European Journal of Medicinal Chemistry 44 (2009) 1463–1470

Contents lists available at ScienceDirect

European Journal of Medicinal Chemistry

journal homepage: http://www.elsevier.com/locate/ejmech

Original article

Quantitative structure–activity relationship study of serotonin (5-HT$_7$) receptor inhibitors using modified ant colony algorithm and adaptive neuro-fuzzy interference system (ANFIS)

M. Jalali-Heravi*, M. Asadollahi-Baboli

Department of Chemistry, Sharif University of Technology, P.O. Box 11155-9516, Tehran, Iran

ARTICLE INFO

Article history:
Received 20 July 2008
Received in revised form
17 September 2008
Accepted 29 September 2008
Available online 14 October 2008

Keywords:
Quantitative structure–activity relationship
Modified ant colony
Adaptive neuro-fuzzy interference system
Serotonin receptor
Shuffling cross-validation

ABSTRACT

Quantitative structure–activity relationship (QSAR) approach was carried out for the prediction of inhibitory activity of some novel quinazolinone derivatives on serotonin (5-HT$_7$) using modified ant colony (ACO) method and adaptive neuro-fuzzy interference system (ANFIS) combined with shuffling cross-validation technique. A modified ACO algorithm is utilized to select the most important variables in QSAR modeling and then these variables were used as inputs of ANFIS to predict 5-HT$_7$ receptor binding activities of quinazolinone derivatives. The best descriptors describing the inhibition mechanism are Q_{max}, Se, Hy, PJI3 and DELS which are among electronic, constitutional, geometric and empirical descriptors. The statistical parameters of R^2 and root mean square error are 0.775 and 0.360, respectively. The ability and robustness of modified ACO-ANFIS model in predicting inhibition behavior of quinazolinone derivatives (pIC$_{50}$) are illustrated by validation techniques of leave-one-out and leave-multiple-out cross-validations and also by Y-randomization technique. Comparison of the modified ACO-ANFIS method with two other methods, that is, stepwise MLR-ANFIS and GA-PLS-ANFIS were also studied and the results indicated that the proposed model in this work is superior over the others.

1. Introduction

In humans, serotonin (5-HT) receptor is well documented to play a key role in mood and emotion, and a number of psychiatric conditions including anxiety, depression, schizophrenia, and anorexia nervosa and it is undoubtedly involved in attention, learning, and memory [1,2]. The complex interactions between 5-HT neurons and other neuronal phenotypes, 5HT receptor heterogeneity, and the conflicting results of behavioral experiments have made functional assessments of 5-HT in cognitive processes particularly difficult. Nevertheless, there is now a credible body of evidence suggesting that several of the 5-HT receptor subtypes could indeed serve as therapeutics targets for memory enhancement in some neuropsychiatric disorders [1,2].

There are seven types of 5-HT receptors, 5-HT$_1$–5-HT$_7$, in which the 5-HT$_7$ receptor is recently cloned [3,4]. The 5-HT$_7$ receptor has been identified as important factor in circadian rhythms and sleep and some recent electrophysiologic studies appear to suggest that the receptor could serve as a target for anticonvulsant drugs [5]. This receptor also binds several antidepressants and antipsychotics with high affinity, indicating that this receptor may represent a therapeutic target for schizophrenia and other psychiatric disorders [6]. Since 5-HT$_7$ has been associated with these diseases, its receptor ligand has potential therapeutic value and has been extensively studied in pharmaceutical industry. Some novel quinazolinone derivatives are recently reported as selective inhibitor of 5-HT$_7$ with very good binding affinities [7].

To rationalize the findings and design compounds with enhanced activity, a systematic study of the various substituents on the activity of the analogues is needed. In addition, the growth of computational techniques has accelerated the drug design process. Quantitative structure–activity relationship (QSAR) searches information relating chemical structure to biological and other activity by developing a QSAR model. By using such an approach one could predict the activities of newly designed compounds before a decision is being made whether these compounds should be really synthesized and tested. Building a QSAR model begins with calculating theoretical parameters for the compounds involved. The experimental information associated with biological properties is taken as dependent variables in developing a model. Several descriptors could be generated in QSAR studies, but only some of them are statistically significant in terms of correlation with biological activity for a particular analysis. The challenge, therefore, is to select the subset of descriptors that describe the most critical structural and physicochemical features associated

* Corresponding author. Tel.: +98 21 66165315; fax: +98 21 66012983.
E-mail address: jalali@sharif.edu (M. Jalali-Heravi).

0223-5234/$ – see front matter © 2008 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.ejmech.2008.09.050

with inhibitory activity. Effective descriptor or variable selection is an integral part of the QSAR modeling process [8]. Obtaining a good quality QSAR model depends on many factors, such as the quality of biological data, the choice of descriptors and statistical methods. There have been many variable selection methods the mostly used ones are stepwise regression, evolutionary algorithms [9] and genetic algorithms [10,11] and so on. Recently, a variable selection technique was proposed by Qi Shen and co-workers [12] based on ant colony optimization [13,14]. ACO has emerged recently as stochastic optimization approach, which originated as a simulation of ant colony system. As a novel computational approach, ACO algorithms have attracted attention of researchers in many fields [15,16].

For developing a reliable QSAR model, adaptive neuro-fuzzy interference system (ANFIS) was used. The synergism of fuzzy logic systems and neural network has produced a functional system capable of learning high-level thinking, and reasoning [17]. It is an improved tool for determining the behavior of imprecisely defined complex systems. The purpose of a neuro-fuzzy system is to apply neural learning techniques to identify the parameters and/or structure of neuro-fuzzy systems. These neuro-fuzzy systems can combine the benefits of the two powerful paradigms into a single capsule [18]. The fuzzy systems have fascinated the growing attention and interest in bioinformatics applications, decision-making studies, pattern recognition, and data analysis [19].

In the present work, we have examined modified ACO algorithm and shuffling cross-validation for variable selection and ANFIS for model developing in QSAR analysis of some novel quinazolinone derivatives inhibitory activity. The results of modified ACO-ANFIS were compared to those of stepwise MLR-ANFIS (stepwise multiple linear regression) and GA-PLS-ANFIS (genetic algorithm-partial least square). It has been demonstrated that the modified ACO is a useful tool for variable selection comparable to the stepwise MLR and GA-PLS. Finally, the accuracy of proposed model was illustrated using leave-one-out (LOO), leave-multiple-out (LMO) cross-validations and *Y*-randomization techniques.

2. Theoretical routines

2.1. Modified ant colony optimization

Ant algorithm optimization was first proposed by Dorigo and Gambardella [13,14] as a multi-agent approach for difficult combinatorial optimization problems such as traveling sales man problem (TSP) and the quadratic assignment problem [15]. This algorithm has been inspired by the behavior of real ant. Real ants are capable of finding the shortest path from a food source to their colony. They leave some pheromone on the ground, thus marking the path by the trail of the substance. The pheromone trail can be observed by other ants and motivates them to follow the path [13]. A moving ant will follow the pheromone trail with a probability which is proportional to the number of ants choosing that path as each ant would reinforce the trail by depositing its own pheromone. Therefore, pheromone will be updated after an ant passed a path [20].

From then researchers have applied ACO to many other problems such as variable selection [21,22]. The variable selection problem refers to the task of identifying and selecting a useful subset of features to be used to represent patterns from a larger set of often redundant or possibly irrelevant descriptors. Therefore, variable selection in QSAR is a subset selection problem which is quite different from the ordering problems. In this algorithm called modified ACO, a binary notation was used for a variable selection. In modified ACO, *N* variables are considered as *N*-dimensional search space in which an ant motion is restricted to 0 or 1 on each dimension. State “1” represents the selection of this variable and state “0” represents the reverse.

ACO was applied to the variable selection problem as following: let b_i be the number of ants on variable *i* and let $m = \sum_{i=1}^{n} b_i$ be the total number of ants, and let τ_{i1} and τ_{i0} be the intensity of the pheromone trail on variable *i* corresponding to a dimension taking the value 1 or 0. τ_{i0} and τ_{i1} are updated according to these update rules:

$$\tau_{i0}(\text{new}) = f\tau_{i0}(\text{old}) + \Delta\tau_{i0} \tag{1}$$

$$\Delta\tau_{i0} = \sum_{k=1}^{m} \Delta\tau_{i0}^{(k)} \tag{2}$$

$$\tau_{i1}(\text{new}) = f\tau_{i1}(\text{old}) + \Delta\tau_{i1} \tag{3}$$

$$\Delta\tau_{i1} = \sum_{k=1}^{m} \Delta\tau_{i1}^{(k)} \tag{4}$$

where *f* is a coefficient ($0 < f < 1$) which represents the extent of pheromone retained on the path from previous iteration. $\Delta\tau_{i0}$ and $\Delta\tau_{i1}$ presented the increment of pheromone corresponding to the dimension *i* taking the value 1 or 0 at this iteration ($f = 0.8$ in our study). $\Delta\tau_{i0}^{(k)}$ and $\Delta\tau_{i1}^{(k)}$ showed the amount of pheromone that ant *k* left on the variable *i* at this iteration. These two parameters can be calculated by following equations:

$$\Delta\tau_{i1}^{(k)} = \text{Fitness if variable}i\text{was selected by ant } k \text{ in the current iteration} \tag{5}$$

$$\Delta\tau_{i0}^{(k)} = \text{Fitness if variable } i\text{was not selected by ant } k \text{ in the current iteration} \tag{6}$$

Here *Fitness* is a fitness function. As can be seen from these equations, the amount of pheromone depends on the importance of variable *i* in the model. This fitness function is calculated with selected variables by ants using partial least square method. For calculating the fitness function, multi-objective fitness function was used to account both the residual errors and the number of variables according to Eq. (7):

$$\text{Fitness} = \left(\text{RMSE}_{\text{Cal}} + \text{RMSE}_{\text{Val}} + p^{1/2}\right)^{-1} \tag{7}$$

where RMSE_{Cal} and RMSE_{Val} values are root mean square error of calibration and validation sets, respectively. *p* is the number of variables in the model. The smaller the RMSE of partial least square model is and the fewer variables are involved in the model, the larger the fitness function is and the higher probability that those variables are being selected.

Ant *k* makes decision concerning the variable selection according to the pheromone amount. The moving probability is:

$$p_i^{(k)} = \frac{\tau_{i1}}{\tau_{i1} + \tau_{i0}} \tag{8}$$

In the modified ACO, *m* ants select important variables from all *N* variables according to the probability defined by Eq. (8). After one variable has been selected, the amount of pheromone is updated according to Eqs. (1)–(7). This process is iterated until the minimum error criterion is attained. In this algorithm, the pheromone levels were updated not only by current individual’s information but also by each ant’s previous performance. The algorithm of this method is shown in Fig. 1. The detail of the modified ACO has been described elsewhere [12,21].

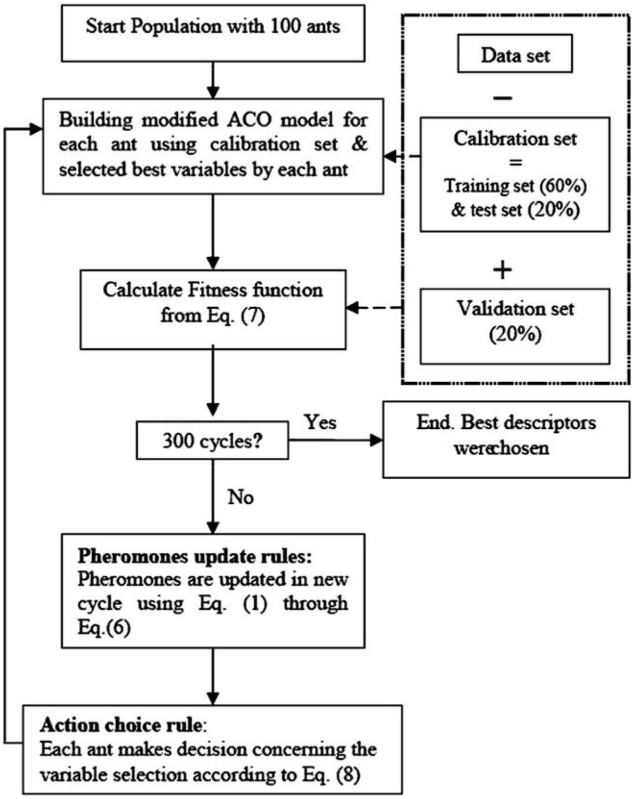
Start Population with 100 ants
Building modified ACO model for each ant using calibration set & selected best variables by each ant
Data set
–
Calibration set
=
Training set (60%) & test set (20%)
+
Validation set (20%)
Calculate Fitness function from Eq. (7)
300 cycles?
Yes
End. Best descriptors werechosen
No
Pheromones update rules: Pheromones are updated in new cycle using Eq. (1) through Eq.(6)
Action choice rule: Each ant makes decision concerning the variable selection according to Eq. (8)

Fig. 1. Scheme of modified ACO algorithm for variable selection.

2.2. Adaptive neuro-fuzzy interference system

The proposed neuro-fuzzy model in ANFIS is a multilayer neural network-based fuzzy system [18,23]. Its topology is shown in Fig. 2, and the system has a total of five layers. In this connectionist structure, the input (*layer 0*) and output (*layer 5*) nodes represent the descriptors and the response, respectively, and in the hidden layers, there are nodes functioning as membership functions (MFs) and rules. This eliminates the disadvantage of a normal feed forward multilayer network, which is difficult for an observer to understand or to modify. ANFIS simulates TSK (Takagi-Sugeno-Kang) fuzzy rule [24] of type-3 where the consequent part of the rule is a linear combination of input variables and a constant. For a Sugeno fuzzy model a common rule set with the fuzzy if-then rule is as following:

$$\text{IF } x \text{ is } A_i \text{ and } y \text{ is } A_i \text{ THEN } f_i = p_i x + q_i y + r_i \quad \text{for} \quad i = 1,2$$

For simplicity, we assume that the examined fuzzy inference system has two inputs x and y and one output. The ANFIS contains five layers as shown in Fig. 2.

2.2.1. Layer 1

The fuzzy part of ANFIS is mathematically incorporated in the form of membership functions (MFs). A membership function $\mu_{A_i}(x)$ can be any continuous and piecewise differentiable function that transforms the input value x into a membership degree, that is to say a value between 0 and 1. The most widely applied membership functions are the generalized bell (gbell MF) or the Gaussian function in Eqs. (9) and (10), which are described by the three parameters, a, b, and c. Therefore, Layer 1 is the *fuzzification* layer in which each node represents a membership:

$$\mu_{A_i}(x) = \frac{1}{1 + \left[\left(\frac{x-c_i}{a_i}\right)^2\right]^{b\_i}} \tag{9}$$

$$\mu_{A_i}(x) = \exp\left[-\left(\frac{x-c_i}{a_i}\right)^2\right] \tag{10}$$

As the values of the parameters {a_i, b_i and c_i} change, the bell-shaped functions vary accordingly, thus exhibiting various forms of membership functions on linguistic label A_i. Parameters in this layer are referred to as premise parameters.

2.2.2. Layer 2

Every node in this layer is a fixed node labeled, whose output is the product of all the incoming signals:

$$O_{2,1} = w_i = \mu_{A_i}(x) \times \mu_{B_i}(y) \quad \text{for} \quad i = 1,2 \tag{11}$$

Every node in this layer computes the multiplication of the input values and gives the product as the output as in the above equation. The membership values represented by $\mu_{A_i}(x)$ and $\mu_{B_i}(y)$ are multiplied in order to find the firing strength of a rule where the variables x and y has linguistic values A_i and B_i, respectively.

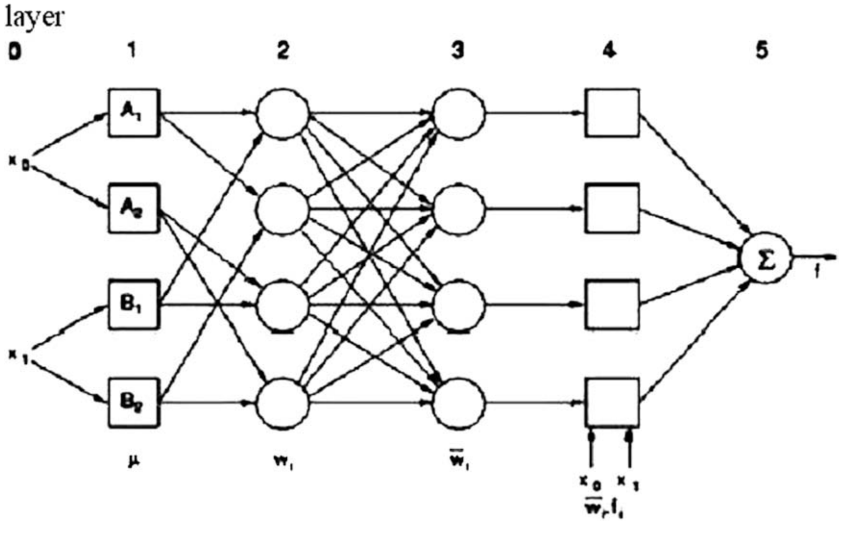

Fig. 2. Basic ANFIS structure.

Table 1
Experimental and calculated inhibitor data using modified ACO-ANFIS model for novel quinazolinone derivatives together with their structures used in QSAR study

No	Subset[a]	*n*[b]	X[c]	Y[c]	R1[c]	R2[c]	Experimental *p*IC$_{50}$	Modified ACO-ANFIS[d]
1	C	0	H	H	H	H	6.027	6.343
2	D	0	H	H	H	*o*-Cl	6.292	6.157
3	B	0	H	H	H	*p*-Cl	6.432	6.072
4	A	0	H	H	H	*p*-Me	6.137	6.389
5	E	0	H	H	H	2,3-Me2	7.022	7.210
6	C	0	H	H	H	2,4-Me2	6.194	6.012
7	B	0	H	H	H	3,4-Me2	5.959	6.533
8	D	0	H	H	H	*o*-OMe	7.097	6.912
9	A	0	H	H	H	*m*-OMe	5.854	5.637
10	E	0	H	H	H	*p*-OMe	6.149	6.315
11	B	0	H	H	H	*o*-OEt	7.347	7.126
12	C	0	H	H	H	*p*-Ac	5.046	5.336
13	E	0	H	H	*p*-F	H	6.167	6.411
14	A	0	H	H	*p*-F	2,4-Me2	6.301	6.024
15	D	0	H	H	*p*-F	2,6-Me2	6.301	6.469
16	B	0	H	H	*p*-F	*p*-OMe	5.602	5.836
17	C	0	H	H	*p*-F	*o*-OEt	5.886	5.674
18	A	0	H	H	*p*-F	*p*-NO2	5.137	5.295
19	E	1	H	H	H	H	6.187	6.050
20	D	1	H	H	H	*o*-F	6.398	6.499
21	B	1	H	H	H	*p*-F	5.620	5.851
22	E	1	H	H	H	*o*-Cl	6.886	6.950
23	C	1	H	H	H	*m*-Cl	6.959	7.282
24	A	1	H	H	H	*p*-Cl	6.347	6.161
25	B	1	H	H	H	3,4-Cl2	6.796	7.084
26	D	1	H	H	H	2,3-Me2	6.337	6.475
27	A	1	H	H	H	2,4-Me2	6.161	6.326
28	E	1	H	H	H	2,5-Me2	6.377	6.230
29	C	1	H	H	H	3,4-Me2	6.699	7.117
30	A	1	H	H	H	*o*-OMe	7.678	7.564
31	B	1	H	H	H	*p*-OMe	5.699	5.415
32	E	1	H	H	H	*o*-OEt	7.585	7.512
33	D	1	H	H	H	*m*-CF3	6.456	6.341
34	E	1	H	F	H	H	6.032	5.863
35	C	1	H	F	H	*o*-F	6.114	5.737
36	E	1	H	F	H	*p*-F	6.114	6.270
37	A	1	H	F	H	*o*-Cl	7.284	7.379
38	D	1	H	F	H	*m*-Cl	6.721	6.420
39	B	1	H	F	H	*p*-Cl	5.658	5.910
40	E	1	H	F	H	3,4-Cl2	6.347	6.013
41	C	1	H	F	H	2,3-Me2	7.125	6.957
42	A	1	H	F	H	2,4-Me2	6.027	6.092
43	D	1	H	F	H	2,5-Me2	5.824	5.721
44	E	1	H	F	H	3,4-Me2	6.229	6.596
45	C	1	H	F	H	*o*-OMe	7.071	7.556
46	B	1	H	F	H	*p*-OMe	5.959	5.511
47	D	1	H	F	H	*o*-OEt	7.721	7.425
48	D	1	H	F	H	*m*-CF3	7.036	7.307
49	E	1	H	F	H	*p*-Ac	6.167	6.430
50	C	1	F	H	H	H	6.328	6.595
51	A	1	F	H	H	*o*-F	6.000	6.327
52	B	1	F	H	H	*p*-F	6.377	6.610
53	D	1	F	H	H	*o*-Cl	6.699	6.914
54	E	1	F	H	H	*m*-Cl	6.886	7.218
55	C	1	F	H	H	*p*-Cl	5.921	5.527
56	A	1	F	H	H	3,4-Cl2	6.569	6.832
57	B	1	F	H	H	2,3-Me2	7.260	7.020
58	E	1	F	H	H	2,4-Me2	6.208	6.454
59	D	1	F	H	H	2,5-Me2	6.699	6.481
60	A	1	F	H	H	3,4-Me2	6.585	6.508
61	C	1	F	H	H	*o*-OMe	6.921	6.875
62	E	1	F	H	H	*m*-OMe	6.036	6.156
63	B	1	F	H	H	*p*-OMe	6.456	6.687
64	D	1	F	H	H	*o*-OEt	7.921	7.836
65	A	1	F	H	H	*m*-CF3	7.131	7.010
66	E	1	F	H	H	*p*-Ac	6.229	6.495
67	C	1	H	H	*o*-OMe	*o*-OMe	6.678	6.464
68	B	1	H	H	*o*-OMe	*o*-OEt	7.538	7.681
69	D	1	H	F	*o*-OMe	*o*-OEt	6.959	6.715
70	A	1	F	H	*o*-OMe	*o*-OMe	7.102	7.427
71	E	1	F	H	*o*-OMe	*o*-OEt	7.319	7.084
72	C	1	H	H	*m*-OMe	*o*-OMe	7.009	7.253
73	D	1	H	H	*m*-OMe	*o*-OEt	7.260	7.023
74	A	1	H	F	*m*-OMe	*o*-OEt	7.041	7.257
75	B	1	F	H	*m*-OMe	*o*-OMe	6.638	6.382
76	E	1	F	H	*m*-OMe	*o*-OEt	7.796	7.851

Table 1 (*continued*)

No	Subset[a]	*n*[b]	X[c]	Y[c]	R1[c]	R2[c]	Experimental pIC_{50}	Modified ACO-ANFIS[d]
77	A	1	H	H	*p*-OMe	*o*-OMe	7.013	6.865
78	C	1	H	H	*p*-OMe	*o*-OEt	7.538	7.698
79	D	1	H	F	*p*-OMe	*o*-OEt	7.310	7.467
80	B	1	F	H	*p*-OMe	*o*-OMe	6.721	6.984
81	C	1	F	H	*p*-OMe	*o*-OEt	7.796	7.481

[a] A, B, C, D and E subsets.
[b] Number of CH_2 group.
[c] X, Y, R1,R2 are substituted groups in quinazolinone derivatives shown in Fig. 3.
[d] The calculated values of quinazolinone derivatives for run 6 in Table 3.

2.2.3. Layer 3

This layer is the normalization layer which normalizes the strength of all rules according to the Eq. (11):

$$O_{3,i} = \overline{w}_i = \frac{w_i}{w_1 + w_2} \text{ for } i = 1,2 \tag{12}$$

where w_i is the firing strength of the *i*th rule which is computed in layer 2. Node *i* computes the ratio of the *i*th rule's firing strength to the sum of all rules' firing strengths. For convenience, outputs of this layer are called normalized firing strengths.

2.2.4. Layer 4

Every node *i* in this layer is an adaptive node with a node function:

$$O_{4,i} = \overline{w}_i f_i = \overline{w}_i (p_i x + q_i y + r_i) \tag{13}$$

where w_i is a normalized firing strength from layer 3 and $\{p_i, q_i, r_i\}$ is the parameter set of this node. Parameters in this layer are referred to as consequent parameters.

2.2.5. Layer 5

The single node in this layer is a fixed node labeled $\sum$, which computes the overall output as the summation of all incoming signals:

$$\text{overall output} = O_{5,i} = \sum_i \overline{w}_i f_i = \frac{\sum_i \overline{w}_i f_i}{\sum_i w_i} \tag{14}$$

Thus we have constructed an ANFIS system that is functionally equivalent to Sugeno fuzzy model, which has been used in the present QSAR study due to its transparency and efficiency.

2.3. Shuffling cross-validation

In this technique, the data set would be divided into several subsets, and variable selection process and model developing would be performed for all combinations of the subsets. Then the most frequent descriptors appeared in the developed models would be selected as most important variables, describing the inhibition behavior. Therefore, these techniques can be used for the feature selection.

In the present work, the data set was randomly divided into five subsets (A, B, C, D and E). Three subsets were used as a training set for the model generation. However, one subset was used as a test set to take care of the overfitting while the ANFIS model is developing. One subset was used as a prediction set to evaluate the generated model. The molecules included in each subset (A, B, C, D and E) are shown in Table 1. Ten different combinations of calibration and validation subsets were used in the present study to develop the modified ACO-ANFIS model. Using cross-validation techniques ensure that the developed model is consistent and reliable and it is not obtained by chance.

3. Experimental

3.1. Data set

The data set consists of 81 molecules of novel quinazolinone derivatives together with their inhibitory activities (pIC_{50}) and is taken from the article recently published by Hyunah Choo and coworkers [7]. The activity parameter IC_{50} is a measure of antiviral potency and refers to the molar concentration of each compound required to reduce the concentration of serotonin (5-HT_7) viral by 50% respect to the level measured in an infected culture. The main skeleton with different functional positions for quinazolinone derivatives is shown in Fig. 3. A list of inhibitory activities is given in Table 1. As mentioned in Section 2.3, the data set was divided into five subsets: 3/5 of subsets as a training set, 1/5 of subsets as a test set and 1/5 of subsets as a prediction set.

Prior to the calculation of the molecular descriptors, the 3D structures of the studied compounds were optimized using semi-empirical quantum-chemical methods of AM1 implemented in Hyperchem computer program [25].

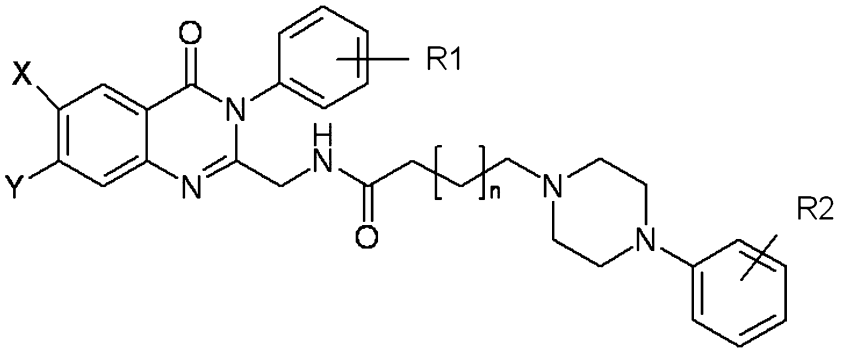

Fig. 3. Main skeleton with different functional positions for quinazolinone derivatives.

3.2. Molecular descriptors

A main step in every QSAR study is choosing and calculating the structural descriptors as numerical encoded parameters representing the chemical structures. In this work, over 100 meaningful descriptors were calculated for each compound, which encoded different aspects of the molecular structures. These descriptors consist of constitutional, topological, electronic, geometric and empirical descriptors. Pairs of descriptors that are highly correlated ($R > 0.90$) encoded similar information, and therefore one of them has been eliminated. Descriptors with constant or almost constant values for all molecules were also eliminated. As a result, only a total of 52 descriptors were remained and were used for further studies, which are listed in Table 2. All these molecular descriptors were generated using Dragon3 software [26]. Table 3 shows 10 different combinations of calibration and validation subsets used for developing the modified ACO-ANFIS model. Table 4 shows the selected descriptors in each combination and the frequency of each descriptor in the developed models. Modified ACO-ANFIS algorithm was written in our laboratory using MATLAB 7.0 [27] and run on a personal computer (Intel Pentium processor 4/1.8 GHz 512 MB RAM).

4. Results and discussion

4.1. Modified ACO-ANFIS strategy for developing QSAR model

For the purpose of the selection of the most important variables, modified ACO was applied using 52 descriptors as independent variables and observed $p\mathrm{IC}_{50}$ as a dependent variable. A modified ACO contained a population of 100 and evolved for 300 cycles. All parts of this process are illustrated in Fig 1. The most important selected variables using the modified ACO algorithm were used as inputs for developing the ANFIS model to predict the value of $p\mathrm{IC}_{50}$.

Table 2
Definition of the descriptors used in this study

Functional families	Descriptors
Constitutional	MW (molecular weight), Se (sum of atomic Sanderson electronegativities), Ms (mean of topological state), Sp (sum of atomic polarizabilities), Sv (mean atomic van der Waals volume), nF and nO (number of nitrogen and oxygen atoms), Hbond acceptor
Topological	ZM1 (Zagreb M1 index), ZM2 (Zagreb M2 index), SIC (structural information content), X1A (average connectivity index chi-2), CIC (complementary information content), S1K (1-path kier shape index), S2K (2-path kier shape index), BIC (bond information content), UNIP (unipolarity), PCR (ratio of multiple path count to path count), PCD (difference of multiple path count to path count), PHI (Kier flexibility index), BLI (Kier Benzene-likeliness index), WA (average Wiener index), SMTI (Schultz Molecular Topological Index), GSI (Gordon–Scantlebury index), *J* (Balaban *J* index)
Electronic	Q_{max} (maximum charge), Q_{min} (minimum charge), Q_{mean} (average charge), SPP (subpolarity parameter), TE2 (topological electronic), JGT (global topological charge index), LDIP (local dipole index), HOMO (highest occupied molecular orbital, LUMO (lowest unoccupied molecular orbital)
Geometric	W3D (3D-Wiener index), AGDD (average geometric distance degree), DDI (distance–distance index), MAXDN (maximal electrotopological negative variation), MAXDP (maximal electrotopological positive variation), DELS (molecular electrotopological variation), SPAN (span R), ESTP (E-state topological parameter), PJI3 (3D Petijean shape index), SPH (spherocity), ASP (asphericity), FDI (folding degree index), SPH2 (average shape profile index of order 2)
Empirical	Ui (unsaturation index), Hy (hydrophilic factor), ARR (aromatic ration), AlogP (log of the partition coefficient), MolRef (molar refractivity)

Table 3
Selecting the important variables using shuffling cross-validation and modified ACO-ANFIS method

Run	Training set	R^2_{Tr}	$RMSE_{Tr}$	Test set	R^2_{Test}	$RMSE_{Test}$	Validation set	R^2_{Val}	$RMSE_{Val}$
1	A+B+C	0.864	0.220	D	0.841	0.298	E	0.791	0.345
2	A+B+D	0.870	0.225	E	0.825	0.332	C	0.804	0.320
3	A+B+E	0.819	0.268	C	0.753	0.351	D	0.748	0.384
4	A+C+D	0.840	0.222	E	0.793	0.339	B	0.780	0.362
5	A+C+E	0.817	0.241	B	0.780	0.341	D	0.745	0.377
6[a]	A+D+E	0.875	0.214	C	0.847	0.295	B	0.807	0.318
7	B+C+D	0.790	0.280	A	0.767	0.330	E	0.746	0.386
8	B+C+E	0.831	0.240	D	0.804	0.322	A	0.782	0.357
9	B+D+E	0.856	0.248	A	0.821	0.320	C	0.777	0.368
10	C+D+E	0.839	0.254	B	0.810	0.325	A	0.769	0.379
Average		**0.840**	**0.241**		**0.804**	**0.325**		**0.775**	**0.360**

[a] The calculated $p\mathrm{IC}_{50}$ values are listed in Table 1 for this run.

ANFIS modeling involves two steps: (a) structure identification and (b) parameter identification. The former is related to finding a suitable number of rules and a proper partition of the feature space. The latter is concerned with the adjustment of system parameters, such as MF (membership function) parameters, linear coefficients, and so on. It is concluded that by increasing the number of MFs per input, the number of rules increases accordingly. For the first stage of ANFIS modeling grid partitioning was used for partitioning the features. The number and type of membership functions was optimized using RMSE as a criterion for the test set. It should be noted that the test set, consisted of 20% of the data set and was used to take care of the overfitting.

The effectiveness of the QSAR model for selecting the variables and to predict the value of $p\mathrm{IC}_{50}$ value was also estimated using the shuffling cross-validation technique. For this purpose the whole modified ACO-ANFIS process was carried out for 10 different training, test and validation sets. Statistical parameters obtained using these models for different sets are shown in Table 3. The selected descriptors in each run and the frequency of each descriptor in modified ACO-ANFIS models are shown in Table 4. Inspection of this table shows that Q_{max}, Se, Hy, PJI3 and DELS descriptors appeared more frequently in 10 runs compare to the other descriptors.

The appearance of maximum charge (Q_{max}) and sum of atomic Sanderson electro negativities (Se) among other descriptors in the model shows the importance of the inhibitors charge on the inhibition mechanism. The presence of some heteroatoms such as nitrogen and oxygen influences the value of Q_{max} and Se. Another important factor in interaction between inhibitor and isosyme is the hydrophilic factor (Hy) of the inhibitor. This factor shows that how much a molecule is charged-polarized and it is capable of hydrogen bonding. 3D Petijean shape index (PJI3) and molecular electrotopological variation (DELS) descriptors show the size and

Table 4
Selected descriptors in each run and the frequency of each descriptor

run	Selected descriptors
1	WA, DELS, Q_{max}, LDIP, Se, Hy
2	PJI3, DELS, LUMO, Se, SPH2
3	PJI3, SPAN, Q_{max}, Se, Hy
4	ESTP, DELS, Q_{max}, MAXDP, Hy, PCD
5	PJI3, LDIP, Q_{max}, Se, SPH2
6	PJI3, DELS, Q_{max}, Se, Hy
7	JGT, WA, DELS, SPP, Se, Hy
8	PJI3, DELS, Q_{max}, Se, Hy
9	PJI3, DELS, Q_{max}, Se, GSI
10	PJI3, ARR, DELS, Q_{max}, Se, Hy

The frequency of Se, Q_{max}, Hy, PJI3 and DELS descriptors appeared in GA-ANN models are 9, 8, 7, 7 and 8, respectively.

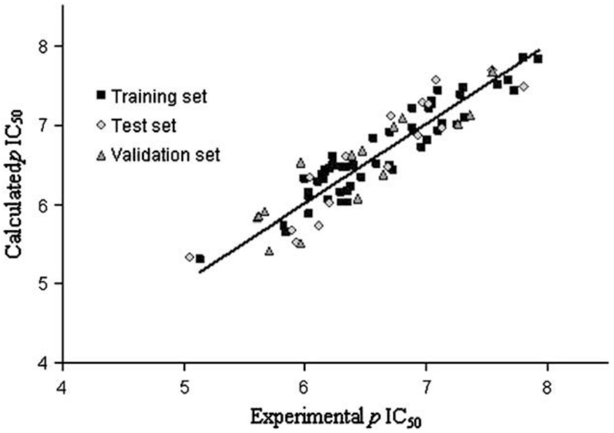

Fig. 4. Plot of the modified ACO-ANFIS calculated pIC$_{50}$ values against the experimental ones for the training, test and validation sets, in run 6.

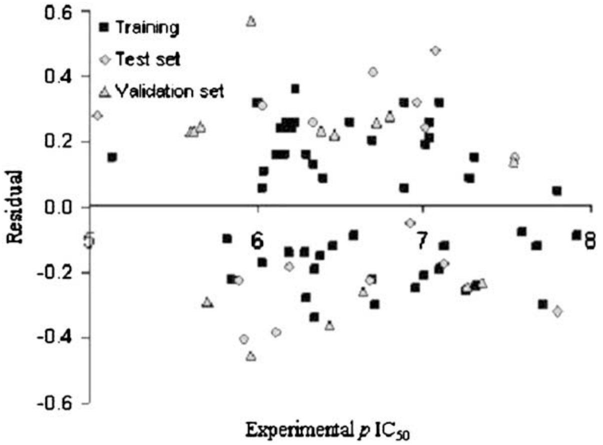

Fig. 5. Plot of residuals versus experimental values of pIC$_{50}$ for the modified ACO-ANFIS model in run 6.

the shape of inhibitors and it confirms that the size of inhibitor plays a major role in the inhibition mechanism. The detailed description of these descriptors was given in Ref. [28]. The statistical parameters of R_p^2 and RMSE$_p$ are 0.775 and 0.360, respectively, using this method. This suggests that the method of variable selection is stable.

The predicted values of quinazolinone derivatives inhibition activity for run 6 are listed in Table 1 (all the five most frequent parameters are used in run 6 and we just presented the results of this run for brevity). The correlation between the experimental and calculated values of pIC$_{50}$ is shown in Fig. 4. The residuals of the calculated values of pIC$_{50}$ are plotted against the experimental ones in Fig. 5. The propagation of the residuals in both sides of zero line indicates that no symmetric error exists in the development of the QSAR model.

4.2. Reliability and robustness of modified ACO-ANFIS model

Next step of this work was investigating the validity of the generated model. The cross-validation techniques of leave-one-out (LOO-CV), leave-multiple-out (LMO-CV) and also Y-randomization were used to prove the consistency of the model. In LOO-CV algorithm, one compound was left in each step as prediction set and the model was developed using the remaining molecules as training set [29]. The accuracy of cross-validation results is extensively acceptable in the literature considering Q_{LOO}^2 value using Eq. (15):

$$Q_{\mathrm{LOO}}^2 = \frac{\mathrm{PRESS}}{\mathrm{SSY}} = 1 - \frac{\sum_{i=1}^{n}(y_{\exp} - y_{\mathrm{pred}})^2}{\sum_{i=1}^{n}(y_{\exp} - \bar{y})^2} \tag{15}$$

In this sense, a high value for the statistical parameter ($Q^2 > 0.5$) is considered as proof of high predictive ability of the model [30]. However, several authors suggest that a high value of Q_{LOO}^2 appears to be necessary but not sufficient [31]. For this reason, we also used LMO-CV and Y-randomization techniques. In the case of LMO, M represents a group of randomly selected data points which would leave out at the beginning and would be predicted by the model which was developed using the remaining data points. So, the M molecules are considered as prediction set. The R_{LMO}^2 can be calculated using Eq. (16):

$$R_{\mathrm{LMO}}^2 = \frac{\mathrm{PRESS}}{\mathrm{SSY}} = 1 - \frac{\sum_{i=1}^{\mathrm{test}}\left(y_{\exp} - y_{\mathrm{pred}}\right)^2}{\sum_{i=1}^{\mathrm{test}}\left(y_{\exp} - \bar{y}_{\mathrm{train}}\right)^2} \tag{16}$$

In this study, we have performed leave-12-out (L12O) and leave-16-out (L16O) cross-validations. A group of 12 and 16 compounds, respectively, was randomly selected from the training set. Then each group was left out and was predicted by the model developed from the remaining observations. This procedure was carried out for 1000 times. Table 5 shows the results for LOO and LMO cross-validations. High values for Q_{LOO}^2 and R_{LMO}^2 indicate the reliability and consistency of the developed model.

In order to assess the robustness of the modified ACO-ANFIS, the Y-randomization test was applied in this contribution. The dependent variable vector pIC$_{50}$ was randomly shuffled and a new QSAR model was developed using the original variable matrix. The new QSAR model is expected to show a low value for R_p^2 and Q_{LOO}^2. Several random shuffles of the y vector were performed for which the results are shown in Table 6. Also for strengthening the robustness of modified ACO-ANFIS model, the p value was calculated in Y-randomization test using Eq. (17):

$$p = \frac{\mathrm{sum}(I(Q_{\mathrm{LOO}}^2 > Q_{\mathrm{LOO}}^2(n)))}{N} \tag{17}$$

Table 5
Statistics using LOO-CV and LMO-CV methods for comparing the results of modified ACO-ANFIS method with the two other methods

Method	LOO		L12O[a]		L16O[a]	
	Q^2	RMSE$_p$	R^2	RMSE$_p$	R^2	RMSE$_p$
Modified ACO-ANFIS	0.751(0.03)[d]	0.394(0.02)	0.734(0.02)	0.436(0.02)	0.745(0.03)	0.420(0.02)
Stepwise MLR-ANFIS[b]	0.520(0.04)	0.881(0.04)	0.487(0.03)	0.935(0.03)	0.470(0.07)	0.954(0.05)
GA-PLS-ANFIS[c]	0.689(0.06)	0.651(0.03)	0.663(0.05)	0.703(0.06)	0.648(0.05)	0.708(0.04)

[a] Calculation of R_{LMO}^2 was based on 1000 random selections of groups of 12 and 16 samples.
[b] Selected variables: Se, Hy, SPH2, PJI3, SMTI and ARR.
[c] Selected variables: $Q_{\max}$, TE2, Se, ESTP, PJI3, BLI and DELS.
[d] Corresponding standard deviation is given in parenthesis.

Table 6
R_p^2 and Q_{LOO}^2 values after several *Y*-randomization tests

Iteration	R_p^2	Q_{LOO}^2
1	0.294	0.236
2	0.240	0.163
3	0.185	0.078
4	0.419	0.341
5	0.089	0.040
6	0.128	0.101
7	0.193	0.143
8	0.178	0.142
9	0.356	0.239
10	0.230	0.165

The *p* value is equal to 1 for 100 iterations using Eq. (17).

where $I()$ is an indicator function that takes value 1 if the inequality in the parenthesis is true; and takes value 0 otherwise. N is the total number of *Y*-randomization test. $Q_{\mathrm{LOO}}^2(n)$ is value of Q_{LOO}^2 in iteration n (where $n = 1,\ldots,N$). The p value is the percentile of Q_{LOO}^2 in the sequence of N $Q_{\mathrm{LOO}}^2(n)$. In our study the p value for 100 iterations is equal to 1.

The poor values for R_p^2 and Q_{LOO}^2 indicate that the good results for the modified ACO-ANFIS model are not due to a chance correlation or structural dependency of the training set.

4.3. *Comparison of modified ACO-ANFIS with stepwise MLR-ANFIS and GA-PLS-ANFIS*

For more investigation, stepwise MLR and GA-PLS techniques are also used to select the most important descriptors in the present work. The theories of these algorithms are discussed elsewhere [32,33]. To find the best model, GA-PLS were run many times with different settings of initial populations. The best models of stepwise MLR and GA-PLS with best fitness were selected. The selected descriptors appeared in these models were used in developing ANFIS model to predict the value of $p\mathrm{IC}_{50}$. The results of Q_{LOO}^2, R_{LMO}^2 and RMSE_p for LOO, L12O and L16O are summarized in Table 5. This table also shows that the best model has six and seven variables for the stepwise MLR and GA-PLS techniques, respectively. It is clear from this table that the results of LOO, L12O and L16O for the modified ACO-ANFIS model are superior compared with those of the stepwise MLR-ANFIS and GA-PLS-ANFIS. However, the results of modified ACO-ANFIS and GA-PLS-ANFIS are somehow comparable, but the modified ACO-ANFIS model uses fewer descriptors for predicting the inhibition activities.

5. Conclusions

The aim of the present work was developing a QSAR model to predict the inhibitory activities of novel quinazolinone derivatives. Using Dragon software one can easily generate a large number of descriptors. However, a very important step in every QSAR studies is selecting suitable descriptors using a variable selection method. In this work, the modified ACO algorithm and shuffling cross-validation technique have been employed for variable selection and satisfactory results have been obtained. It is shown in this work that the modified ACO as variable selection method combined with ANFIS as mapping tool can successfully solve this problem. Selection of five variables of $Q_{\max}$, Se, Hy, PJI3 and DELS by modified ACO algorithm with electronic, constitutional, geometric and empirical characteristics indicates the complexity of inhibition mechanism. The modified ACO-ANFIS has been testified to be an effective method for variable selection and developing model using the cross-validation techniques of leave-one-out, leave-multiple-out and also *Y*-randomization. Comparing the results of stepwise MLR-ANFIS and GA-PLS-ANFIS with those for modified ACO-ANFIS reveals that the latter model selects the best variables to predict the inhibition action of quinazolinone derivatives.

References

[1] H.J. Altman, H.J. Normile, Pharmacol. Biochem. Behav. 28 (1987) 353–359.
[2] M.C. Buhot, S. Martin, L. Segu, Ann. Med. 32 (2000) 210–221.
[3] Y. Shen, F.J. Monsma, M.A. Metcalf, P.A. Jose, M.W. Hamblin, D.R.J. Sibley, Biol. Chem. 268 (1993) 182.
[4] J.L. Plassat, N. Amlaiky, R. Hen, Mol. Pharmacol. 44 (1993) 229.
[5] N.M. Barnes, T. Sharp, Neuropharmacology 38 (1999) 1083–1152.
[6] B. Pouzet, M. Didriksen, J. Arnt, Pharmacol. Biochem. Behav. 71 (2002) 655–665.
[7] A.N. H.Choo, Y.H. Pae, S.H. NaHong, Bio. Org. Med. Chem. 16 (2008) 2570–2578.
[8] T. Ghafourian, M.T.D. Cronin, SAR. QSAR. Environ. Res. 16 (2005) 171–190.
[9] T.L. Brian, J. Chem. Inf. Comput. Sci. 34 (1994) 1279–1288.
[10] A. Yasri, D. Hartsough, J. Chem. Inf. Comput. Sci. 41 (2001) 1218–1227.
[11] S.J. Cho, M.A. Hermsmeier, J. Chem. Inf. Comput. Sci. 42 (2002) 927–936.
[12] Q. Shen, R. Yu, J. Chem. Inf. Model. 45 (2005) 1024–1029.
[13] E. Bonabeau, M. Dorigo, G. Theraulaz, Nature 406 (2000) 39–42.
[14] M. Dorigo, L.M. Gambardella, BioSystems 43 (1997) 73–81.
[15] Y. Wang, J.Y. Xie. The 2000 IEEE Asia-Pacific Conference. (2000), 54–57.
[16] J.F. Gomez, H.M. Khodr, P.M. DeOliveira, et al., IEEE Trans. Power Syst. 19 (2004) 996–1004.
[17] L.A. Zadeh, Fuzzy sets, Inf. Control 8 (1965) 338–353.
[18] Y.L. Loukas, J. Med. Chem. 44 (2001) 2772–2783.
[19] T.E. Exner, J. Brickmann, J. Mol. Model 3 (1997) 321–329.
[20] M.J. Krieger, J.B. Billeter, L. Keller, Nature 406 (2000) 992–995.
[21] W. Shi, Q. Shen, W. Kong, B. Ye, Eur. J. Med. Chem. 42 (2007) 81–86.
[22] S.B. Gunturi, R. Narayanan, A. Khandelwal, Bioorg. Med. Chem. 14 (2006) 4118–4129.
[23] E. Buyukbingol, A. Sisman, M. Akyildiz, F.N. Alparslan, A. Adejare, Bioorg. Med. Chem. 15 (2007) 4265–4282.
[24] M. Sugeno, G.T. Kang, Fuzzy Sets Syst. 28 (1988) 15–24.
[25] Hyperchem, Molecular Modeling System, Hyper Cube, Inc. and Auto Desk, Inc, 1993, Developed by Hyper Cube, Inc.
[26] R. Todeschini, V. Consonni, A. Mauri, M. Pavan, Software Dragon: Calculation of Molecular Descriptors, Department of Environmental Sciences, University of Milano-Bicocca, and Talete, srl, Milan, Italy, 2003.http://disat.unimib.it/chm/Dragon.htm.
[27] MATLAB 7.0 Available from: http://www.mathworks.com/products/matlab/.
[28] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, Wiley/VCH, Weinheim, 2000.
[29] D.W. Osten, J. Chemom. 2 (1998) 39–48.
[30] S. Wold, Quant. Struc.Act. Relat. 10 (1991) 191–193.
[31] A. Golbraikh, A. Tropsha, J. Mol. Graph. Model. 20 (2002) 269–276.
[32] M. Jalali-Heravi, M. Asadollahi-Baboli, P. Shahbazikhah, Eur. J. Med. Chem. 43 (2008) 548–556.
[33] M. Jalali-Heravi, A. Kyani, Eur. J. Med. Chem. 42 (2007) 649–659.