Available online at www.sciencedirect.com
ScienceDirect
Bioorganic & Medicinal Chemistry 14 (2006) 6686–6694
Bioorganic & Medicinal Chemistry

A novel QSAR model for predicting induction of apoptosis by 4-aryl-4*H*-chromenes

Antreas Afantitis,[a,b] Georgia Melagraki,[a] Haralambos Sarimveis,[a,*] Panayiotis A. Koutentis,[c] John Markopoulos[d] and Olga Igglessi-Markopoulou[a]

[a]*School of Chemical Engineering, National Technical University of Athens, Athens, Greece*
[b]*Department of ChemoInformatics, NovaMechanics Ltd, Cyprus*
[c]*Department of Chemistry, University of Cyprus, PO Box 20537, 1678 Nicosia, Cyprus*
[d]*Department of Chemistry, University of Athens, Athens, Greece*

Received 3 February 2006; revised 25 May 2006; accepted 31 May 2006
Available online 16 June 2006

Abstract—A linear quantitative structure–activity relationship (QSAR) model is presented for modeling and predicting induction of apoptosis by 4-aryl-4*H*-chromenes. The model was produced by using the multiple linear regression (MLR) technique on a database that consists of 43 recently discovered 4-aryl-4*H*-chromenes. Among the 61 different physicochemical, topological, and structural descriptors that were considered as inputs to the model, seven variables were selected using the elimination selection-stepwise regression method (ES-SWR). The physical meaning of each descriptor is discussed. The accuracy of the proposed MLR model is illustrated using the following evaluation techniques: cross-validation, validation through an external test set, and Y-randomization. Furthermore, the domain of applicability which indicates the area of reliable predictions is defined.

1. Introduction

Novel medicines are typically developed using a trial and error approach which is time-consuming and costly. The application of quantitative–structure activity relationship (QSAR) methodologies to this problem has the potential to decrease substantially the time and effort required to discover new medicines or improve current ones in terms of their efficacy. QSARs establish mathematical relationships between physical, chemical, biological, or environmental activities of interest and measurable or computable parameters such as topological, physicochemical, stereo chemical or electronic indices.[1–6]

Apoptosis is the vital process by which cells undergo 'programmed cell death' in various biological systems. Diverse groups of molecules are involved in the apoptosis pathway. One set of mediators implicated in apoptosis belongs to the aspartate-specific cysteinyl proteases or caspases.[7–9] Caspases, which are present in all cells as latent enzymes, are a family of proteases that relay a ''doomsday'' signal in a step-wise manner reminiscent of signaling by kinases. Excessive apoptosis is responsible, at least in part, for a variety of diseases, for example, liver disease,[10] brain ischemia,[11] myocardial infraction,[12] Huntington's disease, and Alzheimer's disease.[13]

Recent reports indicate that many clinically useful cytotoxic agents induce apoptosis in cancer cells.[14,15] Compounds that induce apoptosis in cancer cells by targeting the clinically validated tubulin/microtubule system, while retaining activity in multi-drug-resistant tumors, have the potential to offer new treatment options in the field of oncology.[16] The 4-aryl-4*H*-chromenes were found[16] to a be promising series of novel apoptosis inducers that could be developed into new therapeutic anticancer agents.

To our knowledge only three attempts have been made to build QSAR models in the general field of apoptosis. Hansch[17] presented a QSAR study containing a variety of phenolic compounds causing apoptosis and later the same scientific group presented a QSAR of apoptosis

Keywords: Apoptosis; Chromenes; Molecular modeling; QSAR.
* Corresponding author. Tel.: +30 210 772 3237; fax: +30 210 772 3138; e-mail: hsarimv@central.ntua.gr

0968-0896/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2006.05.061

Nomenclature

F	F ratio
h_i	leverage for the ith compound
k	number of descriptors
LOO	leave-one-out
L5O	leave-five-out
n	number of compounds
PRESS	prediction error sum of squares
R^2	correlation coefficient (coefficient of multiple determination)
Q^2	correlation coefficient for cross-validation
$R^2_{\text{cv,ext}}$	external correlation coefficient
RMS	root mean squared error
S_{PRESS}	root mean squared error for cross-validation
SSY	sum of squares of deviations of the experimental values from their mean
x_{i}	the descriptor-row vector for the ith compound
X	the $k \times n$ matrix containing the k descriptor values for each one of the n training compounds
$y_{\text{exp},i}$	experimental output value for the ith compound
$y_{\text{pred},i}$	predicted output value for the ith compound
$\bar{y}$	average value for the output variable

induction in various cancer cells.[18] Selassie et al.[19] investigated apoptosis-inducing effect of 51 substituted caspase-mediated phenols in a murine leukemia cell line (L1210). After a QSAR analysis, they found that the strong dependence of caspase-mediated apoptosis on mostly steric parameters suggests that the process is a receptor-mediated interaction with caspases or mitochondrial proteins being the likely targets.

In this work, a series of 43 4-aryl-4*H*-chromenes[16] with apoptotic activity was studied. Sixty-one physicochemical and topological descriptors were examined in terms of their efficacy to determine and predict the activity of the investigated derivatives. The descriptors were calculated using Topix (www.lohninger.com/topix.html) and ChemSar which is included in the ChemOffice (CambridgeSoft Corporation) suite of programs. Among them, the most statistically significant descriptors were selected using the Elimination Selection-Stepwise Regression (ES-SWR) variable selection method. The result of this study was the development of a new linear QSAR model containing 7 variables. The proposed methodology was validated using several strategies: cross-validation, Y-randomization, and external validation using division of the entire data set into training and test sets. Furthermore, the domain of applicability which indicates the area of reliable predictions was defined.

2. Materials and methods

2.1. Data set

In this QSAR study, 43 biological data from the work of Kemnitzer et al. [16] work were used. The biological activities of these 43 compounds were reported in the same paper.[16] The compounds are shown in Table 1, where the letters *a,b,c,d,e* in the first column correspond to the basic structures of 4-aryl-4*H*-chromenes, depicted in Figure 1. In order to model and predict the specific activity (apoptosis induction), 61 physicochemical constants, topological and structural descriptors (Table 2) were considered as possible input candidates to the model. All the descriptors were calculated using ChemSar and Topix. Before the calculation of the descriptors, the structures were fully optimized using CS Mechanics and more specifically MM2 force fields and the Truncated-Newton-Raphson optimizer, which provide a balance between speed and accuracy (Chemoffice Manual).

2.2. Stepwise multiple regression

As mentioned in the introduction, the ES-SWR algorithm[20] was used to select the most appropriate descriptors. ES-SWR is a popular stepwise technique that combines forward selection (FS-SWR) and backward elimination (BE-SWR). It is essentially a forward selection approach, but at each step it considers the possibility of deleting a variable as in the backward elimination approach, provided that the number of model variables is greater than two.

2.3. Kennard and Stones algorithm

The Kennard and Stones algorithm[21] has gained increasing popularity for splitting data sets into two subsets. The algorithm starts by finding 2 samples that are the farthest apart from each other on the basis of the input variables in terms of some metric, for example, the Euclidean distance. These 2 samples are removed from the original data set and put into the calibration data set. This procedure is repeated until the desired number of samples has been reached in the calibration set. The advantages of this algorithm are that the calibration samples map the measured region of the input variable space completely with respect to the induced metric and that the test samples all fall inside the measured region. According to Tropsha[22] and Wu,[23] the Kennard and Stones algorithm is one of the best ways to build training and test sets.

2.4. Cross-validation technique

Cross-validation is a popular technique used to explore the reliability of statistical models. Based on this technique, a number of modified data sets are created by

Table 1. Apoptosis inducing activity of 4-aryl-4*H*-chromenes in human breast cancer cells T47D

Compound	R^5	R^6	R^7	R^8	A	R′	R″	R‴	EC_{50} (μM) (observed)	$\log(1/EC_{50})$ (observed)	$\log(1/EC_{50})$ (predicted)	St. Res.
1a	H	H	NMe_2	H	—	—	—	—	0.073	1.137	1.0957	0.15
2a	H	H	NH_2	H	—	—	—	—	1.2	−0.079	0.0593	−0.49
3a	H	H	NHEt	H	—	—	—	—	0.33	0.481	0.5424	−0.22
4a	H	H	NEt_2	H	—	—	—	—	0.48	0.319	0.0763	1
5a	H	H	OMe	H	—	—	—	—	0.16	0.796	0.4845	1.08
6a	H	H	OH	H	—	—	—	—	5.8	−0.763	−0.4861	−1.05
7a	H	Me	NHEt	H	—	—	—	—	1.1	−0.041	0.4513	−1.73
8a	H	OCH_2O		H	—	—	—	—	0.21	0.678	0.412	1.02
9a	H	H	NH_2	Me	—	—	—	—	0.31	0.509	0.7245	−0.78
10a	H	H	NMe_2	H	—	—	—	—	0.019	1.721	1.5656	0.56
11b	H	H	NH_2	H	—	—	—	—	0.033	1.481	1.3957	0.29
12b	H	H	NHEt	H	—	—	—	—	0.014	1.854	1.4581	1.42
13b	H	H	OMe	H	—	—	—	—	0.017	1.769	1.7576	0.04
14b	H	H	OEt	H	—	—	—	—	0.064	1.194	1.3757	−0.63
15b	H	H	OH	H	—	—	—	—	0.13	0.886	1.2394	−1.19
16b	H	H	Br	H	—	—	—	—	0.14	0.854	0.5808	1.05
17b	H	H	Cl	H	—	—	—	—	0.16	0.796	1.0936	−1.16
18b	H	H	NH_2	NH_2	—	—	—	—	0.034	1.468	1.0719	1.39
19b	H	H	NH_2	Me	—	—	—	—	0.026	1.585	1.1786	1.4
20b	H	H	Me	Me	—	—	—	—	0.042	1.377	1.2653	0.47
21b	H	H	OH	NH_2	—	—	—	—	0.061	1.215	0.9701	0.85
22b	H	H	OH	OH	—	—	—	—	1.7	−0.230	0.6617	−3.12
23c	—	—	—	—	C	OMe	OMe	OMe	0.026	1.585	1.5246	0.22
24c	—	—	—	—	C	OMe	H	OMe	0.015	1.824	1.4017	1.48
25c	—	—	—	—	C	OMe	H	H	0.052	1.284	1.0919	0.68
26c	—	—	—	—	C	Br	H	H	0.052	1.284	1.4256	−0.5
27c	—	—	—	—	C	Cl	H	H	0.08	1.097	1.3418	−0.86
28c	—	—	—	—	C	NO_2	H	H	0.089	1.051	1.1589	−0.44
29c	—	—	—	—	C	H	H	H	0.36	0.444	0.3342	0.41
30c	—	—	—	—	N	H	H	H	0.17	0.769	0.5477	0.82
31c	—	—	—	—	N	OMe	H	H	0.047	1.328	0.9788	1.22
32d	—	—	—	—	C	OMe	OMe	OMe	0.049	1.310	1.5225	−0.76
33d	—	—	—	—	C	OMe	H	OMe	0.055	1.210	1.4817	−0.78
34d	—	—	—	—	C	OMe	H	H	0.11	0.959	1.2023	−0.85
35d	—	—	—	—	C	Cl	H	H	0.12	0.921	1.3373	−1.49
36d	—	—	—	—	C	NO_2	H	H	0.11	0.959	0.8805	0.32
37e	—	—	—	—	C	Cl	OMe	OMe	0.024	1.620	1.5488	0.25
38e	—	—	—	—	C	I	OMe	OMe	0.049	1.310	1.4411	−0.46
39e	—	—	—	—	C	Br	OH	OMe	0.023	1.638	1.7399	−0.38
40e	—	—	—	—	C	OMe	H	OMe	0.092	1.036	0.9424	0.33
41e	—	—	—	—	C	CN	H	H	0.39	0.409	0.5614	−0.58
42e	—	—	—	—	C	Br	H	H	0.15	0.824	0.7236	0.37
43e	—	—	—	—	C	NO_2	H	H	0.39	0.409	0.1638	1.04

Model predictions using Eq. 9.

deleting in each case one or a small group (leave-some-out) of objects.[24–26] For each data set, an input–output model is developed, based on the utilized modeling technique. The model is evaluated by measuring its accuracy in predicting the responses of the remaining data (the ones that have not been utilized in the development of the model). In particular, the leave-one-out (LOO) and the leave-five-out (L5O) procedures were utilized in this study, which produce a number of models, by deleting one or five objects, respectively, from the training set. The maximum number of models produced by the LOO procedure is equal to the number of available examples n, while for the L5O procedure the maximum number of models is equal to $\frac{n!}{5!(n-5)!}$. Prediction error sum of squares (PRESS) is a standard index to measure the accuracy of a modeling method based on the cross-validation technique. Based on the PRESS and SSY (sum of squares of deviations of the experimental values from their mean) statistics, the Q^2 and S_{PRESS} values can be easily calculated. The formulae used to calculate the aforementioned statistics are presented below:

$$Q^2 = 1 - \frac{\mathrm{PRESS}}{\mathrm{SSY}} = 1 - \frac{\sum_{i=1}^{n}(y_{\mathrm{exp},i} - y_{\mathrm{pred},i})^2}{\sum_{i=1}^{n}(y_{\mathrm{exp},i} - \bar{y})^2} \quad (1)$$

$$S_{\mathrm{PRESS}} = \sqrt{\frac{\mathrm{PRESS}}{n}} \quad (2)$$

2.5. Y-randomization test

This technique ensures the robustness of a QSAR model.[22,27] The dependent variable vector [$\log(1/EC_{50})$] is randomly shuffled and a new QSAR model is developed

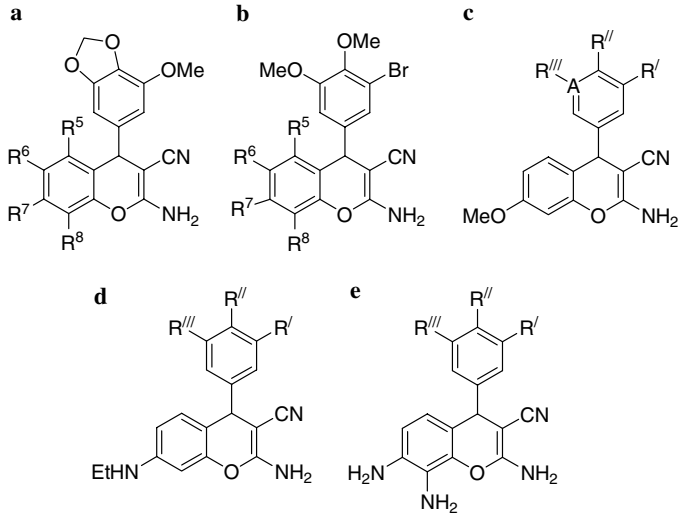

Figure 1. Structures of 4-aryl-4*H*-chromenes.

Table 2. Physicochemical constants, topological and structural descriptors

ID	Description	Notation	ID	Description	Notation
1	Molar refractivity	MR	**2**	Diameter	Diam
3	Partition coefficient (Octanol Water)	C log *P*	**4**	Molecular topological index	TIndx
5	Principal moment of inertia Z	PMIZ	**6**	Number of rotatable bonds	NRBo
7	Principal moment of inertia Y	PMIY	**8**	Polar surface area	PSAr
9	Principal moment of inertia X	PMIX	**10**	Radius	Rad
11	Connolly accessible area	SAS	**12**	Shape attribute	ShpA
13	Connolly molecular area	MS	**14**	Shape coefficient	ShpC
15	Total energy	TotE	**16**	Sum of valence degrees	SVDe
17	LUMO energy	LUMO	**18**	Total connectivity	Tcon
19	HOMO energy	HOMO	**20**	Total valence connectivity	TVCon
21	Balaban Index	BIndx	**22**	Wiener index	Windx
23	Cluster count	ClsC	**24**	Randic 0	Chi0
25	Randic 1	Chi1	**26**	Randic 2	Chi2
27	Randic 3	Chi3	**28**	Randic 4	Chi4
29	Randic information 0	ChiInf0	**30**	Randic information 1	ChiInf1
31	Randic information 2	ChiInf2	**32**	Randic information 3	ChiInf3
33	Randic information 4	ChiInf4	**34**	Kier-Hall 0	Ki0
35	Randic Mod	ChiMod	**36**	Xu1	Xu1
37	Xu2	Xu2	**38**	Xu3	Xu3
39	Balaban Topological	TopoJ	**40**	Topological radius	TopoRad
41	Topological diameter	TopoDia	**42**	Number of clusters	NClusters
43	Number of rings	NRings	**44**	Wiener Dim	Wiener Dim
45	Bertz	Bertz	**46**	AtomCompMean	AtomCompMean
47	AtomCompTot	AtomCompTot	**48**	Zagreb1	Zagreb1
49	Zagreb2	Zagreb2	**50**	Quadratic	Quadr
51	ScHultz	ScHultz	**52**	Kappa1	Kappa1
53	Kappa3	Kappa3	**54**	Kappa2	Kappa2
55	Wiener Distance	WienerDistCode	**56**	Wiener Information	InfWiener
57	DistEqMean	DistEqMean	**58**	DistEqTotal	DistEqTotal
59	InfMagnitDistTot	InfMagnitDistTot	**60**	Polarity	Polarity
61	Gordon	Gordon			

using the original independent variable matrix. The new QSAR models (after several repetitions) are expected to have low R^2 and Q^2 values. If the opposite happens then an acceptable QSAR model cannot be obtained for the specific modeling method and data.

2.6. Estimation of the predictive ability of a QSAR model

According to Tropsha[22] the predictive power of a QSAR model can be conveniently estimated by an external $R^2_{\text{cv,ext}}$

$$R^2_{\text{cv,ext}} = 1 - \frac{\sum_{i=1}^{\text{test}} (y_{\text{exp},i} - y_{\text{pred},i})^2}{\sum_{i=1}^{\text{test}} (y_{\text{exp},i} - \bar{y}_{\text{tr}})^2} \quad (3)$$

where $\bar{y}_{\text{tr}}$ is the averaged value for the dependent variable for the training set. Furthermore the same group[22,28] considered a QSAR model predictive, if the following conditions are satisfied:

$$R^2_{\mathrm{cv,ext}} > 0.5 \tag{4}$$

$$R^2 > 0.6 \tag{5}$$

$$\frac{(R^2 - R^2_{\mathrm{o}})}{R^2} < 0.1 \text{ or } \frac{(R^2 - R'^2_{\mathrm{o}})}{R^2} < 0.1 \tag{6}$$

$$0.85 \leqslant k \leqslant 1.15 \text{ or } 0.85 \leqslant k' \leqslant 1.15 \tag{7}$$

Mathematical definitions of $R^2_{\mathrm{o}}, R'^2_{\mathrm{o}}$, k, and k' are based on regression of the observed activities against predicted activities and vice versa (regression of the predicted activities against observed activities). The definitions are presented clearly in Golbraikh et al.[28] and are not repeated here for brevity.

2.7. Defining model applicability domain

The domain of application[22,29] of a QSAR model must be defined if the model is to be used for screening new compounds. Predictions for only those compounds that fall into this domain may be considered reliable. *Extent of Extrapolation*[29] is one simple approach to define the applicability of the domain. It is based on the calculation of the leverage[30] h_i for each chemical, for which QSAR model is used to predict its activity:

$$h_i = x_i^{\mathrm{T}}(X^{\mathrm{T}}X)x_i \tag{8}$$

In Eq. 8 x_i is the descriptor-row vector of the query compound and X is the $k \times n$ matrix containing the k descriptor values for each one of the n training compounds. A leverage value greater than $3k/n$ is considered large and implies that the predicted response is the result of a substantial extrapolation of the model and may not be reliable.

3. Results and discussion

For the selection of the most important descriptors, the aforementioned stepwise multiple regression technique was used. The seven most significant descriptors according to the ES-SWR algorithm are: the number of clusters (NClusters), the Bertz's complexity index (Bertz), the quadratic (Quadr), the dipole length (DPLL), the LUMO energy (LUMO), the polar surface area (PSAr), and the sum of valence degrees (SVDe).

The linear equation that models the apoptosis-inducing activity of the 4-aryl-4*H*-chromenes in human breast cancer cells T47D and corresponds to the aforementioned seven most significant descriptors is the following:

$$\begin{aligned}\log(1/\mathrm{EC_{50}T47D}) = &-17.6(\pm4.80)\\ &-0.534(\pm0.25)\mathrm{NClusters}\\ &+0.229(\pm0.06)\mathrm{Bertz}\\ &-0.00185(\pm0.0005)\mathrm{Quadr}\\ &+0.294(\pm0.079)\mathrm{DPLL}\\ &+2.31(\pm0.57)\mathrm{LUMO}\\ &-0.0654(\pm0.015)\mathrm{PSAr}\\ &+0.340(\pm0.089)\mathrm{SVDe}\\ \mathrm{RMS} = 0.276,\quad &R^2 = 0.772,\quad F = 16.95,\\ Q^2 = 0.668,\quad &S_{\mathrm{PRESS}} = 0.333,\quad n = 43\end{aligned} \tag{9}$$

The possibility of having included outliers in our data set was investigated by calculating the standard residuals (Table 2). Standardized residuals greater than 2 and less than −2 are usually considered large. Compound with id **22b** has standardized residual −3.12 and can safely be excluded from the data set as outlier. The new linear equation after the rejection of compound **22b** has a better predictive ability and is the following:

$$\begin{aligned}\log(1/\mathrm{EC_{50}T47D}) = &-17.6(\pm4.14)\\ &-0.521(\pm0.22)\mathrm{NClusters}\\ &+0.234(\pm0.053)\mathrm{Bertz}\\ &-0.00186(\pm0.0004)\mathrm{Quadr}\\ &+0.267(\pm0.069)\mathrm{DPLL}\\ &+2.22(\pm0.50)\mathrm{LUMO}\\ &-0.0630(\pm0.013)\mathrm{PSAr}\\ &+0.337(\pm0.077)\mathrm{SVDe}\\ \mathrm{RMS} = 0.237,\quad &R^2 = 0.816,\quad F = 21.60,\\ Q^2 = 0.718,\quad &S_{\mathrm{PRESS}} = 0.294,\quad n = 42\end{aligned} \tag{10}$$

Table 3 presents the correlation matrix, where it is clear that the seven selected descriptors are not highly correlated.

Table 3. Correlation matrix for the seven selected descriptors

	NClusters	Bertz	Quadr	DPLL	LUMO	PSAr	SVDe
NClusters	1						
Bertz	0.856	1					
Quadr	0.736	0.556	1				
DPLL	0.109	0.217	−0.220	1			
LUMO	0.109	−0.063	0.190	−0.455	1		
PSAr	0.609	0.510	0.313	0.457	−0.144	1	
SVDe	0.788	0.582	0.906	−0.028	−0.033	0.621	1

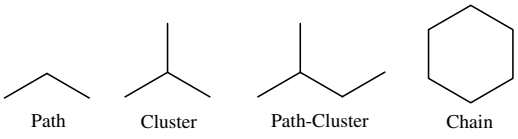

Figure 2. Elementary molecular subgraphs.

A brief explanation of the seven descriptors that were utilized in the produced QSAR model follows next:

Number of clusters (NClusters), the total number of clusters in the module, is one of the four commonly used subgraph types (Fig. 2).[20] It is an attractive descriptor due to its fast and easy calculation.

Bertz's complexity index,[20] the most popular complexity index, takes into account both the variety of kinds of bond connectivities and atom types of H-depleted molecular graph.

Quadratic index[20] is calculated by normalization of the 1st Zagreb index[20] which is based on the vertex degree of the atoms in the H-depleted molecular graph.

Dipole Length[20] (DPLL) is the electric dipole moment divided by the elementary charge. Electric dipole is a vector quantity which encodes displacement with respect to the center of gravity of positive and negative charges in a molecule.

Molecular orbital (MO) surfaces visually represent the various stable electron distributions of a molecule. According to Frontier Orbital Theory, the shapes and symmetries of the highest-occupied and lowest-unoccupied molecular orbitals (HOMO and LUMO) are crucial in predicting the reactivity of a species and the stereochemical and regiochemical outcome of a chemical reaction.

Polar surface area (PSAr)[20] is defined as the part of the surface area of the module associated with oxygens, nitrogens, sulfurs, and the hydrogens bonded to any of these atoms.

Sum of valence vertex degrees (SVDe) is used in order to take into account all valence electrons of the atoms. SVDe is the sum of all δ^v values in a module as is defined by Kier and Hall.[31] δ^v encodes the electronic identity of the atom in terms of both valence electron and core electron counts. It is useful for characterizing heteroatoms and carbon atoms involved in multiple bonds. Different groups that can be used as possible substituents with the respective valence vertex degrees (δ^v) are the following: C_{sp3}, C_{sp2}, and C_{sp} with $\delta^v = 4$, N_{sp3}, N_{sp2}, and N_{sp} with $\delta^v = 5$, O_{sp3}, O_{sp2}, S_{sp3}, and S_{sp2} with $\delta^v = 6$, F, Cl, Br, and I with $\delta^v = 7$.

According to the produced QSAR equation (Eq. 11) a high value of the number of clusters, quadratic index, and polar surface area contributes negatively to the activity. Thus, designing models with fewer or no clusters in the H-depleted molecular graph should improve activity (Fig. 2). With the elimination of large substituents such as Phenyl and *N*-morpholino the quadratic index is reduced and the activity is increased. This remark agrees with the work of Kemnitzer et al.[17] which clearly indicated that small hydrophobic groups are preferred. Polar surface area (PSAr) is related to the hydrogen-bonding ability of the compounds. The presence of oxygens, nitrogens, sulfurs, and the hydrogens bonded to any of these atoms increases PSAr value.

On the other hand, a high value of the Bertz's complexity index, LUMO energy, dipole length, and the sum of valence vertex degrees gives a positive contribution to the activity. Bertz's complexity index is the sum of I_{CPB} and I_{CPA} which are the information contents related to the bond connectivity and the atom-type diversity. Molecular complexity increases with size, branching, vertex, and weights. The term I_{CPB} measures the complexity of a molecule given by the partition of equivalent connections sensitive to branching, rings, and multiple bonds of the module. The atom complexity term I_{CPA} takes into account the presence of heteroatoms in a molecule.

Molecules with low LUMO energy values are more able to accept electrons than molecules with high LUMO energy values.[20] The LUMO energy value is increased

Table 4. Model predictions for 4-aryl-4*H*-chromenes using Eq. 11

Compound	EC_{50} (μM) (observed)	$\log(1/EC_{50})$ (observed)	Training data $\log(1/EC_{50})$ (predicted)	Validation data $\log(1/EC_{50})$ (predicted)
1a	0.073	1.137	1.122	
2a[a]	1.2	−0.079	—	0.179
3a	0.33	0.481	0.551	—
4a	0.48	0.319	0.068	—
5a	0.16	0.796	0.600	—
6a	5.8	−0.763	−0.272	—
7a[a]	1.1	−0.041	—	0.507
8a	0.21	0.678	0.611	—
9a	0.31	0.509	0.877	—
10a[a]	0.019	1.721	—	1.514
11b	0.033	1.481	1.416	—
12b[a]	0.014	1.854	—	1.373
13b	0.017	1.769	1.724	—
14b	0.064	1.194	1.316	—
15b	0.13	0.886	1.294	—
16b	0.14	0.854	0.664	—
17b	0.16	0.796	1.142	—
18b	0.034	1.468	1.124	—
19b	0.026	1.585	1.241	—
20b	0.042	1.377	1.273	—
21b	0.061	1.215	1.089	—
22b	1.7	−0.230	—	—
23c	0.026	1.585	1.556	—
24c	0.015	1.824	1.423	—
25c	0.052	1.284	1.122	—
26c	0.052	1.284	1.381	—
27c[a]	0.08	1.097	—	1.302
28c	0.089	1.051	1.182	—
29c	0.36	0.444	0.363	—
30c	0.17	0.769	0.558	—
31c	0.047	1.328	1.010	—
32d	0.049	1.310	1.454	—
33d	0.055	1.210	1.409	—
34d	0.11	0.959	1.140	—
35d	0.12	0.921	1.219	—
36d	0.11	0.959	0.843	—
37e	0.024	1.620	1.602	—
38e	0.049	1.310	1.496	—
39e[a]	0.023	1.638	—	1.796
40e	0.092	1.036	1.035	—
41e	0.39	0.409	0.612	—
42e[a]	0.15	0.824	—	0.7611
43e	0.39	0.409	0.292	—

[a] Validation set.

with the presence of electron donating groups (EDGs). This remark also agrees with Kemnitzer et al.[17] who recommend the introduction of EDGs such us NMe_2, NH_2, NHEt, and OMe.

Dipole length encodes information about the charge distribution in molecules and is important for modeling polar interactions.[20] Large substituents decrease DPLL value which is not desirable.

Eq. 11 indicates that a high value of the SVDe increases the activity. Halogens have the largest δ^v values compared to other groups. However, with the addition of halogens in the module the activity is reduced, although SVDe increases. This can be explained by noticing that halogens are inductively electron-withdrawing groups (EWGs) and thus lower the LUMO energy. The optimal solution is to use as possible substituents the following: NMe_2, NH_2, NHEt, and OMe. These groups have a large δ^v value and are EDGs.

The predictive ability of the selected descriptors was further explored, by dividing the full data set consisting of 42 4-aryl-4*H*-chromenes into a training set of 35 compounds, and a validation set of 7 compounds. The selection of the combinations in the training set was made according to the Kennard and Stones algorithm.

The combinations that constituted the training and validation sets are clearly presented in Table 4. The validation examples are marked with [a]. The rest of the study will focus on the model which is constructed from the training set and will examine the predictive ability of the produced model. Using the same seven descriptors that were selected by the ES-SWR method, we developed a new MLR equation based on only the 35 training examples:

$$\begin{aligned}\log(1/EC_{50}T47D) = &-17.1(\pm4.40)\\ &-0.461(\pm0.23)\mathrm{NBranch}\\ &+0.215(\pm0.06)\mathrm{Bertz}\\ &-0.00182(\pm0.0005)\mathrm{Quadr}\\ &+0.251(\pm0.072)\mathrm{DPLL}\\ &+2.15(\pm0.512)\mathrm{LUMO}\\ &-0.0602(\pm0.014)\mathrm{PSAr}\\ &+0.327(\pm0.082)\mathrm{SVDe}\\ \mathrm{RMS}=0.222,\quad R^2=0.806,\quad F=16.07,&\\ Q^2=0.658,\quad S_{\mathrm{PRESS}}=0.295,\quad n=35&\end{aligned} \tag{11}$$

Eq. 11 was used to predict the apoptosis-inducing activity for both the training and validation examples. Experimental versus predicted values are shown graphically in Figure 3, where 95% confidence intervals on the predicted values are indicated. The predicted apoptosis-inducing activities are also shown numerically in the two last columns of Table 4. The R^2 statistic for the training set is equal to 0.806 as shown above, while for the validation set the R^2_{pred} statistic is 0.869.

The results illustrated once more that the linear MLR technique combined with a successful variable selection procedure is adequate to generate an efficient QSAR model for predicting the apoptosis-inducing activity of 4-aryl-4*H*-chromenes.

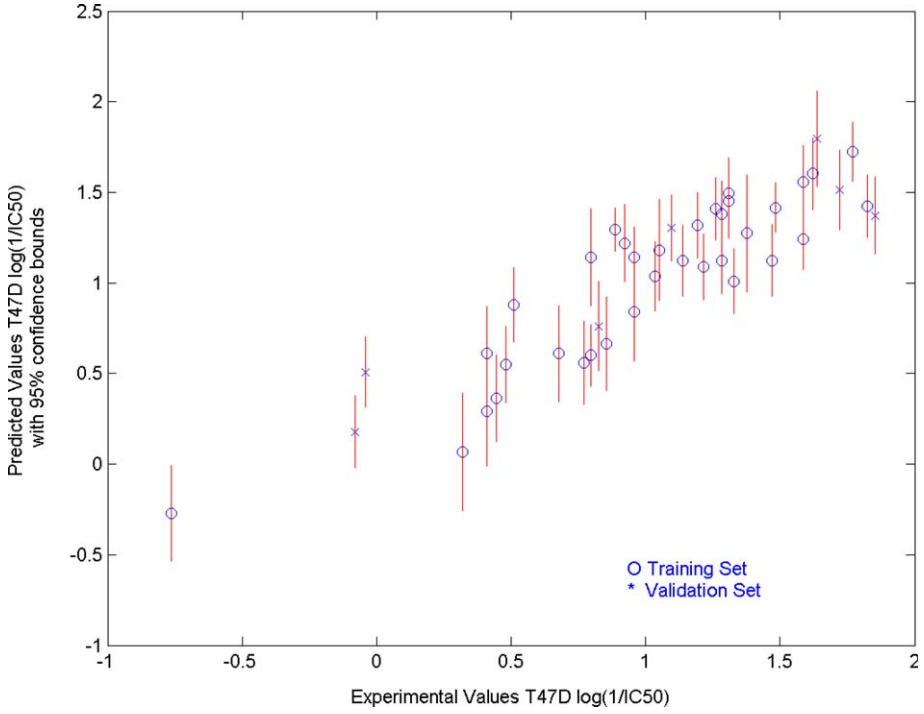

Figure 3. Experimental versus predicted values $\log(1/EC_{50})$ for the training and validation set with 95% confidence bounds.

The proposed model Eq. 11 also passed the rest of the tests that we utilized for illustrating its predictive ability Eqs. 4–7:

$$R^2_{\text{cv,ext}} = 0.819 > 0.5$$

$$R^2 = 0.869 > 0.6$$

$$\frac{(R^2 - R^2_{\text{o}})}{R^2} = -0.3231 < 0.1, \quad \frac{(R^2 - R'^2_{\text{o}})}{R^2} = -0.2810 < 0.1$$

$$k = 1.015 \quad \text{and} \quad k' = 0.922$$

Finally, it was important to note that the model was quite stable to the inclusion–exclusion of compound as measured by LOO and L5O correlation coefficient values, which are presented below:

$$Q^2_{\text{LOO}} = 0.658$$

$$Q^2_{\text{L5O}} = 0.6775$$

Q^2_{LOO} and Q^2_{L5O} are calculated using only the 35 training examples. Calculation of the Q^2_{LOO} statistic was performed using all 35 models that result from excluding one compound each time from the training examples, while calculation of the Q^2_{L5O} statistic was based on 1000 random exclusions of groups of examples.

The model was further validated by applying the Y-randomization test. Several random shuffles of the Y vector were performed and the results are shown in Table 5. The low R^2 and Q^2 values indicate that the good results in our original model are not due to a chance correlation or structural dependency of the training set.

It needs to be emphasized that no matter how robust, significant, and validated a QSAR model may be, it cannot be expected to reliably predict the modeled activity for the entire universe of chemicals. The extrapolation method was applied to the compounds that constitute the test set. The leverages for the compounds **2**, **7**, **10**, **12**, **27**, **39**, and **42** that constitute the validation set are, respectively, equal to 0.184, 0.179, 0.230, 0.219, 0.156, 0.327, and 0.286. None of the 7 compounds fell outside from the domain of the model (warning leverage limit 0.686).

Table 5. R^2 and Q^2 values after several Y-randomization tests

Iteration	R^2	Q^2
1	0.23	0.00
2	0.30	0.00
3	0.35	0.10
4	0.09	0.00
5	0.08	0.00
6	0.15	0.00
7	0.19	0.00
8	0.09	0.00
9	0.11	0.00
10	0.29	0.09

The proposed method, due to the high predictive ability,[22,32] could be a useful aid to the costly and time-consuming experiments for determining the apoptosis-inducing activity of 4-aryl-4*H*-chromenes. The method can also be used to screen existing databases or virtual combinations in order to identify derivatives with desired activity. In this case, the applicability domain will serve as a valuable tool to filter out "dissimilar" combinations.

4. Conclusion

The successful results of this study led to the conclusion that apoptosis-inducing activity can be successfully modeled with physicochemical constants and structural descriptors. The validation procedures utilized in this work (separation of data into independent training and validation sets, Y-randomization) illustrated the accuracy and robustness of the produced QSAR model not only by calculating its fitness on sets of training data, but also by testing the predictive ability of the model. The proposed method, due to the high predictive ability, offers a useful alternative to the costly and time-consuming experiments for determining apoptosis-inducing activity of 4-aryl-4*H*-chromenes.

Acknowledgments

A.A. thanks Cyprus Research Promotion Foundation (Grant No. PENEK/ENISX/0603/05) and A.G. Leventis Foundation for its financial support. G.M. thanks the Greek State Scholarship Foundation for a doctoral assistantship.

References and notes

1. Melagraki, G.; Afantitis, A.; Sarimveis, H.; Igglessi-Markopoulou, O.; Alexandridis, A. *Mol. Div.* **2006**. doi:10.1007/s11030-005-9008-y.
2. Melagraki, G.; Afantitis, A.; Makridima, K.; Sarimveis, H.; Igglessi-Markopoulou, O. *J. Mol. Model.* **2006**, *12*, 297.
3. Melagraki, G.; Afantitis, A.; Sarimveis, H.; Igglessi-Markopoulou, O.; Supuran, C. T. *Bioorg. Med. Chem.* **2006**, *14*, 1108.
4. Afantitis, A.; Melagraki, G.; Sarimveis, H.; Koutentis, P. A.; Markopoulos, J.; Igglessi-Markopoulou, O. *Mol. Div.* **2006**. doi:10.1007/s11030-005-9012-2.
5. Leonard, J. T.; Roy, K. *QSAR Comb. Sci.* **2004**, *23*, 387.
6. Netzeva, T.; Aptula, A. O.; Chaudary, S. H.; Duffy, J. C.; Schultz, T. W.; Schüürmann, G.; Cronin, M. T. D. *QSAR Comb. Sci.* **2003**, *22*, 575.
7. Salvesen, G. S.; Dixit, V. M. *Cell* **1997**, *91*, 443.
8. Villa, P.; Kaufmann, S. H.; Earnshaw, W. C. *Treds Biochem. Sci.* **1997**, *22*, 388.

9. Kothakota, S.; Azuma, T.; Reinhard, C.; Klipel, A.; Tang, J.; Chu, K.; McGarry, T.; Kirschner, M.; Koths, K.; Kwiatkowski, D.; Williams, L. *Science* **1997**, *278*, 294.
10. Hoglen, N.; Chen, L.; Fisher, C.; Hirakawa, B.; Groessl, T.; Contreras, P. *J. Pharmacol. Exp. Ther.* **2004**, *309*, 634.
11. Han, B.; Xu, D.; Choi, J.; Han, Y.; Xanthoudakis, S.; Roy, S.; Tam, J.; Vaillancourt, J.; Colucci, J.; Siman, R.; Giroux, A.; Robertson, G.; Zamboni, R.; Nicholson, D.; Holtzman, D. *J. Biol. Chem.* **2002**, *277*, 30128.
12. Haunstetter, A.; Izumo, S. *Circ. Res.* **1998**, *82*, 1111.
13. Wellington, C.; Hayden, M. *Clin. Genet.* **2000**, *57*, 1.
14. Herr, L.; Debatin, K. M. *Blood* **2001**, *98*, 2603.
15. Rich, T.; Allen, R. L.; Wyllie, A. H. *Nature* **2000**, *407*, 777.
16. Vaillancourt, L.; Charron, S.; Dodd, J.; Attardo, G.; Labrecque, D.; Lamothe, S.; Gourdeau, B.; Tseng, B.; Drewe, J.; Cai, S. X. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 4745.
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. Todeschini, R.; Consonni, V.; Mannhold, R. In Handbook of Molecular Descriptors; Kubinyi, H.; Timmerman, H., Eds.; Wiley-VCH: Weinheim, Germany, 2000; Vol. 11.
21. Kennard, R. W.; Stone, L. A. *Technometrics* **1969**, *11*, 137.
22. Tropsha, A.; Gramatica, P.; Gombar, V. K. *QSAR Comb. Sci.* **2003**, *22*, 1.
23. Wu, W.; Walczak, B.; Massart, D. L.; Heuerding, S.; Erni, F.; Last, I. R.; Prebble, K. A. *Chemometr. Intell. Lab. Syst.* **1996**, *33*, 35.
24. Efron, B. *J. Am. Stat. Assoc.* **1983**, *78*, 316.
25. Efroymson, M. A. *Multiple regression analysis*. In *Mathematical Methods for Digital Computers*; Ralston, A.; Wilf, H. S., Eds.; Wiley: New York, 1960.
26. Osten, D. W. *J. Chemom.* **1998**, *2*, 39.
27. Wold, S.; Eriksson, L.; *Statistical validation of QSAR results*. In *Chemometrics Methods in Molecular Design*; Van de Waterbeemd, H., Ed., Wiley: VCH Weinheim, 1995.
28. Golbraikh, A.; Tropsha, A. *J. Mol. Graph. Mod.* **2002**, *20*, 269.
29. Shen, M.; Beguin, C.; Golbraikh, A.; Stables, J.; Kohn, H.; Tropsha, A. *J. Med. Chem.* **2004**, *47*, 2356.
30. Atkinson, A. *Plots, Transformations and Regression*; Clarendon Press: Oxford, 1985.
31. Kier, L. B.; Hall, L. B. *Molecular Connectivity in Structure Activity Analysis*; Wiley: Chichester, 1986.
32. Aptula, A. O.; Jeliazkova, N. G.; Schultz, T. W.; Cronin, M. T. D. *QSAR Comb. Sci.* **2005**, *24*, 385.