Modeling the Toxicity of Chemicals to *Tetrahymena pyriformis* Using Heuristic Multilinear Regression and Heuristic Back-Propagation Neural Networks

Iiris Kahn, Sulev Sild, and Uko Maran*

Institute of Chemistry, University of Tartu, 2 Jakobi Str., Tartu 51014, Estonia

Received June 29, 2007

During the last years, considerable effort has been devoted to model the toxicity of chemicals to *Tetrahymena pyriformis* for medium and large sized data sets using various artificial neural network (ANN) techniques. Motivation behind this has been to model highly complex relationships with nonlinear character making it possible to describe wide structural diversity within one model. The current work compares the performance of two heuristic methods in developing quantitative structure−activity relationship (QSAR) models: the best multilinear regression (BMLR) approach and the heuristic back-propagation neural networks (hBNN). The modeling is based on a diverse data set of 1371 organic chemicals with toxicity data ($\log(1/IGC_{50})$) collected from the literature. The toxicity values correspond to the static 40-h *Tetrahymena pyriformis* population growth impairment assay. The comparison of the two methods showed that the BMLR approach produces acceptable QSAR models ($R^2 = 0.726$), whereas the hBNN method produced a statistically more significant model ($R^2 = 0.826$) for the given endpoint. The hBNN method was able to relate different descriptors to the toxicity than the BMLR method. Both models were validated with an external prediction set. The descriptors in the models were analyzed and discussed.

INTRODUCTION

QSAR studies have an important part in environmental risk assessment allowing the analysis of toxicology data. Since it was noticed that the properties of compounds depend on their structure, QSARs have been used in elucidating the specific mechanisms underlying the toxic effects. At present, predictive QSARs have been recognized by the regulatory authorities as an affordable and safe alternative for toxicological measurements.[1,2] For the assessment of the environmental impact of toxicants, the unicellular ciliated protozoan, *Tetrahymena pyriformis*, is attractive for its fast growth rates and inexpensive assays. The testing method[3] has been carefully established giving the assurance of very high quality to the data produced. In addition to the environmental safety, toxicity data to this organism have proven useful in estimation of the toxic potencies of compounds to other aquatic organisms.[4−6]

Neural networks methodology has been applied in QSAR development to a variety of biological endpoints, including the toxicity to *Tetrahymena*. Table 1 lists the ANN QSAR models for *Tetrahymena* published in the literature. The studied series of compounds have different constitutions and the most attention being given to the various subsets of aromatic compounds. Panaye et al.[7] and Ren[8] have modeled nitro- and cyanoaromatic compounds (Table 1: #2, 3, 9). Phenols were modeled by Melagraki et al.,[9] Devillers,[10] Yao et al.,[11] and Ivanciuc[12] (Table 1: #1, 4−6, 18). A set of more diverse aromatic compounds was studied by Serra et al.[13] (Table 1: #11, 12), and different sets of substituted benzenes were studied by Burden et al.[14,15] (Table 1: #14−17). Craciun et al.[16] and Gini et al.[17] combined unsupervised and supervised ANNs in clustering a diverse set of 724 compounds and developing individual models for each of the clusters (Table 1: #7 and 8). The most diverse data sets were addressed by Niculescu et al.[18] and Kaiser et al.[19] (Table 1: #13 and 10), who used probabilistic ANNs with a set of fragment descriptors. On the largest data set of 1084 compounds Kaiser et al. achieved prediction statistics, $R^2 = 0.80$ and $s = 0.44$, on the external validation set of 84 compounds.

The performance of ANN models compared to the multilinear as well as rule based systems is discussed for large sets of acute aquatic toxicity data in a review by Kaiser.[20] It has been proven that ANNs are able to model diverse chemical classes with more than one mechanism of toxicity.[15] However, many improvements are needed in the QSAR development with the ANN methodology such as introduction of descriptors with improved information content, effective variable selection algorithms, optimization of the training procedure, and tools for model interpretation, often referred to as "transparency". Care must be taken to provide molecular descriptors with high discrimination ability relevant to the particular endpoint. For receiving a reliable predictive model with high generalization power it is equally important to use an appropriate feature selection algorithm that is consistent with the modeling method. Variable selection in formation of the ANN models listed in Table 1 has been mostly either linear in nature or absent, when all the calculated descriptors were included in the model or when the model descriptors were selected a priori based on subjective reasoning. However, we have previously shown[21] that in the modeling of mutagenicity effective selection of descriptors to the ANN model improves the quality of the prediction and supported this with two additional studies: on a diverse set of mutagenicity data[22] and dielectric constants.[23] This was also shown by Serra et al.[13] who

* Corresponding author phone: +372 7375254; fax: +372 7375264; e-mail: uko.maran@ut.ee.

10.1021/ci700231c CCC: $37.00
Published on Web 11/07/2007

Table 1. List of Recently Published Neural Networks QSPR Models on Toxicity to *Tetrahymena pyriformis* log(L/mmol)[a]

no.	compounds	$N(n_{test})$	h	variable selection	method	model descriptors	R^2	R^2_{test}	s_{test}	year	ref
1	phenols	221(41)	40	−	RBFNN	logK_{OW}, pK_a, E_{LUMO}, E_{HOMO}, N_{hdon}	0.94	0.88	0.24[d]	2006	9
2	nitro- and cyanoaromatics	202(41)	40	−	RBFNN	logK_{OW}, A_{max}	0.73	0.72	0.49	2006	7
3	nitro- and cyanoaromatics	202(41)			GRNN	logK_{OW}, A_{max}	0.72	0.72	0.42	2006	7
4	phenols (polar narcotics)	153	40	−	RBFNN	logK_{OW}, pK_a, E_{LUMO}, E_{HOMO}, N_{hdon}	0.89	−	−	2004	11
5	phenols	250(50)	40	a pruning algorithm	BPNN + CGDA	logD7.35, E_{LUMO}, MW, P_{NEG}	0.64	−	−	2004	10
6	phenols	250(50)	40	−	BPNN + CGDA	logD7.35, E_{LUMO}, MW, P_{NEG}, ABSQon, MaxHp, 1,4-OH, 4-NH_2, 2,6-P	0.83	−	−	2004	10
7	diverse organic compds, organized into 6 clusters	724(20%)	40	−	USNN, SNN	264 molecular descriptors		0.72	0.51	2004	17
8	diverse organic compds, organized into 10 clusters	724(20%)	40	−	USNN, BPNN	264 molecular descriptors		0.79	0.33[c]	2003	16
9	nitro- and cyanoaromatics	203(41)	40	−	NN	logK_{OW}, A_{max}	0.73	0.73	0.45	2003	8
10	diverse organic compds	1084(84)	48−60	−	PNN	41 fragment and atom counts, MW	0.90	0.80	0.44	2002	19
11	aromatic compds	448	48	SA	CNN	9 topol, 1 geom, 1 electrostatic	0.72	−	−	2001	13
12	aromatic compds	448(52)	48	GA	CNN	6 topol, 1 geom, 4 hybrid descriptors	0.88	−	0.59[d]	2001	13
13	diverse organic compds	825(75)	48	−	PNN	33 molecular fragment descriptors no. of occurrences of the functional groups of C, H, Br, Cl, F, I, N, O, and S atoms	0.93	0.89	0.30	2000	18
14	nitro- and cyanoaromatics	277(15%)	40	PCA	BRANN	32 molecular indices	0.86	−	0.08[b]	2000	
15	nitro- and cyanoaromatics	277(15%)	40	ARD	BRANN	27 molecular indices	0.84	−	0.09[b]	2000	14
16	nitro- and cyanoaromatics and phenols	278(56)	40	PCA	BRANN	5 Randic, 5 Kier and Hall, 13 atomistic indices	0.83	−	0.10[b]	2000	15
17	nitro- and cyanoaromatics and phenols	278(56)	40	PCA	BRANN	5 Randic, 5 Kier and Hall, 13 atomistic, 13 fragment indices	0.94	−	0.11[b]	2000	15
18	para-substituted phenols	30(-LOO)	40	−	NN	logK_{OW}, pK_a	0.93	0.91	0.30	1998	12

[a] N − number of compounds in a data set (n − number or % of compounds in the subset); R^2 − Pearson's correlation coefficient; s − standard deviation; BNN− Bayesian neural network; RBFNN − radial basis function neural networks; GRNN − general regression neural networks; BPNN − back-propagation neural network; CGDA − conjugate gradient decent algorithm; USNN − unsupervised neural networks; PNN − probabilistic neural network; CNN − computational neural networks; SA − simulated annealing; GA − genetic algorithms; BRANN − Bayesian-regularized neural network; ARD − automatic relevance determination; PCA − principal component analysis; LOO − leave-one-out; A_{max} − maximum superdelocalizability; K_{OW} − n-octanol/water partition coefficient; P_{NEG} − negatively charged molecular surface area; SsOH − electrotopological state index for the hydroxy group; ABSQon − sum of absolute charges on nitrogen and oxygen atoms; MaxHp − largest positive charge on a hydrogen atom; MW − molecular weight; pK_a − acid dissociation constant; E_{LUMO} − energy of the lowest unoccupied molecular orbital; E_{HOMO} − energy of the highest occupied molecular orbital; N_{hdon} − number of H-donor sites in the molecule. [b] Standard error (data scaled from 0 to 1). [c] MAE − mean absolute error. [d] rms error.

compared nonlinear descriptor selection (by a genetic algorithm coupled to a CNN fitness evaluator) with linear descriptor selection (by the simulated annealing method coupled to a linear fitness evaluator) in the ANN modeling of the toxicity of 448 aromatic compounds to *Tetrahymena pyriformis* (Table 1: #11 and 12) and proved the benefits of consistency between the nature of the relationship and the selection of the descriptors.

The goal of the present investigation is to determine the efficiency of the heuristic method for the variable selection within the framework of back-propagation neural networks (hBNN) in QSAR development for the toxicity of chemicals to *Tetrahymena pyriformis* and to compare their performance with the heuristic multilinear regression analysis. To achieve this goal a data set of 1371 compounds with large structural diversity spanning a variety of mechanisms of toxic action, including narcoses and electrophilic mechanisms, was used. The data set was characterized with a large pool of whole-molecule descriptors (including quantum-chemical descriptors) and those descriptors were used to derive QSAR models.

DATA AND METHODS

Toxicological Data and Data Set Partitioning. Median population growth impairment concentration data (log(1/IGC_{50})) to the ciliate *Tetrahymena pyriformis* for 1371 compounds from multiple literature refs 24−46 were used in the modeling. All experimental values corresponded to the 40-h static design with the population density measured spectrophotometrically as the endpoint (see protocol details

by Schultz[3]). The set of compounds included a large variety of classes according to the functional groups: *aromatic compounds* (nitro- and nitroso- compounds, cyanides, anilines, amides, phenols and thiols, halogenated compounds (Br, Cl, F, I), carboxylic acids, carbonyls, esters, ethers, pyridines, thioureas, quinones, and resorcinols, etc.) and *saturated cyclic compounds* and *aliphatic saturated* and *unsaturated compounds* (alcohols, ethers, carboxylic acids and esters, carbonyls, amides and amines, thioureas, cyanides, isothiocyanates, thiocyanates, phosphates, sulfur-containing compounds, halogenated compounds, oximes, carbamates, hydrazides, etc.). The toxicity (L/mmol) ranged from −2.67 to 3.35 log units.

In order to assess the predictive power of the developed QSAR models, the data set was partitioned into training and external prediction sets by sorting the data by the property values and moving every third compound to the prediction set. The remaining compounds were used as the training set. The sizes of the training and prediction sets were 914 and 457, respectively. In addition to external prediction the BMLR model was also validated with the leave-one-out cross-validation procedure.

Molecular Structures. The chemical structures were optimized using the Merck Molecular Force Field (MMFF)[47,48] followed by conformational search via the Monte Carlo Multiple Minimum (MCMM) search method[49,50] as implemented in MacroModel 8.0.[51] The molecular geometries corresponding to the lowest energy conformer were selected for the further calculations. In the case of stereoisomerism, the lowest energy conformer or the R-isomer was used for modeling. In the case of no cis/trans isomers specified in the literature the lower-energy trans isomer was used for modeling.

Further QSAR model development was carried out in a computational grid environment using the OpenMolGRID[52] system. The molecular structures were refined using an eigenvector following algorithm[53] for geometry optimization and AM1 semiempirical parametrization[54] in the quantum chemical program package MOPAC 7.0.[55] The optimization criterion, gradient norm 0.01 kcal/Å, was specified for each structure.

The OpenMolGRID system is an open computing grid for molecular science and engineering.[56] It provides grid-enabled components, such as a data warehouse for chemical data, software for building QSPR/QSAR models, and molecular engineering tools for generating compounds with predefined chemical properties or biological activities. The examples of its use include the modeling of inhibition of aspartyl protease enzyme,[57] acute toxicity to the fathead minnow,[58] aqueous solubility,[59] and avian oral toxicity.[60]

Molecular Descriptors. The semiempirical MOPAC calculations were followed by the molecular descriptor calculation. The MDC module from CODESSA Pro[61] as integrated in the OpenMolGRID system was used for the calculation of molecular descriptors. On average, around 600 molecular descriptors were calculated for each structure. The calculated set of descriptors was diverse; it contained descriptors from the following categories: constitutional, topological, geometric, electrostatic, and quantum chemical. In addition, logP was calculated using KowWin software[62] and included in the descriptor pool. The experimental logP values were used when available.

Best Multilinear Regression. The statistical models were developed and evaluated by means of the best multilinear regression (BMLR) analysis[63] module MDA of CODESSA Pro[61] as integrated in OpenMolGRID.[52] During the BMLR procedure the pool of descriptors is cleared from insignificant descriptors ($R^2 < 0.1$) and the descriptors with missing values followed by the construction of the best two-parameter regression, the best three-parameter regression, etc. based on the statistical significance and noncollinearity criteria ($R^2 < 0.6$) of the selected descriptors. In BMLR, the descriptor scales are normalized and centered automatically, with the final result given in natural scales. The final model has the best representation of the activity in the given descriptor pool within the given number of parameters. The quality of the models was assessed by the coefficient of determination (R^2), cross-validated (leave-one-out) coefficient of determination (R^2_{CV}), standard error of the estimate (s), and the Fisher's criterion (F).

Heuristic Back-Propagation Neural Networks. In-house artificial neural network software was used to calculate the ANN models. The experimental data and descriptor values were scaled for the ANN treatment. Feed-forward multilayer neural networks[64,65] with input, hidden, and output layers were used to represent nonlinear QSAR models. The sigmoid activation function was used for neurons in the network. The back-propagation algorithm with a momentum term was used to train the ANN models. The overtraining of neural networks was avoided with a cross-validation technique using early stopping.[66] For early stopping, one-third of the training set (304 compounds) was further separated and used as a validation set. The validation set error was monitored to stop the training process when its error started increasing. Dynamics of the validation set error is a good indicator for signs of overtraining, because the validation set is not used for updating the neural network weights.

The significant descriptors for the ANN models were selected by the heuristic forward selection algorithm (hBNN). The hBNN descriptor selection algorithm started with the preselection of molecular descriptors. If two descriptors in the list were very highly intercorrelated, then only the first descriptor was selected, and also descriptors with an insignificant variance were rejected. This procedure helps to speed up the descriptor selection and reduces the risk of including unrelated descriptors into the model. The descriptor selection algorithm started by evaluating ANN models ($1 \times 1 \times 1$) with one descriptor as input. The best models were then selected in the next step, where a new descriptor was added to the input layer, and the number of hidden units was increased by one. Again, the best models were selected, and this stepwise procedure was repeated until the addition of new input parameters did not improve the model significantly. Since ANN models are quite likely to converge to some local minima, each model was retrained 30 times, and the model with the lowest error was selected. To speed up the descriptor selection phase, a number of techniques were applied to the training process of individual ANN models: (1) a higher learning rate was used for more rapid convergence; (2) a more aggressive early stopping criterion; and (3) a limited number of epochs were used since the reduction of errors is more rapid in the early phases of the training. A limited number of final models with the smallest errors were further optimized in more detail.

Table 2. Six-Parameter BMLR Model and Training Set of 914 Compounds[a]

descriptor	X	ΔX	t-test
intercept	−0.39147	0.17073	−2.29
HASA-2/TMSA (AM1) (all)	21.15992	1.35766	15.58
logP (Exp|Calc)	0.33695	0.02335	14.43
WPSA1 (Zefirov PC)	0.00987	0.00083	11.86
final heat of formation	0.00592	0.00062	9.55
HOMO−LUMO energy gap	−0.12961	0.01376	−9.42
number of N atoms	−0.27090	0.04634	−5.85

[a] $R^2 = 0.726$, $R^2_{CV} = 0.721$, $F = 401.1$, $s = 0.551$.

RESULTS AND DISCUSSION

Multilinear Model. The BMLR analysis produced a number of multilinear regression models that involved 2−9 descriptors. The analysis of the statistical criteria of the training and prediction sets suggested the six-parameter model to be optimal. Even though additional descriptors moderately improved statistical parameters of the training set, the leave-one-out cross-validation squared correlation coefficient did not improve comparably. In addition, the statistical parameters of the prediction set started to decrease, indicating that the additional descriptors actually reduce the predictive power of the model. The details of the model are given in Table 2 and in Figure 1.

The model's statistical parameters for the training set of 941 compounds and the external prediction set of 457 compounds, presented in Table 3, indicate very high stability; the respective squared correlation coefficients (and standard deviations) were $R^2 = 0.726$ ($s = 0.556$) and $R^2 = 0.720$ ($s = 0.557$). For the entire data set the corresponding values are $R^2 = 0.724$, $s = 0.555$. The high stability of the model was also confirmed by the leave-one-out cross-validation coefficient, $R^2_{CV} = 0.716$.

The list of descriptors in Table 2 is given in the order of significance according to the t-test. HASA-2/TMSA (AM1) (all) is the square root of the hydrogen-bonding acceptor surface area (HASA) normalized by the total molecular solvent accessible surface area (TMSA)

$$\text{HASA-2} = \sum_A \sqrt{S_A} \qquad (1)$$

where the summation is performed over the surface areas of all possible H-bonding acceptor sites in the molecule (S_A). The solvent accessibility is accounted for by adding 1.5 Å to the van der Waals' radius of the acceptor atom. All heteroatoms in the molecule are considered as the H-bond acceptors: N, O, S, P, and halogens. Limiting conditions considering the coordination number are applied as follows: N atoms < 4, S < 3, and P < 4. AM1 semiempirical parametrization was used to obtain the 3D geometries.

Hydrophobicity, logP (Exp|Calc), is estimated from the partitioning of the chemical between the phases of 1-octanol and water. LogP is essentially a complex empirical descriptor that is beneficial in modeling diverse data sets related to biological properties that include a similar phenomenon to partitioning between the aqueous media and the lipid barrier of the cell membrane. The toxic potency of a large group of polar and nonpolar narcotics depends explicitly on logP as the macromolecular constituents of this barrier are the site of action for narcotic compounds,[67] while the rest of the compounds need to be absorbed into the cell to reach their targets. In the last case membrane penetration may or may not be the limiting phase of the whole process, and this descriptor alone may not be sufficient to model the potential danger to the organism. HASA-2/TMSA and logP can be addressed as descriptors related to interactions that are important for the molecule in reaching the target and binding to it. The coefficients of these descriptors indicate that an increase in toxicity is accompanied by an increasing hydrophobicity, while, at the same time, the molecule's ability to form H-bonds with target biomolecules is essential. In other words, the highest toxic potency range is determined by the balance between the lipophilicity and the H-bond formation capability because, at a higher logP, the value of HASA will start decreasing, which leads to the decrease of toxicity.

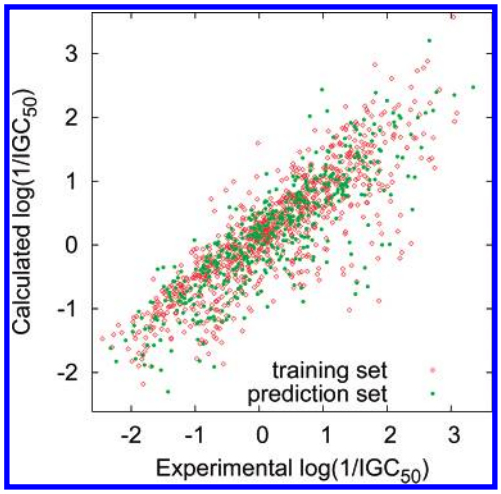

Figure 1. Plot of calculated versus experimental values of the toxicity of compounds, $\log(1/IGC_{50})$, to *T. pyriformis* calculated with the six-parameter BMLR model.

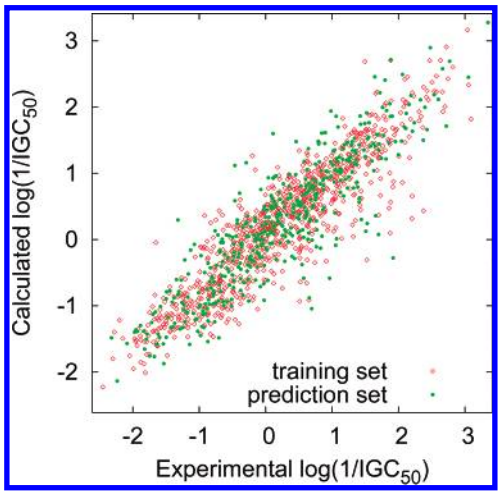

Figure 2. Plot of calculated versus experimental values of the toxicity of compounds, $\log(1/IGC_{50})$, to *T. pyriformis* calculated with the ANN model.

WPSA-1 (Zefirov) − surface weighted charged partial positive surface area is another surface area descriptor that takes into account the solvent accessible positively charged surface area of the molecule

$$\text{WPSA-1} = \sum_P S_P \cdot \frac{\text{TMSA}}{1000} \qquad (2)$$

Table 3. Statistical Parameters for the Best Linear and Nonlinear Models

method	N_{desc}	$n_{train}/n_{test}/n_{valid}$	R^2_{train}	$R^2_{train(CV)}$	s_{train}	R^2_{test}	s_{test}	R^2_{valid}	s_{valid}
BMLR	6	941/−/457	0.726	0.721	0.551	−	−	0.720	0.561
hBNN	6	610/304/457	0.826	−	0.442	0.820	0.449	0.794	0.484

Table 4. Descriptors in the hBNN Model and Training and Test Sets of 610 and 304 Compounds, Respectively[a]

symbol	descriptor
logP	logP (Exp|Calc)
E_{LUMO}	LUMO energy
$ASIC^0$	average structural information content (order 0)
HDPSA(v2)	H-donors PSA (version 2)
Q_{min}	min. net atomic charge
#N/#all	relative number of N atoms

[a] $R^2 = 0.826$, $s = 0.442$.

where the summation is performed over the solvent accessible surface areas (S_P) of all positively charged atoms ($q_P > 0$) in a molecule. In this case the charges have been calculated by Zefirov's electronegativity equalization method.[68] The highest values are possessed by alcohols and esters with long carbon chains. WPSA-1 also describes an aspect of the compounds connected with the readiness of membrane penetration or interacting with it.

The nature of the next descriptors can be associated with the chemical reactivity of the compounds. The final heat of formation, ΔH_f, calculated with the AM1 parametrization is an enthalpy change of a reaction in which the compound is formed from its composite elements in the most stable conformation in the gas phase. In the current data set, a higher heat of formation can be associated with less stability, and, therefore, the higher activity potential of the compound leading to higher toxicity as heat may be given off in the reaction to obtain a more stable lower energy state.

The energy difference between the highest occupied and the lowest unoccupied molecular orbitals, the HOMO−LUMO energy gap, is proportional to the absolute hardness of the molecule that has been defined by Parr and Pearson[69,70] as the second derivative of the energy with respect to the number of electrons, and thus it is related to the electronic chemical potential. The model coefficient for this descriptor suggests that compounds that have lower hardness, which encompasses easier polarizability and higher reactivity, are more likely to express an elevated level of toxicity.

And finally, the last descriptor, the number of nitrogen atoms, takes into account the impact of compounds with nitrogen containing functional groups on the population growth of the ciliate. Many of these compounds are capable of forming strong intermolecular hydrogen bonds or may be subject to ionization depending on the pH of the medium. These abilities make them behave differently from their homologues which do not contain N atoms.[71]

Neural Network Model. The hBNN approach was used to produce QSAR models with up to eight descriptors as input. By analyzing these models, we found that QSAR models with six descriptors having ANN architecture with seven units in a hidden layer ($6 \times 7 \times 1$) are adequate for modeling this data set. The final ANN QSAR model is presented in Tables 3 and 4 and in Figure 2. The squared correlation coefficient of the training set was $R^2 = 0.826$. The external validation set showed a prediction quality of $R^2 = 0.794$. The overall performance of the training, test, and prediction sets together was $R^2 = 0.813$.

The model includes simple and straightforward whole molecule descriptors. Two of them are well-known in predictive toxicology, namely, the logarithm of the octanol/water partition coefficient (logP (Exp|Calc)) that was also present in our BMLR model and the energy of the lowest unoccupied molecular orbital (LUMO energy). A model containing only these descriptors forms the so-called response surface approach[72] where logP describes membrane penetration and interaction with the site of action, and the LUMO energy quantifies the electrophilic potency of the biochemically reactive compounds.

Average structural information content (order 0),[73] the next descriptor selected in the ANN model, is based on Shannon's information theory and is defined as follows

$$ASIC^0 = \frac{-\sum_i \frac{n_i}{n} \log_2 \frac{n_i}{n}}{\log_2 n} \quad (3)$$

where n_i is the number of atoms in the ith class, and n is the total number of atoms in the molecule. The order of the index is determined by the coordination sphere that is dependent on the division of the atoms into different classes. $ASIC^0$ encodes the constitutional diversity of a molecule, such as the content of heteroatoms and unsaturated bonds. The highest values are possessed by small and heterogeneous molecules that are often exceptionally reactive, such as small α-haloactivated carbonyls and nitriles.

H-donors positive surface area (version 2), HDPSA(v2), is the sum of the solvent accessible surface areas over all the hydrogens in the molecule that can be donated (S_D) for H-bond formation with other molecules:

$$HDPSA(v2) = \sum_D S_D \quad (4)$$

The H atoms included in this descriptor originate from the following structural environments: H−O, H−S, H−N, H−C−C═O, and H−C−C≡N.

Minimal net atomic charge is the value of the most negative atomic partial charge in the molecule and is calculated by the Mulliken scheme. The highest absolute values for this descriptor were possessed by four phosphorus-containing compounds and by seven −S═O-containing compounds. The lowest absolute values belong to (poly)-chlorinated compounds, thiocyanates, cyanides, and cyclic carbohydrates. The atomic partial charges are considered as static chemical reactivity indices.[74] The net atomic charges on atoms are nondirectional reactivity indices[75] indicating the strength of a possible reactive center.

The relative number of N atoms differentiates the molecules according to their content of nitrogen atoms relative to the size of the molecule that is accounted for by dividing the number of N atoms by the total number of atoms in the

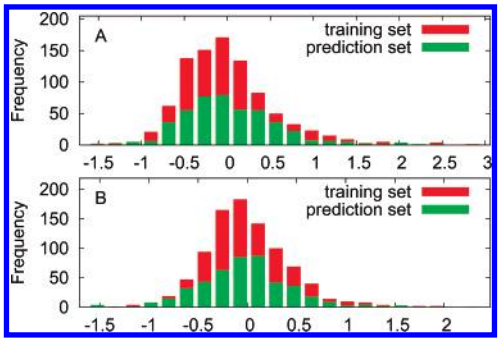

Figure 3. A. Plot of residuals of the toxicity log(1/IGC$_{50}$) calculated with the BMLR model and B. plot of residuals calculated with the 6 × 7 × 1 ANN model.

molecule. In the present data set the value of this descriptor is the highest for small compounds with two nitrogen atoms.

In the hBNN model, LogP is confirmed to be the most influential descriptor in describing the toxicity of heterogeneous chemicals. The HDPSA(v2) represents the H-bond mediated interactions, ASIC^0 takes into account the role of heteroatoms and unsaturated bonds, and the relative number of N atoms brings out the specific nature of nitrogen rich molecules to the growth of *Tetrahymena*. The two quantum chemical descriptors, the LUMO energy and the minimum net atomic charge, describe the intrinsic chemical reactivity of the compounds that lead to elevated levels of toxicity.

As a conclusion to the model building phase we may draw attention to the fact that nonlinear descriptor selection found different descriptors to the model than the linear procedure; the hBNN model is also accompanied by better statistical parameters for the model as well as for the validation of the model (Table 3).

Examination of the plots of the residuals for the BMLR and hBNN models (Figure 3A,B) reveals that the distribution of the residuals of the BMLR model is not symmetric. The observed shift from zero to the negative scale is an indication that the linear method was less successful in modeling the compounds exhibiting elevated levels of toxicity from the baseline defined by the logP (Exp|Calc). In the following discussion about the outliers general comparisons between the linear and nonlinear model were made using the limit of 2*s* (standard deviation eq 5) that equals approximately 5% of the data for both of the models. For the calculation of standard deviation in CODESSA Pro the following formula is used

$$s = \sqrt{\frac{\sum(\text{obs} - \text{pred})^2}{N - k - 1}} \qquad (5)$$

where obs and pred denote the experimental property value and the value calculated using the QSAR model, *N* is the number of data points, and *k* is the number of descriptors in the QSAR model. According to eq 5, the value of 2*s* for the BMLR model is 1.11, and for the hBNN model it is 0.91. For a more detailed evaluation, we considered a deviation of 1.5 units on the log scale from the experimental value.

In the BMLR model, the largest underestimates (Table 5) were the relatively toxic (log(1/IGC$_{50}$) > 1.4) small carbonyl-containing α,*β*-unsaturated compounds (from highest to lowest residual): acrolein (2-propenal), 3-butyn-2-one, ethyl propiolate, 3-buten-2-one, 1-penten-3-one, and 1-hexen-3-one. Extended examination (up to 2*s*) of the outliers of this group shows decreasing values of the residuals with longer carbon chains. On the one hand, the longer alkyl chain decreases reactivity of these compounds, and, on the other hand, hydrophobicity of the compounds is increasing and compensating for the deficiency in the description of the reactivity. The carbonyl-containing α,*β*-unsaturated compounds have been characterized as electrophilic capable of several mechanisms of action within the cell.[45] The carbonyl group in an aldehyde, ketone, or ester is the most common electron-withdrawing group which determines the compound's toxic action: electrophilic activity for aldehydes at the terminal carbonyl and baseline narcoses for the other two. Due to unsaturation in the α,*β*-position conjugated to the carbonyl moiety, a polarized compound is formed, and the reaction is likely to be addition across the carbon−carbon double bond which has reduced electron density. α,*β*-Unsaturation at the terminal position of the molecule is more readily attacked by nucleophiles making these compounds even more reactive. Acrolein is considered unique and forms a subclass of its own, having both a terminal vinyl group and a terminal carbonyl group.

The second significant group of aliphatic chemicals with highly underestimated toxicity values was the α-haloactivated compounds: 1-bromo-3,3-dimethyl-2-butanone, 2-bromoacetamide, 1,3-dichloroacetone, 2,5-dibromo-3,4-hexanedione, 2,3-dibromopropanamide, ethyl-2,3-dibromopropionate, and 1,4-dibromo-2,3-butanedione. In addition to these carbonyl-containing compounds, this subgroup also includes two of the smallest nitriles: bromoacetonitrile and dibromoacetonitrile. The members of this group of compounds are reactive preferably by the S$_N$2 substitution mechanism with the halogen atom as a strong leaving group. Most of the

Table 5. Outliers to the BMLR Model: Compounds with Residuals Higher than 1.5 log(L/mmol)

ID	name	exp	calc	residual
	Training Set			
373	2-propenal (acrolein)	1.873	−0.874	2.747
1225	3-butyn-2-one	1.412	−1.019	2.431
199	methoxyhydroquinone	2.202	−0.207	2.409
871	1-bromo-3,3-dimethyl-2-butanone	2.377	0.027	2.350
427	bromoacetonitrile	2.229	−0.114	2.343
1362	2-bromoacetamide	1.523	−0.574	2.097
192	methylhydroquinone	1.858	−0.209	2.067
875	(2*S*,5*S*)-2,5-dibromo-3,4-hexanedione	2.168	0.214	1.954
1226	3-hexyn-2-one	1.319	−0.462	1.781
872	3-chloro-2,4-pentanedione	1.444	−0.324	1.768
862	ethyl (*R*)-2,3-dibromopropionate	2.208	0.466	1.741
1229	1-octen-3-one	1.914	0.176	1.738
873	1,4-dibromo-2,3-butanedione	1.762	0.029	1.733
55	4-amino-2-methylphenol	1.307	−0.375	1.682
1025	4-nitroaniline	1.880	0.334	1.546
1212	benzyl isothiocyanate	2.745	1.208	1.537
1087	carbon tetrachloride	−0.020	1.597	−1.617
	Prediction Set			
533	ethyl propiolate	1.699	−0.650	2.349
354	3-buten-2-one	1.506	−0.772	2.278
260	1-penten-3-one	1.527	−0.591	2.118
877	1,3-dichloroacetone	2.046	0.009	2.037
1228	1-hexen-3-one	1.656	−0.356	2.012
1366	(*R*)-2,3-dibromopropanamide	1.921	0.002	1.919
428	dibromoacetonitrile	2.398	0.555	1.843
166	4-amino-2,3-dimethylphenol	1.440	−0.147	1.587
1120	2-hydroxyethyl acrylate	0.690	−0.885	1.575

Table 6. Outliers to the hBNN Model: Compounds with Residuals Higher than 1.5 log(L/mmol)

ID	name	exp	calc	residual
	Training Set			
871	1-bromo-3,3-dimethyl-2-butanone	2.377	0.434	1.943
199	methoxyhydroquinone	2.202	0.318	1.884
1362	2-bromoacetamide	1.523	−0.308	1.831
373	2-propenal (acrolein)	1.873	0.096	1.777
	Test Set			
1122	(*R*)-2-hydroxypropyl acrylate	0.650	−0.913	1.563
862	ethyl (*R*)-2,3-dibromopropionate	2.208	0.657	1.551
459	methacrylonitrile	−1.653	−0.046	−1.607
	Prediction Set			
1366	(*R*)-2,3-dibromopropanamide	1.921	−0.276	2.196
1120	2-hydroxyethyl acrylate	0.690	−1.043	1.733
533	ethyl propiolate	1.699	0.078	1.621
512	2,2′,4,4′-tetrahydroxybenzophenone	0.960	−0.581	1.541
730	(*R*)-1-hexyn-3-ol	0.657	−0.880	1.537
1224	3-methyl-2-cyclopenten-1-one	−1.323	0.292	−1.615
401	2,2,2-trichloroethanol	−0.465	1.116	−1.581

compounds contain bromine as the halogen, which has been proved to elicit a higher toxic effect than chlorine in theoretical structure−activity relationship studies by Schultz and co-workers.[46,76−78] Reactivity of the compounds was shown to be highest when more than one halogen was present and when at the same time a halogen atom was at the terminal position of the molecule.

Aromatic compounds were estimated with relatively greater accuracy than the aliphatic ones by the BMLR model. With residuals greater than 1.5 log units were representatives of such chemical classes as hydroquinones (methoxyhydroquinone and methylhydroquinone) and para-substituted phenols (4-amino-2-methylphenol and 4-amino-2,3-dimethylphenol). These compounds have been common outliers in the QSAR modeling of aquatic toxicity.[79] Both, hydroquinones and para-substituted phenols, are susceptible to oxidation to the respective quinone which has a toxicity greater than that of other phenols. These compounds react by free radical formation that initiates a number of competing processes within the cell.[79,34] 4-Nitroaniline is another frequent outlier with QSAR modeling of aquatic toxicity.[29] It has been proposed that this compound's excess reactivity is due to its abiotic transformation.[80]

A relatively harmless compound to the unicellular ciliate compared to the above outliers, carbon tetrachloride, was the only largely overestimated chemical with the residual of 1.617.

The hBNN model has been able to reduce the uncertainties of the estimation of toxicity compared to the values produced by the BMLR model and, in several cases, to model the problematic compounds with acceptable accuracy. Although, the largest outliers of the training set in Table 6 match with those of the BMLR model, further examination of the outliers revealed almost no remarkable trends in the nature of the compounds. Some of the statistical outliers can be classified as structurally quite unique. Others belong to the chemical classes represented in the training set by greater numbers, which suggests a unique mechanism for the outlying toxicant compared to the other compounds in its class. Mention could be made only about a small group of α-halocarbonyls with underestimated toxicities of over 1.5 log units belonging to the classes of ketones, amides, and esters: 1-bromo-3,3-dimethyl-2-butanone (#871), 2-bromoacetamine (#1362), 2,3-dibromopropionate (#862), and 2,3-dibromopropanamide (#1366).

In general, the compounds with reactive mechanisms that appear more toxic to *T. pyriformis* were modeled with moderate accuracy by the multilinear method. Several series of chemicals with strongly underestimated toxicity containing similar chemically reactive structural features were identified. These patterns call for further treatments with additional methods of analysis, such as the preclassification of the compounds. The statistical parameters of the hBNN model and validation are in good agreement with the experimental accuracy that is characteristic of structurally and mechanistically diverse biological data. Despite the relatively small number of easily interpretable molecular descriptors and the high predictive ability of the hBNN model, the difficulty of finding structural regularities among the compounds with higher residuals and that of quantifying the role of each descriptor in the description of the endpoint can be viewed as drawbacks of this modeling method.

CONCLUSIONS

Modeling with neural networks improved the results in comparison with multilinear regression; the R^2 of the training set improved considerably, from 0.726 to 0.826. The heuristic method that was directly incorporated in the ANN modeling algorithm enabled the selection of the descriptors from large descriptor pools taking into account the nonlinear relationship of each descriptor with the response. In the distributed computational grid environment, a large number of descriptors can be easily calculated from molecular structures and then efficiently selected in the model based on the statistical criteria determined by the heuristic algorithm. According to the information contained by the descriptors they describe either the chemicals' interactions with the constituents of the studied organism or their reactivity potential. Considering the very high diversity of the data set, the hBNN model provided excellent prediction with $R^2 = 0.794$ and $s = 0.484$ of the set of 457 compounds not used in model development and, therefore, could be suggested for use in broad screening of the danger posed by industrial chemicals to the aquatic environment.

ACKNOWLEDGMENT

The authors are grateful to Prof. T. Wayne Schultz (The University of Tennessee, College of Veterinary Medicine, Knoxville, TN) for the comments on the *Tetrahymena pyriformis* data. Preparation of the manuscript was supported by the Estonian Science Foundation (Grant Nos. 5805 and 7153) and the EU FP6 program Chemomentum (IST-5-033437).

REFERENCES AND NOTES

(1) Walker, J. D. Applications of QSAR in Toxicology: a US Government Perspective. *J. Mol. Struct. (THEOCHEM)* **2003**, *622*, 167−184.

(2) White paper, COM (2003) Directive of the European Parliament and of the Council: Amending council Directive 67/548/EEC in Order to Adapt it to Regulation (EC) of the European Parliament and of the Council Concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals. 2003/0256/COD, 2003/0257/COD (October 29, 2003).

(3) Schultz, T. W. TETRATOX: *Tetrahymena pyriformis* Population Growth Impairment Endpoint - A Surrogate for Fish Lethality. *Toxicol. Methods* **1997**, *7*, 289−309.

(4) Dimitrov, S. D.; Mekenyan, O. G.; Sinks, G. D.; Schultz, T. W. Global Modelling of Narcotic Chemicals: Ciliate and Fish Toxicity. *J. Mol. Struct. (THEOCHEM)* **2000**, *622*, 63−70.
(5) Seward, J. R.; Hamblen, E. L.; Schultz, T. W. Regression Comparisons of *Tetrahymena pyriformis* and *Poecilia reticulata* Toxicity. *Chemosphere* **2002**, *47*, 93−101.
(6) Kahn, I.; Maran, U.; Benfenati, E.; Netzeva, T. I.; Schultz, T. W.; Cronin, M. T. D. Comparative Quantitative Structure-Activity-Activity Relationships for Toxicity to *Tetrahymena pyriformis* and *Pimephales promelas*. *ATLA-Altern. Laborat. Anim.* **2007**, *35*, 15−24.
(7) Panaye, A.; Fan, B. T.; Doucet, J. P.; Yao, X. J.; Zhang, R. S.; Liu, M. C.; Hu, Z. D. Quantitative Structure-Toxicity Relationships (QSTRs): A Comparative Study of Various Non Linear Methods. General Regression Neural Network, Radial Basis Function Neural Network and Support Vector Machine in Predicting Toxicity of Nitro- and Cyano- Aromatics to *Tetrahymena pyriformis*. *SAR QSAR Environ. Res.* **2006**, *17*, 75−91.
(8) Ren, S. J. Modeling the Toxicity of Aromatic Compounds to *Tetrahymena pyriformis*: The Response Surface Methodology with Nonlinear Methods *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1679−1687.
(9) Melagraki, G.; Afantitis, A.; Makridima, K.; Sarimveis, H.; Igglessi-Markopoulou, O. Prediction of Toxicity Using a Novel RBF Neural Network Training Methodology. *J. Mol. Model.* **2006**, *12*, 297−305.
(10) Devillers, J. Linear versus Nonlinear QSAR Modeling of the Toxicity of Phenol Derivatives to *Tetrahymena pyriformis*. *SAR QSAR Environ. Res.* **2004**, *15*, 237−249.
(11) Yao, X. J.; Panaye, A.; Doucet, J. P.; Zhang, R. S.; Chen, H. F.; Liu, M. C.; Hu, Z. D.; Fan, B. T. Comparative Study of QSAR/QSPR Correlations Using Support Vector Machines, Radial Basis Function Neural Networks, and Multiple Linear Regression. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1257−1266.
(12) Ivanciuc, O. Artificial Neural Networks Applications. Part 4 - Quantitative Structure-Activity Relationships for the Estimation of the Relative Toxicity of Phenols for *Tetrahymena*. *Rev. Roum. Chim.* **1998**, *43*, 255−260.
(13) Serra, J. R.; Jurs, P. C.; Kaiser, K. L. E. Linear Regression and Computational Neural Network Prediction of *Tetrahymena* Acute Toxicity for Aromatic Compounds from Molecular Structure. *Chem. Res. Toxicol.* **2001**, *14*, 1535−1545.
(14) Burden, F. R.; Ford, M. G.; Whitley, D. C.; Winkler, D. A. Use of Automatic Relevance Determination in QSAR Studies Using Bayesian Neural Networks. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1423−1430.
(15) Burden, F. R.; Winkler, D. A. A Quantitative Structure-Activity Relationships Model for the Acute Toxicity of Substituted Benzenes to *Tetrahymena pyriformis* Using Bayesian-Regularized Neural Networks. *Chem. Res. Toxicol.* **2000**, *13*, 436−440.
(16) Craciun, M. V.; Neagu, D. C.; Konig, C.; Bumbaru, S. A Study of Aquatic Toxicity Using Artificial Neural Networks. *Lect. Notes Artif. Intell.* **2003**, *2774*, 911−918.
(17) Gini, G.; Craciun, M. V.; Konig, C.; Benfenati, E. Combining Unsupervised and Supervised Artificial Neural Networks to Predict Aquatic Toxicity. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1897−1902.
(18) Niculescu, S. P.; Kaiser, K. L. E.; Schultz, T. W. Modeling the Toxicity of Chemicals to *Tetrahymena pyriformis* Using Molecular Fragment Descriptors and Probabilistic Neural Networks. *Arch. Environ. Contamin. Toxicol.* **2000**, *39*, 289−298.
(19) Kaiser, K. L. E.; Niculescu, S. P.; Schultz, T. W. Probabilistic Neural Network Modeling of the Toxicity of Chemicals to *Tetrahymena pyriformis* with Molecular Fragment Descriptors. *SAR QSAR Environ. Res.* **2002**, *13*, 57−67.
(20) Kaiser, K. L. E. The Use of Neural Networks in QSARs for Acute Aquatic Toxicological Endpoints. *J. Mol. Struct. (THEOCHEM)* **2003**, *622*, 85−95.
(21) Karelson, M.; Sild, S.; Maran, U. Non-Linear QSAR Treatment of Genotoxicity. *Mol. Simulat.* **2000**, *24*, 229−242.
(22) Maran, U.; Sild, S. QSAR Modeling of Genotoxicity on Non-Congeneric Sets of Organic Compounds. *Artif. Intell. Rev.* **2003**, *20*, 13−38.
(23) Sild, S.; Karelson, M. A General QSPR Treatment for Dielectric Constants of Organic Compounds. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 360−367.
(24) Schultz, T. W.; Cronin, M. T. D.; Netzeva, T. I.; Aptula, A. O. Structure-Toxicity Relationships for Aliphatic Chemicals Evaluated with *Tetrahymena pyriformis*. *Chem. Res. Toxicol.* **2002**, *15*, 1602−1609.
(25) Schultz, T. W. Structure-Toxicity Relationships for Benzenes Evaluated with *Tetrahymena pyriformis*. *Chem. Res. Toxicol.* **1999**, *12*, 1262−1267.
(26) Schultz, T. W.; Netzeva, T. I.; Cronin, M. T. D. The Use of Diversity versus Representativity in the Training and Validation of Quantitative Structure-Activity Relationships. *SAR QSAR Environ. Res.* **2003**, *14*, 59−81.
(27) Dearden, J. C.; Cronin, M. T. D.; Schultz, T. W.; Lin, D. T. QSAR Study of the Toxicity of Nitrobenzenes to *Tetrahymena pyriformis*. *Quant. Struct.-Act. Relat.* **1995**, *14*, 427−432.
(28) Seward, J. R.; Hamblen, E. L.; Schultz, T. W. Regression Comparison of *Tetrahymena pyriformis* and *Poecilia reticulata* Toxicity. *Chemosphere* **2002**, *47*, 93−101.
(29) Cronin, M. T. D.; Manga, N.; Seward, R.; Sinks, G. D.; Schultz, T. W. Parameterization of Electrophilicity for the Prediction of Toxicity of Aromatic Compounds. *Chem. Res. Toxicol.* **2001**, *14*, 1498−1505.
(30) Seward, J. R.; Cronin, M. T. D.; Schultz, T. W. Structure-Toxicity Analyses of *Tetrahymena pyriformis* Exposed to Pyridines - an Examination into Extension of Surface-Response Domain. *SAR QSAR Environ. Res.* **2001**, *11*, 489−512.
(31) Sinks, G. D.; Schultz, T. W. Correlation of *Tetrahymena* and *Pimephales* Toxicity: Evaluation of 100 Additional Compounds. *Environ. Toxicol. Chem.* **2001**, *20*, 917−921.
(32) Bearden, A. P.; Schultz, T. W. Comparison of *Tetrahymena* and *Pimephales* Toxicity Based on Mechanism of Action. *SAR QSAR Environ. Res.* **1998**, *9*, 127−153.
(33) Cronin, M. T. D.; Aptula, A. O.; Duffy, J. C.; Netzeva, T. I.; Rowe, P. H.; Valkova, I. V.; Schultz T. W. Comparative Assessment of Methods to Develop QSARs for the Prediction of the Toxicity of Phenols to *Tetrahymena pyriformis*. *Chemosphere* **2002**, *49*, 1201−1221.
(34) Aptula, A. O.; Roberts, D. W.; Cronin, M. T. D.; Schultz, T. W. Chemistry-Toxicity Relationships for the Effects of Di- and Trihydroxybenzenes to *Tetrahymena pyriformis*. *Chem. Res. Toxicol.* **2005**, *18*, 844−854.
(35) Schultz, T. W.; Netzeva, T. I.; Roberts, D. W.; Cronin, M. T. D. Structure-Toxicity Relationships for the Effects to *Tetrahymena pyriformis* of Aliphatic, Carbonyl-Containing alpha-, beta-Unsaturated Chemicals. *Chem. Res. Toxicol.* **2005**, *18*, 330−341.
(36) Schultz, T. W.; Cronin, M. T. D. Response-Surface Analyses for Toxicity to *Tetrahymena pyriformis*: Reactive Carbonyl-Containing Aliphatic Chemicals. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 304−309.
(37) Schultz, T. W.; Netzeva, T. I.; Roberts, D. W.; Cronin, M. T. D. Structure-Toxicity Relationships for the Effects to *Tetrahymena pyriformis* of Aliphatic, Carbonyl-Containing alpha-, beta-Unsaturated Chemicals. *Chem. Res. Toxicol.* **2005**, *18*, 330−341.
(38) Schultz, T. W.; Yarbrough, J. W.; Woldemeskel, M. Toxicity to *Tetrahymena* and Abiotic Thiol Reactivity of Aromatic Isothiocyanates. *Cell Biol. Toxicol.* **2005**, *21*, 181−189.
(39) Toropov, A. A.; Schultz, T. W. Prediction of Aquatic Toxicity: Use of Optimization of Correlation Weights of Local Graph Invariants. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 560−567.
(40) Netzeva, T. I.; Schultz, T. W. QSARs for the Aquatic Toxicity of Aromatic Aldehydes from *Tetrahymena* data. *Chemosphere* **2005**, *61*, 1632−1643.
(41) Gagliardi, S. R.; Schultz, T. W. Regression Comparisons of Aquatic Toxicity of Benzene Derivatives: *Tetrahymena pyriformis* and *Rana japonica*. *Bull. Environ. Contam. Toxicol.* **2005**, *74*, 256−262.
(42) Schultz, T. W.; Tucker, V. A. Structure-Toxicity Relationships for the Effects of N- and N,N -Alkyl Thioureas to *Tetrahymena pyriformis*. *Bull. Environ. Contam. Toxicol.* **2003**, *70*, 1251−1258.
(43) Cottrell, M. B.; Schultz T. W. Structure−Toxicity Relationships for Methyl Esters of Cyanoacetic Acids to *Tetrahymena pyriformis*. *Bull. Environ. Contam. Toxicol.* **2003**, *70*, 549−556.
(44) Schultz, T. W.; Burgan, J. T. pH-Stress and Toxicity of Nitrophenols to *Tetrahymena pyriformis*. *Bull. Environ. Contam. Toxicol.* **2003**, *71*, 1069−1076.
(45) Schultz, T. W.; Yarbrough, J. W. Trends in Structure−Toxicity Relationships for Carbonyl-Containing a,b-Unsaturated Compounds. *SAR QSAR Environ. Res.* **2004**, *15*, 139−146.
(46) Schultz, T. W.; Yarbrough, J. W.; Koss, S. K. Identification of Reactive Toxicants: Structure-Activity Relationships for Amides. *Cell Biol. Toxicol.* **2006**, *22*, 339−349.
(47) Halgren, T. A. Merck Molecular Force Field. I.−V. *J. Comput. Chem.* **1996**, *17*, 490−519, 520−552, 553−586, 587−615, 616−641.
(48) Halgren, T. A. MMFF VII. Characterization of MMFF94, MMFF94s, and Other Widely Available Force Fields for Conformational Energies and for Intermolecular-Interaction Energies and Geometries. *J. Comput. Chem.* **1999**, *20*, 730−748.
(49) Chang, G.; Guida, W. C.; Still, W. C. An Internal-Coordinate Monte Carlo Method for Searching Conformational Space. *J. Am. Chem. Soc.* **1989**, *111*, 4379−4386.
(50) Saunders, M.; Houk, K. N.; Wu, Y. D.; Still, W. C.; Lipton, M.; Chang, G.; Guida, W. C. Conformations of Cycloheptadecane. A Comparison of Methods for Conformational Searching. *J. Am. Chem. Soc.* **1990**, *112*, 1419−1427.
(51) *MacroModel, version 8.0*; Schrödinger, Inc.: Portland, OR, 2004.
(52) OpenMolGRID − Open Computing GRID for Molecular Science and Engineering. www.openmolgrid.org (accessed July 29, 2007).

(53) Baker, J. An Algorithm for the Location of Transition States. *J. Comput. Chem.* **1986**, *7*, 385−395.
(54) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.
(55) Stewart, J. J. MOPAC: a Semiempirical Molecular Orbital Program. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 1−45.
(56) Sild, S.; Maran, U.; Lomaka, A.; Karelson, M. Open Computing Grid for Molecular Science and Engineering. *J. Chem. Inf. Model.* **2006**, *46*, 953−959.
(57) Maran, U.; Sild, S.; Kahn, I.; Takkis, K. Mining of the Chemical Information in GRID Environment. *Future Gen. Comput. Syst.* **2007**, *23*, 76−83.
(58) Maran, U.; Sild, S.; Mazzatorta, P.; Casalegno, M.; Benfenati, E.; Romberg, M. Grid computing for the estimation of toxicity: acute toxicity on fathead minnow (*Pimephales promelas*). In *Grid Computing in Computational Biology*; Dubitzky, W., Schuster, A., Sloot, P. M. A., Schhroeder, M., Romberg, M., Eds.; Springer-Verlag: Berlin, Heidelberg, 2007; Vol. 4360, pp 60−74.
(59) Sild, S.; Maran, U.; Romberg, M.; Schuller, B.; Benfenati, E. OpenMolGRID: using automated workflows in GRID computing environment. In *Advances in Grid Computing*; Sloot, P. M. A., Hoekstra, A. G., Priol, T., Reinefeld, A., Bubak, M., Eds.; Springer-Verlag: Berlin, Heidelberg, 2005; Vol. 3470, pp 464−473.
(60) Mazzatorta, P.; Cronin, M. T. D.; Benfenati, E. A QSAR Study of Avian Oral Toxicity Using Support Vector Machines and Genetic Algorithms. *QSAR Comb. Sci.* **2006**, *25*, 616−628.
(61) Katritzky, A. R.; Karelson, M.; Petrukhin, R. Codessa Pro: http://www.codessa-pro.com/ (accessed July 29, 2007)
(62) Meylan, W. M.; Howard, P. H. Atom/Fragment Contribution Method for Estimating Octanol-Water Partition Coefficients. *J. Pharm. Sci.* **1995**, *84*, 83−92.
(63) Draper, N. R.; Smith, H. *Applied Regression Analysis*; Wiley: New York, 1981; p 407.
(64) Zupan, J.; Gasteiger, J. *Neural Networks in Chemistry and Drug Design: An Introduction*, 2nd ed.; Wiley-VCH: Weinheim, 1999; p 380.
(65) Mitchell, T. M. *Machine Learning*; McGraw-Hill: New York, 1997; p 414.
(66) Borggaard, C.; Thodberg, H. H. Optimal Minimal Neural Interpretation of Spectra. *Anal. Chem.* **1992**, *64*, 545−551.
(67) van Wezel, A. P.; Opperhuizen, A. Narcosis due to Environmental Pollutants in Aquatic Organisms: Residue-Based Toxicity, Mechanisms, and Membrane Burdens. *Crit. Rev. Toxicol.* **1995**, *25*, 255−279.
(68) Zefirov, N. S.; Kirpichenok, M. A.; Izmailov, F. F.; Trofimov, M. I. Scheme for the Calculation of the Electronegativities of Atoms in a Molecule in the Framework of Sanderson's Principle. *Dokl. Akad. Nauk* (*Engl. Transl.*) **1987**, *296*, 883−887.
(69) Parr, R. G.; Pearson, R. G. Absolute Hardness: Companion Parameter to Absolute Electronegativity. *J. Am. Chem. Soc.* **1983**, *105*, 7512−7516.
(70) Pearson, R. G. Absolute Electronegativity and Hardness: Applications to Organic Chemistry. *J. Org. Chem.* **1989**, *54*, 1423−1430.
(71) Sinks, G. D.; Carver, T. A.; Schultz, T. W. Structure-Toxicity Relationships for Aminoalkanols: A Comparison with Alkanols and Alkanamines. *SAR QSAR Environ. Res.* **1998**, *9*, 217−228.
(72) Veith, G. D.; Mekenyan, V. G. A QSAR Approach for Estimating the Aquatic Toxicity of Soft Electrophiles. *QSAR* **1993**, *12*, 349−356.
(73) Basak, S. C.; Harriss, D. K.; Magnuson, V. R. Comparative Study of Lipophilicity *versus* Topological Molecular Descriptors in Biological Correlations. *J. Pharm. Sci.* **1984**, *73*, 429−437.
(74) Franke, R. *Theoretical Drug Design Methods*; Elsevier: Amsterdam, 1984; p 412.
(75) Kikuchi, O. Systematic QSAR Procedures with Quantum Chemical Descriptors. *Quant. Struct.-Act. Relat.* **1987**, *6*, 179−184.
(76) DeWeese, A. D.; Schultz, T. W. Structure-Activity Relationships for Aquatic Toxicity to *Tetrahymena*: Halogen-Substituted Aliphatic Esters. *Environ. Toxicol.* **2001**, *16*, 54−60.
(77) Akers, K. S.; Sinks, G. D.; Schultz, T. W. Structure-Toxicity Relationships for Selected Halogenated Aliphatic Chemicals. *Environ. Toxicol. Pharmacol.* **1999**, *7*, 33−39.
(78) Schultz, T. W.; Ralston, K. E.; Roberts, D. W.; Veith, G. D.; Aptula, A. O. Structure-Activity Relationships for Abiotic Thiol Reactivity and Aquatic Toxicity of Halo-Substituted Carbonyl Compounds. *SAR QSAR Environ. Res.* **2007**, *18*, 21−29.
(79) Netzeva, T. I.; Aptula, A. O.; Chaudary, S. H.; Duffy, J. C.; Schultz, T. W.; Schüürmann, G.; Cronin, M. T. D. Structure-Activity Relationships for the Toxicity of Substituted Poly-hydroxylated Benzenes to *Tetrahymena pyriformis*: Influence of Free Radical Formation. *QSAR Comb. Sci.* **2003**, *22*, 575−582.
(80) Schultz, T. W.; Lin, D. T.; Arnold, L. M. QSARs for Monosubstituted Anilines Eliciting the Polar Narcosis Mechanism of Action. *Sci. Total Environ.* **1991**, *109*, 569−580.

CI700231C