Combining Unsupervised and Supervised Artificial Neural Networks to Predict Aquatic Toxicity

Giuseppina Gini,* Marian Viorel Craciun, and Christoph König

DEI, Politecnico di Milano, Piazza Leonardo da Vinci 31, 20131 Milano, Italy

Emilio Benfenati

Istituto di Ricerche Farmacologiche "Mario Negri", Via Eritrea 62, 20157, Milano, Italy

Received January 20, 2004

Most quantitative structure−activity relationship (QSAR) models are linear relationships and significant for only a limited domain of compounds. Here we propose a data-driven approach with a flexible combination of unsupervised and supervised neural networks able to predict the toxicity of a large set of different chemicals while still respecting the QSAR postulates. Since QSAR is applicable only to similar compounds, which have similar biological and physicochemical properties, large numbers of compounds are clustered before building local models, and local models are ensembled to obtain the final result. The approach has been used to develop models to predict the fish toxicity of *Pimephales promelas* and *Tetrahymena pyriformis*, a protozoan.

INTRODUCTION

In our industrialized society, huge and increasing amounts of chemical substances are used and produced every day. This increasing number of chemicals around us raises the problem of characterization, prediction, and evaluation of their consequences to human health and the environment.

Predictive toxicology is a multidisciplinary science that requires close collaboration between toxicologists, chemists, biologists, statisticians, artificial intelligence (AI), and machine learning researchers. Toxicology provides information about mechanisms, rules, and data characterized by activity levels and defines the safety limits of chemical agents. Chemistry provides knowledge about chemical descriptors and physicochemical properties. Biology studies the mechanisms of action of chemicals on animals and other organisms used for tests. Statistics and AI (machine learning) integrates all these items to analyze the existing data and, especially, extract new knowledge from the data, and generate reliable toxicity predictions for chemical compounds. In this way, the main drawbacks in the study of toxicity, such as the high cost of experiments, the long duration of tests, and the use of animals in scientific experiments,[1] can be surmounted.

Quantitative structure−activity relationships (QSARs) can be developed using continuous (quantitative) data, mostly through a regression process but also through classification.[2,3] This area has been developed in the last 40 years to assess the value of drugs and more recently has been proposed as a way of assessing general toxicity and also to obtain new knowledge from data.[3]

The theoretical basis of classical QSAR is usually expressed through the following postulates.[3]

P1, the molecular structure is responsible for all the activities shown.

P2, similar compounds have similar biological and physicochemical properties.

P3, QSAR is applicable only to similar compounds.

In the field of toxicity prediction and QSAR modeling, various AI techniques have been proposed and developed: artificial neural networks (ANNs),[4−7] statistical learning networks[8] expert systems,[9,10] or hybrid approaches such as neuro-fuzzy models.[11] Combinations of the basic techniques, in either a competitive or a cooperative fashion, have been developed and sometimes preferred to single approaches for constructing the solution.[12,13]

The aim of our investigation was to use ANNs following the QSAR postulates to obtain models that predict the toxicity of large groups of different chemicals using information related to chemical structure, biological and physicochemical properties (**P1**). The first step was to group together similar compounds (**P2**) using unsupervised neural networks; then we built local QSAR models (supervised neural networks) for each group of similar chemicals (**P3**). The idea is represented in Figure 1.

Inspired by the biological nervous system, ANNs are composed of simple elements operating in parallel.[14] They can be supervised (e.g., back-propagation networks) or unsupervised (e.g., self-organization networks).

Self-organization within networks is a fascinating topic in the neural network field. Such networks can learn to detect regularities and correlations in their input data and adapt their future response to that input. The neurons of competitive networks learn to recognize groups of similar vectors and separate dissimilar ones (clustering).[15]

A supervised ANN learns from input−output pair examples to build an external relationship between the input and output. For this purpose, input vectors and the corresponding output vectors are presented to train the network until it manages to approximate a mathematical function between them.

The most commonly used ANN is the fully connected forward network with three layers (input, hidden, output).

10.1021/ci0401219 CCC: $27.50
Published on Web 09/02/2004

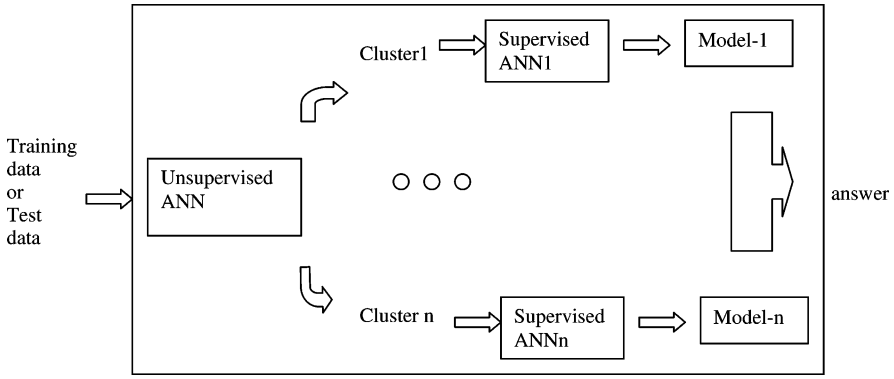

Figure 1. Model building and combination procedure.

The reason for this fact is a theorem[16] which demonstrates that a network of this kind is capable of approximating any continuous function.

ANNs have been used for several years to develop models to predict toxicity within a monolithic approach, i.e., using all molecules in the same system.[6,7] Conversely, in the case of heterogeneous data sets, several studies split the data set according to chemical classes or mode of action (MOA).[13,17,18]

Here we studied a different perspective: to predict toxicity, we used ANN applied to clusters of molecules obtained with self-organized neural networks. Mixtures of experts[19] are a known possible way to improve the prediction capability of a system. Our mixture uses a gating function to combine different individual networks, each built on a subspace of the problem.

MATERIALS AND METHODS

Data Sets. In this paper we used two data sets connected with aquatic toxicity. One is based on the U.S. Environmental Protection Agency (EPA) study referring to acute toxicity in the fathead minnow fish (*Pimephales promelas*).[18] The second set was processed from the TETRATOX database and contains information about the inhibition of growth determined by chemical agents to a protozoan ciliate (*Tetrahymena pyriformis*) [See http://www.vet.uk.edu/TETRATOX/].

The first data set contains 568 different chemicals for which a large number (about 190) of descriptors were calculated using different software systems (Hyperchem 5.0, Hypercube Inc.; CODESSA 2.2.1, SemiChem Inc; Pallas 2.1, ComGenex, Hungary). The descriptors specify each compound in a mathematical way, and according to CODESSA,[20] they can be divided into several groups: quantum chemical, constitutional, topological, geometrical, and physicochemical. The output is the logarithmic value for the lethal concentration for 50% of a population of animals within 96 h: log LC_{50} (96 h) measured in mmol/L.

The second data set has 724 compounds. The output is log IGC_{50} (mmol/L), which means the logarithmic value of the concentration that determined the inhibition of growth for 50% of the ciliates. Descriptors were as described above.

Clustering. Our clustering problem can be formulated in a mathematical way: we want to approximate a function

$$f:\ D \subseteq \mathrm{R}^n \rightarrow \mathrm{R} \tag{1}$$

where $n \in N$ is the number of inputs (descriptors), when we have a set of training pairs

$$(\bar{x}_1, y_1), \ldots, (\bar{x}_m, y_m)$$

where m $\in N$ is the number of training cases, $\bar{x}_i \in D$, for $i = 1, \ldots, m$ are the input vectors and $y_i = f(\bar{x}_i)$ for $i = 1, \ldots, m$ are the corresponding outputs.

The domain of this function can be divided into K subdomains

$$D = D_1 \cup D_2 \cup \ldots \cup D_K \tag{2}$$

where $D_i \subseteq \mathrm{R}^n$ and $D_i \ll D_j = \varnothing$, for $i,j = 1, \ldots, K, I \neq j$ and $f/D_i = f_i$, for $i = 1, \ldots, K$.

According to the divide-and-conquer method, the initial problem was separated into K subproblems. Now we have to approximate K functions with simpler behavior of their domains instead of one function with very complicated behavior.

To solve this problem we performed the following steps.

(1) We built and trained a self-organized neural network, according to the vector quantization algorithm and using Euclidean distance, to split the initial training data into a given number of clusters and obtained training sets for the domains D_i, for $i = 1, \ldots, K$, where K is the number of stated clusters.

(2) For every cluster we built and trained a feedforward neural network to approximate the function $f_i = f/D_i$: $D_i \subseteq \mathrm{R}^n \rightarrow \mathrm{R}$ with the function $F_i = D_i \subseteq \mathrm{R}^n \rightarrow \mathrm{R}$.

(3) We obtained

$$F(x) = \begin{cases} F_1(x),\ x \in D_1 \\ F_2(x),\ x \in D_2 \\ \cdot \\ \cdot\ (\forall) x \in D = D_1 \cup D_2 \cup \ldots \cup D_{\mathrm{K}} \\ \cdot \\ F_{\mathrm{K}}(x),\ x \in D_{\mathrm{K}} \end{cases} \tag{3}$$

where F is the approximation function for f.

Implementation of the Networks. To implement this approach we used inhouse Matlab (Mathworks Inc.) scripts to build, train, and evaluate the neural networks using the data presented and discussed in the second section.

From the initial data sets we excluded any descriptors with missing values and nonvariant descriptors (constant values for all chemicals). After removal, we obtained 156 descriptors.

Every data set was sorted into ascending order according to the output, and every fourth and eighth member from 10 consecutive compounds were extracted to obtain the test set in order to widely cover the whole space. Thus, we had 80% of the chemicals in the training set and 20% in the external test set. Then 10% from the training set were held back to build the validation set which was used to improve the generalization capability of the networks in the early stopping approach.[21]

The values were scaled to the [−1,1] interval to consider all descriptors on the same basis.

Afterward, we built and trained a competitive network to group the training sets into different clusters only for the input space.

Finally, for every cluster, the training data was used to build and train the supervised neural networks with one hidden layer and one output. The transfer function for the hidden layer was sigmoidal ('tansig')[21] and for the output layer a linear function ('purelin'). The training function is 'traingdx', which updates the weights and biases according to the gradient descend momentum and adaptive learning rate.

RESULTS AND DISCUSSION

The considered data sets contain a diversity of compounds with a diversity of structures. Because of the lack of homogeneity, it is hard for a single technique to model these data and obtain good results. In previous research[13] we built a single model with linear regression on the whole data set and obtained a determination coefficient R^2 in 10-fold cross-validation of only 0.55. After reducing the number of descriptors with a wrapper approach, we reached 0.7. Results improved after mixing different experts for different chemical classes.

In this paper, instead of some official, predefined classification, we grouped the chemicals inside the input space (the descriptor space) via a clustering technique using the Euclidian distance function. For clustering we used unsupervised neural networks. Subsequently, we constructed toxicity models for the obtained groups of chemicals. For the toxicity models we used supervised neural network.

In our clustering studies output values were not used. The number of clusters ranged between 2 and 15. The test data was split in the appropriate clusters.

In making the experiments with different numbers of clusters it became clear that this number strongly influences the result. Generally speaking, we expected that more clusters would produce a more detailed and better final model and improve the performance. However, we observed that this improvement was stopped or lowered having reached a particular number of clusters. Furthermore, after a certain point some clusters remain empty (without data), which indicates that the final number of clusters had been reached. Table 1 shows for each data set the distribution of the data for the number of clusters used to construct the final model.

After building all clusters the dimension of the training data was reduced, eliminating the descriptors correlated with R greater than 0.9. The use of a relatively small number of input descriptors can reduce the risk of over-fitting the networks.

Table 1. Distribution of the Training and Test Data in Clusters of the Two Data Sets (number of descriptors in parentheses)

	data set 1 (156) 568 items		data set 2 (264) 724 items	
clusters	train	test	train	test
1	43	12	50	9
2	53	13	119	23
3	46	8	120	38
4	50	10	65	32
5	58	15	109	15
6	35	15	116	28
7	36	10		
8	52	16		
9	81	15		
total	454	114	579	145

Table 2. Performance on Test Data of the Different Models[a]

	data set 1		data set 2	
model	MSE	MAE	MSE	MAE
1	0.45	0.48	0.09	0.23
2	0.38	0.51	0.14	0.27
3	0.05	0.21	0.16	0.30
4	0.13	0.31	0.35	0.47
5	0.15	0.35	0.27	0.42
6	0.16	0.35	0.43	0.49
7	0.46	0.58		
8	0.32	0.41		
9	0.14	0.31		

[a] MSE = mean-squared error. MAE = mean absolute error.

Over-fitting is the main risk, in general, in QSAR, and in particular, in the case of neural networks, even more risk if a high number of inputs (chemical descriptors) is used associated with a low number of molecules. Indeed, "as long as the net structure has enough complexity, a neural net can be trained to produce any desirable error level on the training set. In order to determine a net's ability to generalize, it must be evaluated on a test data set which was not used during the training."[22] For this reason, we tested all models using the external test sets and reduced the number of descriptors.

The cluster of chemicals as we did is homogeneous on the basis of the chemical descriptors. The cluster mimics the classes, which have been used in other modeling studies: classes determined on the basis of the chemical classification or toxic mode of action. However, here we used an automatic way to split the domain of the chemical compounds studied.

Then, for each cluster we modeled toxicity, using supervised neural networks, using as inputs the selected chemical descriptors. The performance of the local models, one for each cluster, can be seen in Table 2. It is clear that not every cluster could be modeled with the same high level of accuracy; an example is cluster 6 of data set 2. However, most of the local models are highly predictive. This positively affects the performance of the final combined model, which assigns the model from the cluster the new chemical belongs to according to the paradigm of competitive strategy in "mixture of experts".[19]

Figures 2 and 3 show for data sets 1 and 2, respectively, the performance of the mixture model on the test set,

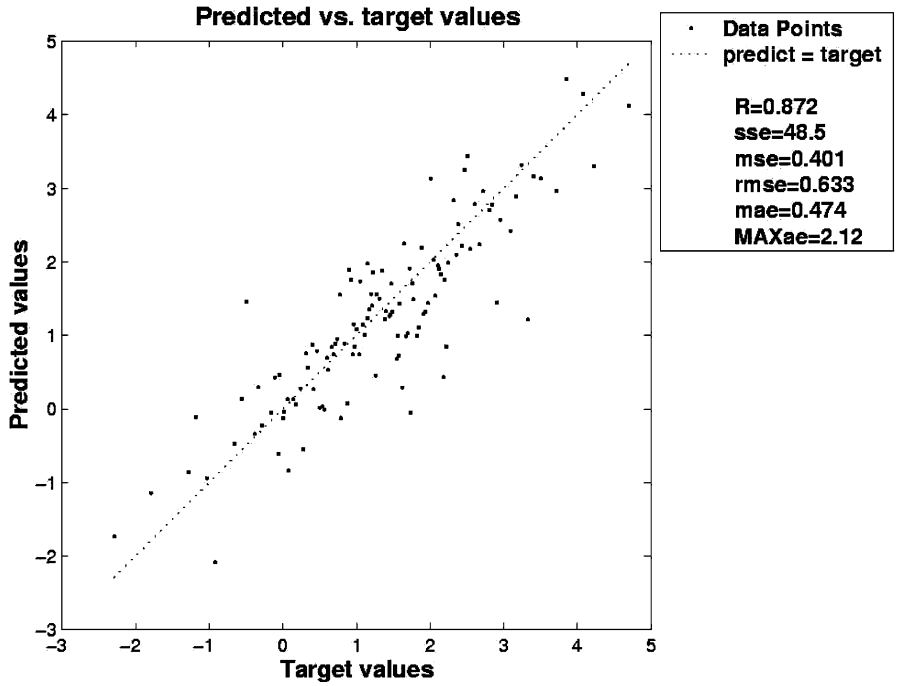

Figure 2. Performance on the external test set for data set 1.

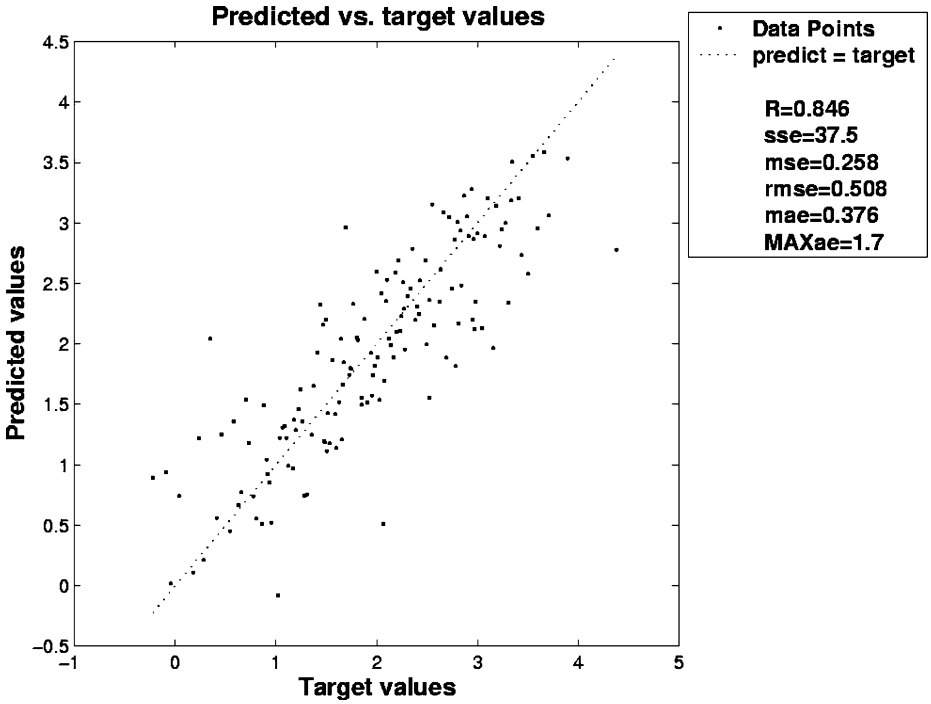

Figure 3. Performance on the external test set for data set 2.

summarizing the data obtained for the data sets. It is worth mentioning that the results on the training data were slightly better than the results on the test sets, which indicate that models were able to generalize their performance. The models achieved low prediction errors and a high correlation between the values predicted by the models and the target values.

The predictive models for toxicity of heterogeneous data sets can be monolithic or split the heterogeneous data sets into subsets to simplify the modeling task, thereby reducing the heterogeneity. Some monolithic models obtained good results on the same or bigger data sets related to fathead minnow.[6,23] On a smaller data set (130 compounds) other studies have been conducted, achieving a value for the determination coefficient R^2 greater than 0.9, using ANNs, but with a more limited test set (10 compounds).[24]

For *Tetrahymena pyriformis*, studies have been performed on a smaller number of compounds and using data sets more homogeneous than ours; for instance, 476 aliphatic compounds[25] and about 200 phenols were used.[17,26]

There have been several studies using neural networks for toxicity prediction and QSAR. Unsupervised systems, such as self-organizing maps (SOM), have been used in QSAR to reduce the number of chemical descriptors or split the chemicals in QSAR sets. For instance, Arciniegas et al. used SOM to reduce the number of chemical descriptors for the

design of novel pharmaceuticals.[22] Similarly, SOM have been employed to obtain the subset of descriptors for the evaluation of toxicity.[27] SOM have been used to select the training and validation sets, maximizing the molecular diversity,[23] in the case of toxicity prediction or in a study on dihydrofolate reductase inhibitors, generating training, cross-validation, and prediction sets.[28]

Instead, in our study unsupervised neural networks have been used to identify homogeneous sets of compounds to be used for successive modeling. Monolithic systems are less flexible and require a complete rebuilding in order to be improved, for instance, to include new chemicals. Conversely, systems that use different submodels are more easily modifiable and flexible and, in principle, can model more accurately the different situations present in heterogeneous sets of molecules.

To overcome the diversity of the data, attempts of subdividing the set using the official chemical classification from EPA were presented; in one case chemical classes were defined by chemists, while in another one chemical classes were assigned by an algorithm.[29,13] After "manual" selection of the chemical class and using multivariate analysis for predictive models, R^2 was about 0.75 with the leave-one-out procedure but for alcohols and ketones/aldehydes.[29] Results for the predictive models obtained after automatic classification of chemical classes gave a value for R^2 of 0.80 on the combined model with a 10-fold cross validation.[13] Both approaches improved the basic R^2 of 0.55 of the simple monolithic model.

The splitting in chemical classes is an apparently simple approach; however, the splitting is performed according to a taxonomy, which has been developed by chemists on the basis of knowledge which is external to toxicology. Hence, it could happen that different chemical classes can act according to the same toxic mechanism and vice versa; some apparently minor chemical differences can produce high differences in toxic activities.[12,18] Furthermore, classification into chemical classes is sometimes ambiguous because more than one functional group could be present in the same molecule; thus, it is possible to classify a compound into many chemical classes.

Another possibility, which has been evaluated, is to split chemicals according to their MOA.[18] The general problem is that the MOA is not known a priori, and thus studies have been performed to predict MOA.[17,18]

The procedure we used here is different. The classification we adopt is not based on chemical classes or a MOA. Actually the nature of the knowledge which is used for the classification is not physical. This knowledge does not rely on laboratory experiments, such as those used to obtain the MOA, and is not related to chemist's knowledge of functional groups. Conversely, the knowledge we use is obtained by an automatic process (the self-organizing networks) that produces virtual knowledge (the clusters), which is used to build up prediction models. For this reason, it is more appropriate to speak about clusters than classes.

CONCLUSIONS

Our results on toxicity prediction are promising and suitable for future research. The predictive capability is comparable to that obtained with the best models published so far on the same data set, and the approach is original. The approach we used was to obtain clusters of chemicals and then produce predictive models. Chemical descriptors were used to afford chemical information, while parallel computing was used for both clustering and predictive models.

Our approach is far from complete, but it has the advantage of being modular, which makes it more flexible than holistic approaches. Monolithic approaches have the advantage of simplicity, but in order to make progress, they require a complete rebuilding of the model. The combination of local experts presented here has some advantages:

- the reliability of the models for different clusters is more clearly recognized and defined;
- it offers a simple opportunity to study the weaker submodels; using a flexible architecture, submodels can be easily modified, introducing better models, or even integrated with new, independent submodels.

ACKNOWLEDGMENT

This work was partially funded by EU contract QLK5-CT-2002-00691, Demetra; Christoph König and Marian Craciun were supported by EU contract HPRN-CT-1999-00015, Imagetox.

REFERENCES AND NOTES

(1) Omenn, G. S. Assessing the risk assessment paradigm. *Toxicology* **1995**, *102*, 23−28.
(2) *Quantitative Structure−Activity Relationships for Pollution Prevention, Toxicity Screening, Risk Assessment, and Web Applications*; Walker, J. D., Ed.; SETAC Press: Pensacola, FL, 2003.
(3) Katritzky, A R.; Petrukhin, R.; Tatham, D.; Basak, S.; Benfenati, E.; Karelson, M.; Maran, U. Interpretation of quantitative structure−property and −activity relationships. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 679−685.
(4) Eldred, D. V.; Weikel, C. L.; Jurs, P. C.; Kaiser, K. L. E. Prediction of Fathead Minnow Acute Toxicity of Organic Compounds from Molecular Structure. *Chem. Res. Toxicol.* **1999**, *12*, 670−678.
(5) Eldred, D. V.; Jurs, P. C. Prediction of Acute Mammalian Toxicity of Organophosphorus Pesticide Compounds from Molecular Structure. *SAR QSAR Environ. Res.* **1999,** *10*, 75−99.
(6) Kaiser, K. L. E.; Niculescu, S. P. Using probabilistic neural networks to model the toxicity of chemicals to the fathead minnow (*Pimephales promelas*): A study based on 865 compounds. *Chemosphere* **1999**, *38*, 3237−3245.
(7) Gini, G.; Lorenzini, M.; Benfenati, E.; Grasso, P.; Bruschi, M. Predictive Carcinogenicity: a Model for Aromatic Compounds, with Nitrogen-containing Substituents, Based on Molecular Descriptors Using an Artificial Neural Network. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1076−1080.
(8) Benfenati, E.; Lemke, F.; Spreafico, M.; Griffiths, E.; Gini, G. Prediction of pesticide toxicity to trout using knowledge extraction technologies. Presented at QSAR2004, Liverpool, U.K., May 9−13, 2004.
(9) Benfenati, E.; Gini, G. Computational predictive programs (expert systems) in toxicology. *Toxicology* **1997**, *119*, 213−225.
(10) Gini, G. Predictive Toxicology of Chemicals: Experience and Impact of AI tools. *AI Mag.* **2000**, *21*, 81−84.
(11) Benfenati, E.; Mazzatorta, P.; Neagu, D.; Gini, G. Combining Classifiers of Pesticides Toxicity through a Neuro-fuzzy Approach. In *MCS2002, Multiple Classifier Systems 2002, Lecture notes in Computer science*; Roli, F., Kittler, J., Eds.; Springler: Berlin, 2002; Vol. 2364, pp 293−303.
(12) Gini, G.; Lorenzini, M.; Benfenati, E.; Brambilla, R.; Malvé, L. Mixing a Symbolic and a Subsymbolic Expert to Improve Carcinogenicity Prediction of Aromatic Compounds. In *Multiple Classifier Systems*; Kittler, J., Roli, F., Eds.; Springler-Verlag: Berlin, 2001; pp 126−135.
(13) König, C.; Gini, G.; Benfenati, E.; Craciun, M. Multi-class classifier from a combination of local experts: toward distributed computation for real-problem classifiers. *Int. J. Pattern Recognit. Artificial Intell.*, in press.

(14) Rumelhart, D. E.; McClelland, J. L. *Parallel Distributed Processing Explanations in the Microstructure of Cognition;* MIT Press: Cambridge, MA, 1986.
(15) Kohonen, T. *Self-Organization and Associative Memory*, 2nd ed.; Springer-Verlag: Berlin, 1997.
(16) Funahashi, K. On the Approximate Realization of Continuous Mappings by Neural Networks. *Neural Networks* **1989**, *2*, 183−192.
(17) Aptula, A. O.; Netzeva, T. V.; Valkova, I. V.; Cronin, M. T. D.; Schultz, T. W.; Kühne, R.; Schüürmann, G. Multivariate Discrimination between Modes of Toxic Action of Phenols. *Quant. Struct.-Act. Relat.* **2002**, *21*, 12−18.
(18) Russom, C. L.; Bradbury, S. P.; Hammermeister, D. E.; Drummond, S. J. Predicting modes of toxic action from chemical structure: acute toxicity in the fathead minnow (*Pimephales promelas*). *Environ. Toxicol. Chem.* **1997**, *16*, 948−967.
(19) Sharkey, A. J. C. *Combining Artificial Neural Nets−Ensembles and modular multi-net systems*; Springer: London, 1999.
(20) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. CODESSA Comprehensive Descriptors for structural and Statistical Analysis, Reference manual; SemiChem: Gainesville, FL, 1994.
(21) http://www.mathworks.com/access/helpdesk/help/toolbox/nnet/nnet.shtml.
(22) Arciniegas, F.; Bennett, K.; Breneman, C.; Embrechts, M. Molecular database mining using self-organizing maps for the design of novel pharmaceuticals. In *Intelligent Engineering Systems through Artificial Neural Networks: Smart Engineering System Design*; Dagli, C. H., Ed.; ASME Press: 2000; Vol. 10, pp 477−482.
(23) Pintore, M.; Piclin, N.; Benfenati, E.; Gini G.; Chrétien, J. R. Predicting toxicity against the fathead minnow by Adaptive Fuzzy Partition. *QSAR Comb. Sci.* **2003**, *22*, 210−219.
(24) Huuskonen, J. QSAR modeling with the electrotopological state indices: predicting the toxicity of organic chemicals. *Chemosphere* **2003**, *50*, 949−953.
(25) Netzeva, T. I.; Schultz, T. W.; Aptula, A. O.; Cronin, M. T. D. Partial least squares modelling of the acute toxicity of aliphatic compounds to *Tetrahymena pyriformis*. *SAR QSAR Environ. Res.* **2003**, *14*, 265−283.
(26) Cronin, M. T.; 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* **2003**, *49*, 1201−1221.
(27) Espinosa, G.; Arenas, A.; Giralt, F. An Integrated SOM-Fuzzy ARTMAP Neural System for the Evaluation of Toxicity. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 343−359.
(28) Guha, R.; Serra, J. R.; Jurs, P. C. Generation of QSAR sets with a self-organizing map. *J. Mol. Graphics Model.*, in press.
(29) Roncaglioni, A.; Colombo, A.; Maran, U.; Karelson, M.; Benfenati, E. Prediction of Acute Aquatic Toxicity to Fish Comparing Different QSAR Approaches. Presented at the 41st Congress of the European Societies of Toxicology, Eurotox 2003, Florence, Italy, Sept 28−Oct 1, 2003; Abstract in *Toxicol. Lett.* **2003**, *144*, s52.

CI0401219