PREDICTIVE TOXICOLOGY WITH OPTIMIZED NEURAL NETWORKS

L.A. Ludwig, F. Berk, A. Grauel
Universität-GH Paderborn, Abteilung Soest
Fachbereich 16, Fachgebiet Mathematik
Steingraben 21, D-59494 Soest, Germany
Phone: +49-2921-378171, Fax: +49-2921-378180
E-mail: ludwig@uni-paderborn.de

ABSTRACT: In this contribution we present investigations about the use of artificial neural networks in the area of predictive toxicology, 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 on different input space vectors are given.

KEYWORDS: predictive toxicology, quantitative structure-activity relationship, artificial neural networks, evolutionary computation

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 multi-layer perceptrons (MLPs). Furthermore, investigations about input space and network optimization by evolutionary computation are discussed.

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 Pesticide 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 that 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 that 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.

MULTILAYER PERCEPTRONS

For modeling the dependency of the toxicity on the significant descriptors a feasible approach is the use of feed-forward ANNs. In a first step we employed MLPs for function approximation using all 157 molecular descriptors as input and $-\log_{10}(LC_{50}/(\text{mmol/l}))$ for trout as target output. The network was 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 $-\log_{10}(LC_{50}/(\text{mmol/l}))$ was linearly transformed to [-0.8, 0.8]. As test set we randomly picked 16 molecules (≈ 10 %) out of the whole set of molecules – the remaining 148 molecules constituted the training set. The results were of course very poor with a maximal correlation on the test set of only $|r| = 0.283$. This is easily to understand, as such big networks with 157 input nodes will not generalize if they are trained with just 148 patterns. The use of only 61 descriptors that are significantly correlated with the output value did not improve the generalization at all.

In the next step we used 57 descriptors that have a significant value in Spearman's rank correlation. The results were very astonishing as we got very low correlations of $|r| = 0.151$ but also very promising values of $|r| = 0.808$ depending on the random choice of training and test subsets. The best network uses 57 linear input nodes, 2 sigmoidal nodes in the first hidden layer, 1 sigmoidal node in the second hidden layer and 1 sigmoidal output node. Even as we took the 57 significantly rank-correlated descriptors or their 8 PCs with an eigenvalue bigger than 1 the results varied that much. 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 Pesticide 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.

OPTIMIZATION BY EVOLUTIONARY COMPUTATION

An algorithm for structure and topology optimization is developed and tested for MLPs and recurrent ANN. GAs 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.

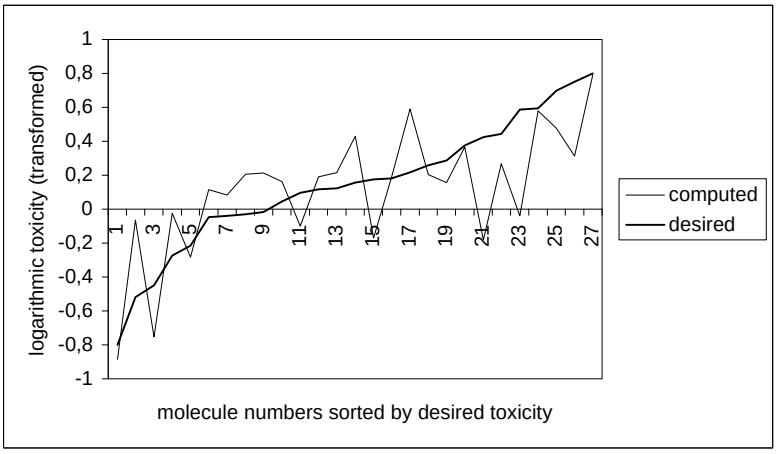

Figure 1: Computed output vs. desired output using an MLP for the organophosphorus class with 2 inputs.

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.

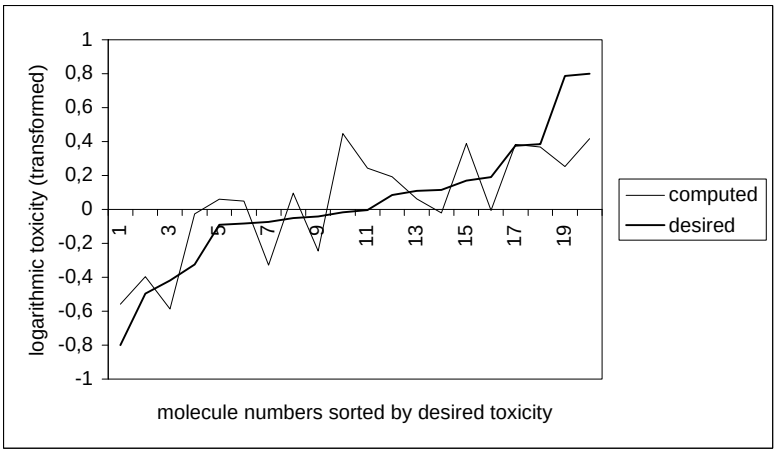

Figure 2: Computed output vs. desired output using an MLP for 20 selected molecules with 2 inputs.

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.

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

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.
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.