ELSEVIER

Journal of Molecular Structure (Theochem) 622 (2003) 85–95

THEO CHEM

www.elsevier.com/locate/theochem

The use of neural networks in QSARs for acute aquatic toxicological endpoints

Klaus L.E. Kaiser*

National Water Research Institute, 867 Lakeshore Road, Burlington, Ont., Canada L7R 4A6

Abstract

This review surveys the applications of neural network (NN) methodologies to the field of Quantitative Structure–Activity Relationships (QSARs) in aquatic toxicology. Several NN methods have been applied to substantial data sets (some involving over 1000 chemicals) for acute and sublethal toxicity endpoints for fish, invertebrate, protozoan and bacterial species. The results clearly demonstrate the methods' general ability to detect and apply non-linear structure–activity relationships for the prediction of the corresponding values for compounds not part of the training sets.

Keywords: Quantitative structure–activity relationships; Neural networks; Aquatic toxicity; Large data sets

1. Introduction

Interest in the application of Quantitative Structure–Activity relationships (QSARs) to the field of aquatic toxicology arose in the late 1970s. Large-scale production and use of certain chemicals generated widespread concern about their environmental and health effects. During that time, a rapidly increasing number of substances were identified as contaminants of water, sediments and aquatic biota, and their bioaccumulation and pathways in the aquatic environment were little understood. Public reaction forced the USA, Canada, several European and Asian countries and international organizations, such as the Organization for Economic Cooperation and Development, to create new environmental protection laws aiming to curb or control substance release. For example, the widespread occurrence of polychlorinated biphenyls (PCBs) in the environment, first discovered by Jensen [1], led to their regulation and eventual ban in Canada [2], the USA [3] and elsewhere [4] during that time.

Also, standardization of certain tests allowing the combination of experimental toxicological data from different laboratories into larger data sets useful for QSAR analysis, was adopted in various jurisdictions and by the OECD [5]. To compensate for the severe lack of compatible toxicological data for aquatic organisms at the time, several research groups initiated systematic measurement programs for quite large sets of compounds, and some of this work is continuing at this date. In terms of aquatic organisms, most of the measured data are for on several species of fish, i.e. fathead minnow (*Pimephales promelas*), guppy (*Poecilia reticulata*), bluegill sunfish (*Lepomis macrochirus*), zebrafish (*Brachydanio rerio*) and rainbow trout (*Oncorhynchus mykiss*), several species of crustaceans (several species of *Artemia*, *Crangon*, and *Daphnia*), protozoa (*Tetrahymena pyriformis*), algae (several species of *Chlorella* and *Scenedesmus*) and bacteria (*Vibrio fischeri*). There is also a considerable amount of data on aquatic insect larvae

* Tel.: +1-905-336-4756; fax: +1-905-336-6430.
E-mail address: klaus.kaiser@cciw.ca (K.L.E. Kaiser).

0166-1280/03/$ - see front matter
PII: S0166-1280(02)00620-6

(*Aedes* and *Culex* mosquito species). Notification requirements for the manufacturing and use of new chemicals typically also stipulate information relevant to environmental protection, such as aqueous solubility [6]. Over the next few years, several tens of thousands of existing substances, including several thousand high-production-volume chemicals must be assessed for their environmental fate and effects. As both testing facilities, and available funds and time are limited, the task of assessing the environmental effects of chemicals is shifting from laboratory work towards the development of comprehensive QSAR expert systems for the prediction of toxic effects of substances based solely on molecular structure and (occasionally) on environmental conditions specific to the experiment.

For a considerable time, the task of modeling the relationship between toxicity endpoints and various measurable characteristics of toxicant molecules had been handled almost exclusively through variations of the linear regression paradigm and its multivariate extensions, including partial least squares (PLS) and principal component analysis [7]. Classical examples of such QSARs are the Hansch-type models, where the toxicity is expressed as function of the octanol/water partition coefficient, or an incremental derivative thereof, such as a substituent's hydrophobic fragment value [8]. The main problem with such models is their extremely limited scope. In other words, they perform well, but only for very narrow classes of compounds. To simultaneously accommodate more complex substances, not sharing the same common molecular backbones, fragments, or characteristics of a congeneric class, a variety of empirical schemes were developed. Typically, they involve rules of thumb on how to handle various classes of chemicals and the selection of which particular relationship to consider for each compound and other workarounds. Most of the attempts to build a toxicity prediction system to calculate the toxicity of chemicals which contain simultaneously more than one substance class result in failure because of their limited scope and the conflicting predictions (often differing by several orders of magnitude) for compounds sharing features common to several classes. An example of such a system is the US EPA's Ecological Structure Activity Relationships (ECOSAR) toxicity estimation program [9], which has been criticized for lack of statistical significance [10,11]. The same criticism is valid for EPA's ASTER expert system, which is based on the same principles [12]. An example of a viable toxicity prediction system using only a few rules of thumb and relying on a few but statistically significant multivariate regression type QSARs is implemented in the program TOPKAT. However, it too suffers from the restrictions imposed by the models on which it is based.

The key to toxic effects of chemicals resides in their molecular structure. In order to identify meaningful relationships between the molecular structure and the effects, the first challenge the expert is faced with is to establish a pool of variables to be considered by the modeling exercise. This pool of variables may include indicators for the presence and/or counts of various atoms and/or fragments of interest, as well as various (more or less subjective) physico-chemical parameters associated with the whole molecule, particular fragments and/or particular atoms of the associated structure, as well as various topological indicators. The second challenge is to generate these data. It must also be kept in mind that in all relationships based on physico-chemical parameters such as log P, kinetic parameters, orbital energies, and so on, each is contributing its own additional uncertainties to the resulting model. This can easily be seen, for example, by comparing the values obtained from different estimation programs for log P with each other and/or the measured values; differences of several orders of magnitude are not uncommon. Therefore, where possible, the best approach is to reduce or entirely eliminate these additional sources of errors from the model by eliminating the intermediate estimation of values from the target relationships, and use instead as input only information reflecting structural features of the studied compounds. The particular choice of parameters constitutes the basis for defining the resulting model's domain. Nevertheless, finding a model's domain boundaries remains an unsolved problem. Only partial and subjective solutions, based on the range of various values associated with the input parameters corresponding to the compounds used to build the model, are available at this moment. With the assumption that the appropriate choice of parameters suited to properly describe the corner of chemical universe subject to study has been made, the

third challenge is to identify the character of the relationship between the input and output parameters of the model. Approximating the target relationship using multilinear regression may generate satisfactory results only for very small and structurally narrow classes of compounds. For the case of large and highly diverse data sets such an approach is unrealistic and an assumption of non-linearity is the natural alternative. On a common basis, because of the complexity and the peculiarities of the available data, it is impossible to verify the mathematical assumptions imposed by most multivariate non-linear statistical methodologies. Consequently, a new approach and new tools are needed for the modeling exercise. In this context, (artificial) neural networks (NN) offer the most affordable alternative.

It should also be mentioned that there has been criticism of the NN approach, particularly from regulatory agencies, as to their 'lack of transparency' and perception as 'black box'. Such views are vehemently being opposed to by specialists in the field and the reader is referred to, for example, the classical work by Zupan and Gasteiger [13]. In addition, it should be noted that there are very different types of NNs, including those that do and those that do not require iterative optimization of the learning phase (e.g. determining the optimum number of training cycles in order to maximize the prediction capability/learning ratio). However, when using identical inputs, and with proper optimization (where such is necessary), NNs will give reproducible results, based on defined mathematical relationships.

2. Neural network based QSARs in aquatic toxicology

Basically, an artificial NN is a computational device consisting of a group of processing elements (neurons) organized in subgroups (layers). Each subgroup may make its independent computations and may pass the results to yet another subgroup. The last subgroup consisting of one or more processing elements determines the output from the network. Practically, at a very simplified level, artificial NNs mimic the way the human brain organizes and processes information, and the way the meaningful part of that information is identified and stored for future purposes. Chemists and toxicologists may get an insight into the field of artificial NN modeling in Ref. [13]. More examples of QSAR studies involving artificial NNs may be found in Refs. [14,15]. In aquatic toxicology artificial NNs may be used mainly for three different purposes; (a) mapping; (b) classification; (c) model dimensionality reduction. The mapping focuses on building QSARs for the estimation of toxicity endpoints, the time required for complete primary and ultimate biodegradation, lipophilicity, etc. Practically, all artificial NNs used for mapping are feed-forward NNs. The best QSAR mapping models were obtained using two different neural network paradigms: (i) back-propagation neural networks (BNNs), based on minimizing distances and maximizing correlations; and (ii) probabilistic neural networks (PNNs), based on recognizing the distributional characteristics of the population. Kohonen neural networks (KNNs) were used to handle classifications based on mode of action [16]. Although not exploited yet, there is an enormous potential to use PNNs (which are in essence implementations of Bayesian classifiers) to handle very diverse toxicant classifications (including mode of action, carcinogenic and/or mutagenic character, etc.). The model dimensionality reduction targets the combination of input parameters corresponding to the best performing NN as identified through the application of a genetic algorithm type (GA) search template to a population of NN models using for fitness evaluation purposes various numerical characteristics associated to the errors generated by the models.

2.1. Neural network models for toxicity to fish species

One of the first attempts in modeling the toxicity of chemicals to fish species making use of artificial NNs is reported in Ref. [17]. The endpoint subject to study was the 96 h LC50 toxicity to the fathead minnow (*P. promelas*) expressed in log(l/mmol). The purpose of the modeling experiment was to investigate the potential of replacing fish as laboratory test species with the luminescent bacterium *V. fischeri*. The input in the model consisted of measured EC50 toxicity to *V. fischeri* (5–30 min), log K_{ow}, the logarithm of the molecular weight, and another 48 molecular descriptors including

the presence and/or counts of numbers of atoms and/or fragments of interest (aromatic rings, acid groups, ester groups, amino groups, etc.). The data set contained information on 419 highly diverse compounds, including both organics and inorganics. The model used a three-layer back-propagation NN with 51 neurons in the input layer, 7 neurons in the hidden layer, and one neuron in the output layer, all in agreement with the geometric pyramid rule (see page 176 in Ref. [18]). Data pre-processing consisted in combinations of Z-transforms, sigmoid logistic, and convenient interval compression functions. The learning involved 400 training cycles, the optimal moment of stopping the training being detected using a complete 20% leave-out cross-validation experiment using random selection. The random uniform distribution in [−0.1, 0.1] was used to seed the training process. The best learning rate was found to be 0.5. On the given 419 compounds set, the correlation between measured and predicted values generated by this model was 0.916, with an average error of 0.158, standard deviation of errors of 0.596 and the average square error of 0.333, altogether indicating very good performance. The experiment proved that for a large number of classes of compounds it is possible to replace the fathead minnow as test species with *V. fischeri*. Ref. [17] is also important, although for a completely different reason: it gives the explicit system of equations describing a trained NN based on real data, clear proof that NNs are not black boxes.

Fine tuning a back-propagation NN is a challenging and time-consuming task. Other NN choices may provide a faster approach. A PNN with Gaussian kernel QSAR model for the same problem, using the same data pre-processing algorithms, and validated through a complete 20% leave-out cross-validation experiment and using random selection, is reported in Ref. [19]. For the given 419 compounds set, the correlation between measured and predicted values generated by the PNN model was 0.941, with an average error of −0.041, standard deviation of errors of 0.530 and the average square error of 0.283, completely confirming the conclusions of the model reported in Ref. [19]. At this stage, it was natural to ask the question if the presence of the measured toxicity value for another species was really necessary for the QSAR model, and the answer was that it was not necessary. Three PNNs with Gaussian kernel QSAR models for the 96 h LC50 toxicity to the fathead minnow, based on the same set of compounds, and using log K_{ow} and the same physico-chemical descriptors but no toxicity values for another species are presented in Ref. [20]. Training corrections were included in the models and the validation was performed through complete 20% leave-out cross-validation experiments and used random selection. The difference between the three models was simply the choice of data pre-processing strategy. For example, applying data pre-processing consisting of combinations of *Z*-transforms and hyperbolic tangent functions, on the given 419 compounds set, the correlation between measured and predicted values was 0.932, with an average error of practically zero, the standard deviation of errors of 0.55 and the average square error of 0.305. Replacing the hyperbolic tangent function with the sigmoid logistic function resulted in a very small improvement in the model. The same was true for data pre-processing based on finite interval transforms.

All previously discussed NN models for the 96 h LC50 toxicity to the fathead minnow included log K_{ow} as input variable. For many substances, the log K_{ow} values are generated by other QSARs interpreting the molecular structure information in their peculiar ways. Therefore, it should be possible to omit log K_{ow} as variable when including in the list of input variables the type of information on which the computation of log K_{ow} relies. Such a model is reported in Ref. [21]. It is based on a larger data set consisting of 865 highly diverse chemicals and uses as input variables 33 molecular descriptors associated with the presence or counts of molecular fragments (aromatic rings, acid groups, nitro groups, cyano groups, ether linkages, etc.), counts of atoms (C, H, As, Br, Cl, F, Fe, Hg, I, K, N, Na, O, P, S, Se, Si, Sn and Zn), the logarithm of the molecular weight, and the logarithm of the ratio of the molecular weight corresponding to all counted atoms over the whole molecular weight. The PNN with Gaussian kernel was used to handle the modeling exercise. The data pre-processing consisted of combinations of *Z*-transforms and hyperbolic tangent function. Training corrections were included in the model. The validation was handled through a complete 20% leave-out cross-validation experiment using random selection. On

the 865 substances data set, the correlation coefficient between measured and values predicted by the model was 0.932, with an average error of practically zero, standard deviation of errors of 0.591, and average square error of 0.349. This is one of the first non-log K_{ow} type models for the 96 h LC50 to the fathead minnow reported in the literature and based exclusively on molecular structure. A practical implementation of this methodology (described in Ref. [22]) has been used to perform substance toxicity screening for almost 2000 compounds from the Canadian Domestic Substances List, with results superior to all other available methodologies, including ECOSAR, ASTER, CNN, TOPKAT, and OASIS [23].

A completely different approach for the prediction of the acute toxicity of organic compounds to the fathead minnow based on the molecular structure is presented in Ref. [24]. The modeling experiment considered a 375 compound subset of the data set used in Ref. [19]. For each compound, a total of 272 descriptors were calculated (counts of atoms of interest, number of double bonds, solvent-accessible surface area, moments of inertia, the charge on the most negative or positive atoms, HOMO and LUMO energies, charged partial surface area descriptors, etc.). Eliminating the featureless descriptors, the ones exhibiting greater than 90% redundancy in their response values, as well as selecting from each subgroup of highly correlated descriptors a representative one, resulted in a reduced pool of 123 descriptors. The 375 substances set was randomly split into three subsets, the largest of them consisting of 287 compounds and used for training purpose, a subset of 44 compounds used to validate the training (e.g. to detect the moment when the training must be stopped in order to avoid over-training), while the remaining compounds were used as external test set. The reduced pool of descriptors was then submitted to a GA feature selection routine incorporating a three-layer 8-6-1 feed forward fully connected neural network (CNN) for fitness evaluation. The 10 best models identified by the GA/CNN combination were reviewed and then used to generate predictions for the compounds in the external test set. The criterion of selecting the final model was its performance on this set. The eight descriptors used by the final model were as follows: (d1) count of the number of chains of length 7 in the molecule; (d2) number of double bonds (excluding those in aromatic rings); (d3) sum of weighted paths originating from oxygen atoms; (d4) ratio of intermediate to shortest geometric axis of the molecule, including hydrogen atoms; (d5) the charge of the most negatively charged atom in the molecule; (d6) second major moment of inertia; (d7) the product of the sum of negatively charged partial surface area and the total molecular surface area; and (d8) the ratio of the sum of charges on acceptor atoms over the number of acceptor atoms. On the given 375 compounds set, the correlation between measured values and the predictions generated by this model was 0.866, with an average error of -0.022, the standard deviation of errors of 0.722 and the average square error of 0.522.

It is possible to expand the scope of the toxicity modeling exercise by including into the model also information which reflects biological response characteristics of the organisms under study as well as information on specific test conditions. There are two studies reporting such models, specifically on the acute toxicity of pesticides to fish species. The first study [25] investigates the acute toxicity to the rainbow trout *O. mykiss* using detailed information of toxicity experiments for 70 pesticides. It is based on a data set consisting of 447 LC50 experimental values expressed as log(l/mmol), together with the associated bioassay information (fish weight, time of exposure, water temperature, pH, and hardness). For modeling purposes, the data set was split into a training set and a test set. The training set grouped all the information concerning 384 of the 447 available endpoints. The test set grouped the information related to the remaining 63 endpoints. The developed model consists of a three-layer 13-6-1 BNN. Additional to the fish weight, time of exposure, water temperature, pH and hardness, the input to the model includes eight autocorrelation descriptors designed from the hydrogen-suppressed graphs of the molecules: H_0 to H_5 (for lipophilicity), HBA_0 (for H-bonding acceptor ability), and HBD_0 (for H-bonding donor ability). For the data pre-processing convenient linear interval compression transforms were applied. A small data set grouping the same information for seven additional endpoints from chemicals categorized as industrials (and containing the same atoms and functional groups found in the selected pesticides) was used to validate the training. Computed from the values reported in

Tables 1 and 2 in Ref. [25], the standard deviation of the errors produced by the model was 0.355 for the training set, and 0.328 for the test set.

The second study [26] targets the acute toxicity to the fish bluegill (*L. macrochirus*). The data incorporate 431 LC50 experimental values for 66 pesticides together with the associated bioassay information (fish weight, time of exposure, water temperature, pH, and hardness). The data set was then split into two subsets, a training set and a test set, with the former comprising 400 of the 431 available endpoints and the latter consisting of the remaining 31 endpoints. This model also uses a three-layer 13-6-1 BNN with the same input variables and same data pre-processing strategy as the previously discussed rainbow trout model. An additional data set of 15 endpoints from chemicals categorized as industrials was used to validate the training. Computed from the values reported in Tables 1 and 2 in Ref. [26], the standard deviations of errors generated by the model were 0.345 for the training set and 0.352 for the test set. Despite limitations in the data sets, both studies reveal that the NN model was able to take into account the influence of the experimental conditions of the toxicity results.

Classification of substances according to their toxic mode of action is one of the important topics of predictive toxicology. Basak et al. [27] report toxic mode of action classification experiments based on learning vector quantization (LVQ) classification networks. The target species was the fathead minnow fish *P. promelas*, and the data consisted on information on 60 molecular topological descriptors for 283 chemicals. The data set was divided into a training set of 220 substances and a test set containing the information on the remaining 63 compounds. Two modeling experiments were described. The first experiment (tier I analysis) focused on discriminating among uncouplers of oxidative phosphorylation, acetylcholinesterase (AChE) inhibitors, neurotoxicants, neurodepressants/respiratory blockers, and a combined group containing narcosis I (baseline narcosis), narcosis II (polar narcosis), and electrophile/proelectrophile reactive compounds. The model consisted of a 60-5-5 (input-Kohonen-output) LVQ classification NN. For the training set the (producer's) classification accuracy was: 98% for the combined narcosis I, narcosis II, and electrophile/proelectrophile class, 80% for uncouplers, 64% for AChE inhibitors, 57% for the neurotoxicants, and 67% for the respiratory blockers/neurodepressants. For the test set, the classification accuracy was: 98% for the combined narcosis I, narcosis II, and electrophile/proelectrophile class, 100% for uncouplers, 67% for AChE inhibitors, 50% for the neurotoxicants, and 0% for the respiratory blockers/neurodepressants. The second experiment (tier II analysis) attempted to discriminate between narcosis I, narcosis II, and electrophile/proelectrophile reactives. That model consisted of a 60-6-3 (input-Kohonen-output) LVQ classification NN. The (producer's) classification accuracy on the training set was 88% for the narcosis I group, 84% for the narcosis II group, and 66% for the electrophile/proelectrophile reactives. On the test set this accuracy was: 77% for the narcosis I group, 83% for the narcosis II group, 78% for the mixed narcosis I/II compounds, and 56% for the electrophile/proelectrophile reactives. These results may indicate the need for additional input-parameters, reflecting physico-chemical properties of the compounds.

It is well accepted that industrial and municipal wastewaters are major sources of contamination involving very complex mixtures of thousands of different types of chemicals. Both the USA and Canada use a standardized rainbow trout *O. mykiss* acute lethality bioassay [28] to assess the toxic properties of such mixtures. The cost associated with repeating such experiments on a regular basis is considerable. Gagné and Blaise [29] investigated the possibility of using a combination of less expensive 5–15 min incubation time chemoluminescent peroxidase (Cl-Per) and Microtox™ tests instead. The first test can detect radical scavengers and enzyme-inhibiting substances. The second test interprets the reduction of light emission by *V. fischeri* bacteria during the exposure as a measure of toxicity. Neural networks were used to model the relationship between the toxicity assessments generated by the rainbow trout bioassay (96 h) and the assessments based on a combination of Cl-Per and Microtox™ tests. The data used are toxic threshold concentrations detected by Cl-Per and Microtox™ tests and the corresponding lethality concentrations for the rainbow trout in 20 wastewater samples. Two three-layer 2-3-1 BNNs were built using as training set the data associated to 10 randomly selected effluent samples

from the 20 available. The rest of the data was used for test purposes. The first model targeted the actual toxic concentration for trout while the second was designed to differentiate between toxic and non-toxic effluents in 96 h exposure tests. The performance of the networks was measured against the following criterion: the number of correct predicted values within ±20% of the real value. On the test set, the BNN for toxic concentrations was successful in 60% of the cases. The predictive accuracy increased to 90% for the model discriminating between toxic and non-toxic samples. Similar accuracy was provided by a Boltzmann machine type NN investigating the relationship between the toxicity of the effluent as identified by the three toxicity assays. The results suggest that the combination of Cl-Per and Microtox™ tests may provide a less expensive alternative for the rainbow trout test in assessment studies targeting effluent toxicity.

2.2. Neural network models for toxicity to aquatic macro-invertebrates

Macro-invertebrates represent another important group of organisms used for the purpose of investigating the adverse effects of chemicals. Neural network QSAR models for aquatic invertebrates target mainly two species: the midge *Chironomus riparius*, and the water flea *Daphnia magna*.

A QSAR neural network model for the toxicity of ten organophosphorus insecticides to the midge larvae is presented in Ref. [30]. The data set consists of 180 measured EC50 24 h toxicity endpoints expressed in log(l/μmol), together with experimental information on sediment presence and water temperature and pH. The data set was divided into a training set of 164 records and a test set of 16 records. The identified model was a 10-5-1 BNN. Seven of the ten input parameters in the model were Boolean descriptors reflecting the bioassay conditions (temperature: 11, 18, or 25 °C; pH value: 6, 7, or 8; and the presence or absence of sediment). The remaining three descriptors were the components H_0, H_2 and H_7 of the lipophilicity autocorrelation vector H computed from the hydrogen-suppressed graph of the molecule. The data pre-processing consisted of a classical min/max transformation, and the training was stopped after approximately 500 learning cycles. Computed from the values reported in Tables 1 and 2 in Ref. [30] the standard deviation of the errors produced by the model was 0.104 for the training set, and 0.189 for the test set.

Various PNN models for the toxicity to *D. magna* are presented in Ref. [31]. The data consist of 48 h LC50 acute toxicity values for a highly diverse set of 776 organic chemicals, expressed in log(l/mmol). A training set was assembled by a random selection of 700 of the available 776 compounds with the remaining 76 substances forming the external test set. Two complete 20% leave-out cross-validation experiments based on random selection are reported. The first targeted exclusively the training set, while the second was performed on all the data. The input variables for all models consists of 40 molecular fragment descriptors associated with the presence or counts of molecular fragments (aromatic rings, acid groups, nitro groups, cyano groups, ether linkages, etc.), counts of atoms (C, H, Br, Cl, F, Fe, Hg, I, Mn, N, Na, O, P, S, Si, Sn and Zn) and the molecular weight. Four models based on the PNN paradigm were built. The first is represented by a basic PNN with Gaussian kernel trained on the whole training set, training correction included. The other three models are various combinations of the five basic PNNs with Gaussian kernel built as part of the 20% leave-out cross-validation experiment performed on the same 700 compounds set. The data pre-processing consisted of combinations of *Z*-transforms and hyperbolic tangent function. All four models were validated through external validation using the 76 compounds external test set. We limit here our discussion to the first model only. On the 700 compound training set, the Pearson's correlation coefficient between measured and values predicted by this model was 0.875, with an average error of practically zero, a standard deviation of errors of 0.558, and an average square error of 0.312. On the 76 compounds external test set the Pearson's correlation coefficient between measured and values predicted by the model was 0.764, with an average error of 0.061, standard deviation of errors of 0.667, and average square error of 0.449. For comparison, the US EPA ECOSAR toxicity assessment expert system based on Hansch-type relationships and rules of thumb was unable to handle three of the compounds in the external test set. For the remaining 73 compounds, the Pearson's

correlation coefficient between measured and ECOSAR predicted values was 0.322, with an average error of −0.442, standard deviation of errors of 1.362, and average square error of 2.051. At least in this example, the superior performance of the NN model is obvious.

2.3. Neural network models for toxicity to protozoa

Among the aquatic unicellular ciliated protozoa, *Tetrahymena* species are very common, have a very short generation time, and are non-pathogenic. They may be easily cultivated with inexpensive media and ambient conditions, and consequently they represent very attractive alternatives for fish or macro-invertebrates in laboratory experiments targeting the environmental impact of man-made or naturally occurring toxicants.

A first attempt to investigate the 48 h IC50 sublethal (growth inhibitory) toxicity to *T. pyriformis* using NNs is presented in Ref. [32]. All models are based on a data set containing information on 825 highly diverse organic chemicals from which 750 randomly selected compounds were used as training set for the modeling exercise and the remaining 75 compounds as external test set. A complete leave-20%-out cross-validation experiment based on random selection was performed on the 750 compounds set. Four models based on basic PNNs with Gaussian kernels were constructed. The first consists of a basic PNN with Gaussian kernel trained on the whole training set, training corrections included. The next three models involve various combinations of the five PNNs identified as part of the cross-validation experiment. All models were validated using the external test set. The input to all models consists of 32 molecular descriptors for the presence or number of occurrences of fragments of interest (aromatic rings, acid groups, ester groups, amino groups, sulfide bridges, etc.), the molecular weight, and the number of individual occurrences of C, H, Br, Cl, F, I, N, O and S atoms. The output was the 48 h IC50 toxicity endpoint expressed in log(l/mmol). Data pre-processing was handled through combinations of *Z*-transforms and hyperbolic tangent function. We limit here our discussion to the first model only. For the 750 compound training set, the Pearson's correlation coefficient between measured and values predicted by this model was 0.936, with an average error of practically zero, a standard deviation of errors of 0.264, and an average square error of 0.070. On the 75 compounds external test set the Pearson's correlation coefficient between measured and values predicted by the model was 0.882, with an average error of 0.045, standard deviation of errors of 0.306, and average square error of 0.096.

A larger and by far more diverse data set containing information on 1084 compounds is analyzed in Ref. [33]. Using random selection, the data were split into a 1000 compound training set and an 84 compound external test set. The basic PNN with Gaussian kernel (training corrections included) was used to generate the model. The input was expanded to 42 molecular descriptors including presence and counts of fragments of interest (various types of rings, various functional groups, longest aliphatic chain, ether linkages, quinone character, etc.), the molecular weight, and the number of individual occurrences of the C, H, Br, Cl, F, I, N, Na, O and S atoms. The output was the 48 h IC50 toxicity endpoint expressed in log(l/mmol). The data pre-processing strategy was the same as used in Ref. [32]. For the 1000 compound training set, the Pearson's correlation coefficient between measured and predicted values was 0.899, with an average error of practically zero, standard deviation of errors of 0.323, and average square error of 0.104. On the 84 compounds external test set, the Pearson's correlation coefficient between measured and predicted values was 0.803, with an average error of 0.022, standard deviation of errors of 0.441, and average square error of 0.195.

A completely different type of relationship between the molecular structure and the 48 h IC50 toxicity endpoint to *T. pyriformis* was targeted in Ref. [34]. A data set of 448 aromatic compounds was used to build the models, while an external test set of an additional 52 compounds was used to validate the final model. These data are a subset of the data sets used in Refs. [32,33]. The 448 compound set was split by random selection into a 287 compounds CNN training set, a 81 compound CNN training validation set, and a 80 compound GA/CNN evaluation set. For each of the 448 substances a number of approximately 200 molecular structure descriptors were computed

based on topological and electronic properties of the molecules. Based on various criteria, this number was reduced to approximately 60 descriptors. The reduced pool of descriptors was then submitted to a GA feature selection routine incorporating a three-layer 11-5-1 CNN for fitness evaluation. The final model was the result of averaging the 10 best models identified by the GA/CNN combination. The input parameters in the model are: (p1) number of path clusters of length 6 with at least one branch point; (p2) number of path chains of length 7 with at least one atom in a ring system; (p3) number of double bonds; (p4) number of sp^2 hybridized carbons bonded to three other carbons; (p5) number of sp^2 hybridized carbons bonded to two other carbons; (p6) sum of E-state values over all hetero atoms; (p7) square root of the gravitational index over all hetero atoms; (p8) difference in partial positive and partial negative molecular charges; (p9) sum of the surface area charge product of the proton acceptor atoms divided by the number of proton acceptor atoms; (p10) sum of the charges on all donatable protons; and (p11) count of proton acceptor atoms. The output is the 48 h IC50 endpoint expressed in log(l/mmol). For the 52 compound external test set, the Pearson's correlation coefficient between the measured and model predicted values was 0.55, with an average error of 0.033, standard deviation of errors of 0.599 and average square error of 0.353.

A comparatively small set of the toxicity of 278 substituted benzenes towards *T. pyriformis* was analyzed by Burden and Winkler [35] with both the PLS and a Bayesian regularized neural network (BRANN) method, using a leave-20%-out cross-validation. The independent parameters for that study were solely derived from the molecular structure of the compounds. The results showed the superiority of the BRANN method. The authors also concluded that the NN methodology appears to be able to model more diverse chemical classes and more than one mechanism of toxicity.

2.4. Neural network models for toxicity to aquatic bacteria

Bacteria are the smallest test organisms used in conducting experiments assessing the toxic effects of chemicals on the environment. For practical reasons, including the very low cost, the luminescent bacteria *V. fischeri* (formerly known as *Photobacterium phosphoreum*) is by far the most popular choice. The most used endpoint of interest is the effective concentration EC50 inducing a 50% reduction of the light emission of the test bacteria in a given time, typically 5, 15, or 30 min.

A first attempt to model the EC50 for *V. fischeri* using NNs is reported in Ref. [36]. A data set consisting of 30 min EC50 endpoints for 604 compounds was used to build the model. A training validation set consisting of 150 representative compounds was isolated from the 604 chemicals data set, while the remaining 454 compounds formed the training set. An external test set consisting in EC50 noisy values for 143 chemicals for which the toxicity was measured after an exposure period different from 30 min has been assembled. For each chemical in the training set, the first 10 components of autocorrelation hydrophobicity and molar refractivity vectors were computed from the hydrogen-suppressed graph of the molecule. The first five components identified by stepwise regression analysis were used as input for a three-layer 5-10-1 BNN. The model output was the EC50 expressed in log(l/mmol). Data pre-processing consisted of a classical min/max interval linear compression into [0.05, 0.95]. The average square error generated by the model was 0.09 for the training set, 0.116 for the training validation set, and 0.212 for the external test set.

A larger and more diverse data set, containing information on 1308 compounds, is analyzed in Ref. [37]. A test subset of 240 compounds was identified using the N2M method [38]. A training set consisting of 2795 various test duration (mostly 30 min) EC50 endpoints for the remaining 1068 chemicals was assembled. A similar set consisting of 385 EC50 endpoints for the 240 selected compounds was used as test set. The final model was a three-layer 36-26-1 BNN. The 36 input parameters consisted of: exposure time, the first 15 components of the autocorrelation vector *H* encoding lipophilicity, the first 15 components of the autocorrelation vector MR representing molar refractivity, the first four components of the vector HBA encoding H-bonding acceptor ability, and the first component of the autocorrelation vector HBD encoding the H-bonding donor ability, all calculated from the hydrogen-suppressed graph of the molecule. The output was

Table 1
Statistical data for the major neural network studies on aquatic organisms

Species[a]	Number of compounds			RMSE[b], Test	r^2, Test	SE, Test	Method	Ref.
	Total	Training	Test					
FHM	865	865 × 0.80	865 × 0.20		0.76		PNN	[21]
FHM	419	419 × 0.80	419 × 0.20				PNN	[17]
FHM	375	287	44	0.74			CNN	[24]
FHM	375	287	88	0.75			MLR	[24]
RBT	447	384	63				BNN	[25]
BGL	431	400	31				BNN	[26]
DM	776	700	76		0.76	0.67	PNN	[31]
TEHY	825	825 × 0.80	825 × 0.20		0.88		PNN	[32]
TEHY	1084	1000	84		0.80	0.31	PNN	[33]
TEHY	448	287	80	0.34			CNN	[34]
TEHY	448	287	52	0.59			CNN	[34]
VF	1308	1068	240	0.36			BNN	[37]
VF[c]	1238	1238 × 0.80	1238 × 0.20	0.76		0.61	PNN	[39]
VF[d]	1238	1238 × 0.80	1238 × 0.20	0.79		0.63	PNN	[39]

[a] Abbreviations used: FHM, fathead minnow; RBT, rainbow trout; BGL, bluegill sunfish; DM, *Daphnia magna*; TEHY, *Tetrahymena pyriformis*, VF, *Vibrio fischeri*.
[b] RMSE: Root mean square error.
[c] Including solubility as independent parameter.
[d] Excluding solubility as independent parameter.

the EC50 endpoint expressed in log(l/mmol). Classical min/max linear interval compression was used for data pre-processing. The training involved approximately 5000 cycles. The average square error produced by the model was 0.133 for the 2795 endpoints training set and 0.151 for the 385 endpoints test set.

Three PNNs with Gaussian kernel QSAR models for the EC50 endpoint to *V. fischeri* built on a 419 compound data set are presented in Ref. [20]. Where available, 30 min test values were used preferentially, and 15 or 5 min were used instead for compounds where no 30 min values were available. The validation was handled through a 20% leave-out cross-validation using random selection, and training corrections were included in the models. The models input consisted of log K_{ow}, the logarithm of the molecular weight, and another 48 molecular descriptors including the presence and/or counts of numbers of atoms and/or fragments of interest (aromatic rings, acid groups, ester groups, amino groups, etc.). The difference between the three models was the choice of data pre-processing strategy. For example, for the case of data pre-processing consisting of classical linear min/max compression functions, on the given 419 compounds set, the correlation between measured and values produced by the corresponding model was 0.919, with an average error of −0.008, the standard deviation of errors of 0.360 and the average square error of 0.129. Similar results were obtained for data pre-processing consisting of combinations of *Z*-transforms with the sigmoid logistic, respectively, hyperbolic tangent function.

In 1998, another PNN-based model for *V. fischeri* was presented, using a much larger set of 1238 compounds [39]. The cross validation experiment was performed with five random leave-20%-out data sets that resulted in very similar submodels. This work also explored the effect of aqueous solubility on the resulting predictions. Including the experimental aqueous solubility with the structural fragments used as independent parameters gave a standard deviation of 0.614, excluding it, resulted in a model with the standard deviation of 0.628.

3. Conclusions

A simple analysis of the performance of the discussed models confirms the ability of NNs to

model large sets of chemicals of different complexity and mode of action. In contrast to the traditional, highly focused and linear QSAR methods, NNs are able to handle non-linear relationships. This ability is of particular importance in the environmental field, where tens of thousands of complex substances have to be assessed in relatively short time. In fact, many of the compounds listed on the DSL are not amenable to prediction with the presently available linear model programs. Furthermore, NNs do not require the use of subjective rules of thumb, the only required information is the molecular structure and (in some cases) other variables of the bioassay conditions. This makes these NN-type QSARs the perfect tool for conducting toxicity assessment studies for very diverse lists of chemicals. Table 1 gives an overview of the statistics of the major models described here.

Acknowledgements

I thank Dr S.P. Niculescu for contributions to this work.

References

[1] S. Jensen, AMBIO 1 (1972) 123.
[2] Canada, http://cbsc.org.
[3] US EPA, 2001,http://www.epa.gov/grtlakes/bns/pcb/steppcb.html.
[4] H. Fiedler, Regulations and Management of PCB in Germany, http://www.chem.unep.ch/pops/stpeter/stpete2b.html.
[5] OECD, OECD Guideline 202, part 1, *Daphnia* sp., Acute Immobilisation Test.
[6] Canadian Environmental Protection Act, 1999, http://laws.justice.gc.ca/en/C-15.31/SOR-94-260/67704.html.
[7] J.M. Sutter, J.H. Kalivas, P.M. Lang, J. Chemom. 6 (1992) 217.
[8] Y.H. Zhao, M.T.D. Cronin, J.C. Dearden, Quant. Struct.-Act. Relat. 17 (1998) 131.
[9] US EPA, 2000, http://www.epa/gov/oppt/newchems/21ecosar.htm.
[10] K.L.E. Kaiser, J.C. Dearden, W. Klein, T.W. Schultz, Water Qual. Res. J. Can. 34 (1999) 179.
[11] J. Devillers, Modélisation en Écotoxicologie, Club CRIN Environnement et société, ECRIN, Paris, 2000, p. 181.
[12] US EPA, 2002, http://www.epa.gov/med/databases/aster_example.html.
[13] J. Zupan, J. Gasteiger, Neural Networks in Chemistry and Drug Design, second ed., Wiley, Weinheim, 1999.
[14] J. Devillers (Eds.), Neural Networks in QSAR and Drug Design, Academic Press, San Diego, 1996.
[15] J. Devillers (Eds.), Genetic Algorithms in Molecular Modeling, Academic Press, San Diego, 1996.
[16] J. Gasteiger, 2001, http://www2.chemie.uni-erlangen.de/presentations/acs_010401/vortrag-31.html#TopOfPage.
[17] K.L.E. Kaiser, S.P. Niculescu, G. Schüürmann, Water Qual. Res. J. Can. 32 (1997) 637 See also p. 855.
[18] T. Masters, Practical Neural network Recipes in C++, Academic Press, San Diego, 1993.
[19] K.L.E. Kaiser, S.P. Niculescu, M.B. McKinnon, in: F. Chen, G. Schüürmann (Eds.), Quantitative Structure–Activity Relationships in Environmental Sciences—VII, SETAC Press, Pensacola, FL, 1997, p. 285.
[20] S.P. Niculescu, K.L.E. Kaiser, G. Schüürmann, Water Qual. Res. J. Can. 33 (1998) 153.
[21] K.L.E. Kaiser, S.P. Niculescu, Chemosphere 38 (1999) 3237.
[22] K.L.E. Kaiser, S.P. Niculescu, National Water Research Institute, Burlington, Canada, Contribution AEP-TN99-001, 1999, p. 39.
[23] D.R.J. Moore, R.I. Breton, D. McDonald, P.W. Harris, R. Chenier, R. Sutcliffe, J. Sanderson, S. Bradbury, C. Russom, P.C. Jurs, SETAC 20th Annual Meeting, Philadelphia, Abstract Book, Abstract 260, SETAC Press, Pensacola, FL, 1999.
[24] D.V. Eldred, C.L. Weikel, P.C. Jurs, K.L.E. Kaiser, Chem. Res. Toxicol. 12 (1999) 670.
[25] J. Devillers, J. Flatin, SAR QSAR Environ. Res. 11 (2000) 25.
[26] J. Devillers, SAR QSAR Environ. Res. 11 (2001) 397.
[27] S.C. Basak, G. Grunwald, G.E. Host, G.J. Niemi, S.P. Bradbury, Environ. Toxicol. Chem. 17 (1998) 1056.
[28] H.L. Bergman, R.A. Kimerle, A.W. Maki, Environmental Hazard Assessment of Effluents, Pergamon Press, New York, 1986.
[29] F. Gagné, C. Blaise, Chemosphere 35 (1997) 1343.
[30] J. Devillers, Toxicol. Meth. 10 (2000) 69.
[31] K.L.E. Kaiser, S.P. Niculescu, Environ. Toxicol. Chem. 20 (2001) 420.
[32] S.P. Niculescu, K.L.E. Kaiser, T.W. Schultz, Arch. Environ. Toxicol. Chem. 39 (2000) 289.
[33] K.L.E. Kaiser, S.P. Niculescu, T.W. Schultz, SAR QSAR Environ. Res. 13 (2002) 57.
[34] J.R. Serra, P.C. Jurs, K.L.E. Kaiser, Chem. Res. Toxicol. 14 (2001) 1535.
[35] F.R. Burden, D.A. Winkler, Chem. Res. Toxicol. 13 (2000) 436.
[36] J. Devillers, S. Bintein, D. Domine, W. Karcher, SAR QSAR Environ. Res. 4 (1995) 29.
[37] J. Devillers, D. Domine, SAR QSAR Environ. Res. 10 (1999) 61.
[38] D. Domine, J. Devillers, D. Wienke, L. Buydens, Quant. Struct.-Act. Relat. 15 (1996) 395.
[39] K.L.E. Kaiser, S.P. Niculescu, in: J.D. Walker (Ed.), Proceedings of the Eighth International Workshop on QSAR in Environmental Sciences, Baltimore, MD, May 1998, SETAC Press, Pensacola, FL, 2002, in press.