J Mol Model (2006) 12: 297–305
DOI 10.1007/s00894-005-0032-8

ORIGINAL ARTICLE

Georgia Melagraki · Antreas Afantitis
Kalliopi Makridima · Haralambos Sarimveis
Olga Igglessi-Markopoulou

Prediction of toxicity using a novel RBF neural network training methodology

Received: 31 March 2005 / Accepted: 25 July 2005 / Published online: 8 November 2005

Abstract A neural network methodology based on the radial basis function (RBF) architecture is introduced in order to establish quantitative structure-toxicity relationship models for the prediction of toxicity. The dataset used consists of 221 phenols and their corresponding toxicity values to *Tetrahymena pyriformis*. Physicochemical parameters and molecular descriptors are used to provide input information to the models. The performance and predictive abilities of the RBF models are compared to standard multiple linear regression (MLR) models. The leave-one-out cross validation procedure and validation through an external test set produce statistically significant R^2 and RMS values for the RBF models, which prove considerably more accurate than the MLR models.

Keywords RBF architecture · Neural network · QSTR · Toxicity · *Tetrahymena pyriformis*

Introduction

Toxicology deals with the quantitative assessment of toxic effects to organisms in relation to the level, duration and frequency of exposure. Various segments of the population come in contact with toxic chemicals due to misuse (e.g., accidental poisoning), but also through manufacturing, drug and food consumption. Additionally, people working in various jobs (e.g., painters and applicators of pesticides) are exposed to toxic substances. In general, exposure to toxic substances is to be avoided [1].

As the experimental determination of toxicological properties is a costly and time-consuming process, it is essential to develop mathematical predictive relationships to theoretically quantify toxicity [2, 3]. Quantitative structure-toxicity relationship (QSTR) studies can provide a useful tool for achieving this goal, given the successful applications of quantitative structure-activity relationships (QSARs) in several scientific areas, such as pharmacology, chemistry and environmental research. Based on a training database containing measured toxicity potencies of compounds and a number of molecular descriptors, QSTRs can be used to predict the toxicity of chemical compounds that are not included in the database [4–6].

For the formal description of relationships between activity measures and structural descriptors of compounds, various statistical techniques can be used. Among them the most frequently used are multiple linear regression (MLR) and partial least squares (PLS). Several other statistical techniques have been used in QSAR, including discriminant analysis, principal component analysis (PCA) and factor analysis, cluster analysis, multivariate analysis, and adaptive least squares [7–9]. Neural network (NN) techniques have also been used successfully in QSAR [10–16]. The NN methodologies are generally used when the relationships cannot be interpreted accurately by linear functions [17].

The goal of the present study is to determine the efficiency of a newly introduced RBF training methodology in predicting the toxicity of compounds. The methodology uses the innovative fuzzy-means clustering technique to determine the number and the locations of the hidden node centres [18]. Compared to traditional training techniques, the method employed in this work is much faster since it does not involve any iterative procedure, utilizes only one tuning parameter and is repetitive, i.e., it does not depend on a random initial selection of centres. The RBF method is applied to a data set of 221 phenols and the results indicate that it can be used as an efficient new technique for predicting toxicity with significant accuracy, using appropriate descriptors as inputs.

G. Melagraki · A. Afantitis · K. Makridima · H. Sarimveis (✉)
O. Igglessi-Markopoulou
School of Chemical Engineering, National Technical
University of Athens, 9 Heroon Polytechniou Str.,
Zografou Campus,
Athens, 15780 Greece
E-mail: hsarimv@central.ntua.gr
Tel.: +30-210-7723237
Fax: +30-210-7723138

Materials and methods

It is essential in order to obtain a successful QSTR that all data used as part of the training and validation procedure are of high quality. High quality data should derive from the same endpoint and protocol and ideally should be measured in the same laboratory [19]. The data set used in this study fulfills this criterion.

Toxicity data

This data set consists of 221 phenols and their corresponding toxicity data to the ciliate *Tetrahymena pyriformis* in terms of log(1/IGC$_{50}$) (mmol/L). The toxicity values were taken from the literature [20] and are shown in Table 1. The phenols are structurally heterogeneous and represent a variety of mechanisms of toxic action. The dataset consists of polar narcotics, weak acid respiratory uncouplers, pro-electrophiles and soft electrophiles.

Molecular descriptors

The molecular descriptors used to derive the model were taken from the literature [20] and include the logarithm of the octanol/water partition coefficient (log K_{ow}), acidity constant (pK_{a}), the energies of the highest occupied and lowest unoccupied molecular orbital (E_{HOMO} and E_{LUMO} respectively) and the hydrogen bond donor number (N_{hdon}). All these descriptors are related to the toxicity effect of the compounds studied.

Statistical analysis (QSAR development)

In this section, we present the basic characteristics of the RBF NN architecture and the training method used to develop the QSAR NN models.

RBF network topology and node characteristics

RBF networks consist of three layers: the input layer, the hidden layer and the output layer. The input layer collects the input information and formulates the input vector **x**. The hidden layer consists of L hidden nodes, which apply nonlinear transformations to the input vector. The output layer delivers the NN responses to the environment. A typical hidden node l in an RBF network is described by a vector $\hat{x}_l$, equal in dimension to the input vector and a scalar width σ_l. The activity $v_l(\mathbf{x})$ of the node is calculated as the Euclidean norm of the difference between the input vector and the node center and is given by

$$v_l(x) = \|x - \hat{x}_l\| \tag{1}$$

The response of the hidden node is determined by passing the activity through the radially symmetric Gaussian function:

$$f_l(x) = \exp\left(-\frac{v_l(x)^2}{\sigma_l^2}\right) \tag{2}$$

Finally, the output values of the network are computed as linear combinations of the hidden layer responses:

$$\hat{y} = g(x) = \sum_{l=1}^{L} f_l(x) w_l \tag{3}$$

where [w_1, w_2,... ,w_L] is the vector of weights, which multiply the hidden node responses in order to calculate the output of the network.

RBF network training methodology

Training methodologies for the RBF network architecture are based on a set of input–output training pairs ($\mathbf{x}(k)$; $\mathbf{y}(k)$) (k = 1, 2,...,K). The training procedure used in this work consists of three distinct phases:

(i) Selection of the network structure and calculation of the hidden-node centers using the fuzzy-means clustering algorithm [18]. The algorithm is based on a fuzzy partition of the input space, which is produced by defining a number of triangular fuzzy sets on the domain of each input variable. The centers of these fuzzy sets produce a multidimensional grid on the input space. A rigorous selection algorithm chooses the most appropriate knots of the grid, which are used as hidden node centers in the RBF network model produced. The idea behind the selection algorithm is to place the centers in the multidimensional input space so that there is a minimum distance between the center locations. At the same time, the algorithm assures that for any input example in the training set there is at least one selected hidden node that is close enough according to a distance criterion. It must be emphasized that, in contrast to both the k-means [21] and the c-means clustering [22] algorithms, the fuzzy-means technique does not need the number of clusters to be fixed before the execution of the method. Moreover, due to the fact that it is a one-pass algorithm, it is extremely fast even if a large database of input–output examples is available. Furthermore, the fuzzy-means algorithm needs only one tuning parameter, which is the number of fuzzy sets that are used to partition each input dimension.

(ii) Following the determination of the hidden-node centers, the widths of the Gaussian activation function are calculated using the P-nearest neighbor heuristic [23]:

$$\sigma_l = \left(\frac{1}{p}\sum_{i=1}^{p} \|\hat{x}_l - \hat{x}_i\|^2\right)^{1/2} \tag{4}$$

where $\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_p$ are the p nearest-node centers to the hidden node l. The parameter p is selected so that many

Table 1 Predicted values [log(1/IGC_{50})] for the training and the test set

A/A	Name	log(1/IGC_{50})	Training set		Validation set	
			RBF $R^2=0.9424$	MLR $R^2=0.6022$	RBF $R^2=0.8824$	MLR $R^2=0.7861$
1	1,3,5-Trihydroxybenzene	-1.26	-1.2577	0.4071		
2	2-(*tert*)-Butyl-4-methylphenol	1.3	1.1624	1.2334		
3	2,3,5-Trichlorophenol	2.37	2.1688	1.4111		
4[a]	2,3,5-Trimethylphenol	0.36			0.5785	0.7671
5	2,3,6-Trimethylphenol	0.28	0.5460	0.7611		
6	2,3-Dichlorophenol	1.28	1.4070	0.8046		
7[a]	2,3-Dimethylphenol	0.12			0.2007	0.3904
8	2,4,5-Trichlorophenol	2.1	1.8325	1.5046		
9	2,4,6-Tribromophenol	2.03	2.3170	1.6470		
10	2,4,6-Tribromoresorcinol	1.06	1.1134	2.5259		
11	2,4,6-Trichlorophenol	1.41	1.3937	1.3193		
12	2,4,6-Trimethylphenol	0.28	0.3515	0.8490		
13	2,4,6-Tris (dimethylaminomethyl) phenol	-0.52	0.5294	0.3641		
14	2,4-Dibromophenol	1.4	1.6666	1.1616		
15	2,4-Dichlorophenol	1.04	1.0157	0.9485		
16	2,4-Difluorophenol	0.6	0.5917	0.4491		
17	2,4-Dimethylphenol	0.07	0.0467	0.4939		
18[a]	2,5-Dichlorophenol	1.13			1.1504	0.9715
19[a]	2,5-Dimethylphenol	0.08			0.0996	0.3404
20	202,6-Di-(*tert*)-butyl-4-methylphenol	1.8	1.7939	2.3411		
21	2,6-Dichloro-4-fluorophenol	0.8	0.9982	1.0697		
22	2,6-Dichlorophenol	0.74	0.6177	0.7097		
23	2,6-Difluorophenol	0.47	0.3470	0.1981		
24	2,6-Dimethoxyphenol	-0.6	0.5510	0.1055		
25	2-Allylphenol	0.33	0.1816	0.3925		
26[a]	2-Bromo-4-methylphenol	0.6			0.8483	0.8478
27	2-Bromophenol	0.33	0.5950	0.4488		
28	2-Chloro-4,5-dimethylphenol	0.69	0.6884	1.0551		
29	2-Chloro-5-methylphenol	0.39	0.6920	0.6840		
30	2-Chlorophenol	0.18	0.3583	0.3040		
31	2-Cyanophenol	0.03	0.2517	0.1132		
32	2-Ethoxyphenol	-0.36	0.1630	0.1940		
33[a]	2-Ethylphenol	0.16			0.3373	0.3690
34	2-Fluorophenol	0.19	0.1022	0.0294		
35[a]	2-Hydroxy-4,5-dimethylacetophenone	0.71			0.5292	0.7995
36	2-Hydroxy-4-methoxyacetophenone	0.55	0.3823	0.4016		
37	2-Hydroxy-4-methoxybenzophenone	1.42	1.4376	1.7424		
38	2-Hydroxy-5-methylacetophenone	0.31	0.3419	0.7916		
39[a]	2-Hydroxyacetophenone	0.08			0.2318	0.3432
40	2-Hydroxybenzylalcohol	-0.95	0.9364	0.5395		
41	2-Hydroxyethylsalicylate	-0.08	0.0845	0.5963		
42	2-Isopropylphenol	0.8	0.7377	1.2005		
43	2-Methoxy-4-propenylphenol	0.75	0.7445	1.2005		
44	2-Methoxyphenol	-0.51	0.5486	0.1344		
45	2-Phenylphenol	1.09	1.1577	1.2855		
46	2-(*tert*)-Butylphenol	1.3	1.3378	0.8191		
47	3,4,5-Trimethylphenol	0.93	0.7390	0.7521		
48	3,4-Dichlorophenol	1.75	1.5232	1.0530		
49	3,4-Dimethylphenol	0.12	0.1552	0.4499		
50	3,5-Dibromosalicylaldehyde	1.64	1.8912	1.5092		
51	3,5-Dichlorophenol	1.57	1.3614	0.9657		
52	3,5-Dichlorosalicylaldehyde	1.55	1.4080	1.3502		
53	3,5-Diiododsalicylaldehyde	2.34	2.2079	1.6881		
54	3,5-Dimethoxyphenol	-0.09	0.1690	0.1163		
55[a]	3,5-Dimethylphenol	0.11			0.3133	0.2588
56	3,5-Di-(*tert*)-butylphenol	1.64	1.6973	1.8331		
57[a]	3-Acetamidophenol	-0.16			0.1873	-0.1212
58[a]	3-Bromophenol	1.15			0.7477	0.5605
59	3-Chloro-4-fluorophenol	1.13	1.0300	0.8618		
60	3-Chloro-5-methoxyphenol	0.76	0.7190	0.5070		
61	3-Chlorophenol	0.87	0.7820	0.4292		
62	3-Cyanophenol	-0.06	0.0908	0.1710		
63	3-Ethoxy-4-hydroxybenzaldehyde	0.02	-0.0307	0.6282		
64	3-Ethoxy-4-methoxyphenol	-0.3	0.2483	0.4874		
65[a]	3-Ethylphenol	0.23			0.3863	0.3287

Table 1 (contd.)

A/A	Name	log(1/IGC$_{50}$)	Training set		Validation set	
			RBF $R^2 = 0.9424$	MLR $R^2 = 0.6022$	RBF $R^2 = 0.8824$	MLR $R^2 = 0.7861$
66	3-Fluorophenol	0.38	0.3626	0.0624		
67[a]	3-Hydroxy-4-methoxybenzylalcohol	-0.99			0.1893	0.2909
68	3-Hydroxyacetophenone	-0.38	0.3606	0.2105		
69[a]	3-Hydroxybenzaldehyde	0.09			0.0073	0.1464
70	3-Hydroxybenzoic acid	-0.81	0.9606	0.5278		
71[a]	3-Hydroxybenzylalcohol	-1.04			0.7287	0.4854
72	3-Iodophenol	1.12	1.1825	0.6973		
73	3-Isopropylphenol	0.61	0.5719	0.5519		
74	3-Methoxyphenol	-0.33	0.3633	0.0317		
75	3-Phenylphenol	1.35	1.2389	1.2931		
76[a]	3-(*tert*)-Butylphenol	0.73			0.9910	0.7758
77	4-(*tert*)-Octylphenol	2.1	2.0342	1.9128		
78[a]	4-(tert)-Butylphenol	0.91			0.9333	0.8211
79	4,6-Dichlororesorcinol	0.97	0.9034	0.9385		
80[a]	4-Allyl-2-methoxyphenol	0.42			0.2407	0.5247
81	4-Benzyloxyphenol	1.04	1.0458	1.2864		
82[a]	4-Bromo-2,6-dichlorophenol	1.78			1.7768	1.3813
83	4-Bromo-2,6-dimethylphenol	1.17	1.3217	1.2670		
84	4-Bromo-3,5-dimethylphenol	1.27	1.1912	1.2015		
85	4-Bromo-6-chloro-2-cresol	1.28	1.4570	1.3406		
86	4-Bromophenol	0.68	0.6965	0.6116		
87[a]	4-Butoxyphenol	0.7			0.7779	1.0973
88	4-Chloro-2-isopropyl-5-methylphenol	1.85	1.7646	1.7180		
89[a]	4-Chloro-2-methylphenol	0.7			0.8504	0.8675
90[a]	4-Chloro-3,5-dimethylphenol	1.2			1.2333	1.1467
91[a]	4-Chloro-3-ethylphenol	1.08			1.2658	1.1233
92	4-Chloro-3-methylphenol	0.8	0.7377	0.8344		
93[a]	4-Chlorophenol	0.55			0.5155	0.5212
94	4-Chlororesorcinol	0.13	0.5804	0.4712		
95	4-Cyanophenol	0.52	0.3434	0.0974		
96	4-Ethoxyphenol	0.01	-0.1385	0.5105		
97	4-Ethylphenol	0.21	0.3014	0.3981		
98	4-Fluorophenol	0.02	-0.0708	0.2526		
99	4-Heptyloxyphenol	2.03	2.1227	1.9979		
100	4-Hexyloxyphenol	1.64	1.5630	1.6922		
101[a]	4-Hexylresorcinol	1.80			1.5525	1.4144
102	4-Hydroxy-2-methylacetophenone	0.19	0.1939	0.4472		
103	4-Hydroxy-3-methoxyacetophenone	-0.12	0.1004	0.3638		
104	4-Hydroxy-3-methoxybenzonitrile	-0.03	0.0216	0.4072		
105	4-Hydroxy-3-methoxybenzylalcohol	-0.7	0.8639	0.4295		
106	4-Hydroxy-3-methoxybenzylamine	-0.97	0.2649	-0.3264		
107[a]	4-Hydroxy-3-methoxyphenethylalcohol	-0.18			0.1069	0.1330
108	4-Hydroxyacetophenone	-0.3	0.0234	0.1133		
109	4-Hydroxybenzaldehyde	0.27	-0.0006	0.1058		
110	4-Hydroxybenzamide	-0.78	0.6458	0.3673		
111	4-Hydroxybenzoic acid	-1.02	0.8670	0.3948		
112	4-Hydroxybenzophenone	1.02	1.0913	1.1405		
113	4-Hydroxybenzylcyanide	-0.38	0.3997	0.4804		
114[a]	4-Hydroxyphenethylalcohol	-0.83			0.6590	0.4298
115	4-Hydroxyphenylacetic acid	-1.5	1.5063	0.2107		
116[a]	4-Hydroxypropiophenone	0.05			0.3086	0.4059
117	4-Iodophenol	0.85	0.95	0.7254		
118[a]	4-Isopropylphenol	0.47			0.6119	0.6148
119	4-Methoxyphenol	-0.14	0.3372	0.1976		
120[a]	4-Phenylphenol	1.39			1.2357	1.4480
121[a]	4-Propylphenol	0.64			0.7181	0.7046
122	4-(*sec*)-Butylphenol	0.98	1.0932	0.9117		
123	4-(*tert*)-Pentylphenol	1.23	1.3335	1.1356		
124	5-Bromo-2-hydroxybenzylalcohol	0.34	0.4247	0.3608		
125	5-Bromovanillin	0.62	0.6049	0.7279		
126	5-Fluoro-2-hydroxyacetophenone	0.04	0.0517	0.7771		
127	5-Methylresorcinol	-0.39	0.4360	0.1271		
128	5-Pentylresorcinol	1.31	1.3376	1.3020		
129	6-(*tert*)-Butyl-2,4-dimethylphenol	1.16	1.1801	1.5907		
130	a,a,a-Trifluoro-4-cresol	0.62	0.6807	0.5816		

Table 1 (contd.)

A/A	Name	log(1/IGC$_{50}$)	Training set		Validation set	
			RBF $R^2 = 0.9424$	MLR $R^2 = 0.6022$	RBF $R^2 = 0.8824$	MLR $R^2 = 0.7861$
131	Ethyl-3-hydroxybenzoate	0.48	0.5352	0.7593		
132	Ethyl-4-hydroxy-3-methoxyphenylacetate	0.23	0.0891	0.2439		
133[a]	Ethyl-4-hydroxybenzoate	0.57			0.7127	0.6494
134	Isovanillin	0.14	0.2235	0.3669		
135[a]	3-Cresol	-0.06			0.0257	0.0559
136[a]	Methyl-3-hydroxybenzoate	0.05			0.2478	0.4859
137	Methyl-4-hydroxybenzoate	0.08	0.2095	0.4817		
138[a]	Methyl-4-methoxysalicylate	0.62			0.6075	0.6973
139	Nonylphenol	2.47	2.4674	2.4774		
140[a]	2-Cresol	-0.3			0.1056	0.0954
141[a]	2-Vanillin	0.38			0.1732	0.4571
142[a]	4-Cresol	-0.18			0.1592	0.2252
143	4-Cyclopentylphenol	1.29	1.2381	0.9981		
144	Phenol	0.21	0.1106	0.3004		
145	Resorscinol	0.65	0.6311	0.2009		
146	Salicylaldehyde	0.42	0.4010	0.2986		
147	Salicylaldoxime	0.25	0.1620	0.3740		
148	Salicylamide	0.24	0.3046	0.0554		
149	Salicylhydrazide	0.18	0.1825	0.1927		
150	Salicylhydroxamic acid	0.38	0.3768	0.2226		
151	Salicylic acid	0.51	0.5072	0.7902		
152	Syringaldehyde	0.17	0.1762	0.3455		
153	Vanillin	0.03	0.0114	0.3303		
154	2,3,4,5-Tetrachlorophenol	2.71	2.6883	1.8520		
155	2,3,5,6-Tetrachlorophenol	2.22	2.2198	1.6755		
156	2,3,5,6-Tetrafluorophenol	1.17	1.2825	0.6360		
157	2,3-Dinitrophenol	0.46	0.5685	0.7861		
158	2,4,6-Trinitrophenol	-0.16	0.1587	0.4653		
159	2,4-Dichloro-6-nitrophenol	1.75	1.8195	1.7045		
160	2,4-Dinitrophenol	1.08	0.9775	0.5527		
161	2,5-Dinitrophenol	0.95	0.9357	1.0017		
162	2,6-Dichloro-4-nitrophenol	0.63	0.6967	1.1545		
163	2,6-Diiodo-4-nitrophenol	1.71	1.7308	1.6515		
164	2,6-Dinitro-4-cresol	1.23	1.0951	1.17		
165	2,6-Dinitrophenol	0.54	0.6098	0.6845		
166	3,4,5,6-Tetrabromo-2-cresol	2.57	2.5622	2.4724		
167	3,4-Dinitrophenol	0.27	0.2449	0.6613		
168	4,6-Dinitro-2-cresol	1.72	1.8385	0.9805		
169	Pentabromophenol	2.66	2.6674	2.5129		
170	Pentachlorophenol	2.05	2.0362	2.1188		
171	Pentafluorophenol	1.64	1.5253	0.9301		
172	1,2,3-Trihydroxybenzene	0.85	0.3641	-0.4575		
173	1,2,4-Trihydroxybenzene	0.44	0.4386	0.1186		
174	2,3-Dimethylhydroquinone	1.41	2.1983	0.4201		
175	2,4-Diaminophenol	0.13	0.1296	-0.1773		
176	2-Amino-4-(tert)-butylphenol	0.37	0.3471	1.0426		
177	2-Aminophenol	0.94	1.0797	0.0342		
178	3,5-Di-(*tert*)-butylcatechol	2.11	2.1032	2.1321		
179	3-Aminophenol	-0.52	0.6763	0.4105		
180	3-Methylcatechol	0.28	0.3889	0.2381		
181	4-Acetamidophenol	-0.82	0.1854	0.0424		
182	4-Amino-2,3-dimethylphenol	1.44	1.3920	0.0618		
183	4-Amino-2-cresol	1.31	1.2952	0.2362		
184	4-Aminophenol	-0.08	0.0292	0.0845		
185	4-Chlorocatechol	1.06	0.8653	0.7061		
186[a]	4-Methylcatechol	0.37			0.6642	0.2757
187	5-Amino-2-methoxyphenol	0.45	0.4527	-0.1456		
188	5-Chloro-2-hydroxyaniline	0.78	0.7809	0.7450		
189	6-Amino-2,4-dimethylphenol	0.89	0.9603	0.4623		
190	Bromohydroquinone	1.68	1.7439	0.8086		
191[a]	Catechol	0.75			0.2268	-0.0938
192	Chlorohydroquinone	1.26	0.8143	0.3379		
193	Hydroquinone	0.47	0.3551	-0.0659		
194	Methoxyhydroquinone	2.20	0.8448	-0.0157		
195	Methylhydroquinone	1.86	1.5627	0.2166		

Table 1 (contd.)

A/A	Name	$\log(1/IGC_{50})$	Training set		Validation set	
			RBF $R^2=0.9424$	MLR $R^2=0.6022$	RBF $R^2=0.8824$	MLR $R^2=0.7861$
196	Phenylhydroquinone	2.01	2.0494	1.4188		
197	Tetrachlorocatechol	1.700	1.6398	2.3871		
198	Trimethylhydroquinone	1.34	1.0404	0.7284		
199	2,6-Dibromo-4-nitrophenol	1.36	1.2960	1.3558		
200	2-Amino-4-chloro-5-nitrophenol	1.17	1.1656	1.3096		
201	2-Amino-4-nitrophenol	0.48	0.5334	1.0231		
202	2-Chloro-4-nitrophenol	1.59	1.4875	0.8898		
203	2-Chloromethyl-4-nitrophenol	0.75	1.0330	0.7947		
204	2-Nitrophenol	0.67	0.8831	0.6586		
205	2-Nitroresorcinol	0.66	0.6898	1.1367		
206[a]	3-Fluoro-4-nitrophenol	0.94	0.3165		0.9997	0.4381
207	3-Hydroxy-4-nitrobenzaldehyde	0.27	0.3165	0.6154		
208	3-Methyl-4-nitrophenol	1.73	1.3591	0.6877		
209	3-Nitrophenol	0.51	0.4308	0.6024		
210	4-Amino-2-nitrophenol	0.88	0.8491	1.0359		
211	4-Chloro-2-nitrophenol	2.05	2.0047	1.3347		
212	4-Chloro-6-nitro-3-cresol	1.64	1.5944	1.6378		
213	4-Hydroxy-3-nitrobenzaldehyde	0.61	0.6226	0.4118		
214	4-Methyl-2-nitrophenol	0.57	0.6544	1.1031		
215	4-Methyl-3-nitrophenol	0.74	0.7122	1.0180		
216	4-Nitro-3-(trifluoromethyl)-phenol	1.65	1.5893	1.0526		
217	4-Nitrocatechol	1.17	1.1431	0.9175		
218	4-Nitrophenol	1.42	1.4467	0.4263		
219	4-Nitrosophenol	0.65	0.5828	0.3104		
220	5-Fluoro-2-nitrophenol	1.13	1.2294	0.7792		
221	5-Hydroxy-2-nitrobenzaldehyde	0.33	0.1427	0.5858		

[a] Compounds used in the validation set

nodes are activated when an input vector is presented to the NN model.

(iii) The connection weights are determined using linear regression between the hidden-layer responses and the corresponding output training set.

Results

In order to evaluate and compare the performance of the RBF training methodology presented in this work, the data set was initially split into a training and a validation set in a ratio of approximately 80:20% (180 and 41 compounds, respectively). For that, the Kennard and Stones algorithm [24] was used. The Kennard–Stones algorithm has gained increasing popularity for splitting data sets into two subsets. The algorithm starts by finding two samples that are the farthest apart from each other on the basis of the input variables in terms of some metric, e.g., the Euclidean distance. These two samples are removed from the original data set and put into the calibration data set. The procedure described 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 variable space completely with respect to the induced metric and that the test samples all fall inside the measured region. The training and validation compounds are clearly indicated in Table 1. Both RBF network and MLR models were developed based on exactly the same training set. The validation set was not involved in any way during the training phase. The results are shown in Table 1, where the predictions of the two models are shown for both the training and the external examples. The same results are shown in a graphical format in Figs. 1, 2, 3 and 4, where the experimental toxicity is plotted against the predictions of the RBF network and the MLR model. In each figure the corresponding coefficients of determination (R^2-value) are presented, which indicate a much higher correlation between experimental and predicted values using the RBF network methodology. The full linear equation for the prediction of toxicity is the following:

$$\log 1/IGC_{50}=0.5617\log K_{ow}+0.0026pK_a-0.8792E_{LUMO}+0.7995E_{HUMO}+0.2734N_{hdon}+6.2044,$$
$$n=180,\ R^2=0.6022,\ RMS=0.5352. \tag{5}$$

To compare the performance of the modeling schemes further, their predictive ability was also evaluated by the leave-one-out (LOO) cross-validation procedure. A number of modified data sets were created by deleting in each case one object from the data. An RBF network and an MLR model were developed in each case based on the remaining data and were validated using the object that had been deleted. Consequently, 221 RBF networks and MLR models were built, by deleting each time one compound from the training set.

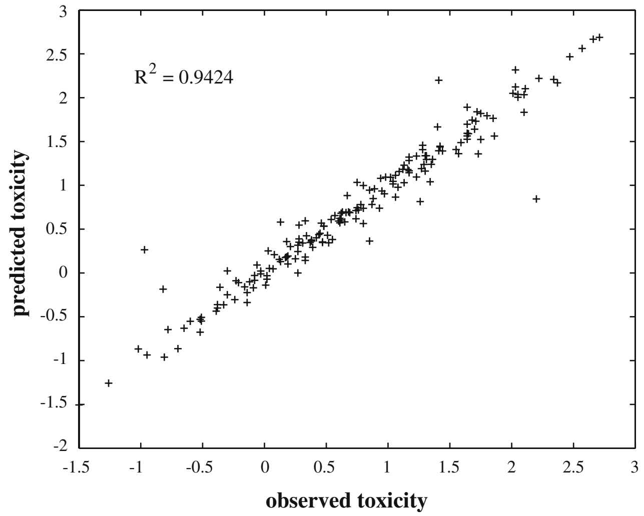

Fig. 1 Experimental versus predicted toxicity using the RBF methodology for the training set (180 compounds)

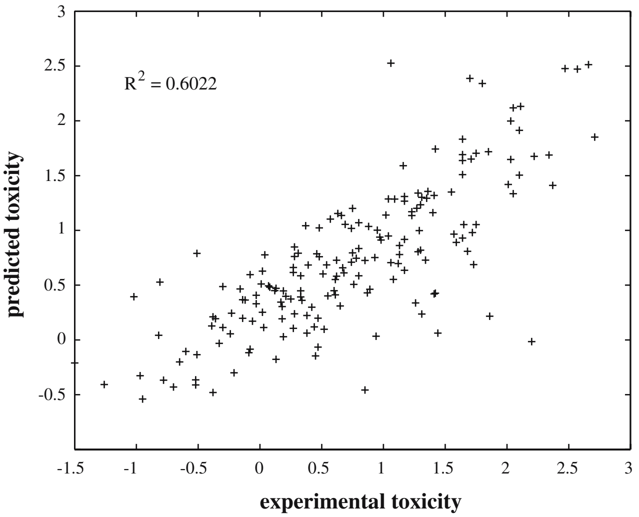

Fig. 2 Experimental versus predicted toxicity using the MLR methodology for the training set (180 compounds)

Figures 5 and 6 show the experimental toxicity versus the predictions produced by the RBF NN models and the multiple regression technique, using the LOO cross validation procedure. The corresponding coefficients of determination R^2_{CV} indicate again that the models derived from the RBF methodology have a higher predictive potential. The comparison between the RBF and the MLR methods is summarized in Table 2. In all cases, the RBF models proved to be remarkably more accurate than the MLR models. The predictive abilities of both modeling techniques can be improved if different models are developed for each one of the several different mechanisms of action, but in this paper we concentrated on building a single model for each methodology that can predict toxicity for the variety of mechanisms that are included in the data set.

It should finally be noted that the MATLAB programming language was used to implement all the training and testing procedures. The computational time required to build the NN models in a Pentium IV 3 GHz processor was always less than 0.2 s. It should also be emphasized that the RBF training method has been developed in-house, so no commercial packages were used to develop the NN models. The complete QSTR models can be made available to the interested readers.

Discussion and conclusions

In this work, we presented a novel QSTR methodology based on the RBF NN architecture. The method was illustrated using a data set of 221 phenols and compared

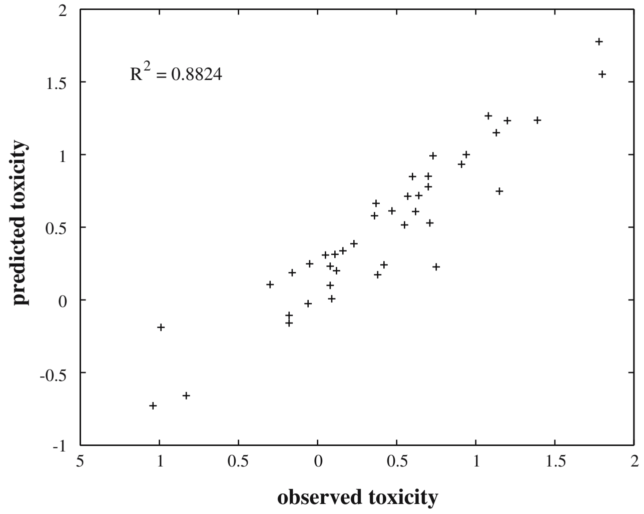

Fig. 3 Experimental versus predicted toxicity using the RBF methodology for the test set (41 compounds)

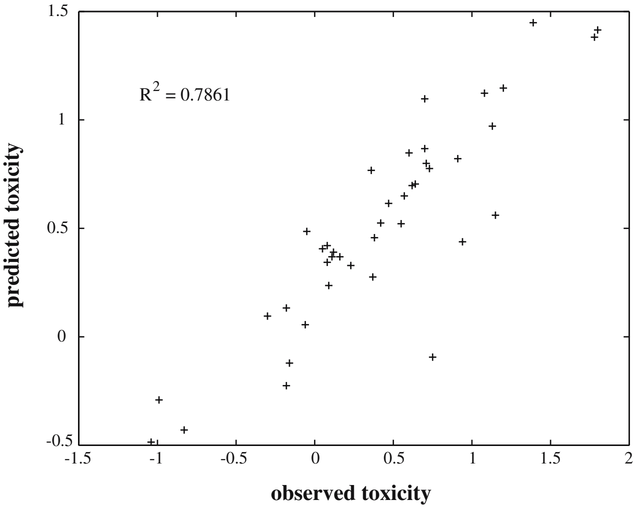

Fig. 4 Experimental versus predicted toxicity using the MLR methodology for the test set (41 compounds)

Table 2 Summary of the results produced by the different methods

Method	Training set	Validation set	R^2_{train}	R^2_{pred}	RMS	RMS_{pred}	Figure
RBF	180	41	0.9424		0.2037		1
MLR	180	41	0.6022		0.5352		2
RBF	180	41		0.8824		0.2398	3
MLR	180	41		0.7861		0.3197	4
RBF LOO	221–*i*	221–*i*		0.7203		0.4350	5
MLR LOO	221–*i*	221–*i*		0.6010		0.5194	6

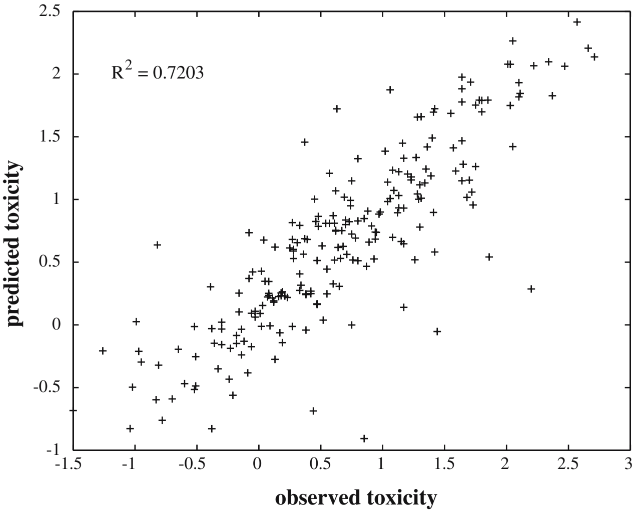

Fig. 5 Experimental versus predicted toxicity with cross validation (RBF methodology)

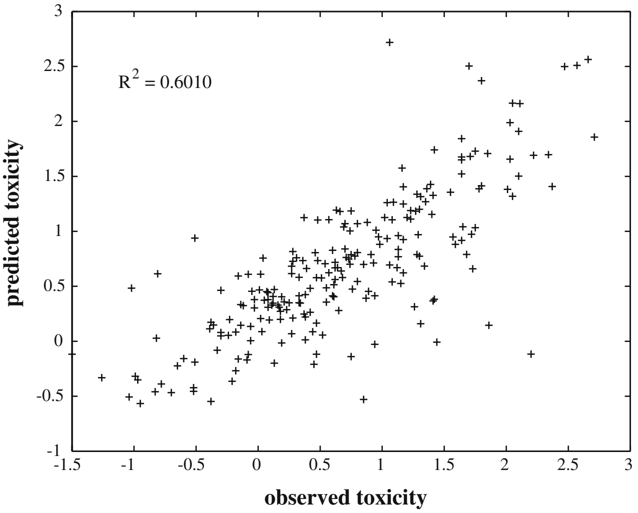

Fig. 6 Experimental versus predicted toxicity with cross validation (MLR methodology)

with standard MLR. Validation of the different QSTR methodologies was based on two evaluation procedures. In the first method the data were split into a training and a validation set and the model generated using the training set was used to predict toxicity in the validation set. The second method was the standard LOO cross-validation procedure. The modeling procedures used in this work illustrated the accuracy of the models produced, not only by calculating their fitness on sets of training data but also by testing the predicting abilities of the models.

The RBF NN models were produced based on the fuzzy-means training method, which is fast and repetitive, in contrast to most traditional training techniques. The model generated for the data set required five descriptors. In terms of the R^2, R^2_{cv} and RMS values, the RBF models proved to have a significant predictive potential. The results obtained illustrated that the RBF NN architecture can be used to derive QSTRs, which are more accurate and have better generalization capabilities compared to linear regression models at the expense of the increased complexity of the model compared to a simple structure of a linear model. The method proposed could be a substitute to costly and time-consuming experiments for determining toxicity.

Acknowledgements G.M. wishes to thank the Greek State Scholarship Foundation for a doctoral assistantship.

References

1. Lu FC, Kacew S (2002) Lu's basic toxicology. Taylor & Francis, London
2. Karcher W, Devillers J (1990) SAR and QSAR in environmental chemistry and toxicology: scientific tool or wishful thinking?. In: Karcher W, Devillers J (eds) Practical applications of quantitative structure-activity relationships (QSAR) in environmental chemistry and toxicology. Kluwer, Dordrecht, pp 1–12
3. Nendza M (1998) Structure-activity relationships in environmental sciences, ecotoxicology series 6. Chapman & Hall, London
4. Schultz TW, Netzeva TI, Cronin MTD (2003) SAR QSAR Environ Res 14:59–81
5. Netzeva TI, Schultz TW, Aptula AO, Cronin MTD (2003) SAR QSAR Environ Res 14:265–283
6. Zahouily M, Rhihil A, Bazoui H, Sebti S, Zakarya D (2002) J Mol Model 8:168–172
7. Cronin MTD, Aptula AO, Duffy JC, Netzeva TI, Rowe PH, Valkova IV, Schultz TW (2002) Chemosphere 49:1201–1221
8. Ren S (2003) Chemosphere 53:1053–1065
9. Bukard U (2003) Methods for data analysis. In: Gasteiger J, Engel Th (eds) Chemoinformatics. Wiley VCH, Weinheim, pp 439–485
10. Devillers J (1996) Neural networks in QSAR and drug design. Academic Press, London
11. Afantitis A, Melagraki G, Makridima K, Alexandridis A, Sarimveis H, Iglessi-Markopoulou O (2005) J Mol Struct: Theochem 716:193–198
12. Devillers J (2004) SAR QSAR Environ Res 15:237–249
13. Kaiser KLE (2003) Quant Struct-Act Relat 22:1–5

14. KaiserKLE (2003) J Mol Struct: Theochem 622:85–95
15. Gasteiger J (2003) Handbook of chemoinformatics: from data to knowledge, vol 3. Wiley VCH, Weinheim
16. Zupan J, Gasteiger J (1999) Neural networks in chemistry and drug design. Wiley VCH, Weinheim
17. Debnath AK (2001) Quantitative structure-activity relationship (QSAR): a versatile tool in drug design. In: Ghose AK, Viswanadhan VN (eds) Combinatorial library design and evaluation: principles, software tools, and applications in drug discovery. Marcel Dekker, New York, pp 73–129
18. Sarimveis H, Alexandridis A, Tsekouras G, Bafas G (2002) Ind Eng Chem Res 41:751–759
19. Lessigiarska I, Cronin MTD, Worth AP, Dearden JC, Netzeva TI (2004) SAR QSAR Environ Res 15:169–190
20. Aptula AO, Netzeva TI, Valkona IV, Cronin MTD, Schultz TW, Kuhne R, Schuurmann G (2002) Quant Struct-Act Relat 21:12–22
21. Darken C, Moody J (1990) Fast adaptive K-means clustering: some empirical results. IEEE INNS Int Joint Conf Neural Netw 2:233–238
22. Dunn JC (1974) J Cybernet 3:32–57
23. Leonard JA, Kramer MA (1991) Radial basis function networks for classifying process faults. IEEE Control Syst 11:31–38
24. Kennard RW, Stone LA (1969) Technometrics 11:137–148