COMPARISON OF NEURO/FUZZY METHODS FOR TOXICITY EVALUATION

A. Grauel, L.A. Ludwig, I. Renners
Universität-GH Paderborn, Abteilung Soest
Fachbereich 16, Fachgebiet Mathematik
Steingraben 21, D-59494 Soest, Germany
Phone: +49-2921-378162, Fax: +49-2921-378180
E-mail: grauel@ibm5.uni-paderborn.de

ABSTRACT: In this contribution we present investigations about the use of neural and neuro-fuzzy methods for toxicity evaluation, in particular for the prediction of aquatic toxicity of pesticides. Therefore different molecular descriptors are computed and the correlation behavior of the different descriptors in the descriptor space is studied. In a first step 164 pesticides are considered and 175 descriptors are taken into account; moreover, subclasses of the whole set of data are investigated. From these data sets results using multi-layer perceptrons and B-spline networks are compared.

KEYWORDS: predictive toxicology, quantitative structure-activity relationship, artificial neural networks, evolutionary computation, neuro-fuzzy systems, B-spline networks

INTRODUCTION

It is of special interest for environmental and health problems to predict the toxicology of chemicals. The accuracy of a prediction is a very sensitive point for critical examinations of toxic activities. Therefore, new thinking in the direction of computational intelligence (CI) is needed for the prediction of toxicity and ecotoxicity. We need a rigorous insight into the analysis mechanisms for a constructive search to establish and perform mathematical operations.

CI can be seen as useful frame besides artificial intelligence (AI) for the formalization of human expert knowledge in a fuzzy logic approach, generalization principles in artificial neural networks (ANNs) and biological optimization principles in evolutionary computation, i.e. genetic algorithms (GAs) and evolution strategies (ESs). In this context hybrid approaches which combine some advantages of the different methods of CI can be considered as a solution-platform alternative to classical mathematical methods.

Recent investigations support the general assumption that macroscopic properties like toxicity and ecotoxicity strongly depend on microscopic features and the structure of the molecule. This assumption enables us to set up quantitative structure-activity relationship (QSAR), quantitative structure-property relationship (QSPR) and quantitative structure-retention relationship (QSRR), which are the basis for the prediction of toxicity from chemical structures of molecules. The further assumption is that these microscopic features and the structures of molecules can be identified and characterized by certain molecular descriptors.

The general objective is to set up a functional dependency of the toxicity to a certain aspect on the selected molecular descriptors. However, it will not be possible to write down this functional dependency in an analytical form. The data in the database considered yield many points in the descriptor space, which can be used to extract unknown functional properties, relations or rules etc.

First of all, fundamental correlation investigations and a principal component analysis as input space transformation to obtain a reduced feature input space are performed (Grauel, Ludwig, Renners, Berk, 1999). The second section deals with a method to find generalized dependencies of the molecular descriptors with genetically optimized multi-layer perceptrons (MLPs). The following section deals with a new type of networks, namely B-spline networks (BSNs). A BSN can be interpreted as Takagi-Sugeno fuzzy controller with remarkable properties. At last, a comparison of the two approaches shall be given.

DATA STATISTICS

The basis for our investigations is a set of 164 pesticides from 7 different chemical classes with data on acute toxicity for rainbow trout and daphnia magna (Benfenati, 1998). The concentrations for this aquatic toxicity – taken from the Pesticides Manual – are given in 2 representations, LC_{50} and $-\log_{10}(LC_{50}/(\text{mmol/l}))$, the latter being the unit usually employed in literature. LC_{50}, the lethal concentration 50 %, is the concentration of the chemical in water at which 50 % of the laboratory animals die after a certain period of time. The correlation of the logarithmic concentration values on rainbow trout and daphnia is $|r| = 0.74$.
175 molecular descriptors such as constitutional and topological descriptors, electrostatic and quantum-chemical descriptors and others, which are partly continuous, partly discrete values, were calculated for each of these 164 pesticides. Since 18 of these descriptors show missing values – mainly those reflecting maximal and minimal partial charges –, these are omitted in the first step of investigation and shall only serve for secondary classification purposes. The correlation of each of the remaining 157 descriptors and the logarithmic concentration is $|r| < 0.595$ for trout and $|r| < 0.538$ for daphnia. However, there are also many totally uncorrelated descriptors with $|r| < 10^{-3}$. A two-sided *t*-test on the zero-hypothesis that the descriptors are uncorrelated to the logarithmic concentration values yields 61 significantly correlated features for trout and 53 for daphnia on the 99 % level. Molecular descriptors are usually classified as substantial, important, likely or specific regarding their correlation to the activity or property to be modeled – Table 1 gives the descriptor classification for the data set chosen.

	correlation	*trout*	*daphnia*
substantial descriptors	$\|r\| \geq 0.99$	0	0
important descriptors	$0.99 > \|r\| \geq 0.80$	0	0
likely descriptors	$0.80 > \|r\| \geq 0.50$	9	5
specific descriptors	$0.50 > \|r\|$	148	152

Table 1: Classification of the molecular descriptors regarding their correlation to the desired output.

In a next step we calculated Spearman's rank correlation, which is based on any monotonous interrelation of the variables, instead of Pearson's correlation, which assumes a linear dependency of the variables. The correlation coefficients are in the same range with $|r| < 0.587$ for trout and $|r| < 0.562$ for daphnia. The *t*-test yields 57 significantly correlated descriptors for trout and 54 for daphnia on the 99 % level.
Since any machine learning system will not succeed in modeling the dependency of the toxicity on all 157 descriptors with only 164 learning examples, we decided to investigate data compression using principal components analysis (PCA). Performing the PCA on the 61 significantly correlated descriptors (Pearson) for trout yields 9 principal components (PCs) with an eigenvalue bigger than 1, which account for more than 91 % of the total variance. For the 53 significantly correlated descriptors for daphnia the PCA yields 7 PCs with an eigenvalue bigger than 1, which account for more than 90 % of the total variance. Performing the PCA on the 57 significant descriptors with respect to Spearman's rank correlation for trout yields 8 PCs with an eigenvalue bigger than 1, which account for more than 91 % of the total variance. For the 54 significantly correlated descriptors for daphnia the PCA yields 7 PCs with an eigenvalue bigger than 1, which account for more than 89 % of the total variance.

OPTIMIZED MULTILAYER PERCEPTRONS

For modeling the dependency of the toxicity on the significant descriptors a feasible approach is the use of feed-forward ANNs, the input space of which is optimized by GAs. First investigations showed that the results depend very much on the random choice of training and test data subsets. A possible explanation may be that the 175 descriptors or the concentration values are not satisfactory or the 164 pesticides are not a representative data set. This assumption is supported by comparison of the 1983 and 1997 Pesticides Manual with the HS database showing that the LC_{50} taken from different databases were identical for some values and very different for others (Benfenati, 1998). Therefore, we decided to use leave-one-out crossvalidation for the following investigations, i.e. each network was trained 164 times with 163 different input values and 1 single output value. This ensures the maximal possible statistical security by testing every output independently from all others. The networks were trained 100000 steps (and more but without any effect) with a learning rate of 0.01, a momentum

term of 0.1 and a uniform random weight initialization in [-1,1]. The input data was standardized to $m = 0, s = 1$ by $\tilde{x} = \left(x - \bar{x}\right)/s$, the output variable -$\log_{10}$($LC_{50}$/(mmol/l)) was linearly transformed to [-0.8, 0.8].

Genetic algorithms are powerful tools for optimization tasks, especially if non-differentiable fitness functions are assumed. They are well-suited for ANN optimization, which can hardly done by other methods since the underlying functions are non-differentiable and even non-continuous. Different optimization goals can be achieved depending on the employed fitness function. On the one hand, the network has to be big enough to model the complex dependency of the output from the inputs. On the other hand, the network has to be as small as possible to show the required generalization abilities. In order to find the best input variables for a small network, we implemented a GA to select a subset of descriptors. Using only the 2 descriptors log D pH 7.4 and LUMO, the MLP reached a correlation of $|r| = 0.642$ on the test set with leave-one-out crossvalidation (Table 2). This network has an astonishingly small structure with 2 linear input nodes, no hidden nodes and 1 tanh output node; results with any number of hidden nodes were worse.

correlation	*input 1*	*input 2*
$\|r\| = 0.642$	log D pH 7.4	LUMO
$\|r\| = 0.634$	log D pH 7.4	relative number of S atoms
$\|r\| = 0.611$	log D pH 7.4	HA dependent HDCA-2/SQRT (TMSA)
$\|r\| = 0.610$	log D pH 7.4	Kier & Hall index (order 2)
$\|r\| = 0.596$	log D pH 7.4	moment of inertia C

Table 2: Top 5 results using MLPs for the total set of molecules with 2 inputs.

With 4 descriptors log D pH 9, moment of inertia A, molecular surface area and RPCG relative positive charge (QMPOS/QTPLUS) the correlation could be slightly improved up to $|r| = 0.654$ (Table 3). The best network has 4 linear input nodes, 2 tanh hidden nodes and 1 tanh output node.

correlation	*input 1*	*input 2*	*input 3*	*input 4*
$\|r\| = 0.654$	log D pH 9	moment of inertia A	molecular surface area	RPCG relative positive charge (QMPOS/QTPLUS)
$\|r\| = 0.652$	log D pH 7.4	HA dependent HDCA-2/TMSA	Kier shape index (order 1)	RPCG relative positive charge (QMPOS/QTPLUS)
$\|r\| = 0.649$	log D pH 9	moment of inertia A	average information content (order 2)	RPCG relative positive charge (QMPOS/QTPLUS)
$\|r\| = 0.648$	log D pH 9	Kier & Hall index (order 2)	RPCG relative positive charge (QMPOS/QTPLUS)	WNSA-1 weighted PNSA (PNSA1*TMSA/1000)
$\|r\| = 0.647$	log D pH 9	HA dependent HDCA-1/TMSA	TMSA total molecular surface area	RPCG relative positive charge (QMPOS/QTPLUS)

Table 3: Top 5 results using MLPs for the total set of molecules with 4 inputs.

MODELING OF SUBCLASSES AND PREDICTIVE TOXICOLOGY

As the data are very inhomogeneous, we decided to investigate more homogeneous chemical subclasses from the total set of molecules. At first, we selected 27 organophosphorus molecules, which is the subclass with the most elements. Small networks using only two input descriptors yielded varying results depending on the inputs selected by the GA. The smallest mean squared crossvalidated test error (MSE) was 0.071 – on the interval of [-0.8, 0.8] – but the correlation was only $|r| = 0.273$, which is very poor (Table 4). The biggest correlation was $|r| = 0.834$ with an MSE of 0.082 (Table 5).

MSE	*correlation*	*input 1*	*input 2*
0.071	$\|r\| = 0.273$	molecular weight	Kier & Hall index (order 3)
0.075	$\|r\| = 0.666$	HA dependent HDCA-1/TMSA	FPSA-3 fractional PPSA (PPSA-3/TMSA)
0.078	$\|r\| = 0.634$	FPSA-3 fractional PPSA (PPSA-3/TMSA)	HA dependent HDCA-2/SQRT(TMSA)
0.079	$\|r\| = 0.358$	molecular weight	randic index (order 2)
0.080	$\|r\| = 0.222$	molecular weight	structural information content (order 1)

Table 4: Top 5 results using MLPs for the organophosphorus class with 2 inputs (sorted by MSE).

correlation	*MSE*	*input 1*	*input 2*
$\|r\| = 0.834$	0.082	HA dependent HDSA-1	log D pH 5
$\|r\| = 0.824$	0.126	number of H atoms	randic index (order 3)
$\|r\| = 0.806$	0.136	DPSA-2 difference in CPSA (PPSA2-PNSA2)	randic index (order 3)
$\|r\| = 0.803$	1.105	randic index (order 3)	molecular volume
$\|r\| = 0.797$	0.099	log D pH 5	minimal partial charge (Q min)

Table 5: Top 5 results using MLPs for the organophosphorus class with 2 inputs (sorted by correlation).

So we realized the fact that good, i.e. high, correlations need not correspond to good errors, i.e. small MSE, and vice versa. Furthermore the reproducibility of the results when training repetitiously the network with the identical data is not very good. A larger MLP with 4 linear input nodes, 2 tanh hidden nodes and 1 tanh output node reached a minimal crossvalidated test MSE of 0.073 with a correlation of $|r| = 0.738$ (Figure 1). This network uses XY shadow, ionization potential, complementary information content (order 0) and number of P atoms as input descriptors.

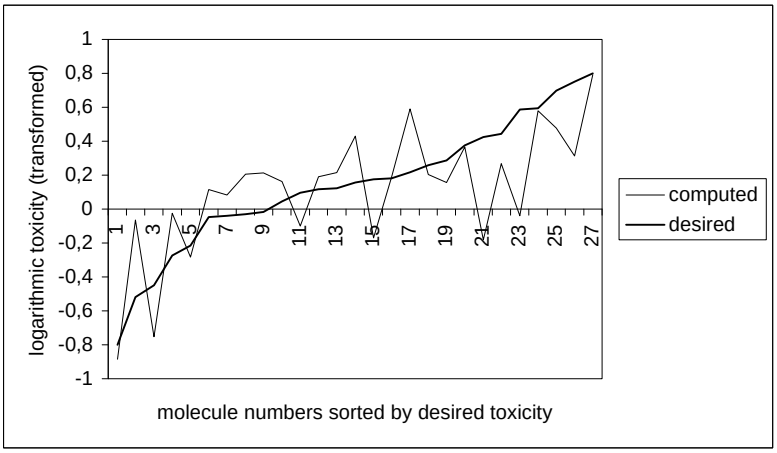

Figure 1: Computed output vs. desired output using an MLP for the organophosphorus class with 2 inputs

The second subclass arises from a comparison of the concentration values in the Pesticide Manual and the HS database. The intersection set of the molecules selected from the Pesticide Manual and the HSDB contains 39 molecules. From these molecules some show very big differences in LC_{50}, others have only small differences and others are absolutely identical. Assuming the values are identical due to the fact that HSDB and Pesticide Manual refer to the same reference of measurement in those cases, these values need not be reliable and accurate. Hence we selected 20 molecules that have small but non-zero differences between HSDB and Pesticide Manual. On these data, a small network with 2 input nodes reached a minimal crossvalidated test MSE of 0.058 with a correlation of $|r| = 0.773$ (Figure 2) using gravitation index (all bonds) and WNSA-2 weighted PNSA (PNSA2*TMSA/1000). The biggest correlation found by the GA was $|r| = 0.940$, however with an MSE of 0.217.

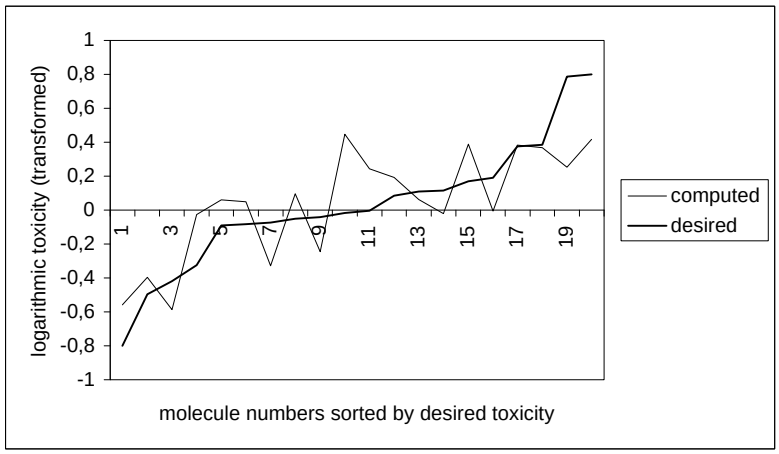

Figure 2: Computed output vs. desired output using an MLP for 20 selected molecules with 2 inputs

B-SPLINE NETWORKS

BSNs (Brown and Harris 1994; Zhang and Knoll 1996) make use of piecewise polynomials as membership functions. Each interval of interest is divided into a number of subintervals, whereby each subinterval is delimited by breakpoints that determine the appearance and position of each B-spline. The activated basis function can be of symmetrical or asymmetrical order. Asymmetrical basis functions in the input space can be seen as clustering method for structuring the descriptor space.
By using BSNs the approximation ability is only limited by the number of knots distributed on each input interval. Furthermore a BSN can be interpreted as a Takagi-Sugeno fuzzy controller, which is useful for the task of rule extraction. Taking into account that most observed real-world data are disturbed to a certain degree, too complex structures of the BSN result in a small training error but possibly in a high test error and thus in a poor correlation. A preferably small model must therefore be found to overcome overfitting to the training data. We carried out a complete permutation search to find the best descriptors for a given small model of 3 uniformly distributed linguistic terms of order 3 on each input variable. Using leave-one-out crossvalidation we achieved a test correlation of $|r| = 0.65$ as best result by finding log D pH 9 and number of P atoms as most promising input descriptors according to $-\log_{10}(LC_{50}/(\text{mmol/l}))$ as desired output (Table 6 and Figure 3).

MSE	*correlation*	*input 1*	*input 2*
0.0661	$\|r\| = 0.6496$	log D pH 9	number of P atoms
0.0685	$\|r\| = 0.6312$	log D pH 7.4	Kier & Hall index (order 3)
0.0692	$\|r\| = 0.6278$	log D pH 9	number of S atoms
0.0698	$\|r\| = 0.6232$	log D pH 7.4	number of S atoms
0.0718	$\|r\| = 0.6091$	log D pH 9	relative number of S atoms

Table 6: Top 5 results using BSNs for the total set of molecules with 2 inputs (sorted by MSE).

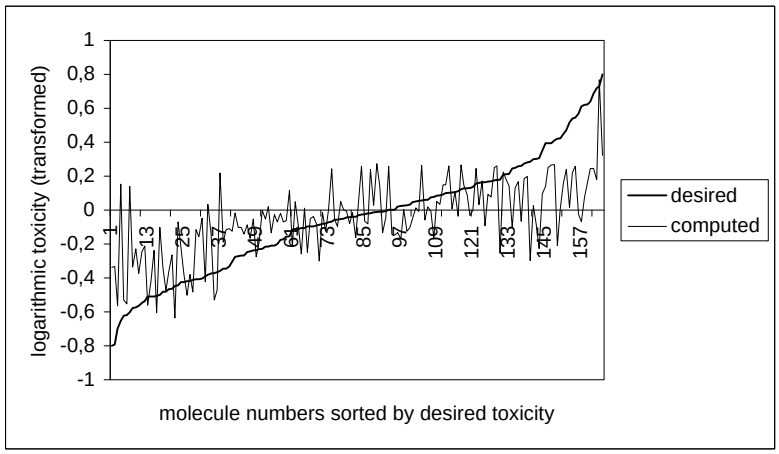

Figure 3: Computed output vs. desired output using a BSN for the total set of molecules with 2 inputs.

For the subclass of the organophosphorus molecules we used again a BSN with two inputs and three uniformly distributed linguistic terms of order 3 on each input interval and computed the test error for all possible input combinations using leave-one-out crossvalidation. The best found input descriptors were log D pH 5 in combination with number of P atoms, which lead to a correlation of $|r| = 0.803$ (Table 7 and Figure 4).

MSE	*correlation*	*input 1*	*input 2*
0.0519	$\|r\| = 0.8026$	log D pH 5	number of P atoms
0.0537	$\|r\| = 0.7930$	log D pH 7.4	number of P atoms
0.0537	$\|r\| = 0.7921$	YZ shadow	number of P atoms
0.0553	$\|r\| = 0.7864$	gravitation index (all pairs)	number of P atoms
0.0560	$\|r\| = 0.7822$	randic index (order 1)	number of P atoms

Table 7: Top 5 results using BSNs for the organophosphorus class with 2 inputs (sorted by MSE).

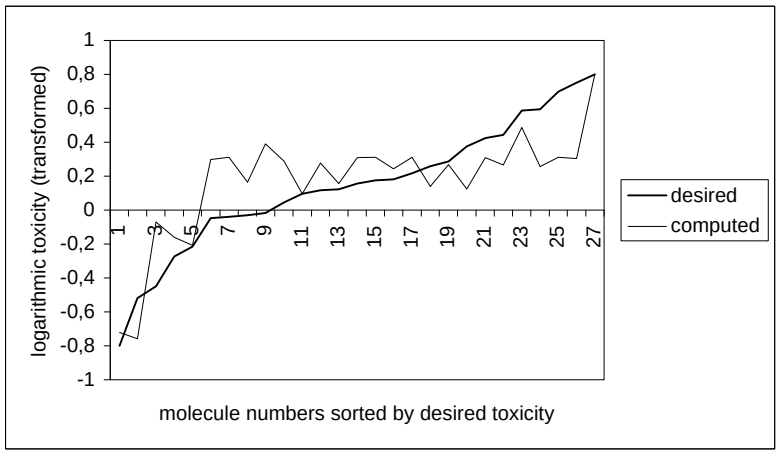

Figure 4: Computed output vs. desired output using a BSN for the organophosphorus class with 2 inputs.

For the 20 selected molecules we achieved a best correlation of $|r| = 0.91$ (Table 8 and Figure 5) using the same BSN approach as described above. The best found input descriptor combination was gravitation index (all pairs) with WPSA-1 weighted PPSA.

MSE	*correlation*	*input 1*	*input 2*
0.0242	$\|r\| = 0.9109$	gravitation index (all pairs)	WPSA-1 weighted PPSA
0.0324	$\|r\| = 0.8779$	molecular volume	relative number of double bonds
0.0352	$\|r\| = 0.8803$	average complementary information content (order 2)	relative number of double bonds
0.0354	$\|r\| = 0.8677$	gravitation index (all bonds)	RPCS relative positive charged SA
0.0371	$\|r\| = 0.8583$	FPSA-3 fractional PPSA	number of Br atoms

Table 8: Top 5 results using BSNs for 20 selected molecules with 2 inputs (sorted by MSE).

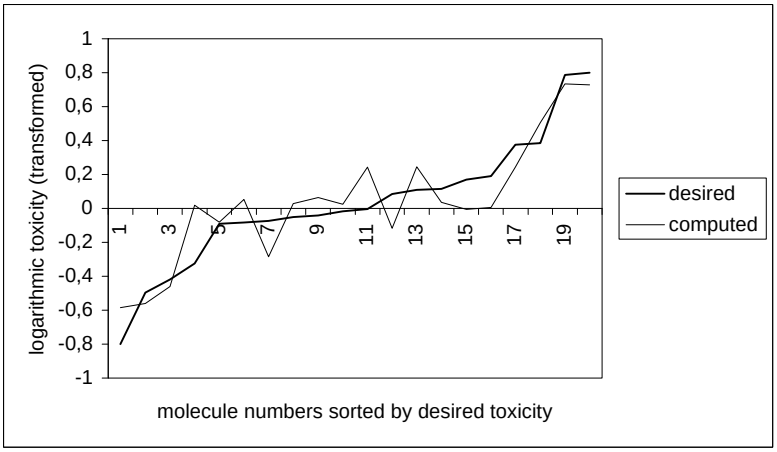

Figure 5: Computed output vs. desired output using a BSN for 20 selected molecules with 2 inputs.

CONCLUSIONS

In 1998 the number of chemicals registered with the Chemical Abstract Service (CAS) rose to over 19 million (CAS, 1999). The chemicals registered increased more than 3 million between 1996 and 1998 (Basak, et. al, 1999). Therefore, it is of special interest to have computational models alternative to traditional test methods for the prediction of toxicity. In this sense the methods of CI can be considered a solution-platform to predict the toxicity of chemicals for environmental and health problems. In a first approach we applied these methods to a restricted class of chemicals, namely pesticides.

The task of QSAR for the acute toxic activity of pesticides is very difficult in our case. To model the complex dependency of the aquatic toxicity on the molecular descriptors, we have only 164 molecules with 175 descriptors. Furthermore, the total set of molecules is very inhomogeneous – it consists of 7 different chemical classes with the biggest subset of a single class containing 27 organophosphorus – and the concentration values LC_{50} are not very reliable compared to other databases.

Any system modeling the QSAR must be big enough to capture its complex behavior, however, it must be small enough to have generalization abilities and not just "learn" the 164 given patterns. In view of the small data set, the above results must be treated with caution and they should be verified on a statistically totally independent test set of acceptable size.

The comparison of different ANNs shows that MLPs have a good performance; therefore MLPs can be seen as prototype for function approximation and non-linear interpolation. The preliminary results indicate that BSNs, which have elements of fuzzy systems, neural networks and clustering algorithms, are very promising for our task of toxicity prediction. It seems that the BSN together with the complete permutation search in the descriptor space yield better results than the MLPs with the genetic search in the descriptor space. However, BSNs are a more complex approach than MLPs and additionally the full enumeration takes longer than the genetic search. Further investigations are necessary to be on the safe side with respect to the accuracy.

ACKNOWLEDGEMENT

This work has been supported by the Commission of the European Communities under the Program "Environment and Climate", Project "COMET", Contract No. ENV4-CT97-0508.

REFERENCES

Andrews, R.; Diederich, J.; Tickle, A.B., 1995, "A survey and critique of techniques for extracting rules from trained artificial neural networks", Konnektionismus und Neuronale Netze, pp. 187–210, GMD, Sankt Augustin.

Basak, S.C., et al., 1999, "Use of Statistical and Neural Net Methods in Predicting Toxicity of Chemicals: A Hierarchical QSAR Approach", Proc. AAAI Spring Symposium Series, March 22–24, 1999, Stanford University (CA), USA.

Benfenati, E., 1998, "Personal Communications", Istituto di Ricerche Farmacologiche Mario Negri, Milan, Italy.

Brown, M.; Harris, C., 1994, "Neurofuzzy Adaptive Modeling and Control", Chapter 8, Prentice Hall, New York.

Grauel, A., Ludwig, L.A., Renners, I., Berk, F., 1999, "Computational Intelligence and Predictive Toxicology", Proc. AAAI Spring Symposium Series, March 22–24, 1999, Stanford University (CA), USA.

Zhang, J.; Knoll, A., 1996, "Constructing fuzzy controllers with B-spline models", Proceedings of the IEEE International Conference on Fuzzy Systems, New Orleans.