Available online at www.sciencedirect.com
ScienceDirect
Bioorganic & Medicinal Chemistry 15 (2007) 7746–7754
Bioorganic & Medicinal Chemistry

A novel QSAR model for prediction of apoptosis-inducing activity of 4-aryl-4-H-chromenes based on support vector machine

Mohammad Hossein Fatemi* and Sajjad Gharaghani

Department of Chemistry, Mazandaran University, Babolsar, Iran

Received 29 May 2007; revised 21 August 2007; accepted 28 August 2007
Available online 1 September 2007

Abstract—In this work some chemometrics methods were applied for modeling and prediction of the induction of apoptosis by 4-aryl-4-H-chromenes with descriptors calculated from the molecular structure alone. The genetic algorithm (GA) and stepwise multiple linear regression methods were used to select descriptors which are responsible for the apoptosis-inducing activity of these compounds. Then support vector machine (SVM), artificial neural network (ANN), and multiple linear regression (MLR) were utilized to construct the nonlinear and linear quantitative structure–activity relationship models. The obtained results using SVM were compared with ANN and MLR; it revealed that the GA–SVM model was much better than other models. The root-mean-square errors of the training set and the test set for GA–SVM model are 0.181, 0.241 and the correlation coefficients were 0.950, 0.924, respectively, and the obtained statistical parameters of cross validation test on GA–SVM model were $Q^2 = 0.71$ and SRESS = 0.345 which revealed the reliability of this model. The results were also compared with previous published model and indicate the superiority of the present GA–SVM model.

1. Introduction

Apoptosis, or programmed cell death, is an innate mechanism by which unwanted, defective, or damaged cells are rapidly and selectively eliminated from the body.

Apoptosis was originally described by Kerr and Wyllie,[1] who observed the unique changes in cells undergoing organized cell death by electron microscopy.

Apoptosis occurs during tissue remodeling, embryonic development, and immune regulation,[2–6] and is the principle mechanism employed by the immune system and chemotherapeutic drugs in eradicating tumor cells. Resistant tumor cells evade the action of anticancer agents by increasing their apoptotic threshold. This has spurred the development of novel chemical compounds capable of inducing apoptosis in chemo/immune-resistant tumor cells. Therefore apoptosis has received a huge amount of attention in recent years.[2] Recent reports indicate that many clinically useful cytotoxic agents induce apoptosis in cancer cells.[7,8] Since many medicines are typically developed using a trial and error approach which is costly and time-consuming, therefore the development of theoretical methods as alternative tools for predicting medical activities of chemicals has been the subject of many intensive studies. As we know, the properties of a chemical are all derived from, and related to, the unique molecular structure of that chemical, because these properties are all originated from the molecular structure of the chemical. It follows that these relationships also exist between structure and activities of chemical. These principles form the underlying basis for the prediction of activity from chemical structure. Among the theoretical methods, the quantitative structure–activity relationships (QSAR) have been successfully established to predict different important biopharmaceutical properties, including genotoxicity,[9–11] toxicity,[12] oral bioavailability,[13] carcinogenicity,[14,15] and mutagenicity,[16] etc. To our knowledge, only four attempts have been made to build QSAR models in the general field of apoptosis. Hansch et al.[17] presented a QSAR model to study the apoptosis including activities of simple phenols, estradiol, bisphenol A and diethyl stilbesterol on L1210 leukemia cells and later they presented a QSAR model for investigation of apoptosis-induction in various cancer cells.[18] They considered the effect of phenolic compounds on Ramos cells (non-Hodgkins B-cell

Keywords: Quantitative structure–activity relationship; Apoptosis; Support vector machine; Molecular modeling.
* Corresponding author. Tel.: +98 1125242931; fax: +98 1125242002; e-mail: mhfatemi@umz.ac.ir

0968-0896/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2007.08.057

lymphoma); the effect of *O*-8-thapsigargin analogous on human prostate cancer cells (Tsu-Pr-1), and the induction of apoptosis of a complex set of congeners on human fibro sarcoma cells (HT 1080). Selassie et al.[19] investigated the apoptosis-inducting effect of 51 substituted caspase-mediated phenols in a murine leukemia cell line (L1210). They determined the concentration needed to induce caspase activity by 50% (I_{50}) and utilized those data to develop a QSAR model using steric terms and hydrophobic character of the substituents on the phenolic ring. Kemnitzer et al. found 4-aryl-4H-chromenes[20] to be a promising series of novel apoptosis inducers that could be used to develope new therapeutic anticancer agents. In the recent work, Afantitis et al.[21] developed a linear QSAR model using seven descriptors for prediction of apoptosis-induction for the **43** 4-aryl-4H-chromenes. The best global model after rejecting one compound in their model yields the correlation coefficients (R) of 0.806, root-mean-square error of 0.222, and $Q^2 = 0.678$ for training set.

In QSAR studies, there are some techniques which can be applied for construction of model, such as multiple linear regression (MLR) and artificial neural networks (ANN), that were used for inspection of linear and nonlinear relation between interested activity and molecular descriptors, respectively. The flexibility of neural networks enables them to discover more complex nonlinear relationships in experimental data.[22] Neural networks have some problems inherent to its architecture, such as overtraining, overfitting, network optimization, and reproducibility of results, due to random initialization of the networks and variation of stopping criterias.[23] Owing to these reasons there is a tendency to use more accurate and informative techniques in QSAR analysis. The support vector machine (SVM) is a new algorithm developed from the machine learning community.[24] SVM approach automatically controls the flexibility of the resulting classifier on the training data. Accordingly, by the design of the algorithm, the deteriorating effect of the input dimensionality on the generalization ability is largely suppressed. Due to its remarkable generalization performance, the SVM has attracted attention and gained extensive application, such as pattern recognition problems,[23,25] drug design,[26] QSAR,[27] and quantitative structure–property relationship (QSPR) analysis.[28] In the most of these cases, the performance of SVM modeling either matches or is significantly better than that of traditional machine learning approaches. The main aim of the present work was to establish a new QSAR model for predicting apoptosis-inducing activity of the 4-aryl-4H-chromenes in human breast cancer cells (T47D) using SVM techniques. The performance of this model was compared with those obtained by ANN and MLR methods as well as previous work of Afantitis.

2. Materials and methods

2.1. Data set

The data set used in this study was taken from the work of Kemnitzer et al.[20] and is shown in Table 1.This set contains the apoptotic activities of **43** 4-aryl-4H-chromenes compounds which were measured at the same conditions. The basic structures of these compounds are shown in Figure 1.

The specific apoptotic activities of these compounds were expressed as the effective concentration, which causes 50% reduction in cell growth (EC_{50}). The apoptotic activities in logarithmic scale ($\log 1/EC_{50}$) fall in the range of −0.763 for compound No. **1a**, to 1.854 for compound No. **43b**, with a mean value of 0.983.

2.2. Descriptors calculation and selection

The first step to obtain a QSAR model was to encode the structural features of molecules, which were named molecular descriptors. The molecular descriptors used to search for the best model of the apoptosis-induction activity of these compounds were calculated by the Dragon program[29] on the basis of the minimum energy molecular geometries that optimized by the Hyperchem package (Ver. 7.0)[30] based on AM1 semi empirical method. In addition electronic descriptors were calculated by the MOPAC package.[31]

After the calculation of the molecular descriptors, those that stayed constant for all molecules were eliminated and pairs of variables with a correlation coefficient greater than 0.90 were classified as intercorrelated, and one of them in each correlated pair was deleted.

We used nonlinear feature mapping techniques (SVM and ANN) for construction of QSAR models in this work. Since these methods cannot be able to select the more significant descriptors from the pool of calculated molecular descriptors, it would be necessary to use some variable selection methods. In the present work, stepwise multiple linear regression (Stepwise-MLR) and genetic algorithm (GA) variable subset selection methods[32] were used for the selection of the most relevant descriptors from the pool of remaining 180 descriptors. These descriptors would be used as inputs of the SVM and ANN.

2.3. Support vector machine

Support vector machine, developed by Vapnik and Cortes,[33] as a novel type of machine learning method, is gaining popularity due to many attractive features and promising empirical performance. Originally, the SVM was developed for pattern recognition problems.[34] Recently, with the introduction of ε-insensitive loss function, the SVM has been extended to solve nonlinear regression estimation.[35]

In support vector regression (SVR), the basic idea is to map the data x into a higher-dimensional feature space F via a nonlinear mapping Φ and then to do linear regression in this space. Therefore, regression approximation addresses the problem of estimating a function based on a given data set $G = \{(x_i, d_i)\}_{i=1}^{l}$ (x_i is the input vector, d_i is the desired value, and l is the total number of data patterns). SVM approximates the function in the following form:

Table 1. Data set and corresponding observed and GA–SVM calculated values of apoptosis-inducing activity of 4-aryl-4H-chromenes

Compound	R^1	R^2	R^3	R^4	A	R^5	R^6	R^7	log(1/EC_{50}) (Obs.)	log(1/EC_{50}) (Pred.)	Residual
1a	H	H	OH	H	—	—	—	—	−0.763	−0.722	0.041
t2b	H	H	OH	OH	—	—	—	—	−0.230	0.191	0.421
3a	H	H	NH_2	H	—	—	—	—	−0.079	0.274	0.353
4a	H	Me	NHEt	H	—	—	—	—	−0.041	0.308	0.349
5a	H	H	NEt_2	H	—	—	—	—	0.319	0.709	0.390
6e	—	—	—	—	C	CN	H	H	0.409	0.451	0.042
t7e	—	—	—	—	C	NO_2	H	H	0.409	0.391	−0.018
8c	—	—	—	—	C	H	H	H	0.444	0.483	0.041
9a	H	H	NHEt	H	—	—	—	—	0.481	0.520	0.041
10a	H	H	NH_2	H	—	—	—	—	0.509	0.273	−0.236
11a	H	OCH_2O		Me	—	—	—	—	0.678	0.260	−0.418
t12c	—	—	—	—	N	H	H	H	0.769	0.882	0.113
13a	H	H	OMe	H	—	—	—	—	0.796	0.836	0.040
14b	H	H	Cl	H	—	—	—	—	0.796	0.834	0.038
15e	—	—	—	—	C	Br	H	H	0.824	0.864	0.040
16b	H	H	Br	H	—	—	—	—	0.854	0.932	0.078
t17b	H	H	OH	H	—	—	—	—	0.886	0.752	−0.134
18d	—	—	—	—	C	Cl	H	H	0.921	0.836	−0.085
19d	—	—	—	—	C	OMe	H	H	0.959	0.924	−0.035
20d	—	—	—	—	C	NO_2	H	H	0.959	0.919	−0.040
21e	—	—	—	—	C	OMe	H	OMe	1.036	1.076	0.040
t22c	—	—	—	—	C	NO_2	H	H	1.051	1.388	0.337
23c	—	—	—	—	C	Cl	H	H	1.097	1.007	−0.090
24a	H	H	NMe_2	H	—	—	—	—	1.137	0.758	−0.379
25b	H	H	OEt	H	—	—	—	—	1.194	1.474	0.280
26d	—	—	—	—	C	OMe	H	OMe	1.210	1.034	−0.176
t27b	H	H	OH	NH_2	—	—	—	—	1.215	1.244	0.029
28c	—	—	—	—	C	OMe	H	H	1.284	1.137	−0.147
29c	—	—	—	—	C	Br	H	H	1.284	1.324	0.040
30d	—	—	—	—	C	OMe	OMe	OMe	1.310	1.270	−0.040
31e	—	—	—	—	C	I	OMe	OMe	1.310	1.270	−0.040
t32c	—	—	—	—	N	OMe	H	H	1.328	1.477	0.149
33b	H	H	Me	Me	—	—	—	—	1.337	1.571	0.234
34b	H	H	NH_2	NH_2	—	—	—	—	1.468	1.427	−0.041
35b	H	H	NH_2	H	—	—	—	—	1.481	1.541	0.060
36b	H	H	NH_2	Me	—	—	—	—	1.585	1.654	0.069
t37c	—	—	—	—	C	OMe	OMe	OMe	1.585	1.790	0.205
38e	—	—	—	—	C	Cl	OMe	OMe	1.620	1.281	−0.339
39e	—	—	—	—	C	Br	OH	OMe	1.638	1.572	−0.066
40b	H	H	NMe_2	H	—	—	—	—	1.721	1.648	−0.073
41b	H	H	OMe	H	—	—	—	—	1.769	1.728	−0.041
t42c	—	—	—	—	C	OMe	H	OMe	1.824	1.454	−0.370
43b	H	H	NHEt	H	—	—	—	—	1.854	1.814	−0.040

The letters of **a**, **b**, **c**, **d**, and **e** in the first column correspond to the basic structures of 4-aryl-4H-chromenes depicted in Figure 1, and t is referring to test set.

$$y = \sum_{i=1}^{l} w_i \Phi(x_i) + b, \tag{1}$$

where $\{\Phi(x_i)\}_{i=1}^{l}$ are the features of inputs, and $\{w_i\}_{i=1}^{l}$ and b are coefficients. They are estimated by minimizing the regularized risk function $R(C)$

$$R(C) = C\frac{1}{N}\sum_{i=1}^{N} L_\varepsilon(d_i, y_i) + \frac{1}{2}\|w\|^2, \tag{2}$$

where

$$L_\varepsilon(d,y) = \begin{cases} |d-y| - \varepsilon & \text{for} \quad |d-y| \geqslant \varepsilon \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

and ε is a prescribed parameter.

In Eq. 2, $C(1/N)\sum_{i=1}^{N} L_\varepsilon(d_i, y_i)$ is the so-called empirical error (risk), which is measured by ε-insensitive loss function $L_\varepsilon(d,y)$, and indicates that it does not penalize errors below ε. The parameter of ε is called the tube size, and it is equivalent to the approximation accuracy placed on the training data points. The second term, $1/2||w||^2$, is used as a measurement of function flatness. C is a regularized constant determining the trade-off between the training error and the model flatness. Introduction of slack variables 'ξ' leads Eq. 2 to the following constrained function:

$$\text{Minimize } R(w, \xi_i, \xi^*) = 1/2\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{4}$$

$$\begin{aligned} \text{Subject to } & w\Phi(x_i) + b - d_i \leqslant \varepsilon + \xi_i^*, \\ & d_i - w\Phi(x_i) - b \leqslant \varepsilon + \xi_i, \\ & \xi_i + \xi_i^* \geqslant 0. \end{aligned} \tag{5}$$

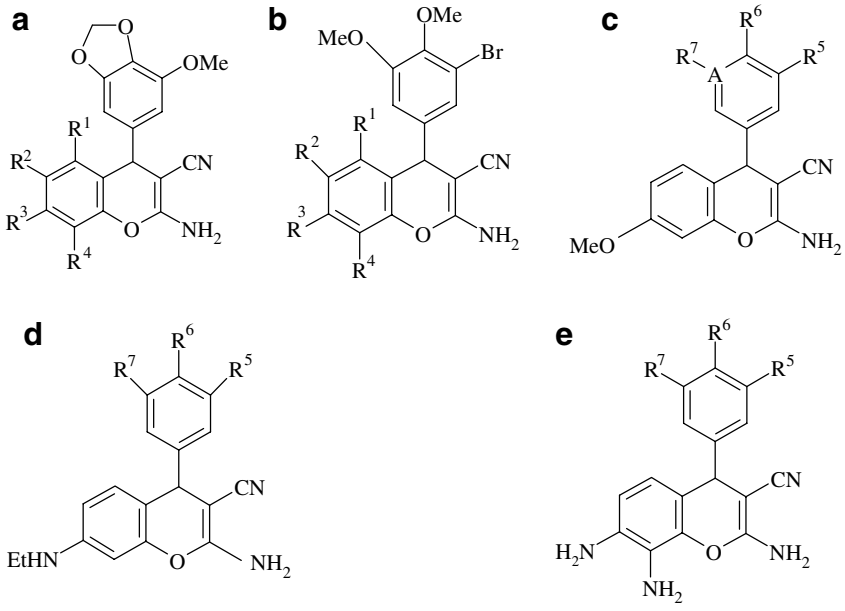

Figure 1. Basic structures of 4-aryl-4H-chromenes.

Thus, decision function of Eq. 1 changes to the following form:

$$f(x,\alpha_i,\alpha_i^*)=\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)K(x,x_i)+b \tag{6}$$

In this equation, $K(x_i,x_j)$ is the kernel function and α_i and α_i^* are the introduced Lagrange multipliers. They satisfy the equality of $\alpha_i\alpha_i^*=0$, $\alpha_i\geqslant 0,\alpha_i^*\geqslant 0$, and are obtained by maximizing the dual form of Eq. 4, which has the following form:

$$\Phi(\alpha_i,\alpha_i^*)=\sum_{i=1}^{l}d_i(\alpha_i-\alpha_i^*)-\varepsilon\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)-(1/2)\times\sum_{i=1}^{l}\sum_{j=1}^{l}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)K(\alpha_i-\alpha_j) \tag{7}$$

with the following constrains

$$\begin{aligned}&0\leqslant\alpha_i\leqslant C\quad i=1,\ldots,l\quad 0\leqslant\alpha_i\leqslant C\\&i=1,\ldots,l\quad\sum_{i=1}^{l}(\alpha_i,\alpha_i^*)=0\end{aligned} \tag{8}$$

Based on the Karush–Kuhn–Tucker (KKT) conditions of quadratic programming, only a number of coefficients $(\alpha_i-\alpha_i^*)$ will be assumed to have nonzero values, and the data points associated with them could be referred to support vectors.

The value of kernel function is equal to the inner product of two vectors x_i and x_j in the feature space $\Phi(x_i)$ and $\Phi(x_j)$, that is, $K(x,x_i)=\Phi(x_i)\ \Phi(x_j)$. The elegance of using kernel function lieds in the fact that one can deal with feature spaces of arbitrary dimensionality without having to compute the map $\Phi(x)$ explicitly. Any function that satisfies Mercer's condition can be used as the kernel function. In the practical study, the type of kernel function and its parameter are defined by the user. In support vector regression, the Gaussian radial basis function (RBF) and the polynomial function are commonly used, and are defined in Eqs. 9 and 10, respectively:

$$K(x_i,x_j)=\exp\left(\frac{-\|x_i-x_j\|^2}{2\sigma^2}\right) \tag{9}$$

$$K(x_i,x_j)=(x_i^Tx_j+1)^d \tag{10}$$

In case of the RBF kernel, the parameter of σ represents the kernel width, and d in Eq. 10 denotes the degree of the polynomial kernel. The kernel parameter and also the earlier mentioned parameters C and ε need to be selected properly by the user, because the generalization performance of the SVR model heavily depends on the right setting of these parameters.

The overall performance of SVM was evaluated in terms of root-mean-square error which was calculated from the following equation:

$$\mathrm{RMSE}=\sqrt{\frac{\sum_{i=1}^{n_s}(y_i-y_o)^2}{n_s}} \tag{11}$$

In this equation y_i is the desired output, y_o is the predicted value by model, and n_s is the number of molecules in data set.

The predictive power of the SVM models developed on the selected training sets is estimated on the predictions of an external test set and also examined by cross validation test to the calculating statistical parameters of Q^2 and SPRESS as follows:

$$Q^2 = 1 - \frac{\sum(y_o - y_i)^2}{\sum(y_i - y_m)^2} \quad (12)$$

$$\text{SPRESS} = \sqrt{\frac{\sum(Y_o - Y_i)^2}{n - k - 1}} \quad (13)$$

In the above expressions, y_m is the mean of dependent variable, n is the number of observations, and k is the number of independent variables in regression equation.

All calculations in this work were carried out by using Matlab (V 7.1, The Mathworks, Inc.) and the SVM toolbox was developed by Gunn.[36,37] The calculations were performed on a 3.4 GHz Intel Pentium IV with 1 GB RAM under windows XP.

3. Result and discussion

As mentioned in the previous section, two linear and nonlinear variable selection methods were used to select the most significant descriptors (Stepwise-MLR and GA). The selected descriptors by these methods were used to construct some linear and nonlinear models by using MLR, ANN, and SVM techniques. Based on the types of variable selection method and also the types of the feature mapping technique, these models can be shown as MLR–MLR, MLR–ANN, MLR–SVM, GA–MLR, GA–ANN, and GA–SVM.

The statistical parameters of these models are shown in Table 2. As can be seen from this table the statistical parameters of GA–SVM model are better than the other models, therefore we only will explain descriptors which were used in this model.

Since the chemical variation of the considered compounds are low, the selection of chemical descriptors, which can encode the small variations between structures of molecules in data set, is very important. In this way, GETAWAY and WHIM descriptors are very informative 3D descriptors that can encode structural features of molecules and they are included in the GA–SVM model. The seven most significant descriptors which were selected by genetic algorithm are: E-state topological parameter (ESTP), maximum partial charge for on carbon atom (PC_{max}), H autocorrelation of lag 3/unweighted (H3u), H autocorrelation of lag 6/weighted by atomic masses (H6m), average information content order 0 (AIC0), R maximal autocorrelation of lag 3/weighted by atomic Sanderson electronegativities ($R3e^+$), and first component accessibility directional WHIM index/weighted by atomic electrotopological states (E1s). The statistical parameters of GA–MLR model constructed by these descriptors are shown in Table 3. The methods for calculations of these descriptors and the meaning of them have been explained in the Handbook of Molecular Descriptors by Todeschini et al.[38] In these 7 descriptors, $R3e^+$, H3u, and H6m are GETAWAY type, ESTP is geometrical, PC_{Max} is electrostatic, E1s is WHIM, and AIC0 is topological descriptor. The correlation matrix between these descriptors is shown in Table 4. As can be seen in this table, the linear correlation between each two descriptors is lower than 0.4, therefore they are independent from each other.

For inspection of the relative importance and contribution of each descriptor in the model to apoptosis-inducing activity, the value of mean effect (MF) was calculated for each descriptor by the following equation and is shown in the last column of Table 3:

$$\text{MF}_j = \frac{\beta_j \sum_{i=1}^{i=n} d_{ij}}{\sum_j^m \beta_j \sum_i^n d_{ij}}, \quad (14)$$

where MF_j is the mean effect for considered descriptor j, β_j is the coefficient of descriptor j and d_{ij} is the value of interested descriptors for each molecule, and m is the number of descriptors in the model. The value of MF revealed the relative importance of a descriptor in comparison with the other descriptors in the model and its sign represented the direction of variation in the values of activities resulted by increasing (or decreasing) the values of this descriptor.

One of the newly developed 3D descriptors is geometry topology and atomic weight assembly (GETAWAY) descriptors that were presented by Consonni et al.[39,40] They encode geometrical information given from influence matrix, topological information given by molecular graph, and chemical information from selected atomic properties. They contain two sets of theoretically closely related molecular descriptors; H-GETAWAY descrip-

Table 2. The statistical parameters of different constructed QSAR models

	Training set			Test set		
	RMSE	R	F	RMSE	R	F
Previous work	0.222	0.844	16	0.319	0.932	33
MLR–MLR	0.261	0.915	165	0.365	0.830	16
MLR–ANN	0.349	0.868	97.7	0.477	0.681	6
MLR–SVM	0.207	0.935	221	0.384	0.811	13
GA–MLR	0.261	0.914	162	0.366	0.859	20
GA–ANN	0.356	0.837	75	0.473	0.657	5
GA–SVM	0.181	0.950	290	0.241	0.924	41

Table 3. Details of the constructed GA–MLR model

Descriptor[a]	Coefficient	MF[b]
ESTP	−0.003(±0.001)	−0.561
PC_{Max}	−73.552(±21.290)	−3.76
H3u	−0.629(±0.283)	−14.41
H6m	1.202(±0.302)	0.21
AIC0	2.702(±0.769)	5.92
$R3e^+$	−9.259(±3.116)	−1.5
E1s	−2.043(±0.736)	−0.76
Constant	2.812(±2.365)	

$n = 34$, $R = 0.914$, RMSE = 0.261, $F = 162$.

[a] The name and chemical meanings of descriptors are explained in the text.

[b] MF refer to the mean effect value.

Table 4. Correlation matrix of descriptors selected by genetic algorithm

	ESTP	PC_{max}	H3u	H6m	AICO(0)	$R3e^+$	E1s
ESTP	1	0.257	−0.022	0.258	0.203	−0.122	0.132
PC_{max}		1	0.182	−0.341	−0.002	−0.395	0.234
H3u			1	0.124	−0.084	−0.123	0.054
H6m				1	0.242	0.210	−0.072
AICO(0)					1	0.100	0.383
$R3e^+$						1	0.317
E1s							1

tors which have been calculated by the molecular influence matrix and R-GETAWAY descriptors which have been formed by the influence/distance matrix, where the elements of this matrix are combined with those of the geometry matrix.

H3u is one of GETAWAY type descriptors, which appeared in the model. This descriptor is related to the size and location of the atom in the molecule. By increasing the size of the atom and the distance between an atom and the center of the molecule the value of this descriptor increases. As shown in Table 3, the mean effect of H3u has negative sign, which indicates that log(1/EC_{50}) is inversely related to this descriptor; therefore, increasing the size of molecules leads to decrease in its activity.

The second descriptor is H6m which was weighted by atomic mass. As shown in Table 3, mean effect of H6m has positive sign, therefore the value of apoptotic activity was varied in the same direction to this descriptor. By considering the value of this descriptor for the whole molecules in data set, it was concluded that by increasing of molecular mass the value of this descriptor increased which caused an increase in its activity.

The third GETAWAY descriptor is $R3e^+$, which was calculated by the multiplication of the leverages between two atoms with topological distance equal to 3 and the maximum value of the respective Sanderson electronegativities, and then it was divided to the geometrical distance between them. Thus increasing the size and the electronegativity of molecule increases its $R3e^+$ value. Mean effect of $R3e^+$ has negative sign, which indicates that an increase in $R3e^+$ leads to decrease in hydrophobicity of the molecule and finally a decrease in its activity.

The second type of descriptors in the model was weighted holistic invariant molecular descriptors (WHIM). WHIM descriptors are the molecular descriptors based on statistical indices calculated on the projections of the atoms along principal axes.[38,41] They are built in such a way to capture relevant molecular 3-dimensional information regarding to the molecular size, shape, symmetry, and atom distribution with respect to invariant reference frames. The WHIM descriptor E1s, calculated from the fourth-order moments of the t_m scores weighted by the electrotopological states, is related to the atom distribution along the first axis for the electrotopological states-weighted scheme. According to this, E1s represents a combination of both electronic and topological characteristics in a mathematically defined zone. Mean effect of E1s has negative sign which indicated that by increasing the size of molecules, its apoptosis-inducing activity was decreased.

E-state topological parameters (ESTP) were derived from applying the *Ivanciuc–Balaban operator* to the *E-state index* values used to characterize molecule atoms.[42]

This descriptor describes the atomic connectivity and branching information in the molecule. The mean effect of ESTP has negative sign, which indicated that by increasing the size of molecule, its activity decreases.

The next descriptor was maximum partial charges on carbon atom and its mean effect had negative sign, which indicated that an increase in partial charge on carbon atom in a molecule leads to a decrease in its hydrophobicity and caused a decrease in its apoptosis-inducing activity.

The average information contents descriptors are defined on the basis of the Shannon information theory and it is inversely related to size of molecule. They can be calculated for different orders of neighborhoods, r ($r = 0, 1, 2, \ldots, \rho$), where ρ is the radius of the molecular graph G. At the zeroth-order level, the atom set is partitioned solely on the basis of its chemical nature; at the level of the first-order topological neighborhood, the atoms are partitioned into disjoint subsets on the basis of their chemical nature and their first-order bonding topology. At the next level, the atom set is decomposed into equivalence classis using their chemical nature and bonding pattern up to the second-order bonded neighbors.[43] In essence, this descriptor gives us information on how many different atoms are in the molecule and how diverse the branching of these atoms is at zeroth valence level (coordination sphere). The mean effect of AIC0 has positive sign which indicates that by decreasing of molecular size the apoptosis activity increases. According to the above discussion. it was concluded that steric parameters as well as electronic interactions can affect the apoptosis-inducing activities of 4-aryl-4H-chromenes. The justification of these descriptors based on mean effect sign showed a good compatibility with the work of Kemnitzer et al.[20]

These GA selected descriptors were used as inputs for the construction of SVM model. Firstly, the kernel function should be determined, which represents the sample

distribution in the mapping space. In this work, the polynomial kernel function was set to 2 ($d = 2$) because it had good general performance. The next step in the construction of SVM model was optimizing of its parameters, including ε and C. The optimization of SVM parameters was performed by systemically changing their values in the training step and calculating the RMSE of the model. The optimal value for ε depended 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 ε, there will be some practical consideration of the number of resulting support vectors. ε-Insensitivity prevents the entire training set to meeting boundary conditions and allows the possibility of sparsity in the dual formulations solution. So, choosing the appropriate value of ε is a critical step. To find an optimal value for ε, the RMSE of SVM models with different ε values was calculated. The variation of RMSE versus the epsilon values is plotted in Figure 2. As it is shown in this figure, the optimal value of ε was 0.04. The other parameter is a regularization parameter C that controls the trade-off between maximizing the margin and minimizing the training error. If C is too small, then insufficient stress will be placed on fitting the training data. On the other hand if C is too large, then the SVM model will overfit on the training data. To find an optimal value of C, the RMSE of SVM models with different C values was calculated. The obtained results revealed that the variation of RMSE in $C > 20$ is small, therefore this value ($C = 20$) was selected as the optimal value of C. Also for inspection of any interactions between C and epsilon, after optimization of the C value, the epsilon value was varied. The results indicated that the value of optimized epsilon was not varied in this stage, which concludes that they are independent from each other.

After optimizing SVM parameters, it was used to calculate the $\log(1/EC_{50})$ of training and test set. The statistical parameters of this model are RMSE = 0.181, $R = 0.950$, and $F = 290$ for the training set, and RMSE = 0.241, $R = 0.924$, and $F = 41$ for the test set. Also the leave-5-out cross validation test was carried out for the evaluation of the prediction power of

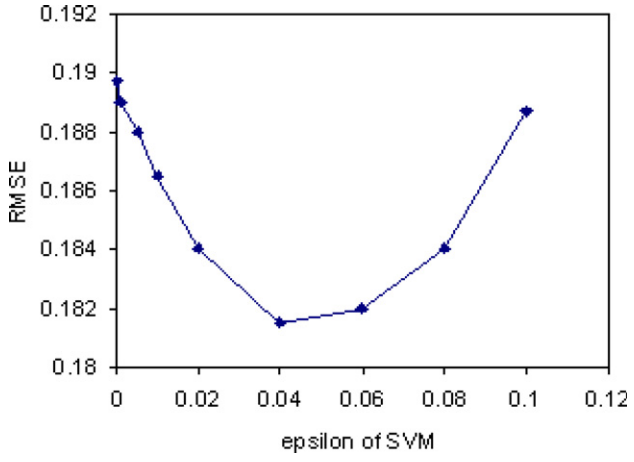

Figure 2. The variation of RMSE versus epsilon ($C = 100$, $d = 2$).

obtained SVM model. The calculated values of Q^2 and SPRESS for cross validation test were 0.710 and 0.345, respectively. The comparison between these values and those obtained by Afantitis et al.[21] ($Q^2 = 0.678$) revealed the superiority of obtained GA–SVM model. In addition, it was noteworthy that in the QSAR model presented by Afantitis et al.,[21] one compound was considered as outlier and rejected during modeling process (No. **2b**), while in the present work all of **43** 4-aryl-4-H-chromenes were considered in modeling.

The predicted GA–SVM values of $\log(1/EC_{50})$ were plotted versus their experimental values in Figure 3, and their residuals were plotted in Figure 4. The random distribution of the residuals around the zero line indicated that there were not any systematic errors in this model. Based on the residuals values (Table 1 and Fig. 4) it was found that the compound No. **2b** is the worst predicted compound (residual = 0.421). This compound has two adjacent OH groups which can make an intermolecular hydrogen bonding that can affect its activity. In contrast, other molecules in data set cannot formed intermolecular hydrogen bonding. Since the predicted value of apoptosis-inducing activity of this compound in the previous work had large residual (0.892), this compound was omitted from their investigation.

In a benchmark test, the results of support vector regression model were compared with several modeling techniques currently used in this field, such as artificial neural network and multiple linear regression. Table 2 shows the statistical parameter of some models constructed by these feature mapping methods and used GA and stepwise multiple linear regression techniques for variable subset selection. In this table, it can be seen that the RMSEs of the GA–SVM model for the training and test set (0.181 and 0.241, respectively) are lower than that those of constructed ANN and the MLR models as well as the previous work. Also through a regression analysis on the results obtained for the prediction of apoptosis-inducing activity by various methods, the values of correlation coefficients (R) and statistical F-value were calculated and shown in Table 2. Comparison between these parameters revealed that the values of R and F for GA–SVM model were higher than the other models for both training and test sets, which revealed the superiority of GA–SVM over other investigated models, as well as previous work.

4. Conclusion

In the present study, two linear and nonlinear variable selection methods were used to select the most significant descriptors, and the multiple linear regression, artificial neural network and the support vector machine were used to construct a quantitative relation between the apoptosis-induction activities of 4-aryl-4-H-chromenes and their calculated descriptors. The obtained results demonstrated that the GA–SVM models produced better results with good predictive ability than

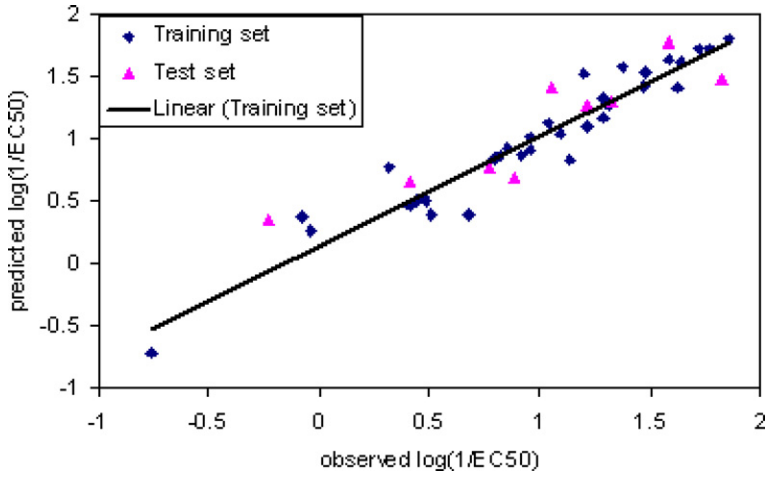

Figure 3. Plot of predicted log(1/EC_{50}) by GA–SVM model versus experimental values.

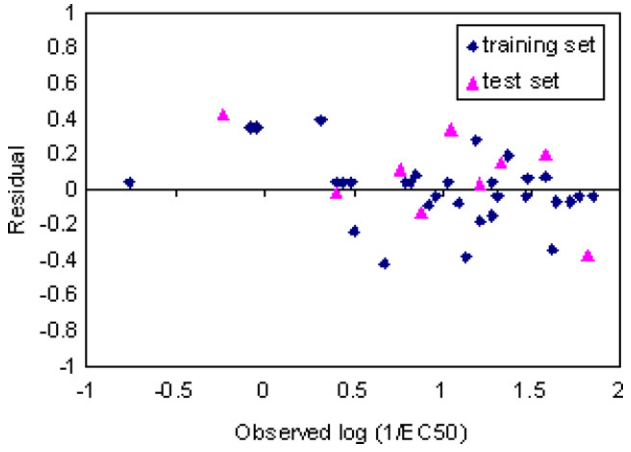

Figure 4. Plot of the residuals versus experimental log(1/EC_{50}).

other methods. These results also indicated that the constructed GA–SVM model predicted the values of log(1/EC_{50}) more accurately than those reported previously by Afantitis et al.[21]

Therefore it was concluded that (1) SVM proved to be a useful tool in the prediction of the drugs activity, (2) nonlinear relationship can describe the relationship between the structural parameters and the apoptosis-induction activities of 4-aryl-4-H-chromenes accurately, (3) the proposed models could identify and provide some insight into what structural features are related to the apoptosis-induction activities of 4-aryl-4-H-chromenes.

References and notes

1. Kerr, J. F.; Wyllie, A. H.; Currie, A. R. *Br. J. Cancer* **1972**, *26*, 239.
2. Lockshin, R. A.; Zakeri, Z.; Tilly, J. L. *When Cells Die*; Wiley-Liss: New York, 1998.
3. Coultas, L.; Strasser, A. *Apoptosis* **2000**, *5*, 491.
4. Ashkenazi, A.; Dixit, V. M. *Science* **1998**, *281*, 1305.
5. Nagata, S. *Cell* **1997**, *88*, 355.
6. Schulze-Osthoff, K.; Ferrari, D.; Los, M.; Wesselborg, S.; Peter, M. *Eur. J. Biochem.* **1998**, *254*, 439.
7. Herr, L.; Debatin, K. M. *Blood* **2001**, *98*, 2603.
8. Rich, T.; Allen, R. L.; Wyllie, A. H. *Nature* **2000**, *407*, 777.
9. Mosier, P. D.; Jurs, P. C.; Custer, L. L.; Durham, S. K.; Pearl, G. M. *Chem. Res. Toxicol.* **2003**, *16*, 721.
10. McElroy, N. R.; Thompson, E. D.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2111.
11. Mattioni, B. E.; Kauffman, G. W.; Jurs, P. C.; Custer, L. L.; Durham, S. K.; Pearl, G. M. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 949.
12. Cronin, M. T. *Curr. Opin. Drug Discov. Dev.* **2000**, *3*, 292.
13. Yoshida, F.; Topliss, J. G. *J. Med. Chem.* **2000**, *43*, 2575.
14. Hemmateenejad, B.; Safarpour, M. A.; Miri, R.; Nesari, N. *J. Chem. Inf. Model* **2005**, *45*, 190.
15. Novak, M.; Rajagopal, S. *Chem. Res. Toxicol.* **2002**, *15*, 1495.
16. Kazius, J.; McGuire, R.; Bursi, R. *J. Med. Chem.* **2005**, *48*, 312.
17. Hansch, C.; Bonavida, B.; Jazirehi, A.; Cohen, J.; Milliron, C.; Kurup, A. *Bioorg. Med. Chem.* **2003**, *11*, 617.
18. Hansch, C.; Jazirehi, A.; Mekapati, S.; Garg, R.; Bonavida, B. *Bioorg. Med. Chem.* **2003**, *11*, 3015.
19. Selassie, C. D.; Kapur, S.; Verma, R. P.; Rosario, M. *J. Med. Chem.* **2005**, *48*, 7234.
20. Kemnitzer, W.; Kasibhatla, Sh.; Jiang, S.; Zhang, H.; Zhao, j.; Jia, Sh.; Xu, L.; Corgan-Grundy, C.; Denis, R.; Barriault, N.; Vaillancourt, L.; Charron, S.; Dodd, J.; Attardo, G.; Labrecque, D.; Lamothe, S.; Gourdeau, H.; Tseng, B.; Drewe, J.; Cai, S. X. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 4745.
21. Afantitis, A.; Melagraki, G.; Sarimveis, H.; Koutentis, P. A.; Markopoulos, J.; Igglessi-Markopoulou, O. *Bioorg. Med. Chem.* **2006**, *14*, 6686.
22. Fatemi, M. H. *Anal. Chim. Acta* **2006**, *556*, 355.
23. Byvatov, E.; Fechner, U.; Sadowski, J.; Schneider, G. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1882.
24. Niani, Ch.; Wencong, L.; Jie, Y.; Gozheng, L. *Support Vector Machine in Chemistry*; World Scientific Publishing Co. Pet. Ltd, 2004.
25. Liu, H. X.; Zhang, R. S.; Luan, F.; Yao, X. J.; Liu, M. C.; Hu, Z. D.; Fan, B. T. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 900.
26. Burbidge, R.; Trotter, M.; Buxton, B.; Holden, S. *Comput. Chem.* **2001**, *26*, 5.
27. Liu, H. X.; Zhang, R. S.; Yao, X. J.; Liu, M. C.; Hu, Z. D.; Fan, B. T. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1288.

28. Liu, H. X.; Zhang, R. s.; Yao, X. J.; Liu, M. C.; Hu, Z. D.; Fan, B. T. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 161.
29. http://www.disat.unimib.it%252Fchm%252F.
30. HyperChem, Release 7.0 for windows, Hypercube, Inc., 2002.
31. Stewart, J. J. P. MOPAC, Semi Empirical Molecular Orbital Program, QCPE, 455, 1983. Research Laboratory, United States Air Force Academy, version 6. **1990**.
32. http://www.models.kvl.dk/source/GAPLS/index.asp.
33. Cortes, C.; Vapnik, V. *Machine Learn.* **1995**, *20*, 231.
34. Burges, C. J. C. http://svm.research.bell-labs.com/SVM-doc.html, **1998**.
35. Vapnik, V.; Golowich, S.; Smola, A. *Adv. Neural Inform. Process. Systems* **1997**, *9*, 281.
36. http://www.isis.ecs.soton.ac.uk/isystems/kernel/.
37. Gunn, S. R. *Support Vector Machines for Classification and Regression*; University of Southampton: UK, 1997.
38. Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, Germany, 2000.
39. Consonni, V.; Todeschini, R.; Pavan, M. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 682.
40. Consonni, V.; Todeschini, R.; Pavan, M. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 693.
41. Todeschini, R.; Gramatica, P. *Quant. Struct. Act. Rel.* **1997**, *16*, 113.
42. Voelkel, A. *Computers Chem.* **1994**, *18*, 1.
43. Basak, S. C.; Balaban, A. T.; Grunwald, G. D.; Gute, B. D. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 891.