyclic ring derivatives using multiple linear regression and Bayesian-regularized neural networks. J Mol Model. 2006;12:168-181.
39. Bose NK, Liang P. Neural network fundamentals with graphs, algorithms, and applications. New York: McGraw-Hill, Inc; 1996.
40. Gupta VK, Khani H, Ahmadi-Roudi B, Mirakhorli S, Fereyduni E, Agarwal S. Prediction of capillary gas chromatographic retention times of fatty acid methyl esters in human blood using MLR, PLS and back-propagation artificial neural networks. Talanta. 2011;83:1014-1022.
41. Jalali-Heravi M, Fatemi M. Simulation of mass spectra of noncyclic alkanes and alkenes using artificial neural network. Anal Chim Acta. 2000;415:95-103.
42. Chang WF, Mak MW. A conjugate gradient learning algorithm for recurrent neural networks. Neurocomputing. 1999;24:173-189.
43. Tashkhourian J, Hormozi-Nezhad MR, Khodaveisi J, Dashti R. Localized surface plasmon resonance sensor for simultaneous kinetic determination of peroxyacetic acid and hydrogen peroxide. Anal Chim Acta. 2013;762:87-93.
44. Saghaie L, Shahlaei M, Fassihi A, Madadkar-Sobhani A, Gholivand MB, Pourhossein A. QSAR Analysis for Some Diaryl-substituted Pyrazoles as CCR2 Inhibitors by GA-Stepwise MLR. Chem Biol Drug Des. 2011;77:75-85.
45. Garkani-Nejad Z, Ahmadvand M. Investigation of linear and nonlinear chemometrics methods in modeling of retention time of phenol derivatives based on molecular descriptors. Separ Sci Technol. 2011;46:1034-1044.
46. Gramatica P, Papa E. QSAR modeling of bioconcentration factor by theoretical molecular descriptors. QSAR Comb Sci. 2003;22:374-385.
47. Atkinson AC. Plots, transformations, and regression: an introduction to graphical methods of diagnostic regression analysis. Oxford: Clarendon Press; 1985.