ICC
Iranian Chemical Communication
Original Research Article
Payame Noor University http://icc.journals.pnu.ac.ir

A Comparative QSAR study of aryl-substituted isobenzofuran-1(3H)-ones inhibitors

Zahra Rostami*, Eslam Pourbasheer

Department of Chemistry, Payame Noor University (PNU), P.O. BOX 19395-3697 Tehran, Iran

Received: 25 May 2017, Accepted: 11 December 2017, Published: 11 December 2017

Abstract

A comparative workflow, including linear and non-linear QSAR models, was carried out to evaluate the predictive accuracy of models and predict the inhibition activity of a series of aryl-substituted isobenzofuran-1(3H)-ones. The data set consisted of 34 compounds was classified into the training and test sets, randomly. Molecular descriptors were selected using the genetic algorithm (GA) as a feature selection tool. Various linear models based on multiple linear regression (MLR), principle component regression (PCR) and partial least square (PLS) and non-linear models based on artificial neural network (ANN), adaptive network-based fuzzy inference system (ANFIS) and support vector machine (SVM) methods were developed and compared. The accuracy of the models was studied by leave-one-out cross-validation ($\boldsymbol{Q^2_{Loo}}$), Y-randomization test and group of compounds as external test set. Six descriptors were selected by GA to develop predictive models. With respect to the linear models, GA-PCR method was more accurate than the reset with statistical results of $R^2_{train} = 0.883$, $R^2_{test} = 0.897$, $R^2_{adj,train} = 0.829$, $R^2_{adj,test} = 0.849$, $F_{train} = 24.07$ and $F_{test} = 34.17$. In case of non-linear models, GA-SVM ($R^2_{train} = 0.992$ and $R^2_{test} = 0.997$) showed high predictive accuracy for the inhibitory activity. It was found that the selected descriptors have the major roles in interpretation of biological activities of the compounds.

Keywords: QSAR; genetic algorithms; global optimization; SVM.

Introduction

Designing new drugs requires screening their estrogenic and biological activities; however, performing these experiments needs biological materials of human and rat trials which are costly, time-consuming and may provide some toxic products. Therefore, it is of prior interest to employ a model for predicting the biological activities of newly designed compounds before synthesis [1]. There has been a growing interest over computational methods to predict the biological activities of compounds, since designing new compounds with higher inhibitory activities cannot be done unless we get aware of their biological features. In this regard, there is a well-known method which could provide useful information based on biological activities and chemical structures of designed molecules [2]. Quantitative structure–activity relationship (QSAR) [3,4] is a widely used method for predicting the biological activities of

*Corresponding author: Zahra Rostami
Tel: +98 (81) 32546721, Fax: +98 (81) 32546722
E-mail: rostami@pnu.ac.ir

compounds using experimental data and chemical structures [5]. QSAR relates a set of physico-chemical properties or molecular descriptors to any activity such as inhibition activity or binding affinity of the chemicals [6]. These responses can be derived rapidly and cost-effectively by QSAR without necessity of performing extra expensive and time consuming laboratory experiments [6]. Recently, some novel inhibitors targeting lymphocyte pore-forming protein perforin which have been synthesized with clear structure activity relationships (SAR) have also been investigated by Spicer and coworkers [7]. Due to the high potency of these inhibitors based on aryl-substituted isobenzofuran-1(3H)-ones, QSAR studies can reveal the chemical features required for subsequent inhibitors to increase the inhibitory activity using this moiety as a core.

Although, the primary goal in QSAR approach is to identify the mechanism of act toward a response by interpreting the relevant molecular features, comparative study of different models can result in more comprehensive understanding of how molecular features correlate to the response. This includes application of various chemical features selection and regressions techniques. Many feature selection techniques such as stepwise regression, simulated annealing and genetic algorithms are available. It has been shown that genetic algorithms (GAs) can be successfully used as a feature selection technique [8]. Leardi [9] demonstrated that GA, after suitable modifications, produces more interpretable results, since the selected variables are less dispersed as compared to other methods. Among the investigation of QSAR, one of the most important factors affecting the quality of the model is the method to build the model. Many multivariate data analysis methods such as Multiple Linear Regression (MLR) [6,10], Partial Least Square (PLS) [6], Principal Component Regression (PCR) [6], Artificial Neural Network (ANN) [11], Adaptive Neuro-Fuzzy Inference Systems (ANFIS) [12], and Support Vector Machine (SVM) [13] have been used in QSAR studies [8].

This work contributes to identification of chemical features that their presence could increase the inhibition activity of inhibitors targeting perforin. The comprehensive study including the application of various regressions tools over the predictive results was also carried out. The models proposed here can be applied for subsequent drugs with aryl-substituted isobenzofuran-1(3H)-ones moiety to predict the inhibition activities with high accuracy. The novelty of our work is to apply the genetic algorithm and global optimization techniques such as Particle Swarm Optimization (PSO) algorithm for neural network training (namely, ANN-GA, ANN-PSO and ANFIS-PSO methods) in order to raise quality and reduce errors in the constructed models.

Methodology

Data set

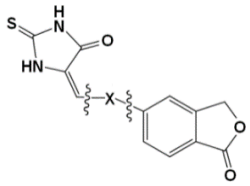
A dataset consists of 34 molecules based on aryl-substituted isobenzofuran-1(3H)-ones derivatives as inhibitors for pore-forming protein perforin was adopted from previously published paper [7]. The chemical structures and the inhibitory activities (pIC_{50}) of these derivatives are listed in Table 1. To reduce the skewness of data set, the IC_{50} values were converted into a logarithmic scale$\left(pIC_{50} = -\log(IC_{50})\right)$. Subsequently pIC_{50}values were used as the dependent variable in the modeling process [6]. Selection of

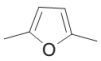
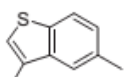
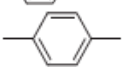
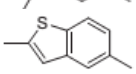
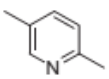
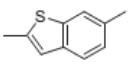
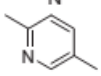
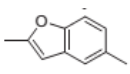
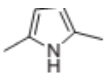
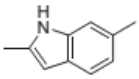
the molecules in the test set (here 20% of whole dataset (7 compounds)) was randomly done in a way to cover fair distribution for the independent variable.

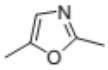
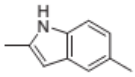
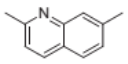
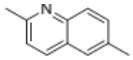
Structure optimization and descriptors generation

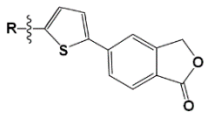
All structures of molecules were drawn in Gauss View 03 software. The best geometries of chemical structures were obtained after performing energy minimization by semi-empirical method (AM1) with the adjusted root mean square gradient of 0.01 kcal mol^{-1} in Gaussian 03 software [1]. Physico-chemical descriptors for each geometrically optimized compound were calculated by DRAGON 5.5 computer software [14]. DRAGON includes various types of molecular descriptors such as 3D-MoRSE, 2nd component shape directional WHIM index, 2D petitjean shape index, radial distribution function etc. The list of molecular descriptors generated for whole compounds was pre-treated by the constant and near constant ones. The remaining molecular descriptors were checked for the existence of collinearity. The threshold was set to 0.9 and the features violating this threshold were removed while presenting lower correlation to the dependent variable comparing to their pairs. Finally, 332 descriptors were remained and preceded to subsequent analyses.

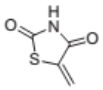
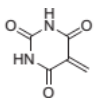
Table 1. Chemical structures and the corresponding experimental and predicted pIC_{50} values by GA-PCR and GA-SVM methods of aryl-substituted isobenzofuran-1(3H)-ones [7]



NO.	X	Exp.	GA-PCR	GA-SVM	NO.	X	Exp.	GA-PCR	GA-SVM
1		5.208	5.354	5.215	7		5.886	5.804	5.914
2		6.108	6.039	6.099	8		5.595	5.626	5.646
3		5.979	5.922	6.002	9		5.842	5.786	5.801
4		5.676	5.577	5.675	10		5.644	5.528	5.647
5		5.745	5.731	5.725	11		6.310	6.212	6.302
6		5.699	5.72	5.75	12		6.444	6.350	6.395
13		6.432	6.314	6.384	19		6.292	6.406	6.296
14		5.611	5.546	5.646	20		6.444	6.355	6.480

15		5.138	4.889	5.232	21		6.086	6.445	6.115
16		5.764	5.717	5.731	22		5.764	6.133	5.737
17		5.851	5.902	5.882	23		6.328	6.683	6.337
18		6.469	6.473	6.503					



NO.	R	Exp.	GA-PCR	GA-SVM	NO	R	Exp.	GA-PCR	GA-SVM
24		5.924	5.737	6.003	30		4.818	4.546	4.798
25		5.951	5.947	5.949	31		6.097	5.852	6.089
26		5.561	5.885	5.556	32		6.398	6.593	6.387
27		5.498	5.512	5.485	33		5.402	5.199	5.451
28		5.157	5.394	5.165	34	CHO	5.200	5.255	5.301
29		5.167	5.138	5.265					

Variable selection

Feature or variable selection is one of the main steps in every QSAR study [6]. Generally, here the problem is to find a suitable group of molecular descriptors from the pool of descriptors presenting the minimum error against the experimental data [1]. In the current study, Genetic Algorithm (GA) as a feature selection tool was used to select the most relevant descriptors with respect to a fitness function [1]. Genetic algorithms are simulated methods based on ideas from Darwin's theory of natural selection and evolution. In GA, a chromosome (or an individual) which can be defined as an enciphered entity of a candidate solution is expressed as a set of variables. GA consists of the following basic steps: (1) A chromosome is represented by a binary bit string and an initial population of chromosomes is created in a random way; (2) A value for the fitness function of each chromosome is evaluated (Here, the fitness function of used genetic algorithm was $\boldsymbol{Q^2_{LOO}}$); (3) Based on the values of the fitness functions, the chromosomes of the next generation are produced by selection, crossover and mutation operations [15].

In principle, any feature selection tool can be coupled with any statistical method of choice for constructing quantitative model [22]. In this study, GA was employed with MLR, PCR,

and PLS as the linear methods and, then, compared to the non-linear methods such as ANN, ANFIS and SVM methods [1]. Feature selection (GA) and nonlinear methods were done in MATLAB 6.5 program [16] and linear methods were done by SPSS software [17].

As already noted in the introduction, the novelty of this work is to apply the genetic algorithm and particle swarm optimization (PSO) algorithm for neural network training in order to improve the model performance in non-linear QSAR models. Particle swarm optimization (PSO) is a population based soft computing technique developed by Eberhart and Kennedy in 1995. PSO technique shares numerous similarities with evolutionary computation techniques such as GA. PSO starts initialization with a population of random solutions and searches for optimum solution by updating generations. However, if we compare PSO with GA then unlike GA, PSO has no operators such as crossover and mutation. In PSO, the possible solutions, called particles, move through the problem space by following the current optimum particles [18]. PSO is a kind of population based metaheuristics which consists of individual of solutions [19]. The solutions in PSO are represented as particle. These particles are randomly initialized with D-dimensional size that consist a vector of position *p*, velocity *v*, personal best fitness (*pbest*) and global best fitness (*gbest*). Along all iterations, the particles are flying through search space and always accelerated towards better solutions. This process can be achieved by updating the velocity *v* of each particle *i* with the following equation:

$$\boldsymbol{v_{i(t+1)}} = \boldsymbol{v_{i(t)}} + \boldsymbol{c_1}\,\boldsymbol{rand_1}\left(\boldsymbol{pbest} - \boldsymbol{x_{i(t)}}\right) + \boldsymbol{c_2}\,\boldsymbol{rand_2}\left(\boldsymbol{gbest} - \boldsymbol{x_{i(t)}}\right) \quad (1)$$

where $\boldsymbol{v_{i(t+1)}}$ is velocity of the *i*th particle at iteration *t*, $\boldsymbol{c_1}\,\boldsymbol{rand_1}\left(\boldsymbol{pbest} - \boldsymbol{x_{i(t)}}\right)$ is cognitive learning for each particle, $\boldsymbol{c_2}\,\boldsymbol{rand_2}\left(\boldsymbol{gbest} - \boldsymbol{x_{i(t)}}\right)$ is social information of the whole particles, $\boldsymbol{c_1}$ and $\boldsymbol{c_2}$ are two positive constants for representing personal and social learning rate respectively and $\boldsymbol{rand_1}$ and $\boldsymbol{rand_2}$ are two separately generated random numbers in the uniform range of [0:1]. In the current study, the PSO algorithm was applied for neural network training in the GA-ANN-PSO constructed model.

In the other section of this work, we were interested to use the novel approach of GA-ANFIS in order to construct the other non-linear model for predicting the inhibition activity of the studied molecules. ANFIS was presented by Jang in 1993 in which he introduced a novel architecture and learning product for fuzzy inference system (FIS) which employs a neural network learning algorithm for building a set of fuzzy if-then rules with proper membership functions from the stipulated input-output pairs. This method of growing a FIS using the structure of adaptive neural network is called an ANFIS [12]. Fundamentally, ANFIS combines the neural network and a fuzzy inference system. The neural networks are used to calculate the parameter of the fuzzy inference system. Figure 1 represents the equivalent ANFIS architecture with 2 typical inputs and 4 rules. According to this Figure, ANFIS structure consists of 5 layers of nodes. Layer 1, every node *i* in this layer is an adaptive node

with a node function. Layer 2, every node in this layer is a fixed node labeled $\boldsymbol{w_i}$, whose output is the product of all the incoming signals. In general, any other t-norm operator that performs fuzzy AND can be used as the node function in this layer. Layer 3, every node in this layer is a fixed node labeled N. The *i*th node calculates 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. Layer 4, every node *i* in this layer is an adaptive node with a node function. Parameters in this layer are referred to as consequent parameters. Layer 5, the single node in this layer is a fixed node labeled $\overline{\boldsymbol{w_i}}$, which computes the overall output as the summation of all incoming signals [20]:

$$\textbf{overall output} = \boldsymbol{O_{5,i}} = \sum_i \overline{\boldsymbol{w_i}}\boldsymbol{z_i} = \frac{\sum_i \boldsymbol{w_i z_i}}{\sum_i \boldsymbol{w_i}} \quad \textbf{(2)}$$

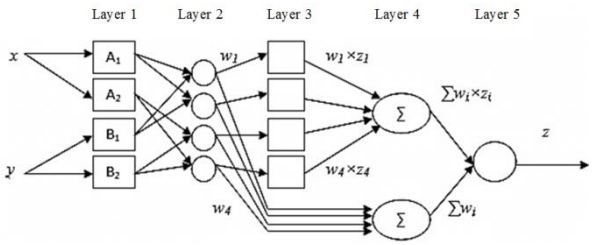


Figure 1. ANFIS structure with 2 inputs and 4 rules

The other novelty of the current work is to apply the PSO algorithm for neural network training in ANFIS procedure in order to raise quality and reduce errors. The corresponding details of these methods (GA-ANFIS and GA-ANFIS-PSO) have been reported in the results and discussion section.

In the final section of this study, GA-SVM model was used to construct the latest nonlinear model based on the same selected descriptors and then the performance of this model was compared to ones obtained by previous models. Support vector machines (SVM) represent an extension to nonlinear models of the generalized portrait algorithm developed by Vapnik and Lerner. The SVM algorithm is based on the statistical learning theory and the Vapnik–Chervonenkis (VC) dimension [21]. Similar to other non-linear models, SVM regression relies on combination of various factors such as kernel function type, capacity parameter C, $\boldsymbol{\varepsilon}$ of $\boldsymbol{\varepsilon}$-insensitive loss function and its corresponding parameters. Kernel function type determines the sample distribution in space. Consequently, the kernel function type should be declared. The radial basis function (RBF) is one of the most popular kernel functions used in this study due to the good performance. The RBF is given as below [1]:

$$\boldsymbol{k(x_i, x_j)} = \textbf{exp}\left(-\boldsymbol{\gamma}\|\boldsymbol{x_i} - \boldsymbol{x_j}\|^2\right) \quad (3)$$

where $\boldsymbol{\gamma}$>0 is one of the kernel parameters that controls the width of RBF function, and $\|\mathbf{x}\|$ is the norm of *x* and is given by $\sqrt{\mathbf{x^T x}}$. The parameter $\boldsymbol{\gamma}$ plays a similar role as the degree of the polynomial kernel in controlling the flexibility of the resulting classifier [22].The next factor in this modeling is parameter C which is to regulate and control the trade-off between maximizing the margin and minimizing the training error [1]. In order to make the learning process stable, a large value should be set up for C (e.g., C = 100). The optimal value for $\boldsymbol{\varepsilon}$ depends

on the type of noise present in the data, which is usually unknown. Even if enough knowledge of the noise is available to select an optimal value for $\boldsymbol{\varepsilon}$, there is the practical consideration of the number of resulting support vectors. $\boldsymbol{\varepsilon}$ -insensitivity avoids the entire training set meeting boundary conditions, and so allows for the possibility of sparsity in the dual formulations solution. Therefore, selecting the suitable value of $\boldsymbol{\varepsilon}$ is important in this model [23].

Validation protocols

To further check the multi-collinearity of descriptors selected by GA, variance inflation factor (*VIF*) analysis was carried out [24]. VIF value is calculated as below:

$$\mathbf{VIF} = \frac{\mathbf{1}}{\mathbf{1} - \mathbf{r^2}} \qquad (4)$$

In this formula, $\mathbf{r}$ is a correlation coefficient of multiple regressions between each variable and the other variables in the constructed QSAR model. *VIF* express different concept when it has different values in different range where if it equals 1.0, it demonstrates that there is not intercorrelation for each variable; if its value falls into the range 1.0–5.0, it exhibits the acceptance of the related model; and if the value of *VIF* becomes larger than 10.0, this denotes that the related model is unstable and recheck is necessary [1]. Also, to investigate the relative importance as well as the contribution of each descriptor in the built model, the value of the mean effect ($\boldsymbol{MF}$) has been calculated for each descriptor. This calculation was performed using the following equation [25]:

$$\boldsymbol{MF_j} = \frac{\boldsymbol{\beta_j} \sum_{\mathbf{i=1}}^{\mathbf{i=n}} \boldsymbol{d_{ij}}}{\sum_{\mathbf{j}}^{\mathbf{m}} \boldsymbol{\beta_j} \sum_{\mathbf{i}}^{\mathbf{n}} \boldsymbol{d_{ij}}} \qquad (5)$$

In this equation, $\boldsymbol{MF_j}$ shows the mean effect for the considered descriptor j, $\boldsymbol{\beta_j}$ represents the coefficient of the descriptor j, $\boldsymbol{d_{ij}}$ indicates the value of the target descriptors for each molecule and *m* is the descriptors number in the model. The $\boldsymbol{MF}$ value demonstrates the relative importance of a descriptor, in comparison with the other descriptors in the model. Its sign reveals the variation direction in the values of the activities resulting from an increase (or a reduction) of this descriptor value [25].

In order to select the most significant linear model, $\boldsymbol{R^2_{test}}$, $\boldsymbol{RMSE_{test}}$, $\boldsymbol{F_{test}}$ and $\boldsymbol{EFF_{test}}$values may be considered. $\boldsymbol{EFF}$ is the efficiency index which was used to evaluate the efficiency of QSAR models. This index was calculated using the following equation:

$$\mathbf{EFF} = \left(\frac{\sqrt{\sum_{\mathbf{i=1}}^{\mathbf{n}} \left(\mathbf{x_{c,i}} - \overline{\mathbf{x_m}}\right)^2}}{\sqrt{\sum_{\mathbf{i=1}}^{\mathbf{n}} \left(\mathbf{x_{m,i}} - \overline{\mathbf{x_m}}\right)^2}} \right)^{\mathbf{2}} \qquad (6)$$

Where $\mathbf{x_{m,i}}$ and $\mathbf{x_{c,i}}$ are the measured and predicted values, respectively and $\overline{\mathbf{x_m}}$ represents the average of measured values. The higher $\boldsymbol{R^2}$, $\boldsymbol{F}$ and $\boldsymbol{EFF}$ values with lower root mean square error ($\boldsymbol{RMSE}$) indicate the predictive ability of the built model.

Further statistical significance of the relationship between activity and the descriptors can be checked by randomization test (Y-randomization) of the models [26]. Y-randomization is a widely used approach to establish the robustness of a given QSAR model. In this approach, dependent variable vector (inhibitory activity) is randomly shuffled and a new QSAR model is built using the original independent variables. If the new QSAR models have lower $\mathbf{R_{MAX}}$ and $\mathbf{R_{CV,MAX}}$ values

for several trials, then the given QSAR model is thought to be robust.

Results and discussion

In the present study, a series of aryl-substituted isobenzofuran-1(3H)-ones inhibitors were examined using GA-MLR, GA-PCR, GA-PLS, GA-ANN, GA-ANN-GA, GA-ANN-PSO, GA-ANFIS, GA-ANFIS-PSO and GA-SVM. Some statistically significance linear and non-linear QSAR based models are derived.

Linear models

For GA-MLR model, the linear equation is as the following (Model 1):

$\mathbf{pIC_{50}}$= +4.520 (±0.626) -371 (±0.153)Mor08u +0.623 (±0.263)Mor29u +0.591 (±0.142)Mor11u – 5.508 (±1.468) P2u +1.631 (±0.685)PJI2 +0.89 (±0.22)RDF070m

$$\boldsymbol{N_{train} = 27,\quad R^2_{train} = 0.861,\quad R^2_{test} = 0.895,\quad R^2_{adj,\ train} = 0.821,\quad R^2_{adj,\ test} = 0.835,\quad F_{train} = 20.76,\quad F_{test} = 28.54}$$

In this model, $\boldsymbol{N_{train}}$ is the number of compounds in training set, $\boldsymbol{R^2}$ is the squared correlation coefficient, $\boldsymbol{R^2_{adj}}$ is adjusted $\boldsymbol{R^2}$ and $\boldsymbol{F}$ is Fisher $\boldsymbol{F}$ statistic.

Table 2. The correlation coefficient of selected descriptors and corresponding MF and VIF values based on GA-MLR model

	Mor08u	**Mor29u**	**Mor11u**	**P2U**	**PJI2**	**RDF070M**	**MF**	**VIF**
Mor08u	1	0	0	0	0	0	-0.112	1.634
Mor29u	-0.193	1	0	0	0	0	0.040	1.450
Mor11u	-0.273	0.396	1	0	0	0	0.061	1.185
P2U	-0.003	0.153	0.241	1	0	0	-0.473	1.330
PJI2	0.225	-0.176	-0.052	0.345	1	0	1.135	1.747
RDF070M	0.556	0.141	-0.167	0.205	0.263	1	0.429	1.411

The selection of genetic algorithm revealed that the descriptors Mor08u, Mor29u, Mor11u, P2u, PJI2 and RDF070m play the most important role in determining aryl-substituted isobenzofuran-1(3H)-ones inhibitors activity. A correlation matrix and corresponding *VIF* and *MF* values for chosen descriptors based on GA-MLR have been calculated and are reported in Table 2. According to this Table, *VIF* values for chosen descriptors are less than 1.8.

The statistical parameters of GA-MLR, GA-PCR and GA-PLS models (Models 1-3) have been reported in Table 3. As the Table 3 shows, in the linear models, the statistically best significant model was obtained by GA-PCR methodology (Model 2) with $\boldsymbol{R^2_{test}}$=0.897, $\boldsymbol{RMSE_{test} = 0.125}$, $\boldsymbol{F_{test} = 34.17}$ and $\boldsymbol{EFF_{test} = 0.963}$ for aryl-substituted isobenzofuran-1(3H)-ones inhibitors; however, the predictive capability for three linear models is acceptable. The predicted activity by GA-PCR method has been reported in Table 1 and the plot of predicted $\mathbf{pIC_{50}}$ values by this method versus experimental $\mathbf{pIC_{50}}$values was displayed in Figure 2.

Table 3. The statistical parameters of various linear and nonlinear models

Method	R^2_{train}	R^2_{test}	$RMSE_{train}$	$RMSE_{test}$	EFF_{train}	EFF_{test}	$RMSE\%_{train}$	$RMSE\%_{test}$
GA-MLR	0.861	0.895	0.164	0.136	0.940	0.957	47%	35%
GA-PCR	0.883	0.897	0.160	0.125	0.950	0.963	42%	38%
GA-PLS	0.846	0.857	0.256	0.249	0.948	0.943	48%	44%
GA-ANN	0.915	0.922	0.150	0.120	0.945	0.953	36%	32%
GA-ANN-GA	0.941	0.938	0.138	0.098	0.965	0.979	32%	26%
GA-ANN-PSO	0.955	0.950	0.100	0.090	0.987	0.983	24%	25%
GA-ANFIS	0.961	0.932	0.094	0.115	0.993	0.975	20%	28%
GA-ANFIS-PSO	0.981	0.990	0.086	0.080	0.983	0.998	17%	12%
GA-SVM	0.992	0.997	0.053	0.046	0.992	0.996	12%	5%

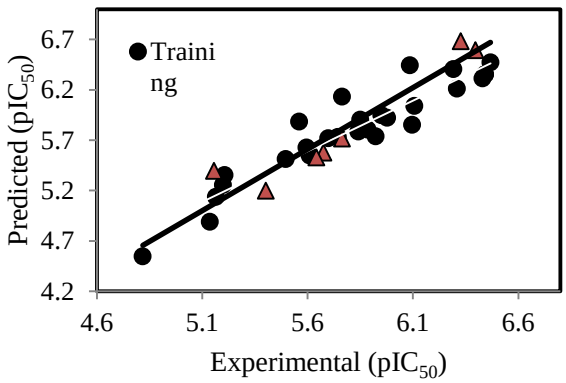


Figure 2. Predicted versus experimental $\mathbf{pIC_{50}}$ values by GA-PCR methodology

The predictive capability of GA-MLR model has been assessed by leaving one out and leaving ten out cross validation techniques ($\boldsymbol{Q^2_{LOO}}$ and $\boldsymbol{Q^2_{LTO}}$), *RMSE* for the LOO procedure ($\boldsymbol{RMSE_{LOO}}$), cross-validated squared correlation coefficient for test and training sets ($\boldsymbol{R^2_{CV,train}}$and $\boldsymbol{R^2_{CV,test}}$), *RMSE* for the cross-validation procedure ($\boldsymbol{RMSE_{CV,train}}$ and $\boldsymbol{RMSE_{CV,test}}$). These obtained statistical parameters for GA-MLR model have been shown in Table 4. This model with a coefficient of determination ($\boldsymbol{Q^2_{LOO} = 0.700}$ and $\boldsymbol{Q^2_{LTO} = 0.826}$) can well approximate the real inhibitor activities. External validation of the model ($\boldsymbol{R^2_{CV,test} = 0.895}$ and $\mathbf{RMSE_{CV,test} = 0.136}$) and *RMSE* for the LOO procedure ($\boldsymbol{RMSE_{LOO} = 0.456}$) showed that GA-MLR model can well predict activity of new aryl-substituted isobenzofuran-1(3H)-ones inhibitors. Again, the GA-MLR model didn't show high $\boldsymbol{R_{MAX}}$ and $\boldsymbol{R_{CV,MAX}}$ (0.536 and 0.267 respectively) even after ten Y-randomizations [27]. This meant that the goodness of this constructed model is not due to the chance [1].

Table 4. The obtained statistical parameters for GA-MLR model

Q^2_{LOO}	Q^2_{LTO}	$RMSE_{LOO}$	$R^2_{CV,train}$	$RMSE_{CV,train}$	$R^2_{CV,test}$	$RMSE_{CV,test}$	R_{MAX}	$R_{CV,MAX}$
0.700	0.826	0.456	0.861	0.164	0.895	0.136	0.536	0.267

Non -linear models

With respect to non-linear methods, GA-ANN, GA-ANN-GA, GA-ANN-PSO, GA-ANFIS, GA-ANFIS-PSO and GA-SVM methods are carried out to develop accurate models to predict inhibition activity.

Model 4 (GA-ANN)

In order to select the first non-linear model, the genetic algorithm (as a feature selection tool) coupled with artificial neural network method (GA-ANN) was applied in MATLAB environment with the following setup: neural network, multi-layered feed forward network; activation functions, sigmoid algorithm; the number of input neurons, 6; the number of output neurons, 1; criteria to determine the number of hidden neurons, mean square error; the number of hidden layers, single; the number of hidden neurons, [1-10]; learning rate, 0.1; the final error, 1e-20, the number of iterations, 5000; momentum rate, 0.1. The obtained values of $\boldsymbol{R^2}$,$\boldsymbol{RMSE}$ and $\boldsymbol{EFF}$ for the training and test sets by GA-ANN method are listed in Table 3. According to this Table, a good predictive accuracy is obtained. The lower $\boldsymbol{RMSE}$, the higher $\boldsymbol{F}$ and $\boldsymbol{R^2}$ values for training and test sets are obtained by GA-ANN in comparison to linear models indicating the superiority of non-linear models over linear models.

Model 5 (GA-ANN-GA)

In this section, we proposed a novel approach to improve the model performance. Despite the use of GA as features selection tool, ANN was trained with optimum internal parameters derived by GA as global optimization method. This method (GA-ANN-GA) was used in MATLAB environment and the corresponding parameters are shown in Table 1s. The obtained values of $\boldsymbol{R^2}$, $\boldsymbol{RMSE}$ and $\boldsymbol{EFF}$ for the training and test sets by GA-ANN-GA are listed in Table 3. Concerning the predictive accuracy, there is a slight improvement comparing to the ones obtained by the GA-ANN method.

Model 6 (GA-ANN-PSO)

In model 6, we used PSO algorithm for neural network training in order to raise quality and reduce errors through a combination of neural network with PSO algorithm. The resulting method (GA-ANN-PSO) was used in MATLAB environment and the corresponding parameters are given in Table 2s.The obtained values of $\boldsymbol{R^2}$,$\boldsymbol{RMSE}$ and $\boldsymbol{EFF}$ for the training and test sets by GA-ANN-PSO method are listed in Table 3. It is revealed that the results had improved comparing to the ones obtained through the GA-ANN and GA-ANN-GA methods.

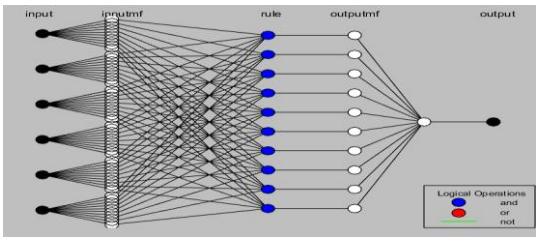
Model 7 (GA-ANFIS)

In this section, the combination of GA and ANFIS was used to construct the other non-linear model. The details of how to develop ANFIS model in this work is listed in Table 3s. Based on the model described in Figure 3, the inputs of designed model by GA-ANFIS were analyzed and some statistic parameters like $\boldsymbol{R^2}$,$\boldsymbol{RMSE}$ and $\boldsymbol{EFF}$ were measured and listed in Table 3. The numerical values in this Table indicated that GA-ANFIS procedure did not produce excellent results for the test set (R^2_{test} = 0.932). Thus, in the next step, we investigated the GA-ANFIS-PSO method and the details of this model were selected for discussion.

Model 8 (GA-ANFIS-PSO)

In model 8, we used GA (as a feature selection tool), ANFIS and PSO algorithm (for neural network training)

in order to improve the model performance. The resulting model (GA-ANFIS-PSO) was used in MATLAB environment and its parameters are presented in Table 4s. The obtained values of R^2, *RMSE* and *EFF* for the training and test sets by this model are listed in Table 3 and indicate the superiority of this model over the other ANN based models.



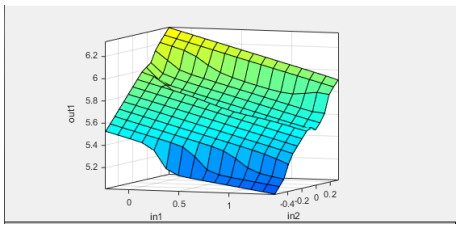


Figure 3.a Adaptive neuro-fuzzy structure and **3.b** A three-dimensional graph between the first and second inputs and output of developed model in this study

Model 9 (GA-SVM)

In this section, we applied GA-SVM model to build the final non-linear model to compare the performance of this model with ones obtained by previous models. The GA-SVM procedure was done using the MATLAB software. The internal parameters for GA-SVM model can be found in Table 5s. The corresponding values of $\boldsymbol{R^2}$, $\boldsymbol{RMSE}$ and $\boldsymbol{EFF}$ for the training and test sets by GA-SVM model are listed in Table 3 which reveals the superiority of this model (with $\boldsymbol{R^2_{test} = 0.997}$, $\boldsymbol{RMSE_{test} = 0.046}$ and $\boldsymbol{EFF_{test} = 0.996}$) over the all of linear and nonlinear models proposed. The correlation plot for experimental and predicted $\mathbf{pIC_{50}}$ based on GA-SVM model is shown in Figure 4.

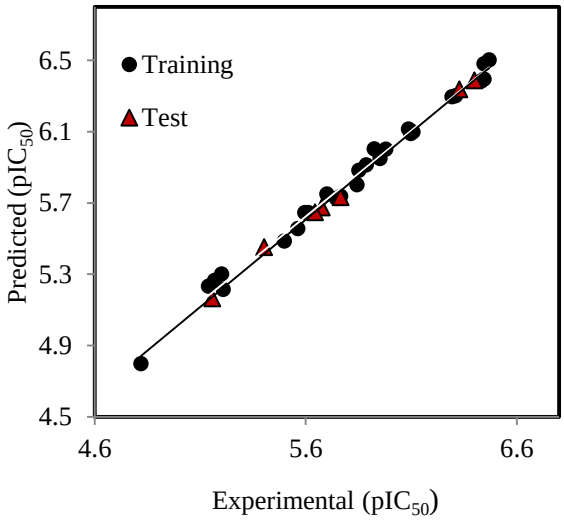


Figure 4. Predicted versus experimental $\mathbf{pIC_{50}}$ values by GA-SVM methodology

Interpretation of descriptors

Beside demonstrating statistical significance, QSAR models should also provide some useful chemical insights to understand the mechanism of inhibition to assist the designing of new drugs presenting higher inhibitory activity. By interpreting the descriptors used in the model, it is possible to gain some insights that are related to the inhibitory activity of a series of aryl-substituted isobenzofuran-1(3H)-ones.

The molecular descriptors selected by the genetic algorithm are reported in Table 2. Mor08u (3D-MoRSE - signal 08 / unweighted) is one of the 3D-MoRSE descriptors that appears in the model. 3D MoRSE descriptors (3D Molecule Representation of Structures based on Electron diffraction) are derived from infrared spectra simulation using a generalized scattering function. This descriptor was proposed as signal 08 / unweighted. As is apparent from Table 2, the mean effect (*MF*) for Mor08u has a negative sign, which indicates that inhibitory activity of aryl-substituted isobenzofuran-1(3H)-ones is reversely related to this descriptor. Therefore, increasing the value of this descriptor leads to a decrease of inhibitory activity.

Mor29u (3D-MoRSE - signal 29 / unweighted) and Mor11u (3D-MoRSE - signal 11 / unweighted) are the second and third descriptors appearing in the model. These are one of the 3D-MoRSE descriptors. The mean effect for Mor29u and Mor11u has a positive sign revealing that the inhibitory activity is directly related to these descriptors. Therefore, increasing the value of these descriptors leads to an increase in inhibitory activity.

The fourth descriptor is P2U (2nd component shape directional WHIM index / unweighted). This descriptor belongs to WHIM descriptors. WHIM descriptors (Weighted Holistic Invariant Molecular descriptors) are geometrical descriptors based on statistical indices calculated on the projections of the atoms along principal axes. WHIM descriptors are built in such a way as to capture relevant molecular 3D information like molecular size, shape, symmetry and atom distribution with respect to invariant reference frames. The mean effect for P2U displays a negative sign, which indicates that the inhibitory activity is inversely related to this descriptor. Hence, by increasing the value of this descriptor the inhibitory activity will be decreased.

The fifth descriptor is PJI2 that appeared in the model. The 2D Petitjean shape index (PJI2) is one of the topological descriptors. This descriptor also called graph-theoretical shape coefficient, is proposed to describe the topological anisometric. This molecular shape descriptor describes the degree of deviation from a perfect cyclic topology. The MF value reveals that PJI2 descriptor has the highest mean effect value with a positive sign. It can be concluded that PJI2 displays a great effect in the model and the value of this descriptor is directly related to inhibitory activity of aryl-substituted isobenzofuran-1(3H)-ones.

The final descriptor is RDF070m (Radial Distribution Function - 7.0 / weighted by atomic masses), which is one of the radial distribution function (RDF) descriptors. RDF in this form meets all the requirements for the 3D structure descriptors. It is independent of the atom number (i.e. the size of a molecule), and is unique regarding the three-dimensional arrangement of the atoms and is also invariant against the translation and rotation of the entire

molecule. The radial distribution function descriptors are based on the distance distribution in the molecule. The radial distribution function of an ensemble of *n* atoms can be interpreted as the probability distribution of finding an atom in a spherical volume of radius *R*. Additionally, the RDF descriptors can be restricted to specific atom types or distance ranges to represent specific information in a certain three-dimensional structure space. The mean effect for RDF070m displays a positive sign, which indicates that $\mathbf{pIC_{50}}$is directly related to this descriptor, and increasing the RDF070m descriptor would result in higher inhibitory activity.

Conclusion

In this study, some aryl-substituted isobenzofuran-1(3H)-ones inhibitors were studied by QSAR based on linear and nonlinear methods. Genetic algorithm method was employed to select a set of molecular descriptors to be used in linear and nonlinear models. The built models were examined and validated for their accuracy, statistical significance and external predictive power. The comparative study of the linear models (GA-MLR, GA-PCR and GA-PLS) and nonlinear models (GA-ANN, GA-ANN-GA, GA-ANN-PSO, GA-ANFIS, GA-ANFIS-PSO, GA-SVM) showed that these models perform equally well in predicting the inhibitory activities of the studied compounds but GA-SVM resulted in highest performance in terms of self-firmness and capability to predict the inhibitory activity of the test set. Interpretation of these descriptors showed that P2U, PJI2 and RDF070M descriptors have the greatest effect on the inhibitory activity. The developed QSAR models revealed some useful structural information associated with inhibitory activity. These models can be useful to predict the inhibition activity for the pore-forming protein perforin by a series of aryl-substituted isobenzofuran-1(3H)-ones and can provide help to design new potent inhibitors.

Acknowledgments

The authors are grateful to Payame Noor University for providing facilities to conduct this study.

References

[1] E. Pourbasheer, R. Aalizadeh, M.R. Ganjali, *Arabian J. Chem.*, http://dx.doi.org/10.1016/j.arabjc.2014.12.021 (**2015**).

[2] A. Habibi-Yangjeh, E. Pourbasheer, M. Danandeh-Jenagharad, *Bull. Korean Chem. Soc.*, **2009**, *140*, 15-27.

[3] E. Pourbasheer, A. Beheshti, H. Khajehsharifi, M.R. Ganjali, P. Norouzi, *Med. Chem. Res*., **2013**, *22*, 4047–4058.

[4] H. Timmerman, *Pharmacochem Libr.*, **1995**, *23*, 413–450.

[5] A. Habibi-Yangjeh, E. Pourbasheer, M. Danandeh-Jenagharad, *Monatsh. Chem.*, **2008**, *29*, 833-841.

[6] A. Khazaei, N. Sarmasti, J.Y. Seyf, Z. Rostami, M.A. Zolfigol, *Arabian J. Chem.*, **2017**, *10*, 801–810.

[7] J.A. Spicer, K.M., Huttunen, Ch.K. Miller, W.A. Denny, A. Ciccone, K.A. Browne, J.A. Trapani, *Bioorg. Med. Chem.*, **2012**, *20*, 1319–1336.

[8] F. Bagheban, A. Niazi, A. Akrami, *J. Mex. Chem. Soc.*, *2015*, *59*, 203–210.

[9] R. Leardi, *J. Chemom.*, **2001**, *15*, 559-569.

[10] Z. Rostami, A. Aminimanesh, L. Samie, *Iranian J. Math. Chem.*, **2013**, *4*, 91–109.

[11] B. Hemmateenejad, M.A. Safarpour, F., Taghavi, *J. Mol. Struct.*, **2003**, *635*, 183-190.

[12] H. Zare-Abyaneh, A. MoghaddamNia, M. BayatVarkeshi, S. Marofi, O. Kisi, *J. Irrig. Drain*, **2011**, *5*, 280–286.
[13] E. Pourbasheer, R. Aalizadeh, M.R. Ganjali, P. Norouzi, *Struct. Chem.*, **2014**, *25*, 355-370.
[14] R. Todeschini, V. Consonni, *Handbook of Molecular Descriptors*, Wiley-VCH, Weinheim, Germany, **2000**.
[15] M. Jalali-Heravi, A. Kyani, *European J. Med. Chem.*, **2007**, *42*, 649-659.
[16] Mathworks, Genetic Algorithm and Direct Search Toolbox Users Guide, The Mathworks Inc., **2005**.
[17] SPSS Base 10.0, Applications Guide, SPSS Inc., Chicago, Ill, USA, **1999**.
[18] M. Hassan, M. Sarfraz, A. Osman, M. Alruwaili, *Int. J. Comput. Sci.*, **2013**, *10*, 55-64.
[19] S. Masrom, Z.Z. Abidin Siti, N. Omar, K. Nasir, *13 Proceedings of the 5th Asian conference on Intelligent Information and Database Systems*, Malaysia, **2013**.
[20] J.K. Jeon, *Fuzzy and Neural Network Models for Analyses of Piles*, Raleigh, North Carolina, USA, **2007**.
[21] O. Ivanciuc, *Rev. Comp. Ch.*, **2007**, *23*, 291-292.
[22] A. Ben-Hur, J. Weston, *Mol. Biol.*, **2010**, *609*, 223-239.
[23] A. Alimoradi, A. Moradzadeh, M.R. Bakhtiari, *J. Min. Environ*, **2013**, *4(1)*:1–14.
[24] V.K. Agrawal, P.V. Khadikar, *Bioorg. Med. Chem.*, **2001**, *9*, 3035–3040.
[25] M. Adimi, M. Salimi, M. Nekoei, E. Pourbasheer, A. Beheshti, *J. Serb. Chem. Soc.*, **2012**, *77*, 639-650.
[26] P.P. Roy, S. Paul, I. Mitra, K. Roy, *Molecules*, **2009**, *14*, 660–1701.
[27] V.H. Masand, K.N. Patil, D.T. Mahajan, R.D. Jawarkar, G.M. Nazerruddin, *Der. Pharma. Chemica*, **2010**, *2*, 22-32.