Molecular Diversity (2006) 10: 213–221
DOI: 10.1007/s11030-005-9008-y

Full-length paper

A novel RBF neural network training methodology to predict toxicity to *Vibrio fischeri*

Georgia Melagraki[1], Antreas Afantitis[1], Haralambos Sarimveis[2,*], Olga Igglessi-Markopoulou[1] & Alex Alexandridis[2]

[1]*Laboratory of Organic Chemistry, School of Chemical Engineering, National Technical University of Athens, 9, Heroon Polytechniou Str., Zografou Campus, Athens 15780, Greece;* [2]*Laboratory of Process Control & Informatics, School of Chemical Engineering, National Technical University of Athens, 9, Heroon Polytechniou Str., Zografou Campus, Athens 15780, Greece*

*(*Author for correspondence, E-mail: hsarimv@central.ntua.gr, Tel.: +30-210-7723237, Fax: +30-210-7723138)*

Received 8 November 2005; Accepted 14 December 2005

Keywords: neural network, QSTR, RBF architecture, toxicity, *Vibrio fischeri*

Summary

This work introduces a neural network methodology for developing QSTR predictors of toxicity to *Vibrio fischeri.* The method adopts the Radial Basis Function (RBF) architecture and the fuzzy means training strategy, which is fast and repetitive, in contrast to most traditional training techniques. The data set that was utilized consisted of 39 organic compounds and their corresponding toxicity values to *Vibrio fischeri*, while lipophilicity, equalized electronegativity and one topological index were used to provide input information to the models. The performance and predictive ability of the RBF model were illustrated through external validation and various statistical tests. The proposed methodology can be used to successfully model toxicity to *Vibrio fischeri* for a heterogeneous set of compounds.

1. Introduction

Toxicology deals with the quantitative assessment of the toxic effects to organisms in relation to the level, duration and frequency of exposure. In general, exposure to toxic substances is to be avoided and thus toxicity assessment of such compounds is vital [1]. Among the bacterial assays, the *Vibrio fischeri* luminescence inhibition assay is the most popular. Bioluminescent bacteria toxicity tests offer a convenient, sensitive and efficient ethical alternative to testing on higher species [2, 3].

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 [4, 5]. Quantitative Structure – Toxicity Relationship (QSTR) studies can provide a useful tool for achieving this goal, that is predicting the toxic potency of untested compounds [6, 7].Apart from serving as predictors of ecological and human health effects, QSTRs are also utilized in the process of designing safer chemicals for commercial use. The use of toxicity data from *Vibrio fischeri* tests in the development of QSTRs is adopted in several publications [8–11].

For the formal description of relationships between activity measures and structural descriptors of compounds various statistical techniques can be used. Among them, the most popular are Multiple Linear Regression (MLR) [12–14] and Partial Least Squares (PLS) [7]. Several other statistical techniques have been used for the same purpose, including discriminant analysis, principal component analysis (PCA) and factor analysis, cluster analysis, multivariate analysis, and adaptive least squares [5, 15]. Neural Network (NN) techniques have also been applied successfully in developing quantitative structure-activity relationships [16–20]. NNs have gained attention due to their ability to describe non-linear relationships with success.

The objective of this work was to investigate the potential of using a special neural network architecture, namely the Radial Basis Function (RBF) networks in the development of a QSTR model for predicting toxicity of compounds to *Vibrio fischeri.* More specifically, a recently introduced training methodology for generating Radial Basis Function (RBF) neural networks was utilized. The method uses the innovative fuzzy means clustering technique to determine the number and the locations of the hidden node centers [21]. The most significant advantages of this method compared to

traditional RBF network training techniques are the following: it is much faster since it does not involve any iterative procedure, utilizes only one tuning parameter and it is repetitive, i.e. it does not depend on an initial random selection of centers. The methodology was applied on a set of 39 compounds and resulted in the development of a successful QSTR model involving only three descriptors that can predict toxicity with significant accuracy. The produced model was compared to QSTRs produced by more conventional modelling techniques, such as Multiple Linear Regression (MLR) and the popular Feedforward Neural Network (FNN) architecture. Various statistical validation techniques illustrated the efficiency of the proposed method.

2. Materials and methods

The proposed methodology was applied on a data set of heterogeneous compounds that are characterized by a narcotic mode of action. The data were taken from the literature [22]. The set is of high quality, since all data were derived from the same endpoint and protocol and were measured in the same laboratory at the Institute of Soil Science, Academia Sinica, Najing [23].

2.1. Data Set

As mentioned above, the toxicity data to *Vibrio fischeri* for the 39 compounds that constituted our data base were obtained from the literature [22]. The toxicities in terms of pEC_{50} ($\log(1/LC_{50})$) are presented in Table 1.

2.2. Descriptors

Three descriptors that give a statistically significant model were collected from the literature and used as input features in the data set, namely $\log P$ as a measure of lipophilicity of the compound, equalized electronegativity χ_{eq} and the topological index ${}^1\chi^v$ which represent the structure of the molecule. In general, all these descriptors are simple and relatively easy to calculate [24, 25].

The first order valence-connectivity index ${}^1\chi^v$ used in this work is representative of the molecule's size, shape, branching, symmetry and heterogenicity and was previously used in QSARs with success [22, 26].

The equalized electronegativity χ_{eq} which accounts for the electronegativity effect of the substituents has also proved to play a dominant role and improve the QSTR models [22, 27]. Charge conservation equation leads to the following expression:

$$\chi_{eq} = N \Big/ \sum (V/\chi) \tag{1}$$

where N = total number of atoms in the species, V is the number of atoms of a particular element in the species and χ is the electronegativity of that element.

Finally, the addition of lipophilicity in terms of $\log P$ was found to improve considerably the efficiency of the produced models. The $\log P$ values of the 39 compounds were taken from the literature [23]. A number of studies have been performed on the relationship between the toxicity and chemical structure using $\log P$. These studies indicate that lipophilicity has emerged as a key parameter for assessing toxicity [28, 29].

2.3. Statistical analysis

In this section we present the basic characteristics of the RBF neural network architecture and the training method that was used to develop the QSTR neural network models.

2.3.1. *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 neural network responses to the environment. A typical hidden node l in an RBF network is described by a vector $\hat{\boldsymbol{x}}_l$, 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:

$$v_l(\mathbf{x}) = \|\mathbf{x} - \hat{\mathbf{x}}_l\| \tag{2}$$

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

$$f_l(\boldsymbol{x}) = \exp\left(-\frac{v_l(\mathbf{x})^2}{\sigma_l^2}\right) \tag{3}$$

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

$$\hat{y} = g(\mathbf{x}) = \sum_{l=1}^{L} f_l(\mathbf{x}) w_l \tag{4}$$

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

2.3.2. *RBF Network Training Methodology*

Training methodologies for the RBF network architecture are based on a set of input-output training pairs $(\mathbf{x}(k); \mathbf{y}(k))$ $(k = 1, 2, \ldots, K)$. 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 [21]. 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

Table 1. True toxicities (pEC_{50}), the values of the input features and the predictions of the three models.

						Training set			Validation set		
A/A	Name	pEC_{50}	χ_{eq}	$^1\chi^v$	log P	(RBF) $R^2 = 0.9403$	(FNN) $R^2 = 0.8756$	(linear) $R^2 = 0.7851$	(RBF) $R^2 = 0.9337$	(FNN) $R^2 = 0.8443$	(linear) $R^2 = 0.8373$
1	4-Chlorobenzyl chloride	5.01	2.8688	2.6322	3.42	4.9997	4.9842	5.4686			
2[a]	4-Chlorobenzaldehyde	4.15	2.4179	2.5150	2.33				3.9869	3.9498	3.8489
3	3-Chlorobenzaldehyde	4.00	2.4179	2.5137	2.33	3.9870	3.9494	3.8483			
4	3,4-Dichloro-benzaldehyde	4.68	2.4820	2.6390	2.96	4.3853	4.6186	4.3615			
5	3,4-Dichlorobenzonitrile	4.22	2.6322	2.5098	2.88	4.5587	4.5934	4.6056			
6	4-Chlorobenzonitrile	4.20	2.5966	1.7751	2.29	4.3270	4.1263	3.8900			
7	4-Chlorobenzyl cyanide	4.73	2.6110	2.4820	2.28	4.2510	4.1554	4.2512			
8[a]	2-Chlorobenzyl cyanide	4.26	2.6119	2.2390	2.28				4.3106	4.1400	4.1388
9	2,4,6-Trichloroaniline	4.51	2.4853	2.5104	3.44	4.6090	4.6945	4.5426			
10	2,6-Dichloroaniline	4.16	2.4225	2.4106	2.71	4.1978	4.2671	3.9956			
11[a]	2,4-Dichloroaniline	4.09	2.4225	2.4040	2.71				4.1975	4.2643	3.9925
12	3,4-Dichloroaniline	4.20	2.4225	2.1540	2.59	4.1472	4.0543	3.8161			
13	3-Chloro-4-uoroaniline	3.28	2.3972	2.4047	2.14	3.8893	3.7958	3.6568			
14[a]	4-Chloroaniline	3.57	2.3630	2.2980	1.91				3.7198	3.7043	3.4160
15	4-Bromoaniline	3.92	2.3541	2.2980	2.06	3.7652	3.6677	3.4688			
16	2-Chloro-4-nitroaniline	3.99	2.5196	2.8030	2.06	3.9478	4.0294	4.0858			
17	2,4-Dinitroaniline	4.16	2.6004	3.2030	1.72	4.2247	4.0735	4.2935			
18[a]	4-Nitroaniline	3.70	2.4638	2.6970	1.26				3.8024	3.9304	3.5178
19	3-Nitroaniline	3.77	2.4638	2.6970	1.26	3.8024	3.9304	3.5178			
20	Diphenylamine	4.88	2.3141	4.3213	3.62	4.8763	4.3354	5.0915			
21	Aniline	3.28	2.3079	2.1990	0.92	2.8663	3.2387	2.7602			
22	Pentachlorophenol	5.69	2.6931	1.9470	4.32	5.6482	5.5685	5.1824			
23	2,4-Dichlorophenol	4.45	2.4757	2.0880	2.96	4.3862	4.4304	4.0874			
24[a]	4-Chlorophenol	4.48	2.4082	2.2320	2.49				4.0770	3.9783	3.7714
25	4-Nitrophenol	4.05	2.4082	2.2390	1.85	3.8102	3.8680	3.4624			
26	2-Methylphenol	3.75	2.5125	2.6320	1.97	3.9298	4.0071	3.9451			
27[a]	Resorcinol	3.00	2.3208	2.5509	0.88				3.0386	3.4487	2.9360
28	Phenol	3.64	2.3474	3.5430	1.48	3.6357	3.4502	3.7571			
29	Hexachloroethane	5.52	2.8571	1.1310	4.61	5.5578	5.4266	5.3150			
30	1,2-Dichloroethane	2.43	2.3760	0.5340	1.46	2.1976	2.4561	2.3951			
31	Tetrachloroethylene	3.94	2.8129	1.0060	3.48	3.8872	4.2128	4.6036			
32	Dichloromethane	1.96	2.4777	0.5346	1.25	2.2055	1.9353	2.5258			
33	1-Octanol	4.90	2.2395	4.0230	2.94	4.8875	4.3382	4.4484			
34	Cyclohexanone	2.95	2.2802	3.6150	0.87	2.9285	2.9793	3.3395			
35[a]	Acetone	0.90	2.3027	1.2041	−0.21				1.2749	1.0819	1.7282
36[a]	Cyclohexane	3.16	2.2183	3.0000	3.35				3.9345	4.3443	4.1178
37	Hexane	3.27	2.2060	2.9140	3.87	3.2958	4.3388	4.3028			
38[a]	Diethyl ether	1.68	2.2569	1.5773	0.87				2.1768	2.2693	2.3260
39	Tetrahydrofuran	1.90	2.2831	2.0773	0.46	2.2818	1.9149	2.4216			

[a] compounds used in the validation set.

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 produced RBF network model. 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 opposed to both the k-means [30] and the c-means clustering [31] 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 utilized 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 neighbour heuristic[32]:

$$\sigma_l = \left(\frac{1}{p} \sum_{i=1}^{p} \| \hat{\mathbf{x}}_l - \hat{\mathbf{x}}_i \|^2 \right)^{1/2} \tag{5}$$

where $\hat{\mathbf{x}}_1$, $\hat{\mathbf{x}}_2, \ldots$, $\hat{\mathbf{x}}_p$ are the p nearest node centers to the hidden node l. The parameter p is selected, so that many nodes are activated when an input vector is presented to the neural network model.

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

2.4. Model validation

2.4.1. *Cross – validation technique*

In order to explore the reliability of the proposed method we used the leave one-out (LOO) and the leave more-out (LMO) cross – validation method [33]. Prediction error sum of squares (PRESS) is a standard index to measure the accuracy of a modeling method based on the LOO cross-validation technique for a number of available examples n. Based on the PRESS and SSY (Sum of squares of deviations of the experimental values from their mean) statistics, the Q^2 and S_{PRESS} values can be easily calculated. The formulae used to calculate all the aforementioned statistics are presented below (Equations (6) and (7):

$$Q^2 = 1 - \frac{\text{PRESS}}{\text{SSY}} = 1 - \frac{\sum_{i=1}^{n} (y_{\text{exp}} - y_{\text{pred}})^2}{\sum_{i=1}^{n} (y_{\text{exp}} - \bar{y})^2} \tag{6}$$

$$S_{\text{PRESS}} = \sqrt{\frac{\text{PRESS}}{n}} \tag{7}$$

2.4.2. *Estimation of the predictive ability of the QSTR model*

According to Tropsha et al. [34] the predictive power of a QSAR model can be conveniently estimated by an external R^2_{CVext} (Equation (8)).

$$R^2_{\text{CVext}} = 1 - \frac{\sum_{i=1}^{\text{test}} (y_{\text{exp}} - y_{\text{pred}})^2}{\sum_{i=1}^{\text{test}} (y_{\text{exp}} - \bar{y}_{\text{tr}})^2} \tag{8}$$

where $\bar{y}_{\text{tr}}$ is the averaged value for the dependent variable on the training set.

Furthermore Tropsha et al. [34–36] considered a QSAR model predictive, if the following conditions are satisfied:

$$R^2_{\text{CVext}} > 0.5 \tag{9}$$

$$R^2_{\text{pred}} > 0.6 \tag{10}$$

$$\frac{\left(R^2_{\text{pred}} - R^2_o\right)}{R^2_{\text{pred}}} < 0.1 \quad \text{or} \quad \frac{\left(R^2_{\text{pred}} - R'^2_o\right)}{R^2_{\text{pred}}} < 0.1 \tag{11}$$

$$0.85 \le k \le 1.15 \quad \text{or} \quad 0.85 \le k' \le 1.15 \tag{12}$$

Mathematical definitions of $R^2_{\text{o}}, R'^2_{\text{o}}, k$ and $k'$are based on regression of the observed activities against predicted activities and the opposite (regression of the predicted activities against observed activities). The definitions are presented clearly in ref. (35) and are not repeated here for brevity.

2.4.3. *Y-Randomization test*

This technique ensures the robustness of the QSPR model [34, 37]. The dependent variable vector (toxicity) is randomly shuffled and a new QSAR model is developed using the original independent variable matrix. The new QSAR models (after several repetitions) are expected to have low R^2 and R^2_{cv} values. If the opposite happens then an acceptable QSAR model cannot be obtained for the specific modeling method and data.

3. Results and discussion

In order to explore the predictive ability of the proposed RBF model, the data set was initially split into a training and a validation set in a ratio of 75%:25% (29 and 10 compounds respectively). The data set was partitioned in a way that we obtained a representative training set and at the same time a diverse test set in terms of molecular structure. The compounds in the dataset included chlorobenzenes, nitrobenzenes, anilines, phenols and others. From each group, we selected at least one representative structure in the test set. The selection was also based on the values of the output parameters so that a wide range of toxicity values was included in both sets. The distribution of the toxicity values for the test set follows the distribution of the toxicity values for the training set. For example, the majority of compounds exhibit toxicity in the range between 3.00 and 5.00 pEC_{50} both in the training and

Table 2. Parameters of RBF neural network model.

	x_1	x_2	x_3	σ	w
1	0.40509	1.1289	0.16029	1.0791	1.1991
2	0.0050893	−0.2711	−0.039709	0.7180	0.92233
3	0.20509	−0.071104	0.16029	0.7087	−0.44248
4	0.20509	0.5289	−0.039709	0.7659	−2.0181
5	0.0050893	0.3289	−0.23971	0.7832	3.2526
6	−0.19491	0.1289	0.16029	0.6532	−3.5499
7	−0.39491	0.3289	0.36029	0.7832	3.2572
8	−0.59491	−0.071104	0.16029	0.6992	0.80551
9	0.60509	−0.4711	0.96029	1.1662	2.8863
10	−0.79491	−0.4711	−0.039709	0.8994	−0.36281
11	1.0051	0.7289	−0.23971	1.1235	1.4784
12	−0.39491	−0.4711	0.56029	0.7572	−1.8571
13	1.0051	1.1289	−0.63971	1.3920	8.9051
14	−0.39491	−0.2711	−1.0397	1.3047	1.522
15	0.60509	0.9289	−0.83971	1.2147	−6.077
16	0.20509	−0.6711	0.76029	0.9238	2.6542
17	−0.79491	−0.6711	0.56029	1.0154	2.1722
18	0.60509	−0.8711	0.36029	1.1450	−1.1002

the validation set. The training and validation compounds are clearly indicated in Table 1.

For the development of the RBF network we scaled the input data and used the fuzzy means procedure that was described in subsection 2.3.2. Several models were developed by altering the key tuning parameter in the fuzzy means methodology, which is the number of sets that are defined in each input dimension. The parameter p in the P-nearest neighbour heuristic method was set to half of the number of hidden nodes, so that multiple hidden node centers are activated when an input example is presented to the network. For the development of the models we used only the 29 training data. The validation set was not involved by any means during the training phase and was used only to test the accuracy of the produced models. The best results were obtained by partitioning each input dimension into 11 sets. This partition produced a network consisting of 18 hidden nodes. The parameters of the RBF model are shown in Table 2.

In order to compare the performance of the produced RBF network we developed more QSTR models using MLR and the FNN architecture, based on exactly the same training and validation data sets. For the development of the FNN model we utilized the MATLAB neural network toolbox. Several models were developed by altering the tuning parameters which are the number of hidden layers and the number of hidden nodes in each layer. We examined two different nonlinear functions, namely the hyperbolic tangent sigmoid function and the log sigmoid function. The Levenberg-Marquardt backpropagation method was utilized as the training procedure. The best FNN model consisted of one hidden layer containing 3 nodes and utilized the hyperbolic tangent sigmoid function.

The development of the MLR model is very simple and can be presented in terms of matrix algebra. Let us assume that $\mathbf{A}$ is the 29×4 dimensional matrix containing the values of the three descriptors for the 29 training compounds in the first three columns, while the fourth column elements are all equal to 1. If $\mathbf{Y}$ is the 29×1 dimensional vector containing the target pEC_{50} values for the training compounds, then the MLR model coefficients are obtained by the following formula:

$$(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{Y} \tag{13}$$

The rest of the MLR model statistics are calculated from statistical functions, included in Microsoft Excel or MATLAB. The MLR model that was obtained for our given training data is the following:

$$\begin{aligned} pEC_{50} &= 0.4878(\pm0.1963)\log P + 2.2905(\pm1.2917)\chi_{eq} \\ &\quad +0.4712(\pm0.2238)^1\chi^v - 4.0110(\pm3.2687) \\ n &= 29,\ R^2 = 0.7851,\ F = 30.45,\ Q^2 = 0.6579, \\ &\quad S_{PRESS} = 0.5250 \end{aligned} \tag{14}$$

The results are presented in Table 1, which contains the predictions of the three models for both the training and the external examples. The same results are shown in a graphical format in Figures 1–3, where the experimental toxicity is plotted against the predictions of the RBF network, the FNN 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 accuracies of all three models in terms of the R^2_{train}, R^2_{pred} and RMS statistics are summarized in Table 3.

Based on the above results and the procedures that were utilized for training RBF networks and FNNs we can state that the FNN methodology is characterized by more tuning parameters and lower prediction accuracy compared to the RBF neural network method. Another disadvantage that has been reported in the literature is that FNN training procedures are more time consuming. This disadvantage was not observed in this study due to the small size of the training data set. A thorough comparison between the two neural network architectures can be found in Ref. [38].

The results that have been presented so far clearly favour the RBF neural network model and prompted us to further explore the predictive ability of this particular model. In order to validate the RBF model, we applied the statistical tests described in subsection 2.4. More specifically, the proposed RBF neural network model passed all the tests for the predictive ability (Equations (9)–(12)):

$$\begin{aligned} &R^2_{CVext} = 0.9641 > 0.5 \\ &R^2_{pred} = 0.9337 > 0.6 \\ &\frac{\left(R^2_{pred} - R^2_o\right)}{R^2_{pred}} = -0.1144 < 0.1 \end{aligned}$$

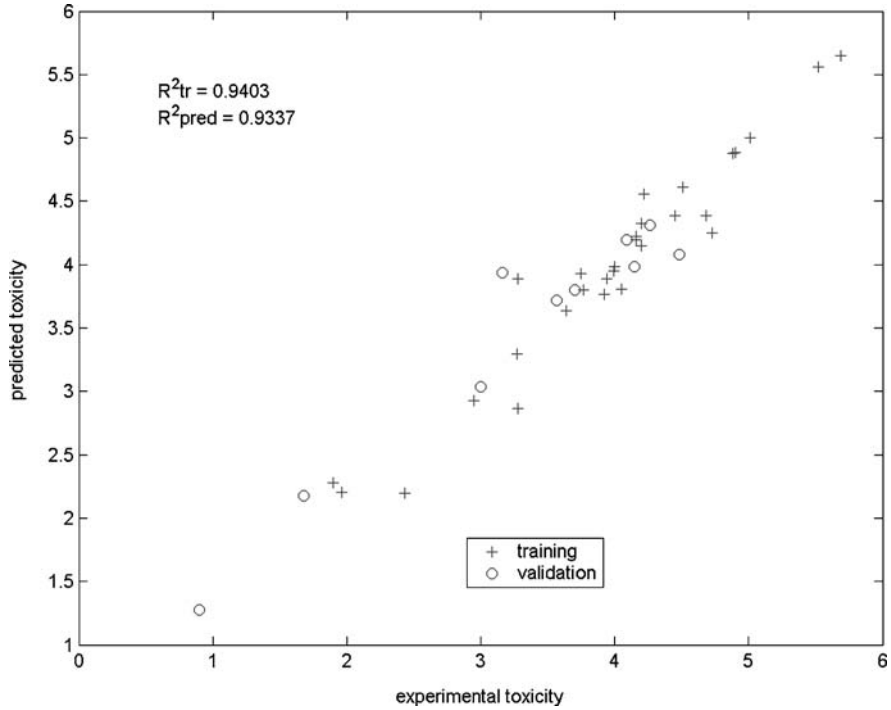

Figure 1. Experimental vs predicted toxicity for the training and validation set (RBF).

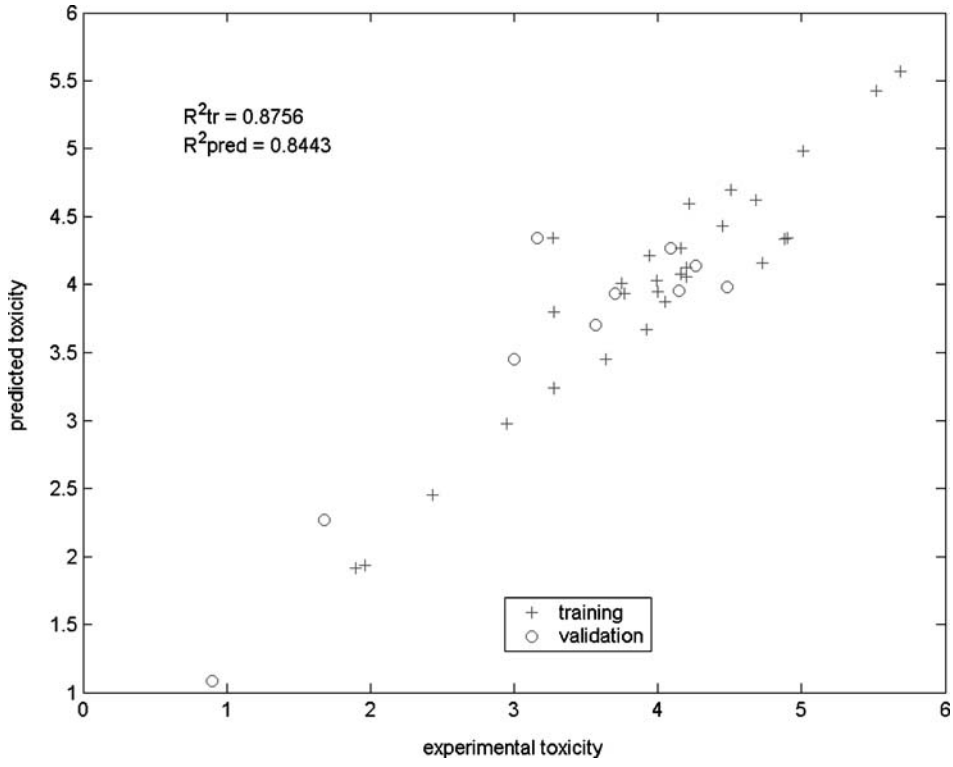

Figure 2. Experimental vs predicted toxicity for the training and validation set (FNN).

or

$$\frac{\left(R^2_{\text{pred}} - R'^2_o\right)}{R^2_{\text{pred}}} = -0.1349 < 0.1$$

$$k = 0.9684 \quad \text{and} \quad k' = 1.0233$$

For a more exhaustive testing of the predictive power of the model, apart from the standard LOO cross-validation technique, we applied a leave-five-out cross validation procedure. From the training set we randomly selected groups of five compounds. Each group was left out and that group

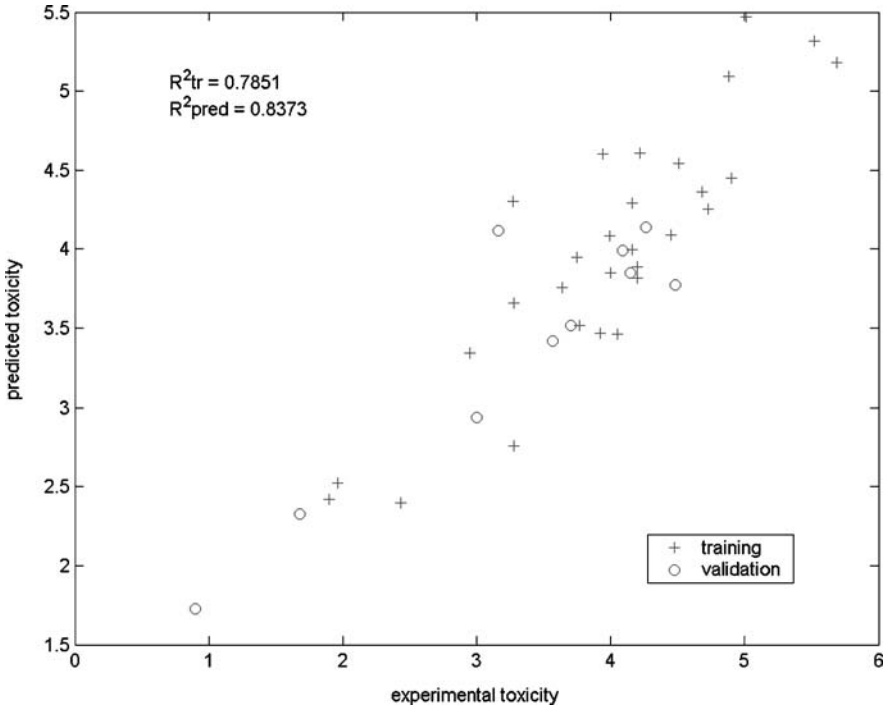

Figure 3. Experimental vs predicted toxicity for the training and validation set (MLR)

was predicted by the model developed from the remaining observations. This process was carried out 20 times.

It is important that the model is quite stable to the inclusion-exclusion of compound as indicated by the LOO and L5O correlation coefficients and Spress values, which are presented below:

$$Q^2_{\mathrm{LOO}} = 0.6712, \quad \mathrm{Spress}_{\mathrm{LOO}} = 0.5022$$
$$Q^2_{\mathrm{L5O}} = 0.7013, \quad \mathrm{Spress}_{\mathrm{L5O}} = 0.4388.$$

The Q^2_{L5O} statistic was calculated as the average R^2 for the prediction subset among the 20 different runs. Standard deviation of the statistic is equal to 0.1692. The results obtained by the LOO and L5O cross validation tests illustrated once more the quality of the obtained model.

Finally the popular randomization of response approach was utilized to establish the RBF model robustness. Based on this test, if all models produced by randomly shuffling the dependent variable present high R^2 or R^2_{CV} values, then this is the result of a chance correlation and the produced model for the given data set is not acceptable. This was not the case for the dataset and the methodology used in this work. Several random shuffles of the **Y** vector (toxicity values) were performed and the results are shown in Table 4. The low R^2 and R^2_{CV} values show that the good results in our original model are not due to a chance correlation or structural dependency of the training set.

It is important to note that the produced QSTR model uses only three descriptors and shows a joint use of lipophilicity and topological indices as molecule descriptors correlates well with the *Vibrio fischeri* toxicity. This is in agreement with previous studies [22, 23]. All three descriptors are well-established, toxicologically relevant and easy to measure.

All the training and testing procedures were implemented using the MATLAB programming language. The computational time required to build the neural network models in a Pentium IV 3GHz processor was always less than 0.2s.

Table 3. Summary of the results produced by the different methods.

Method	Training set	Validation set	R^2_{train}	R^2_{pred}	RMS	Figure
RBF	29	29	0.9403		0.2194	1
FNN	29	29	0.8756		0.3165	2
MLR	29	29	0.7851		0.4160	3
RBF	29	10		0.9337	0.3500	1
FNN	29	10		0.8443	0.4890	2
MLR	29	10		0.8373	0.5195	3

Table 4. Results of the Y-randomization test

Iteration	R^2	R^2_{cv}
1	0.1874	0.03
2	0.3258	0.06
3	0.3484	0.00
4	0.3571	0.00
5	0.2953	0.14
6	0.3268	0.12

4. Conclusions

In this work we presented a novel QSTR methodology based on the RBF neural network architecture. The method was applied on a data set of heterogeneous compounds The RBF neural network models were produced based on the fuzzy means training method, which is fast and repetitive, in contrast to most traditional training techniques. Although a linear QSTR model based on the same data set is also acceptable taking into account the simplicity and ease of interpretation, the RBF model was proven to be significantly more accurate in terms of the R^2_{train}, R^2_{pred} and RMS statistics. The RBF model also outperformed the best model obtained using the FNN architecture. Further validation of the RBF model was based on various evaluation criteria which illustrated that the proposed model has a significant predictive potential.

Acknowledgements

G.M. and A.Al. wish to thank the Greek State Scholarship Foundation for assistanship.

References

1. Lu, F.C. and Kacew, S., LU'S BASIC TOXICOLOGY, Taylor & Francis, London, 2002.
2. Parvez, S., Venkataraman, C. and Mukherji, S., *A review on advantages of implementing luminescence inhibition (Vibrio fischeri) for acute toxicity prediction of chemicals*, Environ. Int., 32 (2006) 265–268.
3. Dawson, D.A., Poch, G. and Schultz, T.W., *Chemical mixture toxicity testing with Vibrio fischeri: Combined effects of binary mixtures for ten soft electrophiles* Ecotox. Environ. Safety (2005) In press.
4. Karcher, W. and Devillers, J., *SAR and QSAR in environmental chemistry and toxicology: Scientific tool or wishful thinking?* In: Karcher, W. and Devillers, J. (Eds.). Practical applications of Quantitative Structure-Activity Relationships (QSAR) in environmental chemistry and toxicology. Kluwer, Dordrecht, The Netherlands, 1990, pp 1–12.
5. Nendza, M., Structure-Activity Relationships in Environmental Sciences, Ecotoxicology Series 6, CHAPMAN & HALL, Great Britain, 1998.
6. Schultz, T.W., Netzeva, T.I. and Cronin, M.T.D., *Selection of data sets for QSARs: Analyses of Tetrahymena Toxicity from aromatic compounds*, SAR QSAR Environ. Res., 14 (2003) 59–81.
7. Netzeva, T.I., Schultz, T.W., Aptula, A.O. and Cronin, M.T.D. *Partial least squares modelling of the acute toxicity of aliphatic compounds to tetrahymena pyriformis*, SAR QSAR Environ. Res., 14 (2003) 265–283.
8. Warne, M.A., Osborn, D., Lindon, J.C. and Nicholson, J.K., *Quantitative Structure-Activity Relationships for halogenated substituted-benzenes to Vibrio fischeri, using atom-based semi-empirical molecular-orbital descriptors*, Chemospere, 38 (1999) 3357–3382.
9. Khadikar, P.V., Mather, K.C., Singh, S., Phadnis, A., Shrivastava, A. and Mandoloi, M., *Study on quantitative structure-toxicity relationships of benzene derivatives acting by narcosis*, Bioorg. Med. Chem., 10 (2002) 1761–1766.
10. Roy, K. and Ghosh, G., *QSTR with extended topochemical indices. Part 5: Modeling of the acute toxicity of phenylsulfonyl carboxylates to Vibrio fischeri using genetic fuction approximation*, Bioorg. Med. Chem., 13 (2005) 1185–1194.
11. Roy, K. and Ghosh, G., *QSTR with extended topochemical atom indices. 4. Modeling of the acute toxicity of phenylsulfonyl carboxylates to Vibrio fischeri using principal component factor analysis and principal component regression analysis*, QSAR Comb. Sci., 23 (2004) 526–535.
12. Melagraki, G., Afantitis, A., Sarimveis, H., Igglessi-Markopoulou, O. and Supuran, C.T., *QSAR study on para-substituted aromatic sulfonamides as carbonic anhydrase II inhibitors using topological information indices*, Bioorg. Med. Chem., 14 (2006) 1108–1114.
13. Afantitis, A., Melagraki, G., Sarimveis, H., Koutentis, P. A., Markopoulos, J. and Igglessi-Markopoulou, O., *A novel simple QSAR model for the prediction of anti-HIV activity using multiple linear regression analysis*, Mol. Diversity, In press (2005).
14. Hansch, C. and Leo, A., Exploring QSAR: Fundamentals and Applications in Chemistry and Biology. ACS, Washington, DC, 1995.
15. Debnath, A.K., *Quantitative structure – activity relationship (QSAR): A versatile tool in drug design*, In: Ghose, A.K. and Viswanadhan, V.N. (Eds.) Combinatorial library design and evaluation: Principles, software tools, and applications in drug discovery, Marcel Dekker, New York, 2001, pp 73–129.
16. Devillers, J., Neural, Networks in QSAR and Drug Design. Academic Press, London, 1996.
17. Kaiser, K.L.E., *Neural Networks for effect prediction in environmental and health issues using large datasets*, Quant. Struct.-Act. Relat., 22 (2003) 185–190.
18. Kaiser, K.L.E., *The use of neural networks in QSARs for aquatic toxicological endpoints*, J. Mol. Str. (Theochem), 622 (2003) 85–95.
19. Afantitis, A., Melagraki, G., Makridima, K., Alexandridis, A., Sarimveis, H. and Igglessi-Markopoulou, O., *Prediction of high-weight polymers glass transition temperature using RBF neural networks*, J. Mol. Str. (Theochem), 716 (2005) 193–198.
20. Melagraki, G., Afantitis, A., Makridima, K., Sarimveis, H. and Igglessi-Markopoulou, O., *Prediction of toxicity using a novel RBF neural network training methodology*. J. Mol. Model., In press (2005).
21. Sarimveis, H., Alexandridis. A., Tsekouras G. and Bafas G., *A Fast and efficient algorithm for training radial basis function neural networks based on a fuzzy partition of the input space*, Ind. Eng. Chem. Res., 41 (2002) 751–759.
22. Agrawal, V.K. and Khadikar, P.V., *QSAR Study on narcotic mechanism of action and toxicity: A molecular connectivity approach to Vibrio fischeri toxicity testing*, Bioorg. Med. Chem., 10 (2002) 3517–3522.
23. Zhao, Y.H., Cronin, M.T.D. and Dearden, J.C., *Quantitative structure-activity relationships of chemicals acting by non-polar narcosis-theoretical considerations*, Quant. Struct.-Act. Relatsh., 17 (1998) 131–138.
24. Todeschini, R. and Consonni, V., Handbook of Molecular Descriptors, Methods and Principles in Medicinal Chemistry, in Series of Methods and Principles of Medicinal Chemistry Vol. 11. Wiley-VCH: Weinheim, Germany, 2000.
25. Hall, L.H. and Kier, L.B., *Issues in representation of molecular structure. The development of molecular connectivity*, J. Mol. Graph. Model., 20 (2001) 4–18.
26. Newsome, L.D., Johnson, D.E., Lipnick, R.L., Broderius, S.J. and Russom, C.L., *A QSAR study of the toxicity of amines to the fathead minnow*, Sci. Total Environ.,109 (1991) 537–551.
27. Khadikar, P.V., Lukovits, I, Agrawal, V.K., Shrivastava, S., Jaiswal, M., Gutman, I., Karmarkar, S. and Shrivastava, A., *Equalized electronegativity and topological indices: Application for modeling toxicity of nitrobenzene derivatives*. Indian J. Chem., 42A (2003) 1436–1441.
28. Zhao, Y.H., Ji, G.D., Cronin, M.T.D. and Dearden, J.C., *QSAR study of the toxicity of benzoic acids to Vibrio fischeri, Daphnia magna and carp*, Sci. Total Environ., 216 (1998) 205–215.

29. Cronin, M.T.D. and Schultz, T.W., *Structure –toxicity relationships for three mechanisms of action of toxicity to Vibrio fischeri*, Ecotox. Environ. Safety, 39 (1998) 65–69.
30. Darken, C. and Moody, J., *Fast adaptive K-means clustering: Some empirical results*. IEEE INNS International Joint Conference On Neural Networks, San Diego, CA, USA, June 17–21, 1990, Proceedings Vol. 2, 1990, 233 – 238.
31. Dunn, J.C., *A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters*, J. Cybernet., 3 (1974) 32–57.
32. Leonard, J.A. and Kramer, M.A., *Radial basis function networks for classifying process faults*, IEEE Control Systems. 11 (1991) 31–38.
33. Osten, D.W., *Selection of oprimal regression models via cross-validati*on J. Chemom., 2 (1988) 39–48.
34. Tropsha, A., Gramatica, P. and Gombar, V.K., *The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models*. Quant. Comb. Sci., 22 (2003) 69–77.
35. Golbraikh, A. and Tropsha, A., *Beware of* q^2!. J. Mol. Graph. Model., 20 (2002) 269–276.
36. Golbraikh, A. and Tropsha, A., *Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection.* Mol. Diversity, 5 (2000) 231–243.
37. Wold, S. and Eriksson, L., *Statistical validation of QSAR results*, in: Van de Waterbeemd, H., (Ed.), Chemometrics Methods in Molecular Design, VCH Weinheim (Germany) 1995, pp. 309–318.
38. Sarimveis, H., *Training algorithms and learning abilities of three different types of neural networks*, Syst. Anal. Model. Simul., 38 (2000) 555–581.