A Genetic Algorithm- Back Propagation Artificial Neural Network Model to Quantify the Affinity of Flavonoids Toward P-Glycoprotein

Jibin Shen[1,2], Ying Cui[*,1,3], Jun Gu[1], Yaxiao Li[1,3] and Lingzhi Li[*,1,3]

[1] *Department of Medicine Chemistry, Logistics College of Chinese People's Armed Police Forces, China*

[2] *Jiangxi Provincial Corps Hospital of Chinese People's Armed Police Force, China*

[3] *Tianjin Key Laboratory of Occupational and Environmental Hazards Biomarkers, Tianjin, China*

Abstract: Flavonoids, the most diverse class of plant secondary metabolites, exhibit high affinity toward the purified cytosolic NBD2(C-terminal nucleotide-binding domain) of P-glycoprotein (P-gp). To explore the affinity of flavonoids for P-gp, quantitative structure-activity relationships (QSARs) models were developed using back-propagation artificial neural networks (BPANN) and multiple linear regression (MLR). Molecular descriptors were calculated using PaDEL-Descriptor, and the number of descriptors was then reduced using a genetic algorithm (GA) and stepwise regression. The MLR (R^2=0.855, q^2=0.8138, R_{ext}^2=0.6916), 14-3-1 BPANN (R^2=0.8514, q^2=0.7695, R_{ext}^2=0.8142), 14-4-1 BPANN (R^2=0.9199, q^2=0.7733, R_{ext}^2=0.8731), and 14-5-1 BPANN (R^2=0.8660, q^2=0.7432, R_{ext}^2=0.8292) models all showed good robustness. While BPANN models exceeded significantly MLR in predictable performance for their flexible characters, could be used to predict the affinity of flavonoids for P-gp and applied in further drug screening.

Keywords: Back-propagation artificial neural networks, flavonoids, genetic algorithm, P-glycoprotein.

INTRODUCTION

The intrinsic or acquired multidrug resistance (MDR) of tumor cells is a major obstacle to successful anticancer treatments in the clinic. Tumor cells can achieve MDR by various mechanisms, including altering the MDR transporter, perturbing the expression of target enzymes, altering drug activation or degradation, and enhancing DNA repair [1]. MDR transporters can be divided into two classes based on their energy source: secondary transporters, which use proton gradients to facilitate an antiporter mechanism, and adenosine triphosphate (ATP) binding cassette (ABC) transporters, which couple ATP hydrolysis to substrate transport across the cell membrane. ABC transporters belong to one of the largest superfamilies of proteins and either import or export a broad range of substrates, including amino acids, lipids, and drugs. The most widely studied MDR-ABC transporter is P-glycoprotein (P-gp), which in humans is known as MDR1/ABCB1, MRP1/ABCC1, BCRP/ABCG2, *etc.* [2]. Overexpression of P-gp appears to be a major mechanism of MDR [3]. Much research has focused on discovering new potent and noncytotoxic compounds that can inhibit P-gp pump function [4-8].

Flavonoids are the most diverse class of plant secondary metabolites and are ubiquitous in fruits, vegetables, flowers, and beverages, even tea and wine. They exhibit several types of biological activity, including antibacterial [9], antifungal [10], antioxidant [11], and anticancer [12] activities. Although many biological actions have been ascribed to these compounds, their ability to modulate P-gp activity has recently attracted much attention. Silibinin [13], quercetin [14, 15], oroxylin A [16], and biochanin A [17] have been shown to strongly inhibit P-gp-mediated efflux transporters through a mechanism that is generally thought to involve the competitive inhibition of substrate binding to the P-gp NBD. The binding affinity of flavonoids to the P-gp NBD is correlated with transport modulation [18-20].

Given the diverse molecular architectures of flavonoids, quantitative structure-activity relationships (QSARs), which correlate physicochemical properties to biological activities, are useful for exploring potential drug compounds. Using molecular descriptors and fitting methods, different QSAR models have been developed for the P-gp modulating activity of flavonoids. Kothandan and coworkers developed receptor-guided CoMFA and CoMSIA models by docking highly active flavones at the proposed P-gp NBD [7]. A linear solvation energy relationship was also employed to model the affinity toward P-gp, permitting an in-depth analysis of the intermolecular interaction forces involved in pharmacokinetic mechanisms [6]. Wang *et al.* built a Bayesian-regularized neural network model to screen flavonoids for the inhibition of P-gp [8].

Multiple linear regression (MLR) is the standard approach for multivariate data analysis. Artificial neural network algorithms constitute a more flexible class of modeling techniques that are naturally able to model complex nonlinear systems in both classification and regression problems [21]. In this work, QSARs were developed for the binding affinity of flavonoids to P-gp using back-propagation artificial neural networks (BPANN) and MLR. Molecular descriptors were calculated with PaDEL-Descriptor [22] and subsequently subjected to variable reduction using a genetic algorithm (GA) and

*Address correspondence to these authors at the Department of Medicine Chemistry, Logistics College of Chinese People's Armed Police Forces, China;
(Ying Cui) Tel: 86-022-60578189; E-mail: flyingyting@aliyun.com;
(Lingzhi Li) Tel: 86-022-60578184; Fax: 86-022-60578184;
E-mail: llzhx@tom.com

1875-5402/14 $58.00+.00

stepwise regression. Then, the selected descriptors were used as inputs for artificial neural networks with different architectures. The results indicate that the BPANN model with the GA-PLS-selected descriptors is superior to the MLR model and comparable to other 3-D models in the literature. BPANN with GA-PLS variable reduction is a reliable method for modeling the affinity of flavonoids to P-gp and could be used in future drug development.

MATERIALS AND METHODS

Dataset

The dataset used in this study was taken from the literatures [23, 24]. The binding affinity of each compound was estimated from the dissociation constant K_d, which was determined using the Grafit program. The negative logarithm of K_d （pKd） was employed to use the data as the dependent variables in the models. The chemical structures of the compounds used in this work are shown in Table **1**. The compounds in the dataset belong to several families: 40 flavones, 1 isoflavone, 22 chalcones, 5 silybins, 14 aurones, and 6 xanthones. The binding affinity dataset was randomly split into training and test sets as shown in Table **1**. All structure types were included in training, test and validation sets.

Molecular Descriptors

Molecular descriptors are quantitative representations of chemical structures and structural or physicochemical properties. First, the molecular structures of the flavonoids in the data set were drawn in ChemBioOffice 2010. Each molecule was then “cleaned up” and exported as a SMILES file, which served as the input for the generation of molecular descriptors by PaDEL-Descriptor, version 6 [22]. Because of the planar structure of these compounds, only 1D and 2D descriptors were calculated. Furthermore, 1D and 2D molecular descriptors, which are based on the molecular formula and connectivity of the molecule, respectively, are easy to interpret.

After eliminating the descriptors with constant values or mostly zero values (>90%), 118 descriptors remained. To improve the prediction performance of the predictors and provide a better understanding of the underlying factors that contributed to the results, GA and stepwise selection and elimination were used for variable selection.

Genetic Algorithm

GA is one of the most popular variable selection methods inspired by Darwin’s evolutionary theory. In this study, the initial population was 100 solutions (chromosomes), and the maximum number of allowed variables in a solution was 30. Each chromosome, which represents a set of variable combinations, was coded by a binary string of numbers. A value of 1 indicates that a particular variable was selected in the model, and a value of 0 indicates that the variable was omitted. The basic steps of GA have been described in the literature [25]. The descriptor values were first autoscaled to range from zero to one, enabling equal weighting of each descriptor regardless of its absolute value. In this work, we used the method implemented by Leardi [26], in which the fitness function is based on a partial least squares (PLS) model. All chromosomes had a mutation probability of 0.01, and two-point crossovers were imposed with a probability of 0.5. To compensate for the volatile nature of chromosomes in GA, the descriptors were selected based on their selection frequencies over all generations. Models based on the descriptors with the highest selection frequencies have been shown to exhibit promising predictivity [27]. After the GA procedure, the selected descriptors were used as the inputs for ANN.

Software

The procedure was performed with MATLAB R2011b (The MathWorks, Inc.) on a personal computer (Intel Pentium Processor E5800 3.2GHz) operating onWindows 7. The SPSS software package (version 17.0, SPSS, Inc.) was used for the standard MLR analysis. The GA [28] and BPANN toolboxes were modified for our study.

Statistics Parameters

The statistical significance of the models was determined by examining the correlation coefficient R^2 between the predictive and experimental values. This coefficient is calculated as follows:

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

where y_i is the experimental activity, $\hat{y}_i$ is the activity predicted by the model, and $\overline{y}$ is the average activity.

The R^2 value increases as the number of variables in the model increases, while the adjusted R^2 value increases only if the new variables improve the model more than expected by chance. Therefore, the adjusted R^2, which is defined below, was also used:

$$R^2_{adj} = R^2 - p(1 - R^2)/(n - p - 1) \quad (2)$$

where p is the total number of regressors in the model, n is the sample size, and R^2 is the correlation coefficient.

To measure the robustness of the model, the leave-one-out cross-validation procedure was used. In this procedure, the predictive squared correlation coefficient (q^2) is defined according to the following expression:

$$q^2 = 1 - \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \Big/ \sum_{i=1}^{n}(y_i - \overline{y})^2 \quad (3)$$

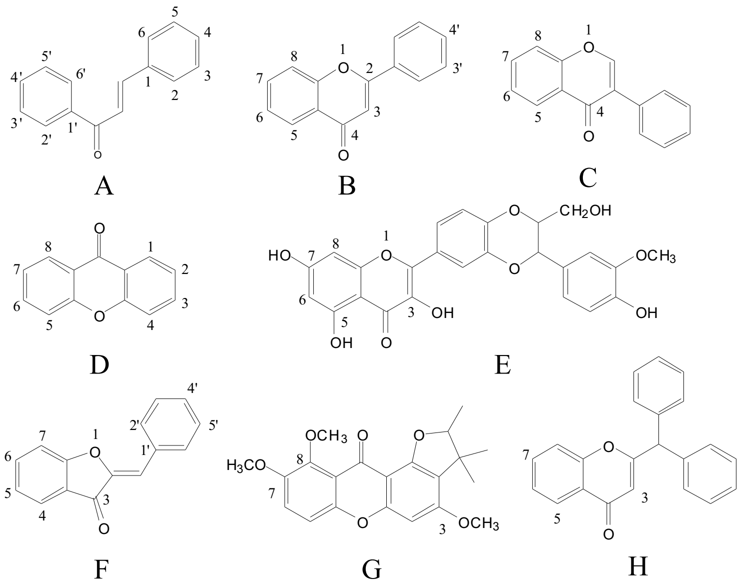
where y_i is the observed activity, $\hat{y}_i$ is the activity estimated by the model, and $\overline{y}$ is the average activity.

RESULTS AND DISCUSSION

Multiple Linear Regression

The first model was generated by MLR, the standard approach for multivariate data analysis, and stepwise selection was employed as the variable selection method. In

Table 1. The Chemical Structures of Flavonoids



ID	Str.	1	2	3	4	5	6	7	8	2’	3’	4’	5’	6’
1	A	-	H	H	H	H	H	-	-	OH	H	OH	H	OH
2	A	-	H	H	OH	H	H	-	-	OH	H	OH	H	OH
3	A	-	H	H	OCH_3	H	H	-	-	OH	H	OH	H	OH
4	A	-	H	H	F	H	H	-	-	OH	H	OH	H	OH
5*	A	-	H	H	Cl	H	H	-	-	OH	H	OH	H	OH
6	A	-	H	H	Br	H	H	-	-	OH	H	OH	H	OH
7**	A	-	H	H	I	H	H	-	-	OH	H	OH	H	OH
8	A	-	H	H	C_2H_5	H	H	-	-	OH	H	OH	H	OH
9**	A	-	H	H	C_3H_7	H	H	-	-	OH	H	OH	H	OH
10	A	-	H	H	C_6H_{13}	H	H	-	-	OH	H	OH	H	OH
11*	A	-	H	H	C_6H_{11}	H	H	-	-	OH	H	OH	H	OH
12	A	-	H	H	C_8H_{17}	H	H	-	-	OH	H	OH	H	OH
13	A	-	H	H	$C_{10}H_{21}$	H	H	-	-	OH	H	OH	H	OH
14	A	-	H	H	$C_{14}H_{29}$	H	H	-	-	OH	H	OH	H	OH
15	A	-	H	H	Prenyl	H	H	-	-	OH	H	OH	H	OH
16	A	-	H	OH	OH	H	H	-	-	OH	H	OH	Prenyl	OH
17	A	-	H	H	H	H	H	-	-	OH	H	H	H	H
18*	A	-	H	H	H	H	H	-	-	OH	H	H	DMA	H
19	A	-	H	H	OH	H	H	-	-	OH	H	H	H	H
20**	A	-	H	Prenyl	H	H	H	-	-	OH	H	H	H	H
21*	A	-	H	H	OCH_3	H	H	-	-	OH	H	H	H	H
22	A	-	H	Prenyl	OCH_3	H	H	-	-	OH	H	H	H	H
23*	B	-	-	OH	-	H	H	H	H	H	H	H	H	H
24	B	-	-	OH	-	OH	H	OH	H	H	H	H	H	H
25**	B	-	-	OH	-	OH	H	OH	DMA	H	H	H	H	H
26*	B	-	-	OH	-	OH	Prenyl	OH	H	H	H	H	H	H

(Table 1) contd…..

ID	Str.	1	2	3	4	5	6	7	8	2’	3’	4’	5’	6’
27	B	-	-	OH	-	OH	H	OH	Prenyl	H	H	H	H	H
28	B	-	-	OH	-	OH	H	OH	H	H	H	OH	H	H
29	B	-	-	OH	-	OH	H	OH	H	H	H	OCH_3	H	H
30**	B	-	-	OH	-	OH	H	OH	DMA	H	H	OCH_3	H	H
31	B	-	-	OH	-	OH	H	OH	H	H	H	F	H	H
32*	B	-	-	OH	-	OH	H	OH	H	Cl	H	Cl	H	H
33	B	-	-	OH	-	OH	H	OH	H	H	H	I	H	H
34**	H	-	-	OH	-	OH	H	OH	H	H	H	H	H	H
35	B	-	-	OH	-	OH	H	OH	H	H	H	C_8H_{17}	H	H
36*	B	-	-	OCH_3	-	OH	H	OH	H	H	H	H	H	H
37	B	-	-	OCH_3	-	OH	H	OCH_3	DMA	H	H	H	H	H
38	B	-	-	OH	-	OH	H	OH	H	H	OH	OH	H	H
39	C	-	-	H	-	OH	H	OH	H	H	H	OH	H	H
40	B	-	-	H	-	H	H	H	H	H	H	H	H	H
41	B	-	-	H	-	H	H	OH	H	H	H	H	H	H
42	B	-	-	H	-	OH	H	OH	H	H	H	H	H	H
43	B	-	-	H	-	OH	CH_3	OH	H	H	H	H	H	H
44*	B	-	-	H	-	OH	H	OCH_3	H	H	H	H	H	H
45	B	-	-	H	-	OH	CH_3	OCH_3	H	H	H	H	H	H
46	B	-	-	H	-	OH	H	OH	H	H	H	OH	H	H
47*	B	-	-	H	-	OH	H	OH	H	H	F	F	H	H
48	B	-	-	H	-	OH	H	OH	H	H	H	I	H	H
49**	B	-	-	H	-	OH	H	O-iPr	H	H	H	H	H	H
50	B	-	-	H	-	OH	iPr	OH	H	H	H	H	H	H
51*	B	-	-	H	-	OH	iPr	O- iPr	H	H	H	H	H	H
52	B	-	-	H	-	OH	IPr	O- iPr	IPr	H	H	H	H	H
53	B	-	-	H	-	OH	Bn	OH	H	H	H	H	H	H
54	B	-	-	H	-	OH	H	OH	Bn	H	H	H	H	H
55*	B	-	-	H	-	OH	Bn	OH	Bn	H	H	H	H	H
56	B	-	-	H	-	OH	H	OBn	H	H	H	H	H	H
57**	B	-	-	H	-	OH	Prenyl	OH	H	H	H	H	H	H
58	B	-	-	H	-	OH	H	OH	DMA	H	H	H	H	H
59*	B	-	-	H	-	OH	H	OH	Prenyl	H	H	H	H	H
60	B	-	-	H	-	OH	Prenyl	OH	Prenyl	H	H	H	H	H
61	B	-	-	H	-	OH	Geranyl	OH	H	H	H	H	H	H
62	B	-	-	H	-	OH	H	OH	Geranyl	H	H	H	H	H
63	B	-	-	H	-	OH	H	OH	DMA	H	H	OH	H	H
64	C	-	-	-	OCH_3	H	OCH_3	H	-	H	H	CN	H	H
65*	C	-	-	-	OCH_3	H	OCH_3	H	-	H	H	$N(CH_3)_2$	H	H
66	C	-	-	-	OCH_3	H	OCH_3	H	-	OCH_3	H	OCH_3	OCH3	H
67	C	-	-	-	OH	H	OCH_3	H	-	H	H	H	H	H
68**	C	-	-	-	OH	H	OCH_3	H	-	H	H	F	H	H

(Table 1) contd…..

ID	Str.	1	2	3	4	5	6	7	8	2’	3’	4’	5’	6’
69	C	-	-	-	OH	H	OCH_3	H	-	H	H	Cl	H	H
70*	C	-	-	-	OH	H	OCH_3	H	-	H	H	Br	H	H
71	C	-	-	-	OH	H	OCH_3	H	-	H	H	I	H	H
72	G	-	-	-	OH	H	OCH_3	H	-	H	H	CN	H	H
73	C	-	-	-	OCH_3	H	OCH_3	H	-	H	H	H	H	H
74**	C	-	-	-	OCH_3	H	OCH_3	H	-	H	H	F	H	H
75	C	-	-	-	OCH_3	H	OCH_3	H	-	H	H	Cl	H	H
76	C	-	-	-	OCH_3	H	OCH_3	H	-	H	H	Br	H	H
77*	C	-	-	-	OCH_3	H	OCH_3	H	-	H	H	I	H	H
78	D	OH	H	OCH_3	H	H	H	OCH_3	OCH_3	-	-	-	-	-
79**	D	O-Prenyl	H	OCH_3	H	H	H	OCH_3	OCH_3	-	-	-	-	-
80	D	-	-	OCH_3	H	H	H	OCH_3	OCH_3	-	-	-	-	-
81*	D	OH	H	OCH_3	H	H	H	OCH_3	OCH_3	-	-	-	-	-
82	D	OH	H	OCH_3	Prenyl	H	H	OCH_3	OCH_3	-	-	-	-	-
83	D	OH	DMA	OCH_3	H	H	H	OCH_3	OCH_3	-	-	-	-	-
84	E	-	-	-	-	-	H	-	H	-	-	-	-	-
85	E	-	-	-	-	-	Prenyl	-	H	-	-	-	-	-
86*	E	-	-	-	-	-	H	-	Prenyl	-	-	-	-	-
87	E	-	-	-	-	-	Geranyl	-	H	-	-	-	-	-
88**	E	-	-	-	-	-	H	-	Geranyl	-	-	-	-	-

*External validation set.
**Internal test set.

the stepwise procedure, the correlations between all the variables and the target activity were calculated. The descriptors that were most closely correlated with the activity were chosen as the original variable. At each step, a new variable that contributed significantly to the model was added, and the contributions of all the variables to the model were then calculated. If the contribution of a new variable decreased or did not increase significantly upon the addition of the new variable, the insignificant variable was removed from the model. The procedure was terminated when all variables in the model were found to be significant and all insignificant variables were removed from the model. The criterion for including a variable in the model was F<=0.05, while the criterion for removing a variable from the model was F>=0.1. As a result of the stepwise selection, the number of descriptors was reduced from 118 to 8. The selected descriptors used in the MLR model are shown in Table **2**. The optimum model is shown as follows:

pK_d=2.711-2.169 ETA_Eta_B_RC-18.548 ETA_dAlpha_B+ 0.162 minsCH3-0.154 SsCH3+0.327 LipoaffinityIndex +0.106 maxdO-0.93 nHaaCH+0.197 SHCsatu.

Table 2. The Selected Set of Descriptors Presented in the MLR Model

Abbreviation	Description
ETA_Eta_B_RC	Branching index EtaB (with ring correction)
ETA_dAlpha_B	A measure of count of hydrogen bond acceptor atoms and/or polar surface area
minsCH3	Minimum atom-type E-State: -CH3
SsCH3	Sum of atom-type E-State: -CH3
LipoaffinityIndex	Lipoaffinity index
maxdO	Maximum atom-type E-State: =O
nHaaCH	Count of atom-type H E-State: :CH:
SHCsatu	Sum of atom-type H E-State: H on C sp3 bonded to unsaturated C

The R^2 value of this model was 0.855, which indicates that the model can account for 85.5% of the variance in pK_d. The regression models were also evaluated using the Fisher ratio (F) in the analysis of variance (ANOVA). With an F value of 45.109 (P=0.000), we suggest that at the 99% level of significance, at least one of the independent variables in the MLR model is useful in predicting the binding affinity of flavonoids to P-gp. The statistical results for the selected descriptors in this model are given in Table **3**. In this model, a Student’s t-test was performed at a confidence level of 95% to confirm the significance of each descriptor. All the P-values of the descriptors were less than 0.05, indicating that the selected descriptors were statistically significant at the 95% level. Moreover, the multicollinearity of the descriptors was evaluated using the variation inflation factor (VIF). A VIF value larger than 10 indicates that a descriptor is highly correlated with one or more of the remaining independent variables. In this model, all the VIF values were less than 4.806, revealing that the descriptors were fairly independent of each other.

Table 3. t-Value and VIF Value of the Descriptors Involved in the MLR Model

Descriptors	t-Value*	Sig	VIF
ETA_Eta_B_RC	7.561	0.000	1.691
ETA_dAlpha_B	-4.630	0.000	1.349
minsCH3	2.046	0.045	2.579
SsCH3	-5.061	0.000	3.694
LipoaffinityIndex	8.303	0.000	3.996
maxdO	6.092	0.000	2.894
nHaaCH	-3.454	0.001	4.806
SHCsatu	2.997	0.004	1.715

*T-test was introduced for compare under the confidence level 95%.

Because high q^2 values appear to be a necessary but not sufficient condition for high predictive power, the predictiveness of the model was further evaluated using an internal validation set and external prediction test set. The robustness, predictiveness, and applicability of the MLR model were demonstrated by a high q^2 value (q^2=0.8138), internal predictive squared correlation coefficient (R_{int}^2=0.7912), and predictive squared correlation coefficient (R_{ext}^2=0.6916) (Table **4**). The flavonoid activities predicted by the MLR model are listed in Table **5**, and Fig. (**1**) shows the plot of experimental activities versus the predicted activities.

Table 4. Performances of MLR and BP-ANN QSAR Models

Models	R^2	R_{adj}^2	q^2	R_{int}^2	R_{ext}^2
MLR	0.855	0.836	0.8138	0.7912	0.6916
14-3-1BPANN	0.8514	0.8492	0.7695	0.7278	0.8142
14-4-1BPANN	0.9199	0.9188	0.7733	0.8638	0.8731
14-5-1BPANN	0.8660	0.8640	0.7432	0.7616	0.8292

Back-Propagation ANN

Because of the poor predictive ability of linear models, non-linear models were also constructed in this study. ANN is one of the most widely used methods for developing non-linear models, and BP, which is a supervised learning technique, is the most popular algorithm [29, 30]. BP-ANN was used as a feature mapping method to construct the non-linear model. More details of ANN theory have been described elsewhere [31]. The ANN developed here was composed of three layers: the input, hidden, and output layers. The number of nodes in the input layer was dependent on the number of descriptors selected in the GA procedure. Because the selection of multiple hidden layers is not reasonable in most cases, only one hidden layer was employed, and the number of nodes in the hidden layer was determined by a parameter ρ, which plays a major role in the optimization of the ANN architecture. The parameter ρ is defined as Equation (4):

$$\rho = \frac{\text{Number of data points in the training set}}{\text{Sum of the number of connections in the ANN}} \qquad (4)$$

When ρ<<1.0, ANN simply memorizes the data. When ρ>>3.0, ANN is unable to generalize. The ρ value in this investigation was maintained between 1 and 3 [32]. One node was used in the output layer. All neurons in the network were fully connected to neighboring-layer neurons through adjustable weights. The initial connection weights were randomly assigned based on a uniform distribution between -0.5 and 0.5. The initial bias values were set to be 1. The inputs were normalized between -1 and 1. A sigmoid function was employed to sum the incoming weights in the hidden layer. Because the performance of ANN is highly dependent on the topology structure, network parameters such as the learning rate and momentum were optimized before training. Furthermore, the mean squared error was used as a criterion to evaluate the ANN performance. To optimize the weights and biases, the network was then trained with the training set using the BP strategy.

The GA-PLS selection procedure led to the selection of 14 descriptors for use in the ANN model (Table **6**). The optimized learning rate and momentum were 0.1 and 0.5, respectively. In this work, we constructed three ANN models with 3, 4, and 5 neurons in the hidden laye, individually. The mean squared error (MSE) was calculated and recorded after every 10 cycles. The network training was stopped when the MSE of the validation set started to increase while that of the training set continued to decrease. The MSE values of the training, test, and validation sets are shown as a function of the number of iterations in Fig. (**2**). For the 14-3-1, 14-4-1, and 14-5-1 BPANN models, the validation set MSEs reached a minimum at 410, 2420, and 760 epochs, respectively, meaning the training of the models was thus completed. Fig. (**1**) shows the plot of the predicted affinities of the flavonoids to P-gp versus the experimental values. The affinity activities predicted by the BPANN models are listed in Table **5**. The statistics of the three BPANN models are in Table **4**. The R^2 values of the three ANN models ranged from 0.8514 to 0.9199 and were comparable to that of the MLR model. A leave-one-out cross-validation (LOO-CV) procedure was also utilized to test the robustness of the three ANN models. Moreover, the robustness and predictive ability of these models were evaluated using internal and external sets of 12 and 18 compounds, respectively. From Table **4**, we can conclude that all of the BPANN models, particularly the 14-4-1 model, are superior to the MLR model in predictive ability. Our models also exhibited better predictivity than some models in the literature. Boccard *et al.* [6] constructed a 3D linear solvation energy VolSurf model exhibited satisfactory internal predictivity (R^2=0.76, q^2=0.71). The ANN approach was also employed by Wang *et al.* [8] to screen flavonoids for the inhibition of P-gp. They developed three models, in which a Bayesian-regularized neural network was found to have the best predictivity with conventional R^2 coefficients of 0.756 and 0.728 for the training and external validation sets, respectively. CoMFA and CoMSIA have also been effectively used to develop

Table 5. Experimental and Predicted Affinities of Flavonoids to the P-gp

ID	Experimental -logKd	Predicted -logKd			
		MLR	14-3-1 BPANN	14-4-1 BPANN	14-5-1 BPANN
1	5.34	5.11	5.21	5.15	5.33
2	5.32	4.99	5.13	5.09	5.18
3	5.64	5.34	5.33	5.50	5.55
4	5.44	5.39	5.34	5.35	5.44
5	5.89	5.71	5.59	5.68	5.77
6	6.24	6.14	6.05	6.18	5.14
7	6.60	6.19	6.25	6.28	6.28
8	5.68	5.83	5.90	5.80	6.02
9	6.00	6.07	6.10	5.98	6.13
10	6.57	6.79	6.76	6.82	6.74
11	6.28	6.77	6.32	6.22	6.30
12	7.70	7.28	7.17	7.25	7.23
13	7.22	7.77	7.55	7.49	7.69
14	4.85	5.08	5.10	4.79	4.84
15	6.28	6.57	6.58	6.37	6.54
16	6.36	6.21	6.37	6.64	6.43
17	5.05	5.41	5.36	5.37	5.56
18	6.36	7.11	6.97	6.27	6.85
19	4.96	5.25	5.24	5.18	5.39
20	6.28	6.83	6.85	6.25	6.75
21	5.74	5.59	5.56	5.65	5.76
22	6.57	6.66	6.73	6.73	6.67
23	5.00	5.28	4.88	4.65	4.89
24	5.23	5.03	5.08	5.02	5.02
25	6.35	6.61	6.43	6.60	6.46
26	6.68	6.37	6.47	6.55	6.50
27	6.66	6.38	6.38	6.54	6.41
28	5.17	4.92	5.02	5.07	4.88
29	5.35	5.26	5.14	5.17	5.29
30	6.70	6.55	6.38	6.68	6.41
31	5.17	5.34	5.23	5.25	5.16
32	5.40	6.12	5.83	5.81	5.95
33	5.96	6.17	6.09	6.00	6.06
34	5.70	6.05	6.10	6.13	6.08
35	7.24	7.17	6.93	7.18	6.94
36	5.05	5.20	5.23	5.29	5.35
37	6.82	6.58	6.67	6.96	6.59
38	5.15	4.81	4.99	5.15	4.78
39	4.58	4.99	5.04	4.95	4.89
40	4.47	5.00	4.90	4.40	4.73
41	4.46	5.25	5.32	4.84	5.19
42	5.05	5.11	5.09	4.95	5.08
43	5.51	5.56	5.54	5.57	5.74

(Table 5) contd…..

ID	Experimental -logKd	Predicted -logKd			
		MLR	14-3-1 BPANN	14-4-1 BPANN	14-5-1 BPANN
44	5.20	5.43	5.35	5.40	5.48
45	5.89	5.62	5.70	5.72	5.80
46	5.00	4.99	5.02	4.95	4.93
47	5.20	5.74	5.42	5.44	5.41
48	5.66	6.17	6.10	6.02	6.06
49	6.00	5.98	5.96	6.12	5.88
50	6.68	5.97	6.12	6.18	6.08
51	6.55	6.83	6.90	7.00	6.89
52	7.48	7.67	7.49	7.58	7.73
53	6.47	6.17	6.36	6.48	6.34
54	6.00	6.19	6.28	6.45	6.28
55	7.44	7.17	7.35	7.58	7.48
56	7.17	6.15	6.15	6.31	6.05
57	6.52	6.45	6.59	6.43	6.56
58	6.70	6.70	6.55	6.46	6.51
59	6.55	6.47	6.49	6.40	6.46
60	7.82	7.44	7.73	7.52	7.52
61	7.35	7.46	7.66	7.53	7.48
62	7.60	7.48	7.57	7.50	7.33
63	6.15	6.59	6.39	6.52	6.40
64	4.70	4.90	4.96	5.12	4.97
65	5.59	5.16	5.10	5.33	5.10
66	4.04	4.37	4.55	4.07	4.22
67	5.88	5.49	5.38	5.45	5.53
68	5.57	5.79	5.47	5.62	5.61
69	6.34	6.07	5.86	5.98	5.94
70	6.82	6.48	6.41	6.42	6.37
71	6.59	6.53	6.62	6.62	6.55
72	5.54	5.54	5.45	5.57	5.58
73	5.15	4.96	5.02	5.34	4.98
74	5.54	5.28	4.95	5.15	5.58
75	6.00	5.31	5.26	5.64	5.39
76	6.09	5.69	5.69	6.14	5.90
77	6.27	5.73	5.82	6.20	6.05
78	4.54	5.00	5.13	4.94	5.04
79	4.86	5.08	5.01	4.89	5.14
80	4.61	4.82	4.93	4.72	4.93
81	6.23	6.06	6.45	6.45	6.33
82	6.41	6.08	6.33	6.47	6.25
83	6.60	6.31	6.56	6.48	6.39
84	5.66	5.89	5.99	5.65	5.77
85	6.43	6.87	6.95	6.43	6.99
86	6.60	6.88	6.85	6.41	6.93
87	6.74	6.85	6.62	6.79	6.80
88	6.92	6.85	6.62	6.80	6.80

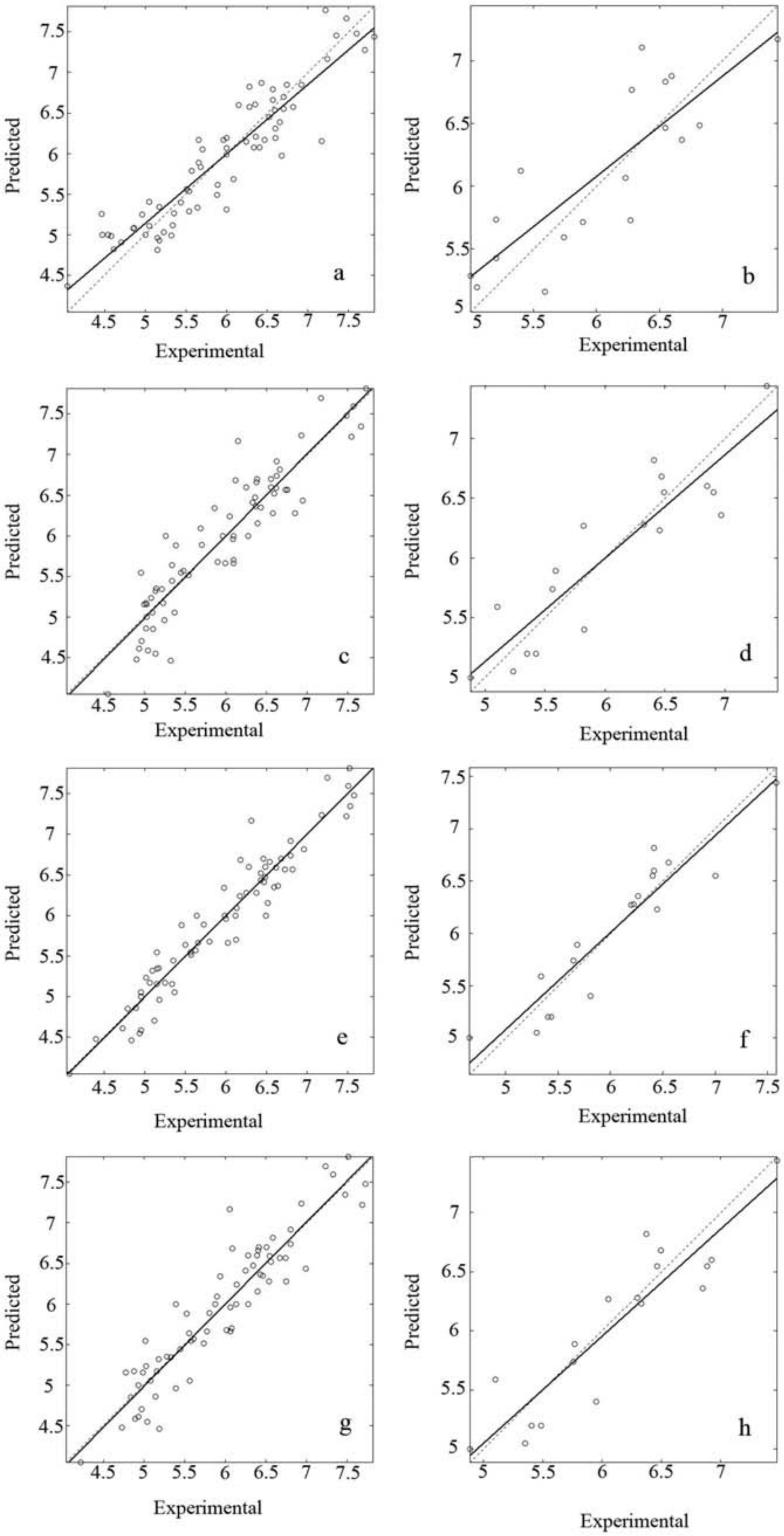


Fig. (1). Plot of predicted pKd values *vs* experimental values in training set and external validation set. **a**. MLR model training set. **b**. MLR model external validation set. **c**. 14-3-1 BPANN model training set. **d**. 14-3-1 BPANN model external validation set **e**. 14-4-1 BPANN model training set. **f**. 14-4-1 BPANN model external validation set. **g**. 14-5-1 model training set. **h**. 14-5-1 BPANN model external validation set.

3D-QSAR models. The models for the affinity of flavonoids to P-gp were constructed by a ligand-based CoMFA (q^2=0.747, R^2_{ext}=0.802) and CoMSIA (q^2=0.810, R^2_{ext}=0.785). In addition, the docking of highly active flavones at the proposed NBD of P-gp was utilized to develop receptor-guided CoMFA (q^2=0.712, R^2_{ext}=0.841) and CoMSIA (q^2=0.805, R^2_{ext}=0.937) models [7]. The molecular descriptors were calculated by PaDEL-Descriptor, which take advantage of the multiple CPU cores to reduce the computational time cost. In addition, this software can work on any platform that supports Java and support most molecular file formats. Our model was easily constructed, and the statistics were comparable to those of the 3D-QSAR models. These results demonstrate the nonlinearity of the correlation between the affinity of flavonoids to P-gp and the selected descriptors, which is not accounted for by simple multiple linear regression.

Table 6. The Selected Set of Descriptors Presented in the ANN Model

Abbreviation	Description
ALogP	Ghose-Crippen LogKow
ATSc2	ATS autocorrelation descriptor, weighted by charges
nssCH2	Count of atom-type E-State: -CH2-
SHaaCH	Sum of atom-type H E-State: :CH:
SssCH2	Sum of atom-type E-State: -CH2-
minsCH3	Minimum atom-type E-State: -CH3
maxHBint4	Maximum E-State descriptors of strength for potential Hydrogen Bonds of path length 4
maxsOH	Maximum atom-type E-State: -OH
maxdO	Maximum atom-type E-State: =O
LipoaffinityIndex	Lipoaffinity index
DELS2	Sum of all atoms intrinsic state differences
ETA_AlphaP	Sum of alpha values of all non-hydrogen vertices of a molecule relative to molecular size
ETA_dAlpha_B	A measure of count of hydrogen bond acceptor atoms and/or polar surface area
MW	Molecular weight

CONCLUSIONS

Overexpression of P-gp is the main mechanism of MDR, a major obstacle to successful anticancer treatment in the clinic. Flavonoids, which are non-transportable inhibitors, can modulate P-gp activity by binding to NBD2 of P-gp without any side effects. In this work, QSAR models were developed for a series of 88 flavonoids using BPANN and MLR. The molecular descriptors were calculated by PaDEL-Descriptor. GA and stepwise regression were then utilized to reduce the number of descriptors to 14 and 8, respectively. To evaluate the robustness and predictive ability of the constructed models, leave-one-out cross-validation and internal and external validation methods were implemented.

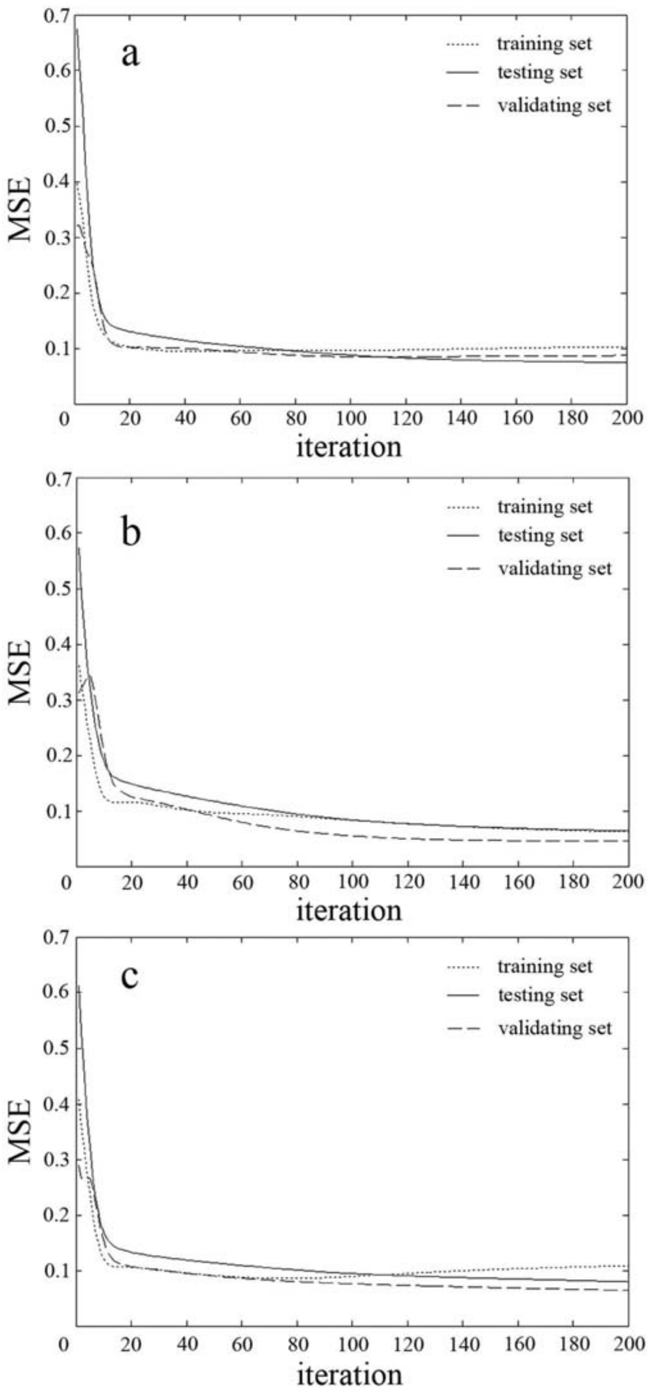


Fig. (2). The MSE value of training set, testing set and validation set versus the iteraction.

The 14-4-1 BPANN model has explained 95.91% of the variance in the affinities of the training-set flavonoids for P-gp and had a R^2 value of 0.8731 for the external validation set, which was superior to those calculated using the other BPANN and MLR models and comparable to those calculated using 3-D models. These results suggest that the 14-4-1 BPANN model constructed with the molecular descriptors selected by GA has the ability to predict the affinity of flavonoids for P-gp NBD2 and could be used in future activity screening.

CONFLICT OF INTEREST

The authors confirm that they do not have any conflicts of interest.

ACKNOWLEDGEMENTS

Declared none.

REFERENCES

[1] Teodori, E.; Dei, S.; Scapecchi, S.; Gualtieri, F. The medicinal chemistry of multidrug resistance (MDR) reversing drugs. *Il Farmaco*, **2002**, *57*(5), 385-415.

[2] Chang, G. Multidrug resistance ABC transporters. *Febs. Lett.*, **2003**, *555*(1), 102-105.

[3] Gottesman, M. M.; Fojo, T.; Bates, S. E. Multidrug resistance in cancer: role of ATP-dependent transporters. *Nat. Rev. Cancer*, **2002**, *2*(1), 48-58.

[4] Boumendjel, A.; McLeer-Florin, A.; Champelovier, P.; Allegro, D.; Muhammad, D.; Souard, F.; Derouazi, M.; Peyrot, V.; Toussaint, B.; Boutonnat, J. A novel chalcone derivative which acts as a microtubule depolymerising agent and an inhibitor of P-gp and BCRP in *in-vitro* and *in-vivo* glioblastoma models. *BMC Cancer*, **2009**, *9*(1), 1-11.

[5] Rao, P. S.; Satelli, A.; Moridani, M.; Jenkins, M.; Rao, U. S. Luteolin induces apoptosis in multidrug resistant cancer cells without affecting the drug transporter function: Involvement of cell line-specific apoptotic mechanisms. *Int. J. Cancer*, **2012**, *130*(11), 2703-2714.

[6] Boccard, J.; Bajot, F.; Di Pietro, A.; Rudaz, S.; Boumendjel, A.; Nicolle, E.; Carrupt, P.-A. A 3D linear solvation energy model to quantify the affinity of flavonoid derivatives toward P-glycoprotein. *Eur. J. Pharm. Sci.*, **2009**, *36*(2–3), 254-264.

[7] Kothandan, G.; Gadhe, C. G.; Madhavan, T.; Choi, C. H.; Cho, S. J. Docking and 3D-QSAR (quantitative structure activity relationship) studies of flavones, the potent inhibitors of p-glycoprotein targeting the nucleotide binding domain. *Eur. J. Med. Chem.*, **2011**, *46*(9), 4078-4088.

[8] Wang, Y.-H.; Li, Y.; Yang, S.-L.; Yang, L. An in silico approach for screening flavonoids as P-glycoprotein inhibitors based on a Bayesian-regularized neural network. *J. Comput. Aid. Mol. Des.*, **2005**, *19*(3), 137-147.

[9] Sivakumar, P. M.; Prabhawathi, V.; Doble, M. Antibacterial activity and QSAR of chalcones against biofilm-producing bacteria isolated from marine waters. *Sar. Qsar. Environ. Res.*, **2010**, *21*(3-4), 247-263.

[10] Sivakumar, P. M.; Muthu Kumar, T.; Doble, M. Antifungal Activity, Mechanism and QSAR Studies on Chalcones. *Chem. Biol. Drug Des.*, **2009**, *74*(1), 68-79.

[11] Sarkar, A.; Middya, T.; Jana, A. A QSAR study of radical scavenging antioxidant activity of a series of flavonoids using DFT based quantum chemical descriptors – the importance of group frontier electron density. *J. Mol. Model.*, **2012**, *18*(6), 2621-2631.

[12] Cabrera, M.; Simoens, M.; Falchi, G.; Lavaggi, M. L.; Piro, O. E.; Castellano, E. E.; Vidal, A.; Azqueta, A.; Monge, A.; de Ceráin, A. L.; Sagrera, G.; Seoane, G.; Cerecetto, H.; González, M. Synthetic chalcones, flavanones, and flavones as antitumoral agents: Biological evaluation and structure–activity relationships. *Bioorgan. Med. Chem.,* **2007**, *15*(10), 3356-3367.

[13] Lee, C.-K.; Choi, J.-S. Effects of silibinin, inhibitor of CYP3A4 and P-glycoprotein *in vitro*, on the pharmacokinetics of paclitaxel after oral and intravenous administration in rats. *Pharmacology*, **2010**, *85*(6), 350-356.

[14] Choi, J.-S.; Piao, Y.-J.; Kang, K. Effects of quercetin on the bioavailability of doxorubicin in rats: Role of CYP3A4 and P-gp inhibition by quercetin. *Arch. Pharm. Res.*, **2011**, *34*(4), 607-613.

[15] Borska, S.; Chmielewska, M.; Wysocka, T.; Drag-Zalesinska, M.; Zabel, M.; Dziegiel, P. *In vitro* effect of quercetin on human gastric carcinoma: Targeting cancer cells death and MDR. *Food Chem. Toxicol.*, **2012**, *50*(9), 3375-3383.

[16] Go, W. J.; Ryu, J. H.; Qiang, F.; Han, H.-K. Evaluation of the Flavonoid Oroxylin A as an Inhibitor of P-Glycoprotein-Mediated Cellular Efflux. *J. Nat. Prod.*, **2009**, *72*(9), 1616-1619.

[17] Zhang, S.; Sagawa, K.; Arnold, R. D.; Tseng, E.; Wang, X.; Morris, M. E. Interactions between the flavonoid biochanin A and P-glycoprotein substrates in rats: *In vitro* and *in vivo*. *J. Pharm. Sci.*, **2010**, *99*(1), 430-441.

[18] Conseil, G.; Baubichon-Cortay, H.; Dayan, G.; Jault, J.-M.; Barron, D.; Di Pietro, A. Flavonoids: A class of modulators with bifunctional interactions at vicinal ATP- and steroid-binding sites on mouse P-glycoprotein. *Proc. Natl. Acad. Sci. USA*, **1998**, *95*(17), 9831-9836.

[19] Pérez-Victoria, J. M.; Pérez-Victoria, F. J.; Conseil, G.; Maitrejean, M.; Comte, G.; Barron, D.; Di Pietro, A.; Castanys, S.; Gamarro, F. High-Affinity Binding of Silybin Derivatives to the Nucleotide-Binding Domain of a Leishmania tropicaP-Glycoprotein-Like Transporter and Chemosensitization of a Multidrug-Resistant Parasite to Daunomycin. *Antimicrob. Agents. Chemother.*, **2001**, *45*(2), 439-446.

[20] Václavíková, R.; Boumendjel, A.; Ehrlichová, M.; Kovář, J.; Gut, I. Modulation of paclitaxel transport by flavonoid derivatives in human breast cancer cells. Is there a correlation between binding affinity to NBD of P-gp and modulation of transport? *Bioorgan. Med. Chem.*, **2006**, *14*(13), 4519-4525.

[21] Prakash, O.; Khan, F.; Sangwan, R. S.; Misra, L. ANN-QSAR Model for Virtual Screening of Androstenedione C-Skeleton Containing Phytomolecules and Analogues for Cytotoxic Activity Against Human Breast Cancer Cell Line MCF-7. *Comb. Chem. High Throughput Screen.*, **2013**, *16*(1), 57-72.

[22] Yap, C. W. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.*, **2011**, *32*(7), 1466-1474.

[23] Boumendjel, A.; egrave; ne; Beney, C.; Deka, N.; Mariotte, A.-M.; Lawson, M. A.; Trompier, D.; Baubichon-Cortay, H.; eacute; Pietro, A. D. 4-Hydroxy-6-methoxyaurones with High-Affinity Binding to Cytosolic Domain of P-Glycoprotein. *Chemical and Pharmaceutical Bulletin*, **2002**, *50*(6), 854-856.

[24] Boumendjel, A.; Di Pietro, A.; Dumontet, C.; Barron, D. Recent advances in the discovery of flavonoids and analogs with high-affinity binding to P-glycoprotein responsible for cancer cell multidrug resistance. *Med. Res. Rev.*, **2002**, *22*(5), 512-529.

[25] Gonzalez, M. P.; Teran, C.; Saiz-Urra, L.; Teijeira, M. Variable Selection Methods in QSAR: An Overview. *Curr. Top. Med. Chem.*, **2008**, *8*(18), 1606-1627.

[26] Leardi, R. Application of genetic algorithm–PLS for feature selection in spectral data sets. *J. Chemometr.*, **2000**, *14*(5-6), 643-655.

[27] Sadat Hayatshahi, S. H.; Abdolmaleki, P.; Ghiasi, M.; Safarian, S. QSARs and activity predicting models for competitive inhibitors of adenosine deaminase. *FEBS Lett.*, **2007**, *581*(3), 506-514.

[28] Houck, C. R.; Joines, J. A.; Kay, M. G. A genetic algorithm for function optimization: a Matlab implementation. *NCSU-IE TR*, **1995**, *95*(09).

[29] Wu, J.; Mei, J.; Wen, S.; Liao, S.; Chen, J.; Shen, Y. A self-adaptive genetic algorithm-artificial neural network algorithm with leave-one-out cross validation for descriptor selection in QSAR study. *J. Comput. Chem.*, **2010**, *31*(10), 1956-1968.

[30] Fatemi, M.; Baher, E. A novel quantitative structure–activity relationship model for prediction of biomagnification factor of some organochlorine pollutants. *Mol. Divers.*, **2009**, *13*(3), 343-352.

[31] Zou, J.; Han, Y.; So, S.-S. In: *Artificial Neural Networks: Methods and Applications*; Livingstone, D. J., Ed. 2008; Vol. 458, pp. 14-22.

[32] Douali, L.; Villemin, D.; Cherqaoui, D. Exploring QSAR of Non-Nucleoside Reverse Transcriptase Inhibitors by Neural Networks: TIBO Derivatives. *Int. J. Mol. Sci.*, **2004**, *5*(2), 48-55.

Received: March 19, 2013 Revised: August 13, 2013 Accepted: August 27, 2013