Available online at www.sciencedirect.com
SCIENCE DIRECT®
Bioorganic & Medicinal Chemistry 13 (2005) 6598–6608
Bioorganic & Medicinal Chemistry

QSAR studies on 1-phenylbenzimidazoles as inhibitors of the platelet-derived growth factor

Alan R. Katritzky,[a,*] Dimitar A. Dobchev,[a,b] Dan C. Fara[a] and Mati Karelson[b]

[a]*Center for Heterocyclic Compounds, Department of Chemistry, University of Florida, Gainesville, FL 32611, USA*
[b]*Department of Chemistry, Tallinn University of Technology, Ehitajate tee 5, Tallinn 19086, Estonia*

Received 9 May 2005; revised 29 June 2005; accepted 30 June 2005
Available online 17 October 2005

Abstract—This work is devoted to the development of quantitative structure–activity relationship (QSAR) models of the biological activity of 123 1-phenylbenzimidazoles as inhibitors of the PDGF receptor. The molecular features are represented by chemical descriptors that have been calculated on geometrical, topological, quantum mechanical, and electronic basis by using CODESSA PRO. The obtained models, linear (multilinear regression) and nonlinear (artificial neural network), are aimed to link the structures to their reported activity $\log 1/IC_{50}$. The former model can be used for physico-chemical interpretation, while the latter possesses a superior predictive ability.

1. Introduction

The call for the discovery of less toxic, more selective, and more effective agents to treat cancer becomes ever more urgent. Inhibition of angiogenesis continues to be one of the mainstreams in current cancer drug activity. Insights into the biology of tumor angiogenesis have led to the identification of various molecules that promote tumor development. Of particular interest are such factors as the platelet-derived growth factor (PDGF), which plays a major role as a regulator of cell growth.[1,2] This factor and its corresponding receptor tyrosine kinases (PDGFR) have become important targets for inhibition of the proliferation of endothelial cells, the main component of blood vessels. Their ability to promote cell proliferation and migration and to induce changes in the pattern of protein synthesis and secretion has made growth factors candidates for therapeutic approaches to treat pathophysiological conditions in which growth factors seem to be missing or inhibited. Binding of PDGF to its transmembrane receptor (PDGFR) results in tyrosine phosphorylation of active substrates in various biochemical pathways, including the involvement of phosphatidylinostiol 3-kinase.

Various groups of compounds have been reported as selective inhibitors of PDGFR.[3] The 3-arylquinolines are one such class[4,5] that display diverse IC_{50} values for inhibition of autophosphorylation of PDGFR derived from vascular smooth muscle cells, acting by inhibition of ATP binding. Some 3-arylquinolines show an IC_{50} of 300 nM for inhibition of autophosphorylation of PDGFR in a 3T3 cell line;[6] others are reported to inhibit PDFG-mediated signaling and are used in clinical trial for the treatment of glioma.[7]

Phenylaminopyramidines,[8,9] which comprise another class of inhibitors, are capable of inhibiting cellular PDFG-receptor autophosphorylation at nanomolar concentration. Potential clinical application of these inhibitors includes their use as anticancer agents, as well as drugs for the treatment of the conditions characterized by inappropriate fibroblast and vascular intimal hyperplasia.

1-Phenylbenzimidazoles[10,11] were recently reported as promising selective inhibitors of PDGF, with clear evidence of the relationship between their molecular features and their inhibitory activity.

Few structure–activity relationship studies involving 1-phenylbenzimidazoles have been published[12–15] and the QSAR models reported were not completely satisfactory. We have now applied quantitative structure–property

Keywords: QSAR; Neural network; CODESSA PRO; PDGF receptor.
* Corresponding author. Tel.: +1 352 392 0554; fax: +1 352 392 9199; e-mail: katritzky@chem.ufl.edu

0968-0896/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2005.06.067

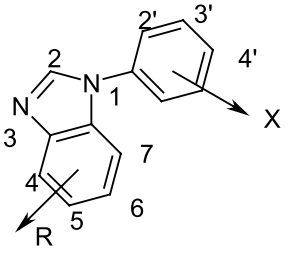

Figure 1. Main skeleton with the different functional positions for 1-phenylbenzimidazoles.

activity relationship (QSPR/QSAR) methodology[16] to these inhibitors.

Shen et al.[17] proposed several QSAR models for the estimation of the inhibitory activities of 1-phenylbenzimidazoles as PDGFR inhibitors. They considered compounds with various substituents on the benzimidazole ring (see Fig. 1). Several correlation equations were obtained with R^2 in the range 0.66–0.73 for the data set of 75 compounds. Further QSAR models were obtained by Zhong et al.,[18] who extended this set by taking into account functional groups in the 1-phenylbenzimidazoles used in Ref. 17. Zhong et al. employed variable connectivity indices as the main descriptors involved in their main QSAR model ($R^2 = 0.78$ for the training set of 55 1-phenylbenzimidazole derivatives). This work also used an external test set of 24 compounds and satisfactory results were obtained in terms of $R^2 = 0.75$.

Artificial neural networks (ANNs)[19–21] have become an important modeling technique for QSAR and QSPR, and artificial neural network (ANN) modeling has been applied in numerous application areas of chemistry and pharmacy.[22–24] The mathematical adaptability of ANN commends them as a powerful tool for pattern classification and building predictive models. A particular advantage of ANNs is their inherent ability to incorporate nonlinear dependencies between the dependent and independent variables without using an explicit mathematical function.

Guha and Jurs[25] recently investigated the biological activity of 79 derivatives of 4-piperazinylquinazoline PDGFR inhibitors using linear models and neural networks. For a NN model 7-3-1, they obtained R^2 for the training and prediction sets of 0.93 and 0.61, respectively. The NN model had better predictive ability than the linear model.

Methodology for a general QSAR/QSPR approach has been developed and coded as the CODESSA PRO software package. CODESSA PRO enables the calculation of numerous quantitative descriptors solely on the basis of molecular structural information (Hansch-type approach).[26,27] Research using CODESSA PRO has successfully correlated and predicted various physical properties,[28] including gas chromatographic properties, melting and boiling points, solvent scales, and refractive indexes.[29] Recent examples include QSPR treatments of (i) the binding energies for 1:1 complexation systems between various organic guest molecules and β-cyclodextrin,[30] (ii) the in vitro minimum inhibitory concentration (MIC) of 3-aryloxazolidin-2-one antibacterials to inhibit the growth of *Staphylococcus aureus*,[31] and (iii) partition coefficients of drugs between human breast milk and plasma.[32]

The present study involves two main treatments of the logarithms of the effectiveness of 1-phenylbenzimidazoles as inhibitors of the platelet-derived growth factor, $\log(1/IC_{50})$: (i) QSAR modeling, by multilinear regression performed with the CODESSA PRO program that applies up to 800 different constitutional, geometrical, topological, electrostatic, quantum chemical, and thermodynamic molecular descriptors and (ii) nonlinear modeling, performed using artificial neural networks (ANNs) with backpropagation learning algorithm and sigmoid activation function developed in-house. In both these treatments, all descriptors used are derived solely from a molecular structure and do not require experimental data or expensive theoretical calculations to be obtained.

Here, we show that a combination of these two different approaches (multilinear and nonlinear) leads to pertinent QSAR models, and their joint application improves the robustness of predictions.

2. Experimental data

Experimental data for the limiting concentrations to inhibit the phosphorylation of a random glutamate/tyrosine (4:1) copolymer by PDFGR proteins (IC_{50}) were taken from Refs. 11, 12 and 13. The combined data set consists of 123 1-phenylbenzimidazoles and their IC_{50} values with reported experimental errors of 15%. Logarithmic $1/IC_{50}$ was used in the QSAR treatment. All 1-phenylbenzimidazoles possess the skeleton of Figure 1. Table 1 lists the substituents and their positions for all 123 compounds.

3. Molecular modeling

Three-dimensional conversions and pre-optimization were performed using the molecular mechanics (MM+) implemented in the HyperChem 7.5 package.[33]

Final geometry optimization of the molecules was carried out using the semi-empirical quantum-mechanical AM1 parameterization.[34] The optimized geometries were loaded into CODESSA PRO software.[35] Overall, more than 800 theoretical descriptors were calculated. These descriptors can be classified into several groups: (i) constitutional, (ii) topological, (iii) geometrical, (iv) thermodynamic, (v) quantum chemical, and (vi) charge-related descriptors.

4. Multilinear and nonlinear approaches—general algorithms

4.1. Multilinear regression modeling

An important stage of the multilinear regression QSAR methodology is the search for the best multilinear

Table 1. Positions of the substituents

No	R	X
1	H	H
2	4-OMe	H
3	4-OH	H
4	5-Me	H
5	5-OMe	H
6	5-OH	H
7	5-Cl	H
8	5-COOH	H
9	5-COOMe	H
10	5-$CONH_2$	H
11	5-NO_2	H
12	5-COMe	H
13	5-CHO	H
14	5-OC_3H_7	H
15	5-OC_2H_5	H
16	5-$OCH(Me)_2$	H
17	5-OC_4H_9	H
18	5-OCH_2CHCH_2	H
19	5-$O(CH_2)_4OH$	H
20	5-OCH_2(oxiranyl)	H
21	5-$OCH_2CH(OH)CH_2OH$	H
22	5-$O(CH_2)_2OH$	H
23	5-$O(CH_2)_2N(Me)_2$	H
24	5-$O(CH_2)_3N(Me)_2$	H
25	5-$O(CH_2)_4N(Me)_2$	H
26	5-$O(CH_2)_2$N morph	H
27	5-$O(CH_2)_3$N morph	H
28	5-$O(CH_2)_4$N morph	H
29	5-SH	H
30	5-SMe	H
31	5-$OCSN(Me)_2$	H
32	6-Me	H
33	6-OMe	H
34	6-OH	H
35	6-Cl	H
36	6-COOH	H
37	6-COOMe	H
38	6-$CONH_2$	H
39	6-NO_2	H
40	6-NH_2	H
41	7-OMe	H
42	4,5-DiOH	H
43	4-OH, 5-OMe	H
44	4-$CH_2CH(Me)O$-5	H
45	5,6-$(OH)_2$	H
46	5,6-Me_2	H
47	5-OCH_2O-6	H
48	5-OMe, 6-Me	H
49	5-OH, 6-Me	H
50	5-OMe, 6-COOH	H
51	5-OH, 6-COOH	H
52	5-OMe, 6-COOMe	H
53	5-OMe, 6-CH_2OH	H
54	5-OMe, 6-CHO	H
55	5-NH_2	H
56	5-Aza	H
57	7-Aza	H
58	H	3′-Me
59	H	3′-OMe
60	H	3′-OH
61	H	3′-Cl
62	H	3′-NO_2
63	H	3′-NH_2
64	H	3′-COMe
65	H	3′-CHO
66	H	4′-OMe
67	H	4′-OH
68	H	4′-Cl
69	H	4′-COOMe
70	H	4′-$CONH_2$
71	H	4′-NO_2
72	H	4′-NH_2
73	H	4′-COMe
74	H	4′-CHO
75	H	4′-CN
76	H	4′-Aza
77	5-OMe	2′-Thienyl
78	5-OMe	3′-Thienyl
79	5-OMe	4′-NH_2
80	4-COOH	H
81	4-COOMe	H
82	4-$CONH_2$	H
83	4-NO_2	H
84	4-NH_2	H
85	7-Me	H
86	7-OH	H
87	7-Cl	H
88	7-COOH	H
89	7-COOMe	H
90	7-$CONH_2$	H
91	7-NO_2	H
92	7-NH_2	H
93	4-OMe, 5-OH	H
94	4,5-DiOMe	H
95	4-Br, 5-OH	H
96	4-Br, 5-OCH_2CHCH_2	H
97	4-CH_2CHCH_2, 5-OH	H
98	5-$S(CH_2)_3$N morph	H
99	4-Me	H
100	4-Cl	H
101	2-Me	H
102	2-OH	H
103	2-NH_2	H
104	H	2′-Me
105	H	2′-OMe
106	H	2′-OH
107	H	2′-Cl
108	H	2′-COOH
109	H	2′-COOEt
110	H	2′-$CONH_2$
111	H	2′-NO_2
112	H	2′-NH_2
113	H	2′-COMe
114	H	2′-CHO
115	H	2′-CN
116	H	3′-COOH
117	H	3′-COOEt
118	H	3′-$CONH_2$
119	H	3′-CN
120	H	4′-Me
121	H	4′-COOH
122	H	2′-Aza
123	H	3′-Aza

equation among a given pool of descriptors. In other words, Eq. 1 correlates the best inhibitory activity (A) with a certain number n of molecular descriptors (D_i) weighted by the regression coefficients b_i:

$$A = b_0 + \sum_{i=1}^{n} b_i D_i. \quad (1)$$

The best multilinear regression method (BMLR),[36,37] encoded in CODESSA PRO software, was used to select significant descriptors for building multilinear QSAR models. The treatment started with a reduction in the number of molecular descriptors. If two descriptors were highly correlated, then only one descriptor was selected; those descriptors with insignificant variance were also rejected. This helps to speed up the selection of descriptors and reduce the probability of including by chance any unrelated descriptors.

The strategy used to develop physically meaningful multilinear QSAR equations from a very large pool of descriptors is a combination of the multilinear regression and forward selection procedures. This strategy involved the following steps:

(1) Detection of all orthogonal pairs of descriptors i and j from the given descriptor space. Pairs of descriptors with a correlation coefficient $R_{ij}^2 > 0.5$ were considered intercorrelated and such pairs were eliminated at this stage.
(2) From a complete set of all two-parameter regression equations of orthogonal pairs, only the 400 possessing the highest R^2 value two-parameter equations were used.
(3) Search for superior multiparameter regression equations: for each descriptor pair, retained in the previous step, additional non-collinear descriptor vectors were successively added, and the appropriate $(n + 1)$-parameter regression treatment was carried out. When the Fisher criterion F (or cross-validation coefficient R_{cv}), obtained for any of these correlations, was lower than for the best correlation of the previous rank (n), the latter was designed as the final result and the search was given up. Otherwise, the descriptor sets with the highest coefficient of determinations were stored and the current step was repeated with the number of parameters (descriptors) increased by one $(n + 2)$.

The final result had therefore the maximum value of the Fisher criterion and the highest cross-validated coefficient of determination.

A major decision in developing successive QSAR is when to stop adding descriptors to the model during the stepwise regression procedure. A simple technique to control the model expansion is the so-called 'breaking point' in the improvement of the statistical quality of the model, by analyzing the plot of the number of descriptors involved in the obtained models versus squared correlation coefficient values corresponding to those models. Frequently, improvement of the statistical quality of the regression model is less significant ($\Delta R^2 < 0.02$) after a certain number of independent variables in the model ('breaking point'). Consequently, the model corresponding to the breaking point is considered the best/optimum model.

To validate the models internally, the parent data set was divided into three subsets (a, b, c): the first, fourth, seventh, etc., data points go into the first subset (a), the second, fifth, eighth, etc., into the second subset (b), and the third, sixth, ninth, etc., into the third subset (c). Then, three training sets A, B, and C were prepared as combinations of two subsets (a and b), (a and c), and (b and c), respectively. The remaining subsets (c, b, and a, respectively) become the corresponding test sets.

For each of the training sets, the correlation equation was derived with the same descriptors. Then, the equation obtained was used to predict $\log(1/IC_{50})$ values for the compounds from the corresponding test set.

The efficiency of QSAR models to predict $\log 1/IC_{50}$ value was estimated using the cross-validation (leave-one-out method).[38]

4.2. Nonlinear modeling

An artificial neural network (ANN) is a biologically inspired computer program designed to simulate the way in which the human brain processes information. ANNs are composed of a number of single processing elements (PEs) or units (nodes). Each PE has weighted inputs, transfer function, and one output. PEs are connected with coefficients (weights) and are organized in a layered topology as follows: (i) the input layer, (ii) the output layer, and (iii) the hidden layers between them. The number of layers and the number of units in each layer determine the functional complexity of the ANN.

In this work, a backpropagation network[39,40] 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 2. 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).

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. 2:

$$I_i = \frac{D_i - D_i(\min) + 0.1}{D_i(\max) - D_i(\min) + 0.1}, \qquad (2)$$

where D_i is the value of the ith descriptor, $D_{i(\max)}$ and $D_{i(\min)}$ are its maximum and minimum values, respec-

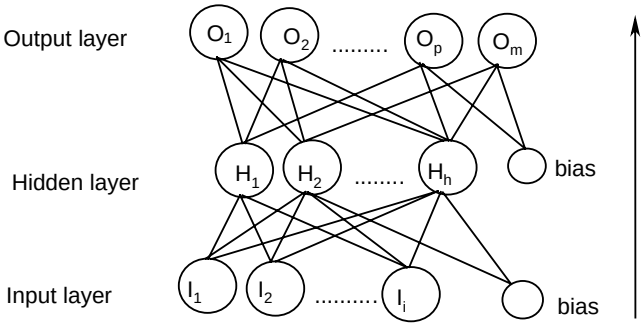

Figure 2. Three layer backpropagation neural network.

tively. The state H_h of each hidden unit is calculated by the squashing (sigmoid, logistic) function:

$$H_h(\varphi_h) = \frac{1}{1+e^{-\varphi_h}}, \tag{3a}$$

$$\varphi_h = \sum_i w_{hi} I_i + \theta_h, \tag{3b}$$

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

$$O_m(\varphi_h) = \frac{1}{1+e^{-\varphi_m}}, \tag{4a}$$

$$\varphi_m = \sum_h W_{mh} H_h + \theta_m, \tag{4b}$$

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].

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

$$E = \sum_p E_p = \frac{1}{2} \sum_p \sum_m (a_{pm} - O_{pm})^2, \tag{5}$$

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 compound; a_{pm} corresponds to the experimentally measured value of the mth dependent variable, in this case it is the IC_{50}. These values were also scaled in the same manner as in Eq. 2.

One of the standard algorithms for minimizing E is the delta rule.[39,40] 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 Eqs. 6a and 6b:

$$W_{mh}^{n+1} = W_{mh}^{n} - \eta \frac{E}{W_{mh}} \tag{6a}$$

$$w_{hi}^{n+1} = w_{hi}^{n} - \eta \frac{E}{W_{hi}} \tag{6b}$$

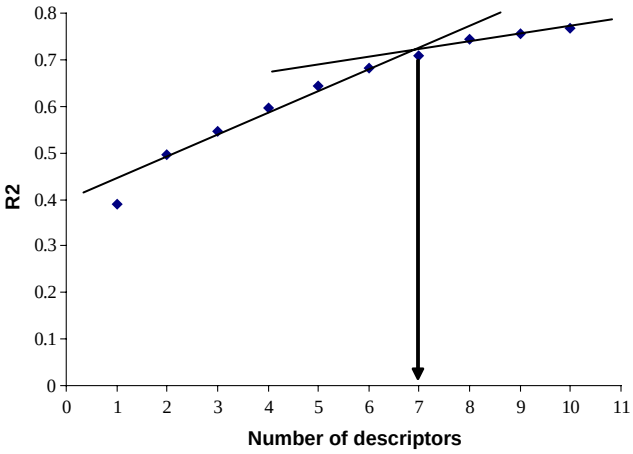

Figure 3. 'Breaking point' for choice of descriptors involved in the model in Table 2.

In Eqs. 6a and 6b, 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.

5. Results and discussion

5.1. Multilinear QSAR model

A selection of reliable data for multilinear regression analysis was required first. In the data taken for regression, for 45 out of 123 compounds, only the upper limit of the IC_{50} was reported; consequently, only the remaining 78 compounds with precise experimental values were utilized for building the MLR QSAR model.

Next, by using 'the best multilinear' method encoded in CODESSA PRO, up to ten multilinear models were obtained. The best model for all 123 compounds that was found according to the breaking rule is shown in Table 3. As can be seen from Figure 3, the optimum number of descriptors is seven. The coefficient of determination of the model for all 78 compounds is $R^2 = 0.71$ ($R^2_{cv} = 0.65$) and the Fisher criterion is $F = 24.33$. The method allows constraints for the intercollinearity coefficient to be set. In our case, the intercollinearity coefficient was set to be less than $R^2_{IC} = 0.70$ to extend the search space of the descriptors to find better QSAR models. For all descriptors in the model, the coefficients are $R^2_{IC} < 0.62$.

Table 2. The main multilinear QSAR model obtained for 78 1-phenylbenzimidazoles ($R^2 = 0.71$, $F = 24.33$)[a]

No.	X	$\pm X$	t test	R^2	R^2_{cv}	S^2	Descriptor
0	−40.187	8.469	−4.745				Intercept
1	0.011	0.001	7.181	0.391	0.367	0.341	PPSA2 Total charge weighted PPSA (Zefirov PC), D_1
2	0.190	0.023	8.176	0.497	0.465	0.530	Min e–e repulsion for bond C–C, D_2
3	−13.373	2.549	−5.246	0.545	0.504	0.260	RNCG charge (QMNEG/QTMINUS) (MOPAC), D_3
4	−0.504	0.112	−4.477	0.597	0.546	0.234	Min exchange energy for bond C–C, D_4
5	1.488	0.386	3.84	0.643	0.589	0.210	Tot entropy (300 K)/n atoms, D_5
6	−10.645	3.086	−3.449	0.681	0.624	0.191	Relative number of N atoms, D_6
7	8.061	3.100	2.600	0.708	0.651	0.176	Max bonding contribution of one MO, D_7

[a] All descriptor definitions are given in Supplementary data.

Table 3. Experimental and predicted log(1/IC_{50}) for the multilinear QSAR model for the data that consist of 78 diverse values and 45 values given with their upper limit

No	Exp.	MLR model
1	5.03	4.69
2	4.3	4.76
3	4.85	5.17
4	5.36	5.35
5	6.37	5.41
6	6.36	5.6
7	5.4	4.96
8	5.03	5.31
9	6.08	5.55
10	4.8	4.98
11	4.8	4.57
12	6.07	5.21
13	6.37	5.78
14	6.6	6.34
15	6.62	6.24
16	5.51	6.1
17	5.89	5.95
18	6.22	6.19
19	6.35	6.88
20	6.5	5.56
21	6.51	6.37
22	6.19	6.21
23	5.82	5.9
24	6.82	6.6
25	6.8	6.68
26	6.14	6.36
27	6.77	6.7
28	6.57	6.68
29	5.48	5.51
30	6.13	6.14
31	5.34	4.88
32	4.4	5.39
33	5.19	5.17
34	5.68	5.39
35	5.27	4.93
36	4.3	5.17
37	4.89	4.65
38	4.6	4.52
39	4.3	4.62
40	4.64	5.11
41	4.43	4.88
42	4.6	4.95
43	5.15	5.19
44	4.54	4.82
45	5.64	5.45
46	5.92	5.25
47	5.66	5.81
48	6	5.56
49	5.6	6.1
50	4.68	5.11
51	5.37	5.44
52	6.06	5.49
53	6.43	6.58
54	6	5.98
55	5.57	5.42
56	5	4.82
57	4.55	4.93
58	4.55	5.26
59	4.6	5.01
60	5.42	5.25
61	4.33	4.55
62	4.8	4.66
63	5.44	5
64	4.72	4.6
65	5.17	5.2
66	4.89	5.38
67	5.75	5.45
68	4.3	4.79
69	5.14	4.84
70	4.64	4.61
71	4.52	4.53
72	5.25	5.38
73	4.62	4.96
74	4.89	5.31
75	4.8	5.14
76	4.92	4.59
77	5.6	5.79
78	6.16	6.35
79	<4.3	6.19
80	<4.3	5.1
81	<4.3	4.69
82	<4.3	3.88
83	<4.3	4.18
84	<4.3	4.14
85	<4.3	4.87
86	<4.3	4.76
87	<4.3	4.32
88	<4.3	4.81
89	<4.3	4.36
90	<4.3	4.22
91	<4.3	4.16
92	<4.3	4.51
93	<4.3	4.68
94	<4.3	5.17
95	<4.3	4.15
96	<4.3	6.43
97	<4.3	5.38
98	<4.3	6.22
99	<4.3	5.38
100	<4.3	4.23
101	<4.3	4.1
102	<4.3	4.25
103	<4.3	4.99
104	<4.3	5.14
105	<4.3	4.07
106	<4.3	4.37
107	<4.3	4.32
108	<4.3	4.21
109	<4.3	4.01
110	<4.3	4.39
111	<4.3	4.4
112	<4.3	4.6
113	<4.3	4.48
114	<4.3	4.37
115	<4.3	4.28
116	<4.3	4.69
117	<4.3	4.81
118	<4.3	4.42
119	<4.3	4.67
120	<4.3	4.56
121	<4.3	4.85
122	<4.3	4.39
123	<4.3	4.73

Table 2 also shows an increase in R^2 and R^2_{cv} with an increase in the number of descriptors. The cross-validation coefficient R^2_{cv} (leave-one-out method) is a measure of relative predictivity of the current data. In addition,

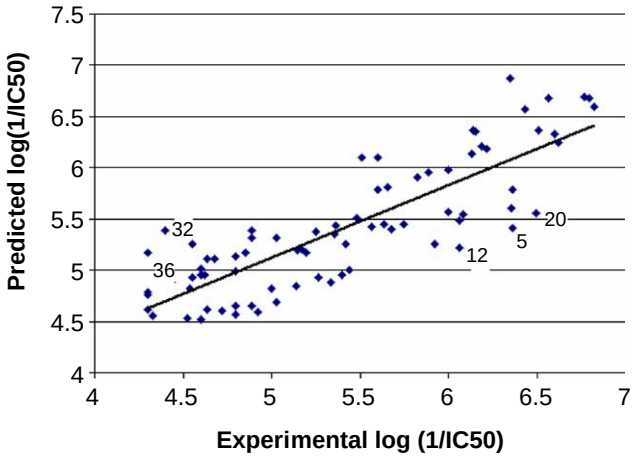

Figure 4. Experimental versus predicted log(1/IC50) for the seven-parameter multilinear QSAR model of Table 2.

all the descriptors possess a t-value >3.00, except for descriptor D_7. However, D_7 increases the model R^2 value by 0.02 and thus agrees with our breaking rule.

The linear plot of experimental and predicted log(1/IC_{50}) of the model from Table 2 is given in Figure 4. The predicted values are given in Table 3 along with the experimental ones.

This model includes seven descriptors (D_1–D_7) relating the activity of the compounds to PDGFR inhibition. The most statistically significant descriptor, according to the t test in the table is, min e–e repulsion for bond C–C, D_2. This is a quantum mechanical descriptor that characterizes the energy of the electrostatic interactions between the chemically bonded atoms. It can be of importance in determining the conformational change within a molecule, which, in turn, may affect a inhibitory activity. In addition, the same conjecture is valid for descriptor D_4. Descriptors D_1 and D_3 are charge-related descriptors that could be connected to the charge distribution in the molecule, which is likely to govern the electrostatic interactions with the PDFG receptor. The next two descriptors D_5 and D_6 are total entropy (300 K)/number of atoms and relative number of N atoms, respectively. The latter descriptor is not unexpected, since all the compounds include N atoms. D_5 is related to the conformational changes in the molecule. Therefore, the latter two descriptors could also be addressed to the processes of the inhibitory activity phenomenon. The last descriptor D_7 relates the quantum chemically calculated bond orders that contain information on the molecular stability.

Five outliers were detected, according to the error of the model. Two of the biggest outliers are compounds **32** and **36** that were predicted as active compounds though their experimental values show them to be inactive. The reason for this faulty prediction is that the values for the descriptor D_1 seem to be apparently overestimated.

However, most of the values predicted by the model are in a good agreement with experiment.

The efficiency of the QSAR model to predict log IC_{50} value was also estimated using the internal cross-validation. The correlation coefficients and standard deviations of linear correlations between experimental and predicted for test sets of log 1/IC_{50} values were also calculated and the results are shown in Table 4. The average values of R^2 (Fit) and R^2 (Pred) are very close (0.684 and 0.698, respectively), which suggests a relatively stable predictivity of the model for these data. In addition, as can be seen from Table 4, the average sample variances are close to each other.

Another validation challenge was used for the MLR model, that is, we predicted values of those 45 inactive compounds that are given with their upper limits of the log 1/IC_{50}. The resulted predictions of the log 1/IC_{50} are given in Table 3. The overall prediction of the log 1/IC_{50} values is quite satisfactory having in mind the experimental error. However, there were still three compounds (**79**, **96**, and **98**) that were predicted as active. A possible reason is that the descriptor D_2 was somewhat overestimated for these compounds. Also, five compounds (**80**, **94**, **97**, **99**, and **104**) were predicted with values larger than 5.00.

5.2. Nonlinear QSAR model

In this study, we used ANN methodology for classification of the log(1/IC_{50}). The experimental values (123 data points) of log(1/IC_{50}) were divided into eight classes according to the experimental error ± 0.31 (15% for IC_{50}) and the range (4.30–6.82) of all values. Thus, the first class includes compounds for which the experimental values are less than 4.61, second class 4.61–4.92, and so on, up to eighth class, as shown in Table 5. The higher the class the more active the compound.

In the first stage, before the neural network treatment began, both experimental classes and descriptor values were normalized to a range 0–0.9 (see Eq. 2). All networks had one input layer, one hidden layer, and one output layer. The next stage of ANN modeling is the selection of significant descriptors from a large descriptor pool and the division of the available data into

Table 4. Internal validation of the QSAR model

Training set	N	R^2 (Fit)	R^2_{cv} (Fit)	S^2 (Fit)	Test set	N	R^2 (Pred)	S^2 (Pred)
A + B	52	0.672	0.638	0.223	C	26	0.716	0.207
A + C	52	0.653	0.613	0.238	B	26	0.694	0.218
B + C	52	0.728	0.711	0.169	A	26	0.683	0.111
Average		0.684	0.654	0.210			0.698	0.179

Table 5. Classification of the log(1/IC_{50}) according to the ranges

Range	Class
<4.3–4.61	Class 1
4.61–4.92	Class 2
4.92–5.23	Class 3
5.23–5.54	Class 4
5.54–5.85	Class 5
5.85–6.17	Class 6
6.17–6.48	Class 7
6.48–6.82	Class 8

training and validation sets. First, one-third (41) of the compounds were randomly selected and used as validation set for the networks to avoid over-training of the models. Second, values calculated of all CODESSA PRO descriptors (811) were examined for intercorrelations. Descriptors with high intercorrelations ($r^2 > 0.5$), small variances ($<10^{-6}$), and descriptors for which no values available for all structures were excluded from further treatment. Thus, the descriptor pool was reduced approximately by 80% to 161 descriptors. From this reduced pool of descriptors were excluded 72 descriptors since they showed random variations with the property, as shown by exploring the scatter plots. The final descriptor pool was reduced to 89 descriptors for which sensitivity-stepwise analysis was performed by building the ANN models (with 1-1-1 architecture) for each relevant descriptor. Those descriptors (around 10) that showed the lowest prediction error at the ANN output were chosen for building the optimum ANN model.

Before the training process started, the weights of the network were initialized with random values between −0.5 and 0.5. During the training stage, the weights were adjusted, according to the output prediction error by using the backpropagation algorithm. The validation set error (also R^2) was monitored to avoid over-training of the ANN and to stop the training process.

We found that a five descriptor model (5-4-1) was appropriate for the log (1/IC_{50}) property. The root-mean-squared (RMS) error for the training and validation data is 0.77 and 1.54, respectively. In addition, an exploration of the standard deviations of the neural network models with different numbers of hidden units was performed. The six descriptor models (6-4-1) did not show any significant improvement over the five-descriptor models (RMS = 1.11). The same result was found for the 6-5-1 models with increased hidden units (RMS = 0.95). The predicted classes of log(1/IC_{50}) obtained are given in Table 6.

In Figure 5 shows the confusion matrix for the training set. As can be seen, the most predicted values lie on the left diagonal showing a good accordance for the training set. The percentage of correct predicted classes (as the ratio between the whole number of compounds in a certain class and the exactly predicted ones) was as follows: class 1—78%, class 2—40%, class 3—40%, class 4—66%, class 5—80%, class 6—57%, class 7—71%, and class 8–100%. It is interesting that class 8 is predicted as 100% for the two compounds that fall in this class. The first

Table 6. Predicted and experimental classes for the training and validation sets of compounds by using ANN models 5-4-1

Compound	Exp. Class	Pred. Class	Compound	Exp. Class	Pred. Class
Training set					
1	3	1	54	6	5
2	1	3	56	3	2
3	2	2	60	4	3
4	4	2	61	1	2
5	7	6	62	2	2
6	7	6	63	4	2
7	4	2	64	2	1
8	3	3	65	3	2
9	6	5	66	2	5
10	2	3	67	5	4
12	6	4	68	1	2
13	7	5	69	3	1
14	8	7	71	1	1
15	8	7	72	4	4
16	4	7	73	2	2
18	7	5	74	2	4
19	7	7	77	5	7
20	7	6	79	8	7
21	8	7	81	1	1
22	6	7	82	1	1
25	8	8	83	1	1
26	6	6	84	1	2
27	8	8	86	1	2
28	8	8	87	1	1
29	4	4	88	1	1
31	4	3	89	1	1
32	1	1	90	1	1
33	3	3	92	1	1
34	5	5	95	1	1
35	4	2	96	1	2
36	1	1	98	1	1
37	2	1	101	1	1
38	1	1	105	1	1
40	2	2	106	1	1
41	1	1	107	1	2
42	1	1	110	1	1
43	3	3	111	1	1
45	5	5	113	1	1
49	5	5	114	1	1
50	2	3	118	1	1
52	6	6	120	1	2
Validation set					
11	2	1	85	1	2
17	6	8	91	1	1
23	5	7	93	1	4
24	8	8	94	1	5
30	6	6	97	1	2
39	1	1	99	1	2
44	1	3	100	1	3
46	6	2	102	1	2
47	5	1	103	1	1
48	6	5	104	1	2
51	4	4	108	1	1
53	7	7	109	1	1
55	5	2	112	1	1
57	1	1	115	1	2
58	1	2	116	1	1
59	1	2	117	1	1
70	2	1	119	1	2
75	3	3	121	1	1
76	2	1	122	1	1
78	6	7	123	1	1
80	1	1			

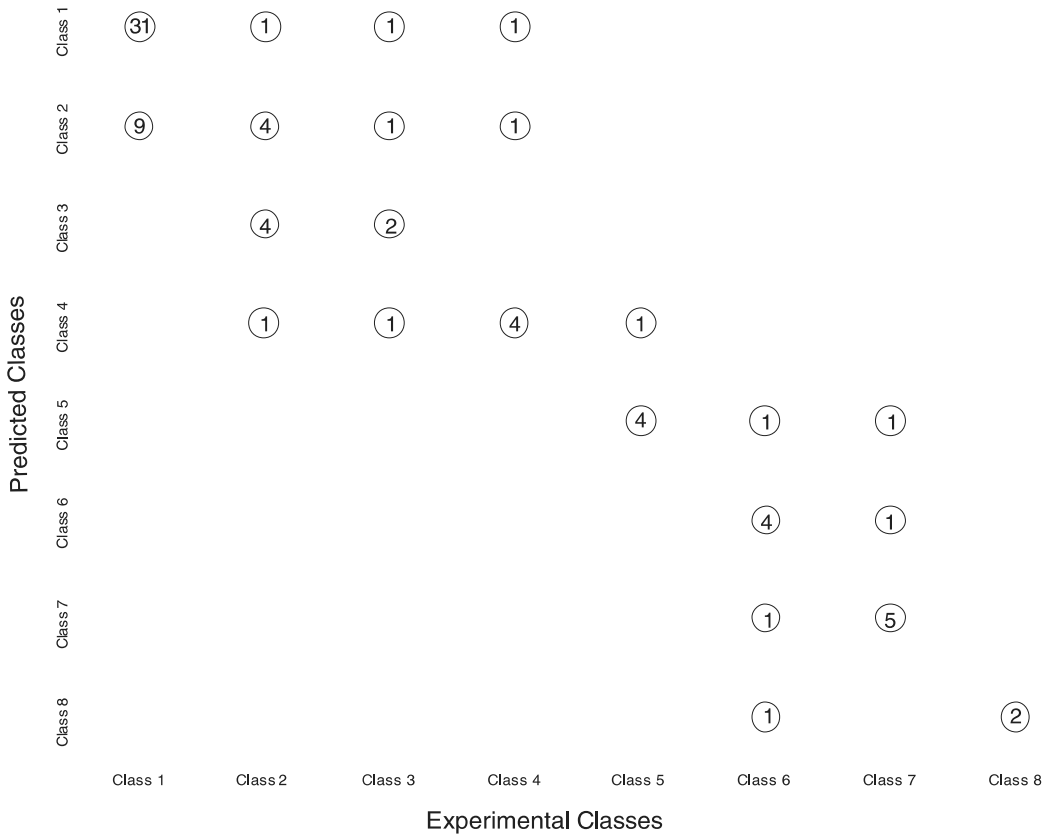

Figure 5. Distribution of predicted versus experimental activity classes for the main five descriptor NN model for the training set.

class comprises the largest number of compounds (40). Thirty-one are correctly predicted as class 1, and all the others as class 2. In contrast, prediction for the classes 2, 3, 4, 6, and 7 was spread out as more than one class. However, the number of confused compounds is much smaller than that correctly predicted.

Table 6 also gives the predicted classes of the validation set for the ANN model from which the maximum possible prediction for the descriptors involved in the model was obtained.

The confusion matrix for the validation set is given in Figure 6. The same pattern is also seen here. The most accurately predicted classes lie on the left diagonal. The percentage of the accurate predicted classes are: class 1—81%, class 2—66%, class 3—100%, class 4—100%, class 5—33%, class 6—25%, class 7—100%, class 8—100%. It is important to note that, for only 8 compounds out of 39 (20%) does the predicted class differ by more than one from the correct experimental class. Therefore, the prediction that a compound belongs within one class was achieved for 80% of compounds, which is a remarkably good result. Since, the RMS of the validation set is bigger than the training set, the accuracy is less. As can be noted from Figure 6, classes 5 and 6 were predicted correctly to within two classes, except for single compounds in classes 1 and 2, respectively. However, the class 8 (the most active) consists of only one compound that was predicted correctly. A feature, that also holds for classes 3 and 4.

The ANN model included the following descriptors used as inputs: moment of inertia B, D_8, min e–e repulsion for C–C bond, D_2, DPSA1 Difference CPSAs (PPSA1-PNSA1) Zefirov, D_9, difference (Pos–Neg) in charged part of charged surface area, D_{10}, RNCG relative negative charge (QMNEG/QTMINUS) (MOPAC PC), D_3. Most of these descriptors are charge-related descriptors. A comparison among the descriptors between the linear model in Table 2 and the nonlinear shows that there are two descriptors (min e–e repulsion for C–C bond, relative negative charge (QMNEG/QTMINUS) (MOPAC PC)) that appear in both models. However, the nonlinearities of the ANN model compared (in terms of the coefficient of determinations of the training set NN and the linear models 0.81 and 0.71, respectively), with the linear one seem to improve the prediction of data. In addition, the selection of descriptors for the two models led to a different number of descriptors related to two different approaches. The BMLR method selects descriptors mainly based on criteria as $R^2$and F parameters of the models, whilst the ANN descriptor selection is based on variables that possess the lowest output error. Also, our selection aimed to generalize of the models is based on the least possible number of descriptors.

6. Conclusions

Our present attempt to correlate the log ($1/IC_{50}$) with theoretically calculated molecular descriptors has led

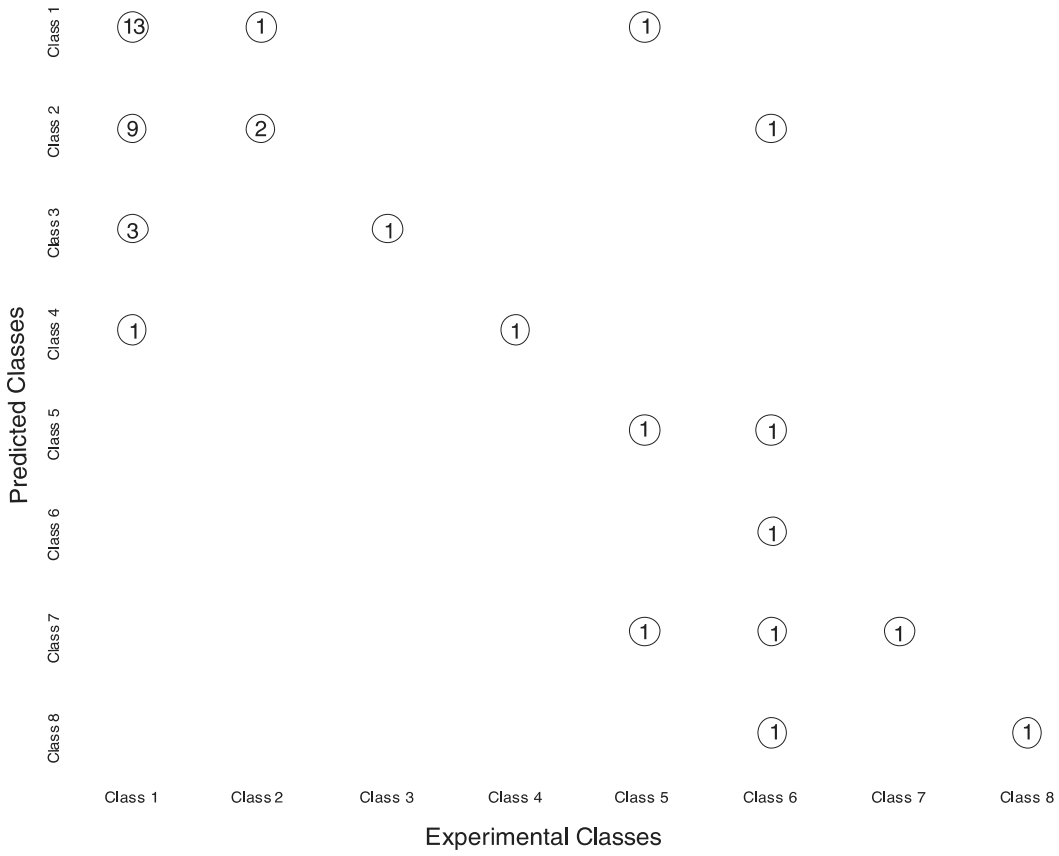

Figure 6. Distribution of predicted versus experimental activity classes of the main five-descriptor NN model for the validation set.

to a relatively successful QSAR model that relates this complex molecular property to structural characteristics of the molecules. Notably, all descriptors appearing in the seven-parameter regression equation and the ANN model have been derived from theoretical molecular calculations. The current computational power available for chemical research allows such calculations for large data sets in realistic time. Thus, in principle, the QSAR models developed in our present work can be used for the prediction of inhibitory activity. The descriptors appearing in these models can be related to the essential electrostatic and conformational interactions between the inhibitory compounds and PDGF receptor.

The results obtained for this work indicate that the regression and ANN models exhibit reasonable prediction capabilities. Though the linear model was developed mainly for the purpose of structure-activity interpretation, the ANN model was primarily developed for predictive ability and classification.

In summary, this work was based on 2D QSAR modeling that should be able to provide prediction of analogous compounds for their log ($1/IC_{50}$) values.

Acknowledgment

The Estonian Science Foundation Grant No. 4548 is acknowledged for partial support of this work.

Supplementary data

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.bmc.2005.06.067.

References and notes

1. Claesson-Welsh, L. *Int. J. Biochem. Cell Biol.* **1996**, *28*, 373.
2. Mayer-Ingold, W.; Eichner, W. *Cell Biol. Int.* **1995**, *19*, 389.
3. Fry, D. W. *Annu. Rep. Chem. Med.* **1996**, *31*, 151.
4. Maguire, M. P.; Sheets, K. R.; McVety, K.; Spada, A. P.; Zilberstein, A. *J. Med. Chem.* **1994**, *37*, 2129.
5. Dolle, R. E.; Dunn, J. A.; Bobko, M.; Singh, B.; Kuster, J. E.; Baizman, E.; Harris, A. L.; Sawutz, D. G.; Miller, D.; Wang, S.; Faltynek, C. R.; Xie, W.; Sarup, J.; Bode, C. E.; Pagani, E. D.; Silver, P. J. *J. Med. Chem.* **1994**, *37*, 2627.
6. Kovalenko, M.; Gazit, A.; Bohmer, A.; Rosman, C.; Ronnstrand, L.; Heldin, C. H.; Waltenberg, J.; Bohmer, F. D.; Levitski, A. *Cancer Res.* **1994**, *54*, 6106.
7. Malkin, M. G.; Mason, W. P.; Liebermann, F. S.; Hannah, A. L. *Proc. Am. Soc. Clin. Oncol.* **1997**, *16*, 385a.
8. Buchdunger, E.; Zimmermann, J.; Mett, H.; Muller, M.; Regenass, U.; Lydon, L. B. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 2558.
9. Zimmermann, J.; Buchdunger, E.; Mett, H.; Meyer, T.; Lydon, N. B.; Traxler, P. *Bioorg. Med. Chem. Lett.* **1996**, *11*, 1221.

10. Palmer, B. D.; Smaill, J. B.; Boyd, M.; Boschelli, D. H.; Doherty, A. M.; Hamby, J. M.; Khatana, S. S.; Kramer, J. B.; Kraker, A. J.; Panek, L. R.; Lu, G. H.; Dahring, T. K.; Winters, R. T.; Showalter, H. D. H.; Denny, W. A. *J. Med. Chem.* **1998**, *41*, 5457.
11. Palmer, B. D.; Kraker, A. J.; Hartl, B. G.; Panopoulos, A. D.; Panek, R. L.; Batley, B. L.; Lu, G. H.; Trumpp-Kallmeyer, S.; Showalter, H. D. H.; Denny, W. A. *J. Chem. Med.* **1999**, *42*, 1373.
12. Oblak, M.; Randic, M.; Solmajer, T. *J. Chem. Inf. Sci.* **2000**, *40*, 4098.
13. Ducrot, P.; Legraverend, M.; Grierson, D. *J. Med. Chem.* **2000**, *43*, 4098.
14. Zhu, L. L.; Hou, T. J.; Chen, L. R.; Xu, X. J. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1032.
15. Kurup, A.; Garg, R.; Hansch, C. *Chem. Rev.* **2001**, *101*, 2573.
16. Katritzky, A. R.; Fara, D. C.; Petrukhin, R. O.; Tatham, D. B.; Maran, U.; Lomaka, A.; Karelson, M. *Curr. Top. Med. Chem.* **2002**, *2*, 1333.
17. Shen, Q.; Lu, Q.-Z.; Jiang, J.-H.; Shen, G.-L.; Yu, R.-Q *Eur. J. Pharm. Sci.* **2003**, *20*, 63.
18. Zhong, C.; He, J.; Xue, C.; Li, Y. A. *Bioorg. Med. Chem.* **2004**, *12*, 4009.
19. Goll, S.; Jurs, P. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1081.
20. Tetteh, J.; Suzuki, T.; Metcalfe, E.; Howells, S. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 491.
21. Zupan, J.; Gasteiger, J. *Neural Networks for Chemists: An Introduction*; VCH- Verlag: Weinheim, 1993, pp 213–228.
22. Burns, J. A.; Whitesides, G. *Chem. Rev.* **1993**, *93*, 2583.
23. Svozil, D.; Kvasnicka, V.; Pospichal, J. *Chem. Intell. Lab. Sys.* **1997**, *39*, 43.
24. Agatonovic Kustrin, S.; Ling, H. L.; Tham, S. Y.; Alany, R. G. *J. Pharmac. Biomed. Anal.* **2002**, *29*, 103.
25. Guha, R.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2179.
26. Katritzky, A. R.; Karelson, M.; Lobanov, V. *Pure Appl. Chem.* **1997**, *69*, 245.
27. Katritzky, A. R.; Lobanov, V. S.; Karelson, M. *Chem. Soc. Rev.* **1995**, *24*, 279.
28. Karelson, M.; Maran, U.; Wang, Y.; Katritzky, A. R. *Coll. Czech. Chem. Commun.* **1999**, *64*, 1551.
29. http://ark.chem.ufl.edu/pages/Research/qspr_2000/qspr_files/frame.htm.
30. 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.
31. Katritzky, A. R.; Fara, D. C.; Karelson, M. *Bioorg. Med. Chem.* **2004**, *12*, 3027.
32. Katritzky, A. R.; Dobchev, D. A.; Fara, D. C.; Karelson, M. *Bioorg. Med. Chem.* **2005**, *13*, 1623.
33. Hyperchem v. 7.5; Hypercube Inc.; Gainesville, FL.
34. Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902.
35. CODESSA PRO Software, University of Florida, 2002.
36. Katritzky, A. R.; Mu, L.; Lobanov, V. S.; Karelson, M. *J. Phys. Chem.* **1996**, *100*, 10400.
37. Karelson, M. *Molecular Descriptors in QSAR/QSPR*; Wiley-Interscience: New York, 2000.
38. Stone, M. J. R. *Stat. Soc.* **1977**, *38*, 44.
39. Haykin, S. *Neural Networks. A Comprehensive Foundation*: Pearson, 2nd ed.; Education: New Jersey, 1999; Vol. 1, pp 156–256.
40. Masters, T. *Practical Neural Network Recipes in C++*; Academic Press, 1993, pp 77–116.