Available online at www.sciencedirect.com

ScienceDirect

Bioorganic & Medicinal Chemistry 14 (2006) 7490–7500

Bioorganic & Medicinal Chemistry

QSAR study of antiplatelet agents

Alan R. Katritzky,[a,*] Liliana M. Pacureanu,[a] Svetoslav Slavov,[a] Dimitar A. Dobchev[a,c] and Mati Karelson[b,c]

[a] *Center for Heterocyclic Compounds, Department of Chemistry, University of Florida, Gainesville, FL 32611, USA*
[b] *Institute of Chemistry, Tallinn University of Technology, Ehitajate tee 5, Tallinn 19086, Estonia*
[c] *Chemistry Department, University of Tartu, 2 Jakobi Street, Tartu 51014, Estonia*

Received 2 May 2006; revised 5 July 2006; accepted 7 July 2006
Available online 30 August 2006

Abstract—A QSAR methodology that involves multilinear (Hansch-type) and nonlinear (ANN backpropagation) approaches was developed to correlate the antiplatelet activity of 60 benzoxazinone derivatives against factor Xa. The statistical characteristics provided by multilinear model ($R^2 = 0.821$) indicated satisfactory stability and predictive ability, while the ANN predictive ability is somewhat superior ($R^2 = 0.909$). The multilinear model provided insight into the main factors that modulate the inhibitory activity of the investigated compounds.

1. Introduction

Platelets play a central role in normal hemostasis and are key participants in pathologic thrombosis due to their capacity to adhere to injured blood vessels and to accumulate at the place of injury.[1,2] Major stimuli able to induce platelet activation including shape change (into spiny spheres), secretion and aggregation include: collagen, thrombin, and adenosine diphosphate (ADP). These items can bind to fibrinogen, aggregate, and release the contents of their intracellular granules including ADP and serotonin; ADP and arachidonic acid (AA) metabolite act as endogenous platelet activators for thromboxane A_2 (Tx A_2), intensify the extent of platelet aggregation, and provide positive feedback.[4] Platelet adhesion and activation is a normal physiological response to the accidental rupture of blood vessels, but the uncontrolled such activity may cause thromboembolic artery occlusion, acute coronary syndrome, and ischemic stroke.

The process of platelet adhesion to extracellular matrix consists in binding to various glycoprotein (GP) receptors mediated by von Willebrand factor (vWf) and collagen.[5]

Factor Xa is a trypsin-like serine protease situated at the convergence of the surface-activated intrinsic and factor-activated extrinsic coagulation pathways. The prothrombinase complex is formed by factor Xa on the phospholipid surface with factor Va and calcium; it catalyzes the proteolysis of prothrombin to thrombin (factor IIa). Thrombin is the main, final enzyme in the phospholipid coagulation system that leads to fibrin formation. It provides positive and negative feedback regulatory signal in the normal hemostasis, while in pathological conditions factor Xa provides catalytic activation of thrombin. Thus, the inhibition factor Xa affects the coagulation but not the platelet function. Therefore, the inhibition factor Xa may provide a novel, effective antithrombotic drug that provides no risk of bleeding. Recently, the inhibition factor Xa has been intensely investigated in order to replace the existing therapies in the treatment or prevention of thromboembolic disorders.[7,8]

A number of synthetic organic compounds have been evaluated as platelet inhibitors including phenyl quinolones derivatives,[9,10] benzo[*d*]isothiazol-3-one derivatives,[11] oxime- and methyloxime-containing flavone and isoflavone derivatives,[12] linoleic acid isomers,[13] 5*H*[1]benzopyrano[4,3-*d*]pyrimidin-5-amine derivatives,[14] polycyclic pyrimidine derivatives,[15] and chalcone derivatives.[16]

Due to the significant limitations of existing antiplatelet drugs, the search for new antiplatelet aggregation agents is of great interest.

Keywords: Antiplatelets; QSAR; Benzoxazinones; CODESSA PRO; ANN.

* Corresponding author. Tel.: +1 352 392 554; fax: +1 352 392 9199; e-mail: katritzky@chem.ufl.edu

0968-0896/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2006.07.022

There are several QSAR studies on platelet inhibition, mainly 3D QSAR and neural network approach, that used a reduced number of datapoints, but a general model for the platelet inhibition is still missing.[6,7,17–19]

Dudley et al. developed a complex methodology to design novel benzoxazinone derivatives that involved SAR and molecular modeling software (GASP) against FXa. To select the substituents, they used Topliss tree approach that accounts for electronic, steric, and lipophilic fields. The novel benzoxazinones were designed, synthesized, and then undergo biological assay that confirmed the higher inhibitory potential of the new molecules from 27 μM to 0.0030 μM.[6]

3D QSAR approach (using Gold) performed on several benzoxazinone derivatives confirmed the previously published X-ray structure binding mode in the following manner: the benzamidine moiety interacts with Asp 189 and amidine moiety is retained in the binding pocket by π-cation interaction with Tyr 99, Phe 174, and Trp 215.[7]

A comparative molecular field analysis (CoMFA) model was developed by G. Roma et al. for a series of 2-amino-4*H*-pyrido[1,2*a*]pyrimidin-4-ones divided into training set consisting of 73 molecules and prediction set of 10 molecules which relates their biological activities to their steric and electrostatic fields. The resulted models involved six components and displayed good statistics and predictive ability for the training set $q = 0.682$, $R^2 = 0.874$, $s = 0.242$, while for the full set the results were similar $q = 0.662$, $R^2 = 0.866$, $s = 0.245$.[17]

Artificial neural network backpropagation methodology (1-3-1 configuration) was employed to correlate the biological activities of a small set of 21 2-substituted phenyl and benzimidazolyl-5-methyl-4-(3-pyridyl)imidazole derivatives that supplied a QSAR model with good statistical characteristics ($R^2 = 0.860$, $s = 0.183$).[18]

In 1994, Tanaka evaluated the antiplatelet activities of 2-substituted phenyl and benzimidazolyl-5-methyl-4-(3-pyridyl)imidazoles using a classical Hansch 2D approach by means of one descriptor that accounts for hydrophobicity, micelle–water partition coefficient ($\log P_{\mathrm{mw}}$), which provided good results ($R^2 = 0.772$, $s = 0.257$, $F = 11.07$, $n = 18$).[19]

Recently, our group was also involved in the investigation of the inhibition of the PDGF receptor.[20] Two good QSAR models were developed based on multilinear and nonlinear (ANN) techniques.

The current paper summarizes our attempt to build linear and nonlinear QSAR models based on a large dataset of benzoxazinone derivative inhibitors factor Xa using CODESSA PRO software and Artificial Neural Network backpropagation algorithm.

2. Data set

The present study comprises 60 experimental IC_{50} values for platelet inhibition against factor Xa, determined in the following references.[3,7,8] The substitution patterns of platelet activating inhibitors are given in Table 1. The data points were converted into molar $\log IC_{50}$ value units which were used instead of IC_{50}s to improve the normal distribution of the experimental data points.

All the data are collected in Table 1 which include the following: (i) the substitutional pattern of benzoxazinones (second, third, and fourth columns), (ii) the pIC_{50} values taken from the original references and converted to logarithm of molar units (fifth column), (iii) the predicted $\log IC_{50}$ values obtained from multilinear model (sixth column) and from nonlinear model (ANN backpropagation) (seventh column).

3. Results and discussions

3.1. Multilinear models

In the current study, a QSAR model is presented for $\log IC_{50}$ of 60 benzoxazinone antiplatelet agents involving theoretical descriptors, which have been calculated from molecular structure.

The distribution of experimental data, the logarithm of inhibitory activity (IC_{50}) for the Full Set of 60 benzoxazinone derivatives is shown in Figure 1.

The BMLR algorithm[21] was used in order to avoid overcorrelation of the regression equations by monitoring the increase of R^2 in the equations with successive number of descriptors involved. The procedure is called the 'break point' technique as illustrated in Figure 2 that shows the breakpoint (the change in the slope) in the plot of R^2 versus number of descriptors added. The procedure was stopped when the difference between R^2 of the two consequent regression equations was less than or equal to 0.02.

The results presented herein are obtained by applying BMLR algorithm to the structures and experimental $\log IC_{50}$. The statistical characteristics and the descriptors involved in the multilinear regression equation are listed in Table 2. A graphical presentation of the relationship between the experimental and predicted $\log IC_{50}$ values is shown in Figure 3.

The types of molecular descriptors involved in the current QSAR models are of molecular orbital (D_3, D_5) and quantum chemical (D_1, D_2, D_4) origin.

The statistical significance of the selected descriptors in the QSAR model of Table 2 was assessed by application of the *t*-test criterion that gives the following order of importance: $D_1 > D_3 > D_2 > D_5 > D_4$.

The maximum electron–nuclear attraction for atom N (D_1) is given by the following formula:

Table 1. The substitutional pattern of benzoxazinones, experimental and predicted $\log IC_{50}$ values

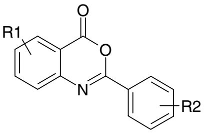
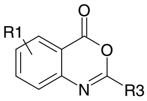
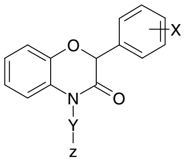
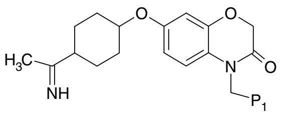
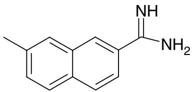
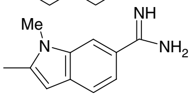
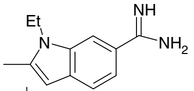
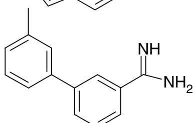
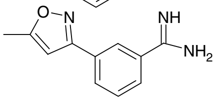
Compound	Substitutional pattern			Experimental[a] $\log IC_{50}$	Predicted $\log IC_{50}$	
					Multilinear model	Nonlinear ANN model
	 4*H*-3,1-Benzoxazin-4-ones					
1	H	2-F		−3.860	−4.200	−4.303
2	6-CF_3	2-F		−4.432	−3.966	−4.338[b]
3	5-Cl,8Cl	2-F		−3.745	−4.770	−4.547
4	5-COOMe	2-F		−3.860	−4.638	−4.556
5	5-NO_2	2-F		−4.585	−4.653	−4.501
6	5-Cl	2-F,6-F		−5.310	−5.001	−4.774
7	5-NO_2	2-F,6-F		−5.201	−4.665	−4.549
8	5-Cl,8-Cl	2-F,6-F		−5.201	−4.893	−5.022[b]
9	6-Me	2-F,6-F		−4.658	−4.395	−4.385
10	6-CF_3	2-F,6-F		−4.796	−3.935	−5.028[b]
11	7-NO_2	2-F,6-F		−3.951	−3.711	−4.876[b]
12	5-F	2-F,6-F		−5.222	−4.764	−4.297
13	6-NO_2	2-F,6-F		−4.721	−4.358	−4.460
14	7-CF_3	2-F,6-F		−3.760	−3.575	−4.244[b]
15	6-OMe	2-F,6-F		−3.813	−4.097	−4.307
16	6-NHAc	2-F,6-F		−4.456	−3.740	−4.015
17	6-NH_2	2-F,6-F		−3.813	−4.491	−4.215
18	5-COOMe	2-F,6-F		−3.951	−4.609	−4.246
19	5-Me	2-F,6-F		−4.051	−4.586	−4.451
20	8-CF_3	2-F,6-F		−4.097	−4.135	−4.321
21	6-Me	2-F,6-F		−5.000	−4.470	−4.589[b]
22	5-F	2-F,6-F		−5.222	−4.846	−4.589
23	6-Me	2-F,6-F		−3.959	−4.467	−4.385
24	6-I	2-Cl		−4.824	−4.584	−4.407
25	6-Me	2-Cl,6-Cl		−4.201	−4.503	−4.416
26	5-NO_2	2-OMe		−5.237	−5.015	−4.537
27	H	2-OMe,5-Cl		−4.523	−4.903	−4.675[b]
28	5-NO_2	2-OCOMe		−4.585	−4.997	−4.661
29	6-NO_2	2-OCOMe		−4.310	−4.716	−4.569
30	6-Cl	2-Br		−5.222	−4.110	−4.872[b]
	 4*H*-3,1-Benzoxazin-4-ones					
	R1	R3				
31	5-NO_2	2-Cl–3-pyridyl		−4.009	−3.932	−4.493[b]
32	6-NO_2	2-Cl–3-pyridyl		−3.854	−4.021	−4.622[b]
	 3,4 dihydro-2H-1,4-benzoxazin-3-one					
	X	Y	Z			
33	H	$(CH_2)_5$	*cis*-2,6-DiMe-piperidinyl	−4.569	−4.780	−4.675
34	3,4-Cl	$(CH_2)_5$	*cis*-2,6-DiMe-piperidinyl	−4.585	−4.510	−4.575[b]
35	4-Cl	$(CH_2)_5$	*cis*-2,6-DiMe-piperidinyl	−4.638	−4.625	−4.600[b]

Table 1 (*continued*)

Compound		Substitutional pattern		Experimental[a] $\log IC_{50}$	Predicted $\log IC_{50}$	
					Multilinear model	Nonlinear ANN model
36	2-Cl	$(CH_2)_5$	*cis*-2,6-DiMe-piperidinyl	−4.796	−5.222	−4.295
37	4-CH3	$(CH_2)_5$	*cis*-2,6-DiMe-piperidinyl	−4.553	−4.749	−4.649
38	4-OCH3	$(CH_2)_5$	*cis*-2,6-DiMe-piperidinyl	−5.310	−4.623	−4.589
39	4-C(=S)NH_2	$(CH_2)_5$	*cis*-2,6-DiMe-piperidinyl	−4.469	−5.391	−4.987
40	4-C(=NH)NHOH	$(CH_2)_5$	*cis*-2,6-DiMe-piperidinyl	−4.181	−5.156	−4.788[b]
41	4-C(=NH)NH_2	$(CH_2)_5$	*cis*-2,6-DiMe-piperidinyl	−5.102	−6.733	−6.984
42	3-C(=S)NH_2	$(CH_2)_5$	*cis*-2,6-DiMe-piperidinyl	−5.770	−5.474	−5.287
43	3-C(=NH)NHOH	$(CH_2)_5$	*cis*-2,6-DiMe-piperidinyl	−5.252	−5.302	−4.856[b]
44	3-C(=NH)NH_2	$(CH_2)_5$	*cis*-2,6-DiMe-piperidinyl	−8.523	−6.807	−7.242
45	3-CH_2–NH_2	$(CH_2)_5$	*cis*-2,6-DiMe-piperidinyl	−6.076	−6.247	−5.975
46	3-C(=NH)NHCH3	$(CH_2)_5$	*cis*-2,6-DiMe-piperidinyl	−5.553	−5.640	−5.164[b]
47	3-C(=NH)NHmorpholinyl	$(CH_2)_5$	*cis*-2,6-DiMe-piperidinyl	−5.824	−5.901	−5.267[b]
48	3-C(=NH)NH_2	$(CH_2)_6$	*cis*-2,6-DiMe-piperidinyl	−7.168	−6.860	−7.370
49	3-C(=NH)NH_2	$(CH_2)_4$	*cis*-2,6-DiMe-piperidinyl	−7.921	−6.709	−6.942
50	3-C(=NH)NH_2	$(CH_2)_5$	*cis*-2,6-DiMe-pyrrolidinyl	−7.319	−6.704	−6.935
51	3-C(=NH)NH_2	$(CH_2)_5$	Piperidinyl	−7.276	−6.814	−7.048
52	3-C(=NH)NH_2	$(CH_2)_5$	morpholinyl	−6.620	−6.476	−6.085
53	3-C(=NH)NH_2	$(CH_2)_5$	NH_2	−5.201	−6.099	−5.004
54	3-C(=NH)NH_2	$(CH_2)_5$	Diisopropylamino	−6.921	−6.776	−7.266[b]
55	3-C(=NH)NH_2	$(CH_2)_5$	Dihexylamino	−5.796	−6.567	−6.347
	P1	 1,6-substituted-1,4 dihydro-2H-1,4-benzoxazin-3-one				
56				−8.222	−8.351	−8.607[b]
57				−8.699	−8.213	−8.616*
58				−8.699	−8.272	−8.383
59				−8.398	−8.532	−8.417
60				−7.284	−7.868	−7.667[b]

[a] Taken from references as follows: compounds **1**–**32** from Ref. 8, compounds **33**–**55** from Ref. 6, and compounds **56**–**60** from Ref. 7.

[b] Validation set.

$$E_{\rm ne}({\rm A}) = \sum_{\rm B}\sum_{\mu,\nu} P_{\mu\nu}\left\langle \mu\left|\frac{Z_{\rm B}}{Z_{\rm iB}}\right|\nu\right\rangle \qquad (1)$$

where A is the given atomic species; $P_{\mu\nu}$ are the density matrix elements over atomic basis $\{\mu\nu\}$; $Z_{\rm B}$ are the charges of the atomic nuclei, B; $R_{\rm iB}$ is the distance between the electron i and atomic nucleus B; and $\left\langle \mu\left|\frac{Z_{\rm B}}{Z_{\rm iB}}\right|\nu\right\rangle$ is the electron–nuclear attraction integral on atomic basis $\{\mu\nu\}$. The first summation is performed over all atomic nuclei in the molecule (B), while the second summation is carried out over all atomic orbital at the given atom (A). This energy describes the strength of the nuclear–electron attraction at a particular atom and can be related to the atomic reactivity of the nitrogen atom.[22]

The minimum exchange energy for bond C–O (D_2) descriptor accounts for the sum of the electronic repulsion energy, electron–nuclear attraction energy, and nuclear repulsion energy between C and O atoms (Eq. 2).[22] In the equation of Table 2 descriptor D_2 has a small positive influence. The additive energy between two atoms can be expressed as:

$$E({\rm AB}) = E_{\rm ee}({\rm AB}) + E_{\rm ne}({\rm AB}) + E_{\rm nn}({\rm AB}) \qquad (2)$$

where $E_{\rm ee}$(AB) is the electronic repulsion energy between two atoms, $E_{\rm ne}$(AB) is the electron–nuclear attraction

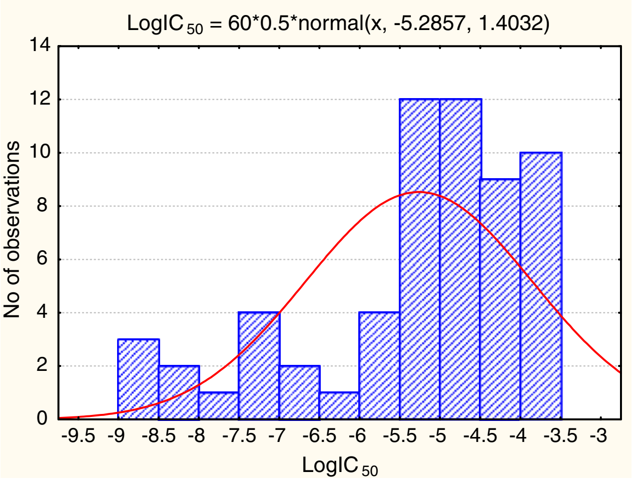


Figure 1. The distribution of the experimental log IC_{50} values.

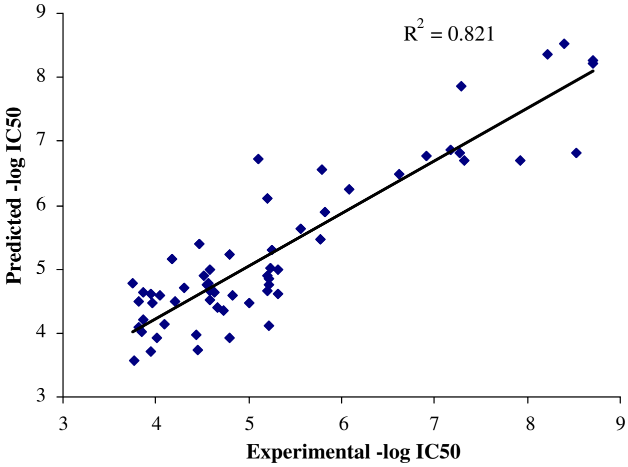


Figure 3. Graphical plot of experimental versus predicted $\log IC_{50}$ according to the model in Table 2.

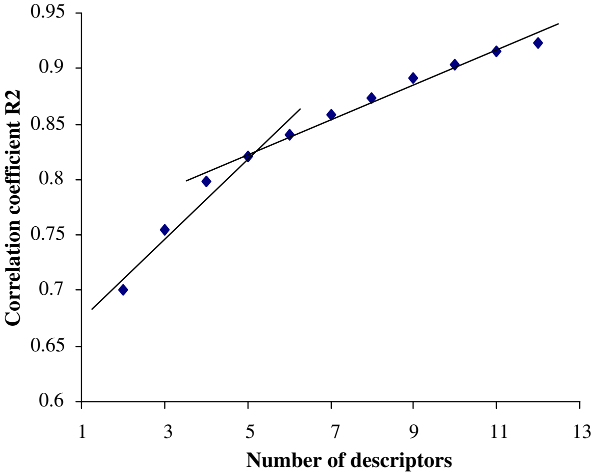


Figure 2. The plot of correlation coefficient R^2 versus the number of descriptors.

energy between two atoms, and $E_{nn}(AB)$ is the nuclear repulsion energy between two atoms. The presence of these descriptors indicates the importance of the C–O bond reactivity in determining the antiplatelet activity of compounds.

Frontier orbital energies are parameters that can characterize the susceptibility of the molecule toward the attack of electrophilic or nucleophilic reagents. The HOMO − 1 energy (Eq. 3) can be related to the reactivity of a compound as a nucleophile or a radical.[22]

$$E_{\mathrm{HOMO}-1} = \langle \phi_{\mathrm{HOMO}-1} | \hat{F} | \phi_{\mathrm{HOMO}-1} \rangle \tag{3}$$

The average nucleophilic reactivity index for atom N (D_4) is directly related to the HOMO energy and to the coefficient of the N atom in the HOMO by the relationship:[22]

$$N_A = \sum_{i \in A} c^2_{i\mathrm{HOMO}} / (1 - \varepsilon_{\mathrm{HOMO}}) \tag{4}$$

The presence of this descriptor in the current multilinear QSAR model provides information about the possibility of hydrogen bonding formation that involves the nitrogen atom of benzoxazinone derivatives.

The min (>0.1) bond order for atom H descriptor D_5. Mulliken bond orders can be calculated as the sum over the occupied molecular orbitals using occupation numbers (n_i) of the molecular orbitals and the atomic contributions ($c_{i\mu}, c_{i\nu}$) in the molecular orbital.[22]

Table 2. The molecular descriptors and statistical characteristics for the best QSAR model

Descriptor	Symbol	b[a]	Δb[b]	t[c]	R^2[d]	R^2_{cv}[e]	F[f]	s^2[g]
					0.821	0.773	49.55	0.621
Intercept		58.695	22.816	2.572				
Max e–n attraction for atom N	D_1	−0.229	0.036	−6.370				
Min exchange energy for bond C–O	D_2	7.297	1.629	4.480				
HOMO − 1 energy	D_3	−1.136	0.246	−4.613				
Avg nucleoph. react. index for atom N	D_4	103.079	29.614	3.481				
Min (>0.1) bond order for atom H	D_5	−43.093	10.971	−3.928				

[a] b are the regression coefficients of the multilinear model.
[b] Δb are the standard errors for these regression coefficients.
[c] t represents the t-values for each selected descriptor.
[d] R^2 is the correlation coefficient.
[e] R^2_{cv} is the cross-validated correlation coefficient.
[f] F is Fisher criterion.
[g] s^2 represents the squared standard deviations of the model.

$$P_{AB} = \sum_{i=1}^{occ} \sum_{\mu \in A} \sum_{\nu \in B} n_i c_{i\mu} c_{j\nu} \quad (5)$$

In our case the minimum value of the bond order for hydrogen atom accounts for its acidic character, and for its ability to act as an hydrogen bonding donor, for example, with an oxygen atom of the biological receptor factor Xa (trypsin-like serine protease).

The descriptors D_2, D_4, and D_5 suggest the importance of the N atom and CO group for the interaction between inhibitor and biological receptor.

For this model (Table 2), we registered two outliers: entries 41 and 44 from Table 1; they display the largest difference between experimental and predicted $\log IC_{50}$ values. Compound **44** presents the lowest value of $\log IC_{50}$ and is the most active compound in the whole series of benzoxazinone inhibitors. Therefore, it is a potential candidate for a parenteral antithrombotic.[6]

3.2. The model validation

3.2.1. Leave-one-out validation. The first technique applied for the validation of the proposed models was based on leave-one-out algorithm. The corresponding squared cross-validated correlation coefficient (R^2_{cv}) for all selected models, which is calculated automatically by the validation module implemented in CODESSA PRO package. The cross-validated correlation coefficient $R^2_{cv} = 0.774$ is pretty close to the correlation coefficient $R^2 = 0.821 (\Delta R^2 = R^2 - R^2_{cv} = 0.047)$ that suggests a good predictive ability of the best multilinear model. The results obtained by us were compared with those reported by Roma et al. for several data sets (73–94 compounds) that consist mainly in 2-amino-4*H*-pyrido[1,2*a*]pyrimidin-4-one derivatives (CoMFA-PLS). The difference between correlation coefficient and cross-validated correlation coefficient for these datasets is between 0.158 and 0.204. It can be easily observed that our multilinear regression equation is better in terms of stability and predictive ability with a lower difference $R^2 - R^2_{cv}$.

3.2.2. Internal validation. As mentioned in Section 5, our internal validation predicts the property values for each one-third of the compounds with the model fitted for the remaining two-thirds of the compounds. The general algorithm of the internal validation includes the following steps:

(i) division of the data set to be analyzed into three sets: CA (the 1st, 4th, 7th, etc., entries), CB (the 2nd, 5th, 8th, etc., entries) and CC (the 3rd, 6th, 9th, etc., entries);
(ii) in each of three combinations, two of the sets are combined into one and the correlation equation with the same descriptors, as in the QSAR model to be validated, is derived;
(iii) the equation developed in the step (ii) is used to predict the property values for the remaining set;
(iv) a comparison of the average of squared correlation coefficients for the fitted and predicted sets is made at the end.

The results of the internal validation applied to our data are listed in Table 3. As can be seen from Table 3 the average R^2 of the both sets (fit and predict) are relatively close showing satisfactory prediction.

3.3. Neural network models

In addition to the standard multilinear QSAR approach, an artificial neural network (ANN) methodology was also applied. The used method implements the feed-forward backpropagation algorithm as one of the most complex and powerful realizations of the ANN.

Since CODESSA PRO software is able to calculate hundreds of descriptors for a given compound, the selection of a good combination of descriptors is needed. A sensitivity analysis was performed over the most important descriptors. This was done by building the 1-1-1 NN models and the first 5 descriptors that showed lowest error at the output were selected. There was no surprise that the same descriptors as in the case of multilinear QSAR approach appeared as the most important for the observed biological effect. That was a favorable situation that could give us the possibility to compare the results obtained by means of both methods.

Before starting the actual calculation for the model, we divided the data set into training (40 compounds) and validation (20 compounds) sets. The following steps were performed in order to cover the same ranges of activities and the same distributions for the training and test datasets. The main idea at this stage was to keep closer to the QSAR rule that a sample should be representative of the general population (i.e., to follow the same distribution law).

1. All compounds were ordered in ascending order of their $\log IC_{50}$ values.
2. Every 3rd compound was selected to become a part of the test set.
3. All the remaining compounds formed the training set.

As can be seen from Figures 4 and 5, the above given procedure ensures that the training and test datasets have similar data distribution.

Several different architectures of ANN models were built. In the search for optimal ANN architecture we tried to use the lowest possible number of neurons to follow the common principle of generality of the ANN prediction.[23] The optimal architecture of the selected

Table 3. Internal validation of the models—statistical characteristics

Set to fit	R^2_{fit}	s^2_{fit}	Set to predict	R^2_{pred}	s^2_{pred}
A + B	0.817	0.431	C	0.806	0.401
A + C	0.832	0.339	B	0.790	0.569
B + C	0.837	0.396	A	0.724	0.521
Average	0.829	0.389		0.773	0.497

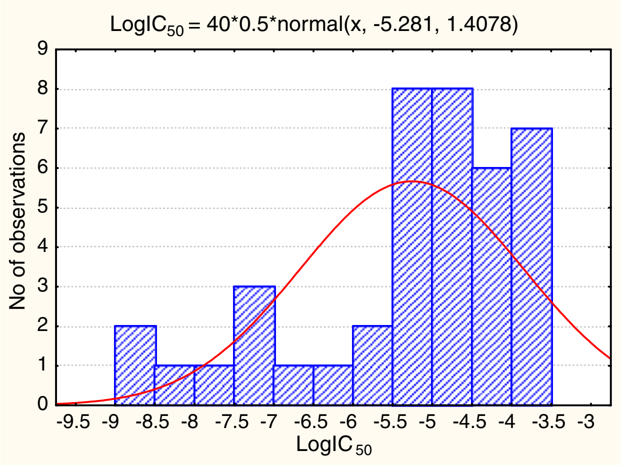


Figure 4. The distribution of the experimental data points for the training set.

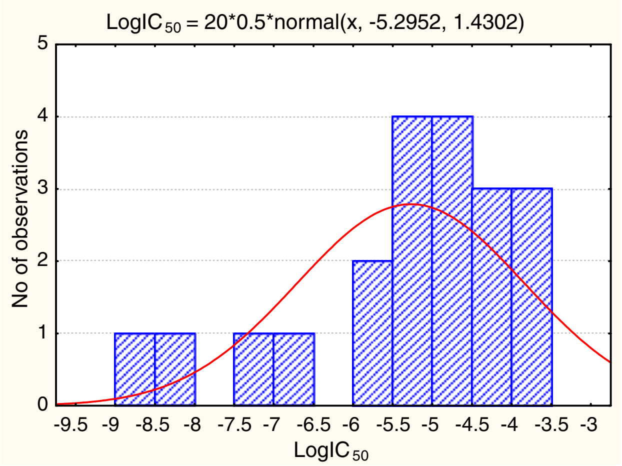


Figure 5. The distribution of the experimental data points for the validation set.

NN model was 5-15-1, which means that the model had five input neurons in the input layer (the selected descriptors), 15 hidden neurons in the hidden layer, and one neuron in the output layer representing the $\log IC_{50}$ values of the studied compounds.

During the training stage the weights were adjusted according to the output prediction error (see Eq. 10). The validation set error (and also the coefficient of determination—R^2) was monitored in order to avoid the over-training of the ANN and to stop the training process.

After the training procedure was successfully completed, the generated ANN rules (optimized weights) were applied to the test set in order to evaluate the quality of the model. The squares of the coefficients of determination (corresponding to R^2 in the case of MLR modeling) for the training and test datasets were 0.825 and 0.909, respectively. The corresponding root-mean-squared (RMS) errors were 0.360 and 0.211. The predicted $\log IC_{50}$ values for the training set are given in Table 1. The graphical presentation of the linear fit between the experimental and predicted values is given in Figure 6.

The comparison of the results by different treatments reveals that there is no significant improvement of the provided statistical parameters for the ANN training set over the BMLR results for the whole dataset (0.825 vs 0.821). However, when the ANN results for training set were compared with the multilinear model obtained for the same descriptors and the same set, it was found that the MLR model has lower quality ($R^2 = 0.796$). The ANN prediction of the data for the validation set (that was not used to train the ANN model) displayed also significantly better statistics ($R^2 = 0.909$) (see Fig. 7). This indicated that the predictive power of the ANN is superior over the multilinear model.

The ability of the ANN model, to predict correctly the $\log IC_{50}$ values, was further analyzed as follows. The

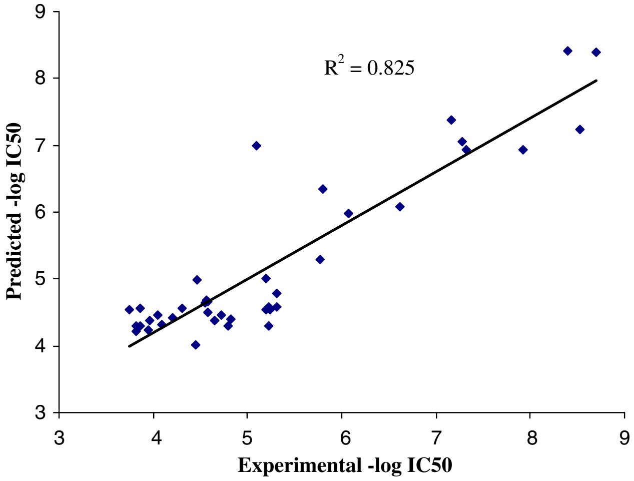


Figure 6. Scatter plot of experimental versus predicted $-\log IC_{50}$ for the training set according to the ANN model.

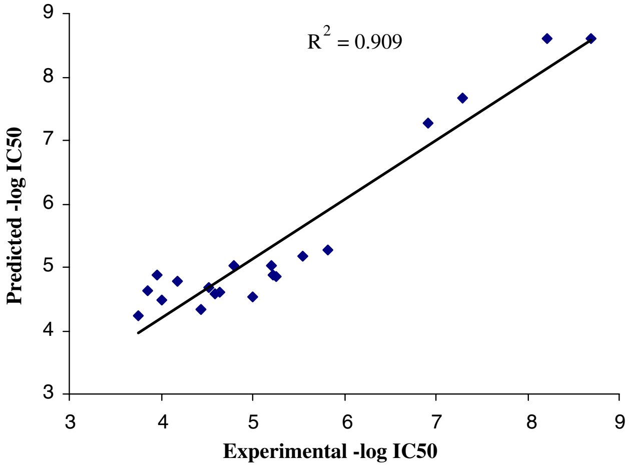


Figure 7. Scatter plot of experimental versus predicted $-\log IC_{50}$ for the validation set.

experimental and predicted ranges of values were separated into 5 subintervals (formally 5 classes—low, moderate, middle, good, and high activity). The confusion matrices based on this classification, which represent visually the ability of the ANN to predict exactly a certain class of antiplatelet activity (calculated as N/N_s where the N the number of antiplatelet agents in a certain class and N_s is the number of the successive 100% predictions), are given in Figure 8.

The confusion matrix in Figure 8 shows that for the training set the compounds with lower or higher activity are predicted with the same accuracy. However, there are several compounds belonging to both categories that are not well predicted. In the case of the most active compounds we had a less number of accurately predicted compounds since their number is smaller than the compounds with lower activity.

Figure 7 shows the linear fit between the experimental and predicted log IC_{50} values for the validation set. The ANN model showed better predictive ability over the multilinear QSAR models regarding the training set and significantly improved quality for the predicted values of biological activity of compounds (Fig. 9).

In this case, the confusion matrix of Figure 9 indicated that the compounds with higher biological activity are better predicted with respect to those that display lower biological activity. Therefore, the ANN backpropagation model can be successfully used to predict biological activities of benzoxazinone derivative platelet inhibitors.

4. Conclusions

A data set involving 60 benzoxazinone derivative antiplatelet agents was investigated to relate their log IC_{50} values to the molecular structure. A QSAR modeling of the in vitro log IC_{50} inhibitory concentration that reduces 50% of the serine-like protease Factor Xa was carried out using CODESSA PRO technique.

The multilinear regression equations displayed moderate statistical characteristics. The studies gave an insight into the forces that modulate the ligand–receptor interaction according to the descriptor that appeared in the MLR model. The interaction of benzoxazinones with Factor Xa is suggested to involve two corresponding N–H C–O bond pairs.

The ANN backpropagation technique supplied a model of a better quality in terms of statistical characteristics for the prediction set of 20 compounds. The ANN models are superior over the multilinear models in sense of prediction.

Since little QSAR work has been previously published for the inhibition of platelet aggregation, the present work contributes to the elucidation of molecular forces that occur in the substrate–inhibitor interactions. The design of drugs with favorable pharmacology is of interest in modern medicinal chemistry and the search for them should be performed using appropriate computational methods.

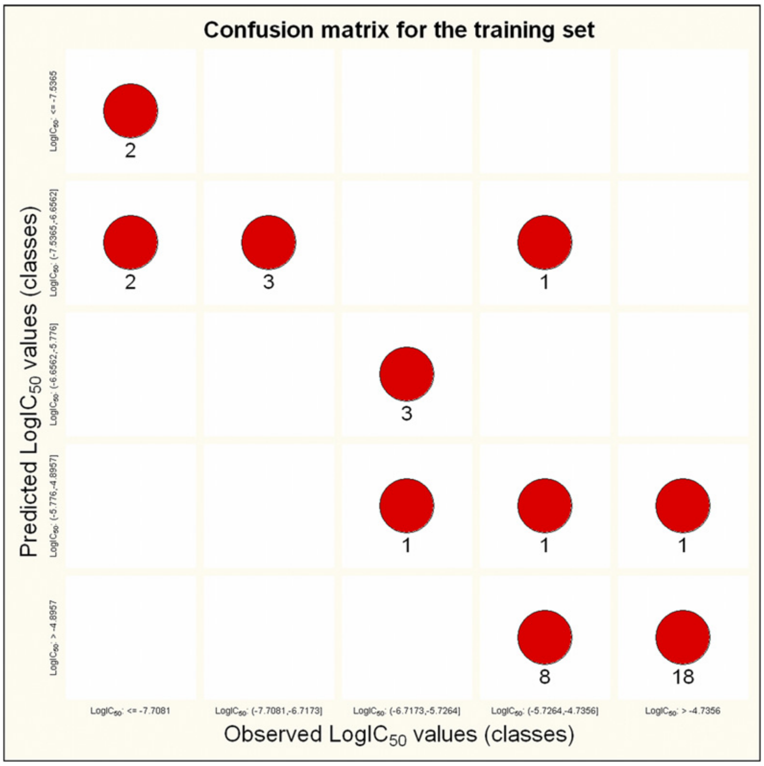


Figure 8. Confusion matrix for the training set.

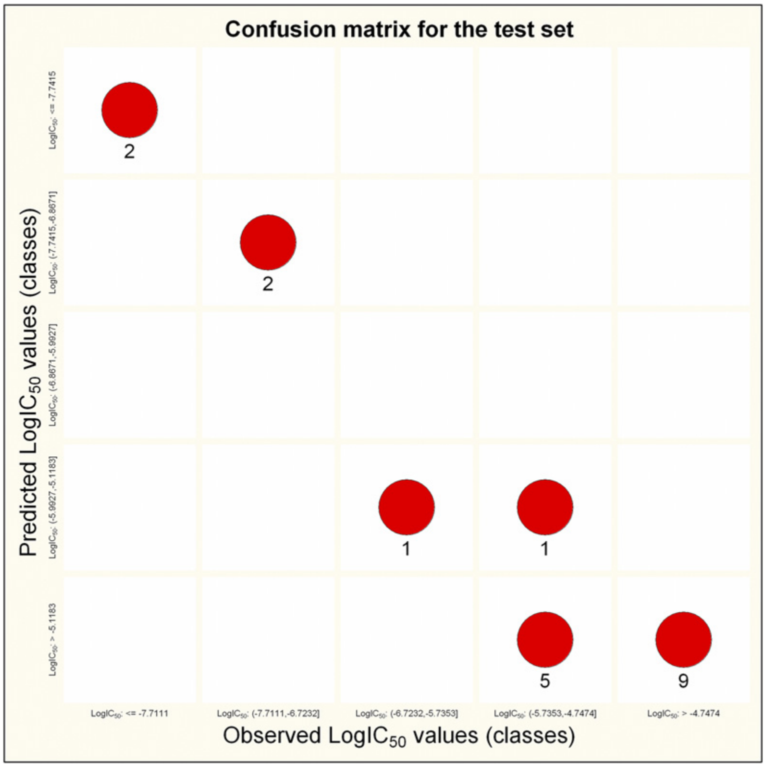


Figure 9. Confusion matrix for the validation set.

5. Methodology

5.1. Multilinear models

The 2D structures of compounds were drawn using Hyperchem 7.5 software[24] and their geometries were preoptimized using the molecular mechanics force field (MM+) available in the Hyperchem 7.5 package. The final refined equilibrium molecular geometries were obtained using the semiempirical methods AM1 (Austin Model-1) without any symmetry constraint[25] applying a gradient norm limit of 0.01 kcal/Å as a stopping criterion for optimized structures.

These geometries were used to calculate up to 700 molecular descriptors, classified as (i) constitutional, (ii) topological, (iii) geometrical, (iv) charge-related, (v) semiempirical, and (vi) thermodynamical, using CODESSA PRO package.[26]

The best multilinear regression analysis searches for the optimal regression coefficients (a_i, b_i) of the linear equation:

$$y_i = a_i + \sum b_i x_i \tag{6}$$

where x and y are the elements of the input and, respectively, the output data.

During the BMLR procedure the descriptor scales are normalized, centered automatically, and the final result is given in natural scales. This result has the best representation of the property in the given descriptors' pool. The final correlation equations are selected on the basis of partial and standard tests of significance and the highest multiple correlation coefficients.

CODESSA PRO software correlated and predicted successfully a series of physico-chemical and biological activities for diverse classes of organic compounds including boiling points,[27] partition coefficients,[28] cyclodextrin complexation free energy,[29] solvent scales,[30] rat blood:-air, saline:air, olive oil:air, blood:air,[31] tissue: air partition coefficients,[32] genotoxicity of aromatic amines and chlorinated aromatic compounds,[33a,b] inhibitory activity of 3-aryloxazolidin-2-one antibacterials against *Staphylococcus aureus*,[34] drug transfer into breast milk,[35] etc.

The QSAR models obtained were validated by the leave-one-out method and by internal validation whereby the compounds are divided into three subsets and each subset is predicted by a model derived from the remaining 2/3 of the compounds. The corresponding squared cross-validated correlation coefficient (R^2_cv) for all selected models is calculated automatically by the validation module implemented in CODESSA PRO package.

5.2. Nonlinear artificial neural network (ANN) models

In this work, a backpropagation ANN[23,36,37] was developed and used to obtain a nonlinear QSAR model. Topologically, it consists of input, hidden, and output layers of neurons or units connected by weights as shown in Figure 10. Each input layer node corresponds to a single independent variable (molecular descriptor) with the exception of the bias node. Similarly, each output layer node corresponds to a different dependent variable (property under investigation).

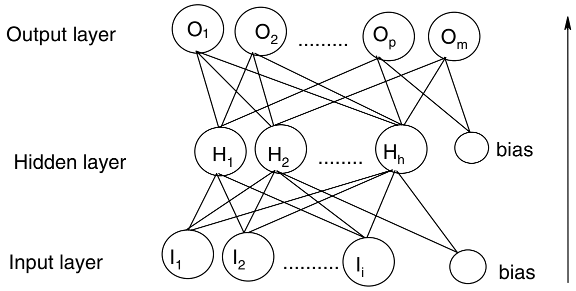


Figure 10. Three-layer backpropagation neural network.

Associated with each node is an internal state designated by I_i, H_h, and O_m for the input, hidden, output layers, respectively. Each of the input and hidden layers has an additional unit, termed a bias unit, whose internal state is assigned a value of 1. The input layer's I_i values are related to the corresponding independent variables by the scaling Eq. 7:

$$I_i = \frac{D_i - D_{i(\min)} + 0.1}{D_{i(\max)} - D_{i(\min)} + 0.1} \qquad (7)$$

where D_i is the value of the ith descriptor, $D_{i(\max)}$ and $D_{i(\min)}$ are its maximum and minimum values, respectively. The state H_h of each hidden unit is calculated by the squashing (sigmoid, logistic) function:

$$H_h(\varphi h) = \frac{1}{1 + \mathrm{e}^{-\varphi_h}} \qquad (8a)$$

$$\varphi_h = \sum_i w_{hi} I_i + \theta_h \qquad (8b)$$

where w_{hi} is the weight of the bond that connects hidden unit h with input unit i and θ_h is the weight connecting hidden unit h to the input layer bias unit. The state O_m of output unit m is calculated by,

$$O_m(\varphi_h) = \frac{1}{1 + \mathrm{e}^{-\varphi_m}} \qquad (9a)$$

$$\varphi_m = \sum_h W_{mh} H_h + \theta_m \qquad (9b)$$

where W_{mh} is the bond that connects output unit m to hidden layer bias unit. The network calculated O_m values are within the range [0,1].

The training of the neural network is achieved by minimizing an error function E with respect to the bond weights $\{w_{hi}, W_{mh}\}$

$$E = \sum_p E_p = \frac{1}{2} \sum_p \sum_m \left(a_{pm} - O_{pm}\right)^2 \qquad (10)$$

where E_p is the error of the pth training pattern, defined as the set of descriptors and activity corresponding to the pth data points, or chemical compounds; a_{pm} corresponds to the experimentally measured value of the mth dependent variable, in this case the IC_{50}. These values were also scaled in the same manner as in Eq. 7.

One of the standard algorithms for minimizing E is the delta rule.[23,36,37] The algorithm is based on an iterative procedure for updating the weights of the neural network from their initially assigned random values. The equations for updating the weights are given below in (11a) and (11b):

$$W_{mh}^{n+1} = W_{mh}^{n} - \eta \frac{\partial E}{\partial W_{mh}} \qquad (11a)$$

$$w_{hi}^{n+1} = w_{hi}^{n} - \eta \frac{\partial E}{\partial w_{hi}} \qquad (11b)$$

In (11a, 11b) the superscript n indicates the consecutive iterations in the minimization procedure and η is the learning rate with values typically less than 1. Similar equations are used for θ_h and θ_m.

References and notes

1. Patrono, C.; Bachmann, F.; Baigent, C.; Bode, C.; De Caterina, R.; Charbonnier, B.; Fitzgerald, D.; Hirsh, J.; Husted, S.; Kvasnicka, J.; Montalescot, G.; Garcia Rodriguez, L. A.; Verheugt, F.; Vermylen, J.; Wallentin, L. *Eur. Heart J.* **2004**, *25*, 166.
2. Hsieh, P.-W.; Hwang, T.-L.; Wu, C.-C.; Chang, F.-R.; Wang, T.-W.; Wu, Y.-C. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 2786.
3. George, J. N. *Lancet* **2000**, *355*, 1531.
4. Nurtjahja-Tjendraputra, E.; Ammit, A. J.; Roufogalis, B. D.; Tran, V. H.; Duke, C. C. *Thromb. Res.* **2003**, *111*, 259.
5. Eriksson, A. C.; Whiss, P. A. *J. Pharmacol. Toxicol.* **2005**, *52*, 356.
6. Dudley, D. A.; Bunker, A. M.; Chi, L.; Cody, W. L.; Holland, D. R.; Ignasiak, D. P.; Janiczek-Dolphin, N.; McClanahan, T. B.; Mertz, T. E.; Narasimhan, L. S.; Rapundalo, S. T.; Trautschold, J. A.; Van Huis, C. A.; Edmunds, J. J. *J. Med. Chem.* **2000**, *43*, 4063.
7. Huang, W.; Zhang, P.; Zuckett, J. F.; Wang, L.; Woolfrey, J.; Song, Y.; Jia, Z. J.; Clizbe, L. A.; Su, T.; Tran, K.; Huang, B.; Wong, P.; Sinha, U.; Park, G.; Reed, A.; Malinowski, J.; Hollenbach, S. J.; Scarborough, R. M.; Zhu, B. Y. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 561.
8. (a) Jakobsen, P.; Ritsmar Pedersen, B.; Persson, E. *Bioorg. Med. Chem.* **2000**, *8*, 2095; (b) Costi, M. P. Second Joint Italian-Swiss Meeting on Medicinal Chemistry (ITCH-MC2005) Modena, Italy, Sep 12–16, 2005; *ARKIVOC* **2006**, 1.
9. Huang, L.-J.; Hsieh, M.-C.; Teng, C.-M.; Lee, K.-H.; Kuo, S.-C. *Bioorg. Med. Chem.* **1998**, *6*, 1657.
10. Ko, T.-C.; Hour, M.-J.; Lien, J.-C.; Teng, M.-C.; Lee, K.-H.; Kuo, S.-C.; Huang, L.-J. *Bioorg. Med. Chem. Lett.* **2001**, *11*, 279.
11. Vicini, P.; Amoretti, L.; Ballabeni, V.; Tognolini, M.; Barocelli, E. *Bioorg. Med. Chem.* **2000**, *8*, 2355.
12. Wang, T.-C.; Chen, I.-L.; Lu, P.-J.; Wong, C.-H.; Liao, C.-H.; Tsiao, K.-C.; Chang, K.-M.; Chen, Y.-L.; Tzeng, C.-C. *Bioorg. Med. Chem.* **2005**, *13*, 6045.
13. Truitt, A.; McNeill, G.; Vanderhoek, J. Y. *Biochim. Biophys. Acta* **1999**, *1438*, 239.
14. Bruno, O.; Brullo, C.; Schenone, S.; Ranise, A.; Bondavalli, F.; Barocelli, E.; Tognolini, M.; Magnanini, F.; Ballabeni, V. *Farmaco* **2002**, *57*, 753.

15. Bruno, O.; Schenone, S.; Ranise, A.; Bondavalli, F.; Barocelli, E.; Ballabeni, V.; Chiavarini, M.; Bertoni, S.; Tognolini, M.; Impicciatore, M. *Bioorg. Med. Chem.* **2001**, *9*, 629.
16. Lin, C.-N.; Hsieh, H.-K.; Ko, H.-H.; Hsu, M.-F.; Lin, H.-C.; Chang, Y.-L.; Chung, M.-I.; Kang, J.-J.; Wang, J.-P.; Teng, C.-M. *Drug Dev. Res.* **2001**, *53*, 9.
17. Roma, G.; Di Braccio, M.; Carrieri, A.; Grossi, G.; Leoncini, G.; Signorello, M. G.; Carotti, A. *Bioorg. Med. Chem.* **2003**, *11*, 123.
18. Ghoshal, N.; Achari, B.; Ghoshal, T. K. *Bioorg. Med. Chem. Lett.* **1997**, *7*, 877.
19. Tanaka, A.; Nakamura, K.; Nakanishi, I.; Fujivara, H. *J. Med. Chem.* **1994**, *37*, 4563.
20. Katritzky, A. R.; Dobchev, A. D.; Fara, D. C.; Karelson, M. *Bioorg. Med. Chem.* **2005**, *13*, 6598.
21. Katritzky, A. R.; Ignatchenko, E. S.; Barcock, R. A.; Lobanov, V. S.; Karelson, M. *Anal. Chem.* **1994**, *66*, 1799.
22. Karelson, M. In *Molecular Descriptors in QSAR/QSPR*; J. Wiley & Sons: New York, 2000.
23. Haykin, S. In *Neural Networks*; Pearson Education: Delhi, 2004.
24. <www.hyper.com/>.
25. Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902.
26. <www.codessa-pro.com/>.
27. Katritzky, A. R.; Lobanov, V. S.; Karelson, M. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 28.
28. Katritzky, A. R.; Tämm, K.; Kuanar, M.; Fara, D. C.; Oliferenko, A.; Oliferenko, P.; Huddleston, J. G.; Rogers, R. D. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 136.
29. Katritzky, A. R.; Fara, D. C.; Yang, H.; Karelson, M.; Suzuki, T.; Solov'ev, V. P.; Varnek, A. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 529.
30. Katritzky, A. R.; Fara, D. C.; Kuanar, M.; Hur, E.; Karelson, M. *J. Phys. Chem.* **2005**, *109*, 10323.
31. Katritzky, A. R.; Kuanar, M.; Fara, D. C.; Karelson, M.; Acree, W. E. *Bioorg. Med. Chem.* **2004**, *12*, 4735.
32. Katritzky, A. R.; Kuanar, M.; Fara, D. C.; Karelson, M.; Acree, W. E.; Solov'ev, V. P.; Varnek, A. *Bioorg. Med. Chem.* **2005**, *13*, 6450.
33. (a) Maran, U.; Karelson, M.; Katritzky, A. R. *Quant. Struct.-Act. Relat.* **1999**, *18*, 3; (b) Beteringhe, A.; Balaban, A. T. *ARKIVOC* **2004**, 163.
34. Katritzky, A. R.; Fara, D. C.; Karelson, M. *Bioorg. Med. Chem.* **2004**, *12*, 3027.
35. Katritzky, A. R.; Dobchev, D. A.; Hür, E.; Fara, D. C.; Karelson, M. *Bioorg. Med. Chem.* **2005**, *13*, 1623.
36. Zupan, J.; Gasteiger, J. In *Neural Networks in Chemistry and Drug Design*, 2nd ed.; Wiley-VCH: Weinheim, 1999.
37. Masters, T. In *Practical Neural Network Recipes in C++*; Academic: Boston, 1993.