B-SPLINE NEURO-FUZZY SYSTEMS FOR TOXICITY PREDICTION

I. Renners, A. Grauel, L.A. Ludwig
University of Paderborn
Campus Soest, Department of Mathematics
Steingraben 21, D-59494 Soest, Germany
Tel.: ++49-2921-378173, Fax: ++49-2921-378180
E-mail: renners@ibm16.uni-paderborn.de

ABSTRACT: In this contribution we present investigations about the use of computational intelligence for toxicity prediction of pesticides. 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 157 descriptors are taken into account. From this set of data preliminary results using a special type of neuro-fuzzy system, namely B-spline networks, are given.

KEYWORDS: predictive toxicology, quantitative structure-activity relationship, B-spline networks

INTRODUCTION

Predicting the toxicology of chemicals is of special interest for enviromental and health problems. Recent investigations support the general assumption that macroscopic properties like toxicity and ecotoxicity strongly depend on microscopic features and the structure of molecules. This assumption enables us to set up **Q**uantitative **S**tructure-**P**roperty **R**elationship (QSPR), **Q**uantative **S**tructure-**A**ctivity **R**elationship (QSAR) and **Q**uantative **S**tructure-**R**etention **R**elationship (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 structure of molecules can be identified and characterized by certain molecular descriptors. The objective is to set up a functional dependency of the toxicity to a certain aspect on the selected molecular descriptors. We used **B**-**S**pline **N**etworks (BSNs), a network type which can be interpreted as a Takagi-Sugeno neuro-fuzzy system, to have the possibilty of applying rule extraction methods. Furthermore BSNs offer some notable properties described below.

DATA STATISTICS

We used a set of 164 pesticides from 7 different classes with data on acute toxicity for rainbow trout and daphnia taken from the Pesticide Manual [Benfenati,1998]. The concentration for this aquatic toxicity is given by $-\log_{10}(\frac{LC_{50}}{mmol/l})$ where LC_{50} is the **L**ethal **C**oncentration which leads to the death of 50 percent of individuals after a certain amount of time. The molecules are described by molecule attributes called descriptors. 175 descriptors were computed for each molecule and after deleting 18 of these descriptors because of missing values we obtained a dataset of 157 decriptors for each of the 164 pesticides. The input data was standardized to $\mu = 0, \sigma = 1$ by $\tilde{x} = \frac{(x-\overline{x})}{s}$ and the output variable $-\log_{10}(\frac{LC_{50}}{mmol/l})$ was linearly transformed to $[-0.8, 0.8]$ and is now called $-\log_{10}^{\text{trans}}(\frac{LC_{50}}{mmol/l})$.

B-SPLINE NETWORKS

B-splines have been employed as surface-fitting algorithms for computer aided design tasks. A B-spline function is a piecewise polynomial mapping formed from a linear combination of weighted basis functions. Because multivariate B-splines are defined on a lattice, a BSN is considered as belonging to the class of associative memory networks.

DEFINITION AND BASIC PROPERTIES OF B-SPLINES

The B-spline $N_{i,k}$ of order k with knots $\lambda_i, \ldots, \lambda_{i+k}$ is defined as:

$$N_{i,k}(x) = (\lambda_{i+k} - \lambda_i) \sum_{p=0}^{k} \frac{(\lambda_{i+p} - x)_+^{k-1}}{\prod_{\binom{h=0}{h\neq p}}^{k} (\lambda_{i+p} - \lambda_{i+h})}. \tag{1}$$

Therefore l B-splines of order k are defined over a knot-vector consisting of $l+k$ knots (Fig. 1). Using these l B-splines as linguistic terms, the minimum input value of a BSN is determined by $a = \lambda_k$ and the maximum input value by $b = \lambda_{l+1}$, constituting the valid input interval to $[a, b]$ in which the linguistic terms built up a partition of unity.

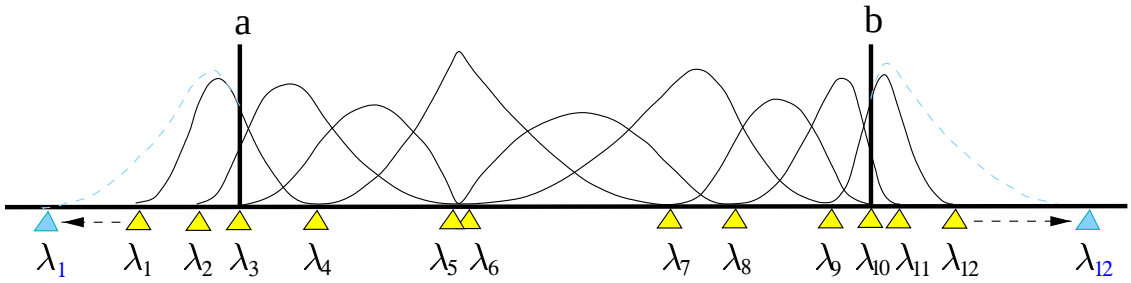

Figure 1: Nine asymmetric B-splines of order 3 defined over 12 non-uniformly distributed knots.

The most important properties of B-splines, with respect to neuro-fuzzy modelling are:

- Recursion: $N_{i,k+1}(x) = \frac{x-\lambda_i}{\lambda_{i+k-1}-\lambda_i} N_{i,k}(x) + \frac{\lambda_{i+k}-x}{\lambda_{i+k}-\lambda_{i+1}} N_{i+1,k}(x)$

  $$N_{i,1}(x) = \begin{cases} 1 & \text{, if } x \in [\lambda_i, \lambda_{i+1}) \\ 0 & \text{, otherwise} \end{cases}$$

- Positivity: $N_{i,k} \geq 0$ for all x
- Local support: $N_{i,k} = 0$ if $x \notin [\lambda_i, \lambda_{i+k}]$
- Partition of unity: $\sum_{i=1}^{l} N_{i,k}(x) = 1, \quad x \in [\lambda_i, \lambda_{i+1}].$

Considering n inputs (in our case n descriptors) a BSN with $l_d (d = 1, \ldots, n)$ linguistic terms covering each input interval $[a_d, b_d]$ forms an n-dimensional hypercube with

$$r = \prod_{d=1}^{n} l_d \tag{2}$$

n-dimensional receptive fields. All r n-dimensional receptive fields $(n > 1)$ of the lattice are covered by multivariate B-splines $N_k^j(x) (j = 1, \ldots, r)$ which are formed by multiplying the corresponding n univariate B-splines (Fig. 2):

$$N_k^j(x) = \prod_{d=1}^{n} N_{i_d,k_d}(x), \tag{3}$$

where k is an n-dimensional integer vector composed of the orders k_d of the n univariate basis functions and $i_d = 1, \ldots l_d$. The output of a **S**ingle **I**nput **S**ingle **O**utput (SISO) B-spline network is the summation of the activations of all univariate B-splines multiplied by a corresponding weight w_i assigned to each univariate B-spline:

$$y = \sum_{i=1}^{l} w_i N_{i,k}(x). \tag{4}$$

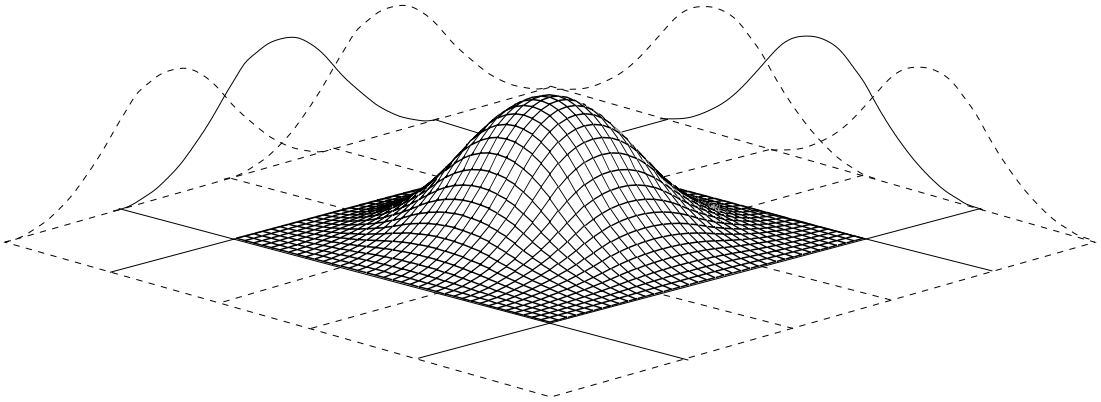

Figure 2: A bivariate B-spline built up by two univariate B-splines of order 3.

Whereas the output of a **M**ultiple **I**nput **S**ingle **O**utput (MISO) B-spline network is the summation of the activations of all multivariate B-splines multiplied by a corresponding weight w_j assigned to each multivariate B-spline:

$$y = \sum_{j=1}^{r} w_j N_k^j(x). \tag{5}$$

Due to the local support property of B-splines the number of activated B-splines for each input is given by

$$t = \prod_{d=1}^{n} k_d, \tag{6}$$

and therefore only t B-spline activation values must be calculated to compute the output value.

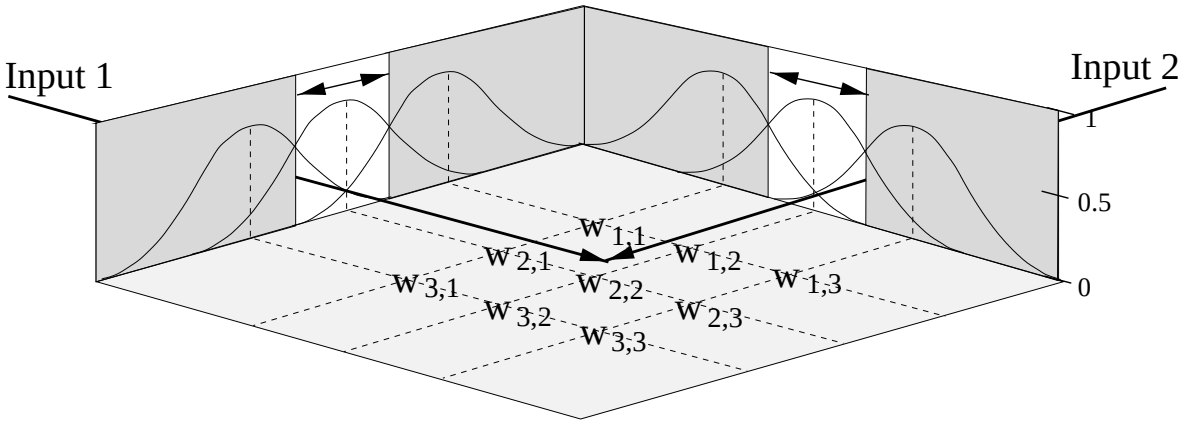

Figure 3: A two-dimensional BSN with three linguistic terms of order 3 on each input interval with r weigths $w_{p,q} (p = 1, \ldots, l_1 \, ; q = 1, \ldots, l_2)$.

WEIGHT DETERMINATION OF B-SPLINE NETWORKS

There are two methods to determine the weights of a BSN. The first one adjusts the weights *iteratively* by using gradient information given by the error-surface [Zhang,Knoll,1996] (normally used to model *on-line* tasks). The second one generates the weights by matrix inversion [Dierckx,1993]. In the following we always used the method based on matrix inversion. Only in cases were not all receptive fields were covered by at least one data point (which leads to a rank deficient of the observation matrix) we used gradient descent.

NUMERICAL RESULTS

MODELLING THE COMPLETE DATASET

Unfortunaly the concentration values $-\log_{10}(\frac{LC_{50}}{mmol/l})$ are not very reliable compared to other databases. Calculating a linear regression shows that the best descriptor (*logD pH 7.4*) is correlated to the logarithmic concentration values with $\mid r \mid = 0.595$ which is classified as *likely descriptor* (Tab. 1).

classification	correlation
substantial descriptors	$\mid r \mid \geq 0.99$
important descriptors	$0.99 > \mid r \mid \geq 0.80$
likely descriptors	$0.80 > \mid r \mid \geq 0.50$
specific descriptors	$0.50 > \mid r \mid$

Table 1: Classification of the molecular descriptors.

To find better correlation values we propose a small model which does not just ”learn” the 164 given patterns while performing a non-linear multi-dimensional regression. In this investigastion we used a BSN with two inputs and three uniformly distributed linguistic terms of order 3 on each input interval (Fig. 3). For all 12246 possible input combinations (157 descriptors and trout $-\log_{10}^{\text{trans}}(\frac{LC_{50}}{mmol/l})$ as output) we computed the 164-fold crossvalidated test error to achieve the maximal possible statistical security by testing every output independently from all other. The best found input descriptor was *logD pH 9* in combination with *number of P atoms* which lead to a correlation of $\mid r \mid = 0.65$ as shown in Tab. 2 and an output as illustrated in Fig. 4.

Input 1	Input 2	crossvalidated test MSE	crossvalidated test correlation
logD pH 9	Number of P atoms	0.066079	0.649574
logD pH 7.4	Kier&Hall index (order 3)	0.068510	0.631180
logD pH 9	Number of S atoms	0.069194	0.627828
logD pH 7.4	Number of S atoms	0.069810	0.623174
logD pH 9	Relative number of S atoms	0.071782	0.609148

Table 2: Numerical results (top 5) of the descriptor search on the complete dataset.

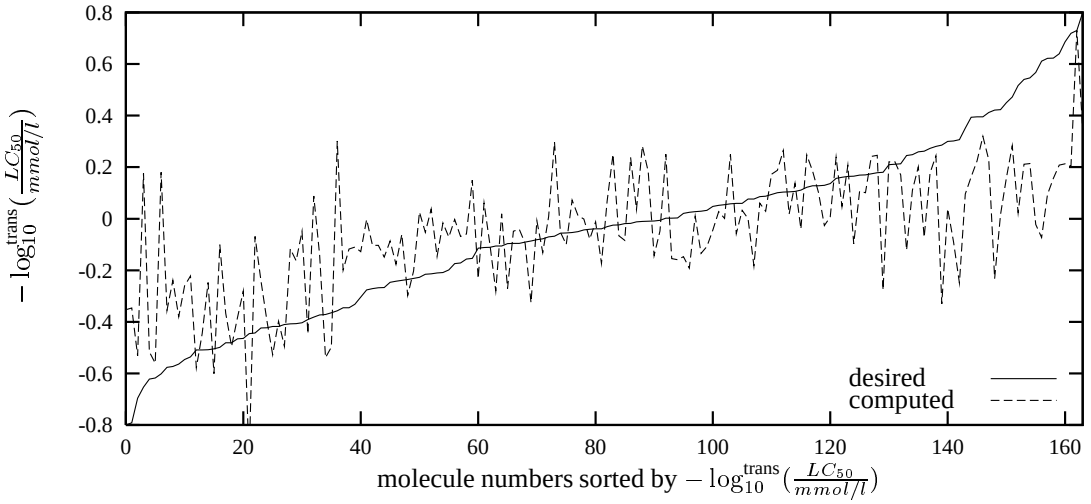

Figure 4: Computed output vs. desired output of the complete data set.

MODELLING THE SUBCLASS ORGANOPHOSPHORUS

As the data are very inhomogeneous, we decided to investigate the chemical subclass *organophophorus* containing 27 molecules (subclass of the given data with the most elements). Calculating a linear regression shows that the best descriptor (*Moment of inertia A*) is correlated to the logarithmic concentration values with $\mid r \mid = 0.75$. Again we used a BSN with two inputs and three uniformly distributed linguistic terms of order 3 on each input interval (Fig. 3) and computed the 27-fold crossvalidated test error for all possible input combinations (now 11325 because some descriptors have an equal value for all molecules of this subclass). The best found input descriptor was *logD pH 5* in combination with *number of P atoms* which lead to a correlation of $\mid r \mid = 0.8$ as shown in Tab. 3 and an output as illustrated in Fig. 5.

Input 1	Input 2	crossvalidated test MSE	crossvalidated test correlation
logD pH 5	Number of P atoms	0.051879	0.802620
logD pH 7.4	Number of P atoms	0.053715	0.792967
YZ Shadow	Number of P atoms	0.053730	0.792131
Gravitation index (all pairs)	Number of P atoms	0.055254	0.786384
Randic index (order 1)	Number of P atoms	0.056000	0.782189

Table 3: Numerical results (top 5) of the descriptor search on the subclass of *organophophorus*.

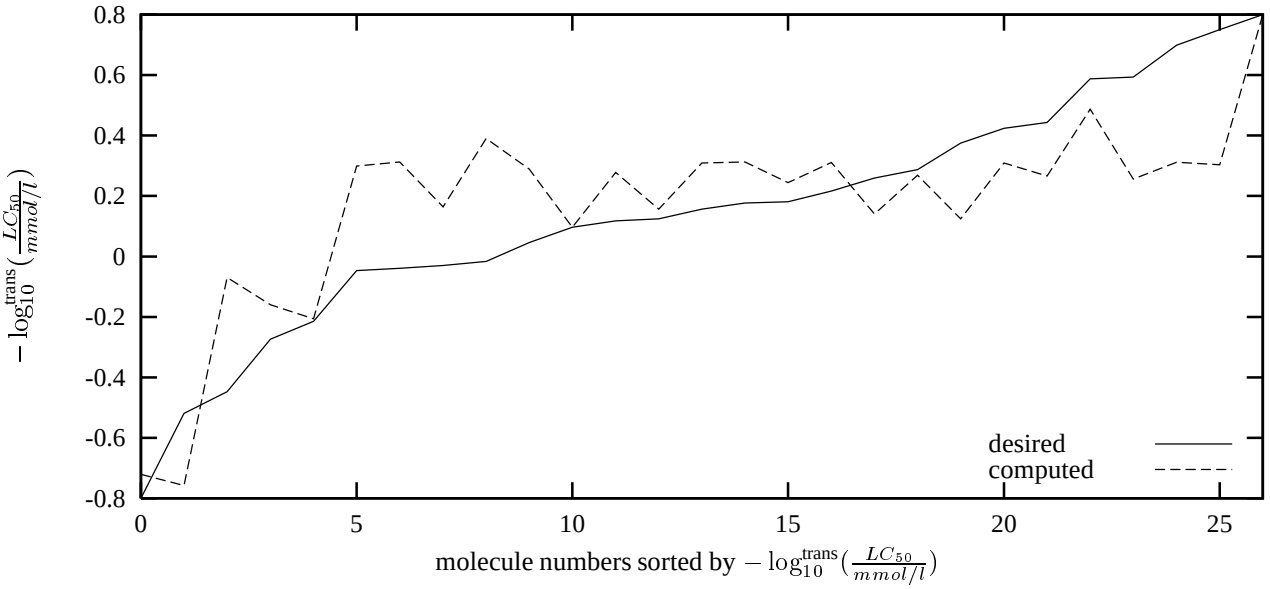

Figure 5: Computed output vs. desired output of subclass *organophosphorus*.

MODELLING A SUBCLASS OF 20 SELECTED DATA

This second subclass arises from a comparison of the concentration values in the Pesticide Manual and the HS database. The intersection of both contains 39 molecules with partly big differences in $-\log_{10}(\frac{LC_{50}}{mmol/l})$. On the other hand some values are absolutely identical which leads to the assumption that for these values the HSDB and Pesticide Manual refer to the same reference of measurement. Hence we selected 20 molecules that have small but non-zero differences between HSDB and Pesticide Manual. Calculating a linear regression shows that the best descriptor (*Gravitation index (all pairs)*) is correlated to the logarithmic concentration values with $\mid r \mid = 0.826$. Using the same BSN structure as described above and performing again a total permutation search we achieved a best correlation of $\mid r \mid = 0.91$ as shown in Tab. 4 and an output as illustrated in Fig. 6. The best found input descriptor combination was *Gravitation index (all pairs)* with *WPSA-1 Weighted PPSA*.

Input 1	Input 2	crossvalidated test MSE	crossvalidated test correlation
Gravitation index (all pairs)	WPSA-1 Weighted PPSA	0.024166	0.910889
Molecular volume	Rel. number of double bonds	0.032426	0.877851
Average Comp. Inf. content (order 2)	Rel. number of double bonds	0.035184	0.880269
Gravitation index (all bonds)	RPCS Rel. pos. charged SA	0.035434	0.867700
FPSA-3 Fractional PPSA	Number of Br atoms	0.037093	0.858296

Table 4: Numerical results (top 5) of the descriptor search on the subclass of 20 selected molecules.

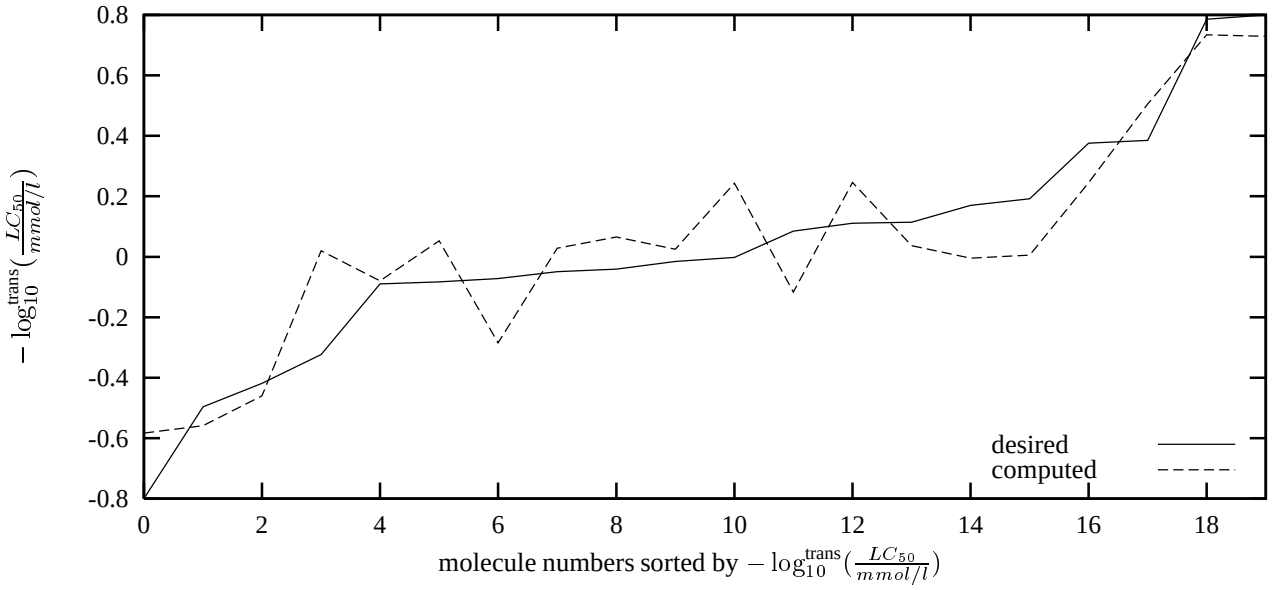

Figure 6: Computed output vs. desired output of the subclass of 20 selected molecules.

CONCLUSIONS

To model the complex dependency of the aquatic toxicity we have only a small very inhomogeneous and unreliable dataset of 164 molecules, consisting of 8 different chemical classes. Nonetheless we achieved on the biggest chemical subclass containing 27 *organophosphorus* a correlation of $\mid r \mid = 0.8$ and on the subclass containing 20 selected molecules a correlation of $\mid r \mid = 0.91$, increasing the output accuracy to the magnitude of *important descriptors*. Both submodels use only two descriptors and therefore using more descriptors in a sparse optimized architecture promises good results. Furthermore investigations are made to increase the available dataset.

ACKNOWLEDGEMENT

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

REFERENCES

[Benfenati, 1998] Benfenati E. 1998. *Personal Communications*, Instituto di Ricerche Farmacologiche Mario Negri, Milan, Italy.

[Dierckx, 1993] Dierckx P. 1993. *Curve and Surface Fitting with Splines*, Oxford Science Publications.

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