Neural networks for effect prediction in environmental and health issues using large datasets

Klaus L. E. Kaiser*

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

Abstract

Neural network methodologies allow the modeling of non-linear relationships. This makes them useful tools for the analysis of larger data sets of non-congeneric compounds with unknown or varying modes of action. This brief review describes recent advances and their applications to sets of several hundred to over 1 000 compounds, modeling acute toxicity data for several aquatic species, including fish, ciliate, bacteria, and non-acute toxicity data for a mammalian species endpoint, i.e. estrogen receptor binding assay data.

1 Introduction

The Domestic Substances List (DSL) tabulates approximately 25 000 substances which are in current use in Canada. An additional 50 000 substances are on the Non-Domestic Substances List, comprising those compounds which are in use either below the threshold limit of 100 kg/year or only used elsewhere. Worldwide, the estimate of substances in use is in excess of 100 000. Based on the recognition of bioaccumulation, toxicity, and persistence in the environment as the three most important properties determining the pathways and effects of chemicals in the environment, countries belonging to the Organization for Economic Cooperation and Development (OECD) have begun to undertake research and regulatory initiatives to collate information on and to assess the potential harmful effects of these substances on the environment and human health.

One critical issue in this endeavor is the lack of physico-chemical and toxicological data for many, in fact most of these substances. Measurements of these properties are both time consuming and costly and avenues to reliably estimate such properties are required. Since the mid-20th century, the field of quantitative structure-activity relationships (QSARs) has developed into a highly useful science, particularly for the development of new and potent pharmaceuticals and the optimization of desirable effects by QSAR-driven variations of the basic chemical structures. In contrast, in the environmental arena, the application of QSAR was adopted much later in the 1980's, when Rachel Carson's *Silent Spring* lead to a widespread recognition for the need to research and regulate this field.

Numerous mathematical models have been developed to aid in the analysis of effect data in relation to the structures of the chemical agents involved and many models are in use to predict hundreds of different types of effects, from acute lethality to chronic and sub-lethal effects, to effect threshold concentrations. In addition, a variety of physico-chemical properties, such as hydrophilicity and hydrophobicity parameters, ionization constants, solubility, melting and boiling points, and other molecular properties are being estimated by an increasing number of computer programs and with increasing accuracy. Until recently, successful QSARs, using linear modeling techniques, were more or less predicated on small data sets of chemicals with an uniform mode of action, and congeneric chemical frameworks.

The development of non-linear technologies, artificial intelligence-based algorithms opened the field to the concurrent analysis of a wider variety of structures with (potentially) varying modes of action and non-congeneric chemicals. This possibility is especially useful in the effect modeling of the wide variety of substances entering the environment from human activity and enterprise. This brief review will look at some of the issues, results and recent attempts in this field.

* To receive all correspondence. Dr. K. L. E. Kaiser, TerraBase Inc., 1063 King St. West, Suite 130, Hamilton, Ontario L8S 4S3, Canada, phone: 905-802-0154, fax: 905-527-0263, email: Klaus < mail@terrabase-inc.com >

Key words: neural networks, toxicity, fish, *Daphnia, Tetrahymena, Vibrio,* steroid receptor binding

Abbreviations: ANN, Artificial Neural Network; BNN, Back-propagation Neural Network; BRANN, Bayesian Regularized Artificial Neural Network; PNN, Probabilistic Neural Network; DSL, Domestic Substances List; FHM, fathead minnow; RBA, relative binding affinity

  0931-8771/03/010■-0001 $ 17.50+.50/0

2 Acute toxicity of chemicals to aquatic organisms

Over the last two decades, three substantial toxicity data sets were developed in this field; they are (i) the fathead minnow lethal concentration (96-hr LC50) data for approximately 900 substances, measured largely by the U.S. Environmental Protection Agency's Research Laboratory at Duluth, MN, and available through AQUIRE [1]; (ii) the 40 to 72-hr growth-inhibitory (ICG50) values for the ciliate *Tetrahymena pyriformis*, largely determined at the University of Tennessee, Knoxville, TN, published values in the literature encompass approximately 1 200 substances, available from T.W. Schultz; and (iii) the 5 to 30-min effective concentration (EC50) values for the luminescent marine bacterium *Vibrio fischeri* (formerly known as *Photobacterium phosphoreum*), covering approximately 2 200 substances, available in substructure-searchable electronic format from TerraBase Inc. [2]. In addition, smaller data sets, covering from one hundred to several hundred substances are available for the acute toxicities of individual chemicals to several fish species, including the rainbow trout (*Oncorhynchus mykiss*), the zebrafish (*Brachydanio rerio*), the red killifish (*Oryzias latipes*), the guppy (*Poecilia reticulata*), the goldorfe (*Leuciscus idus melanotus),* the goldfish (*Carassius auratus*), the channel catfish (*Ictalurus punctatus*), and the bluegill *(Lepomis macrochirus*), as well as the to waterflea (*Daphnia magna)* and the algae *Chlorella* sp. These data are also available from both commercial and non-commercial sources, e.g. [1, 2].

2a QSARs of fathead minnow LC50 data

Numerous reports can be found evaluating various subsets of the fathead minnow (FHM) 96-hr LC50 data. In addition to the models proposed in the literature, the computer program ECOSAR [3], available from U.S. government web sites and others, which derives 96-hr LC50 FHM estimates from measured or calculated octanol/water (log K_{ow}, or logP) partition coefficients by applying these to over one hundred individual linear regressions. However, many of the underlying equations in ECOSAR lack any degree of statistical significance (as many are based on one or two substances only) and are further compromised by the need to know the use of the substance in question, as pointed out in [4]. Due to the variety of structures of the chemicals in the FHM data set, a number of different modes of action have been identified. Some modes of action, apparent from the behavior of the fish upon exposure, are also known as fish acute toxicity syndromes and have been predicted on the basis of the chemicals' structures [5].

With the increasing availability of neural network methods, several studies have been published on the entire FHM data set or large sub sets, using these methods. Without going into details of these reports, the following should be mentioned. One of the first investigations on the use of neural network methodology with aquatic toxicity data was the application of the feed forward back-propagation network (BPN) to approximately 400 FHM data by Kaiser *et al.* [6]. Jurs and coworkers [7] studied a subset of that, comprising 375 substances with an artificial neural network (ANN) method. Kaiser and Niculescu [8] used a probabilistic neural network (PNN) to analyze essentially the full set of 865 FHM values for organic substances. These models were sufficiently successful to be used for the quantitative estimation of FHM values for a large number of DSL substances. Furthermore, a very detailed study employing a variety of statistical measures by Moore *et al.* [9], and using an external test set of 130 substances derived from other sources, showed that the PNN results were superior in almost all aspects when compared to the other models' predictions, including those of ECOSAR [3], ASTER [10], and TOPKAT [11]. These results clearly demonstrate both the general ability of neural network methods to be capable of modeling such diverse data sets and the specific ability of the PNN in doing so. Moreover, only the ECOSAR and PNN methods were able to provide estimates for all compounds in the testing data set.

Quite recently, TerraBase Inc. [2] released a commercially available stand-alone fathead minnow 96-hr LC50 toxicity estimation program under the name TerraQSAR™-FHM, which is also based on the PNN methodology. It has not yet been part of any comparative study, such as the one by Moore *et al.* [9], however a graph showing the estimated and measured values for the training data set is provided in the manual of that program and indicates a high degree of correlation.

2b QSARs of Daphnia magna LC50 data

The waterflea, *Daphnia magna*, is used worldwide as an aquatic test organism. Tests are usually performed in static systems of several liters size and a recommended test protocol has been developed (OECD Test Guideline 202, Part I, under revision) by the Organization for Economic Cooperation and Development (OECD). This test is also a requirement for the assessment of environmental fate and effects of high-production-volume chemicals in OECD countries. Kaiser and Niculescu [12] published the first large-data-set analysis of the acute (48-hr LC50) toxicity of over 700 compounds of *Daphnia magna* using the PNN approach. *Daphnia magna* acute toxicity values are also being estimated by several computer programs, including ECOSAR, ASTER and TOPKAT.

The PNN model [12] of the acute toxicity of 700+ chemicals to *Daphnia m.* was fully cross-validated using a 20% -leave-out procedure, i.e. using five random subsets of 80% each of the training set data. The resulting five models showed very similar statistics, indicating the validity of the main model. This was further tested by applying the main model to an external test set of 76 compounds. The values predicted for the external test set were in good agreement with the experimental data. In comparison to the estimates produced by ECOSAR for the same external test data, the PNN model estimates were found to be much

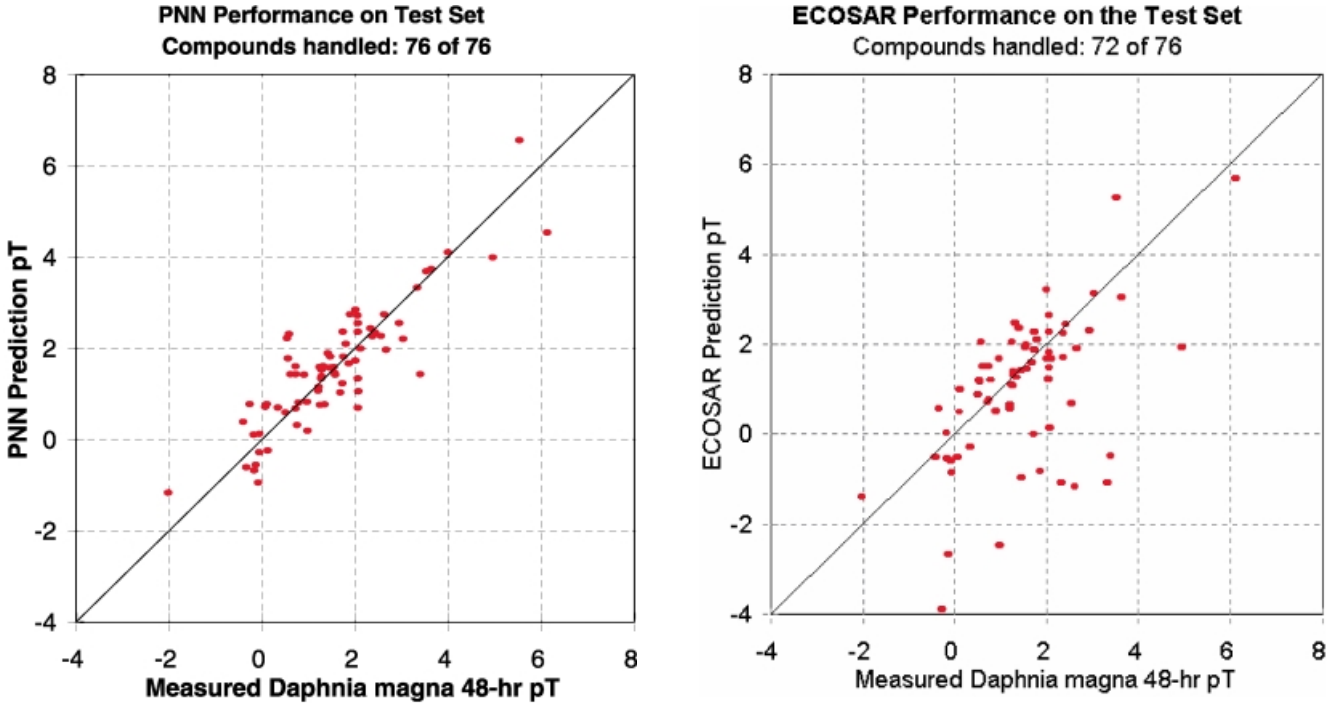

Figure 1. Measured vs. predicted 48-hr LC50 data for Daphnia magna for an external testing set of 76 compounds; (A) prediction results from a probabilistic neural network (PNN, 76 compounds predicted), (B) prediction results from ECOSAR (72 compounds predicted). All values in log(L/mmol). Reproduced with permission from *Environmental Toxicology and Chemistry 20, 420–431* (2001), copyright SETAC, Pensacola, Florida, USA.

superior [12]. These plots of measured vs. predicted values for the 76 compounds of the external test set obtained from both the PNN and ECOSAR models are reproduced in Figure 1; the Pearson's r^2 values for the test sets are 0.76 (PNN) and 0.32 (ECOSAR), respectively.

2c QSARs of Tetrahymena ICG50 data

Over a period of two decades, the results for approximately 1200 individual test compounds have been published by T. W. Schultz, University of Tennessee, and coworkers for a 40 to 72-hr growth inhibition assay (ICG50) with the ciliate *Tetrahymena pyriformis*. These data have also been analyzed with QSAR methods in numerous publications by Schultz *et al.*, e.g. [13], but mostly in small sub sets, typically in the range of 20 to 30 substances each. Schultz and coworkers also identified several mechanistic principles and modes of toxic action of groups of substances. Many of these highly significant relationships, however, cover less variation in chemical structure and properties when compared with models for the much larger fish and *Vibrio* data sets and, therefore, can be expected to perform better on such selections than models dealing with larger sets and more diverse, non-congeneric structures.

More recently, larger sets of the *Tetrahymena* ICG50 data have also been modeled with several neural network methods. They include ANN study of approximately 400 compounds by Jurs and coworkers [14], a PNN study using 825 compounds by Niculescu *et al.* [15], and another PNN study with 1110 substances by Kaiser *et al.* [16]. Burden *et al.* [17] used a much smaller subset of these data, comprising 278 compounds, and a Bayesian Regularized Artificial Neural Network (BRANN) for their investigation of *Tetrahymena*. The results of the PNN work in [16] were cross-validated with a *leave-20%-out* procedure and an external test set of 75 compounds. All sub-models gave very similar results with Pearson's r^2 values of 0.88+ for the test sets. In study [17], comprising 1000 compounds for the training and 84 compounds for the test sets, the Pearson's r^2 values were 0.89 and 0.80, respectively. In the BRANN study with 278 compounds [17], using atomistic, topological, connectivity and fragment parameters as the molecular descriptors, very similar results were obtained for the test set of 56 compounds. In summary, all of the tested neural network methods showed good performance for both training and test sets.

2d QSARs of Vibrio fischeri EC50 data

The Microtox® test uses the marine luminescent bacterium *Vibrio fischeri* (formerly known as *Photobacterium phosphoreum*) and it became a commercially available standardized test, approximately two decades ago. Originally developed for quick testing of the toxicity of effluents, it was rapidly adopted by a number of groups to test individual

compounds for their light-diminishing effect on the bacterial light emission (EC50). Since the test's appearance, well over two thousand chemicals have so been tested and the data are available in monograph [18], as well as electronic, chemical structure-fragment-searchable formats [2]. Various subsets of the available EC50 values for approximately 2200 chemicals have been studied with a variety of QSAR methods, including neural network algorithms. Among the larger sets studied are those dealing with 604 chemicals by Devillers and coworkers [19], and another set of 1068 compounds, including values for three different exposure times using the BNN methodology [20]. Another study using 1200 compounds is using the PNN algorithm and structural fragment descriptors as independent parameters [21].

Without going into details of these studies, it can be generalized that the neural network methodologies applied to these large, non-congeneric data sets have proven successful in modeling the available data, typically spanning up to ten orders of magnitude in the compounds' range of toxic effects observed. Devillers [22] reviewed the use of neural networks for large heterogeneous sets of molecules, including the studies mentioned above, and concluded that "... these models can often compete favorably with classical regression models ...".

3 Sub-acute effects of chemicals to aquatic and non-aquatic organisms

There is a rapidly rising volume of literature on the application of neural network methods to specific non-lethality endpoints for many types of effects, enzymes and organisms. Except for physico-chemical properties, such as the octanol/water partition coefficient and aqueous solubility and certain non-specific effects, such as carcinogenicity and mutagenicity, the data sets in these fields rarely surpass 100 compounds at a time. Moreover, such data sets usually comprise compounds with limited structural variation of a base molecule. Many of these groups of data are generally well suited for analysis by more traditional QSAR methods, including classical QSAR, principal component analysis and related methods. Therefore, neural network methods have only begun to be employed with such data when the other methods failed to provide satisfactory results. The following will give just a few examples of such cases.

3a Steroid receptor binding affinities

There is a considerable interest in the assessment of the DSL substances vis-a-vis their potential endocrine disrupting

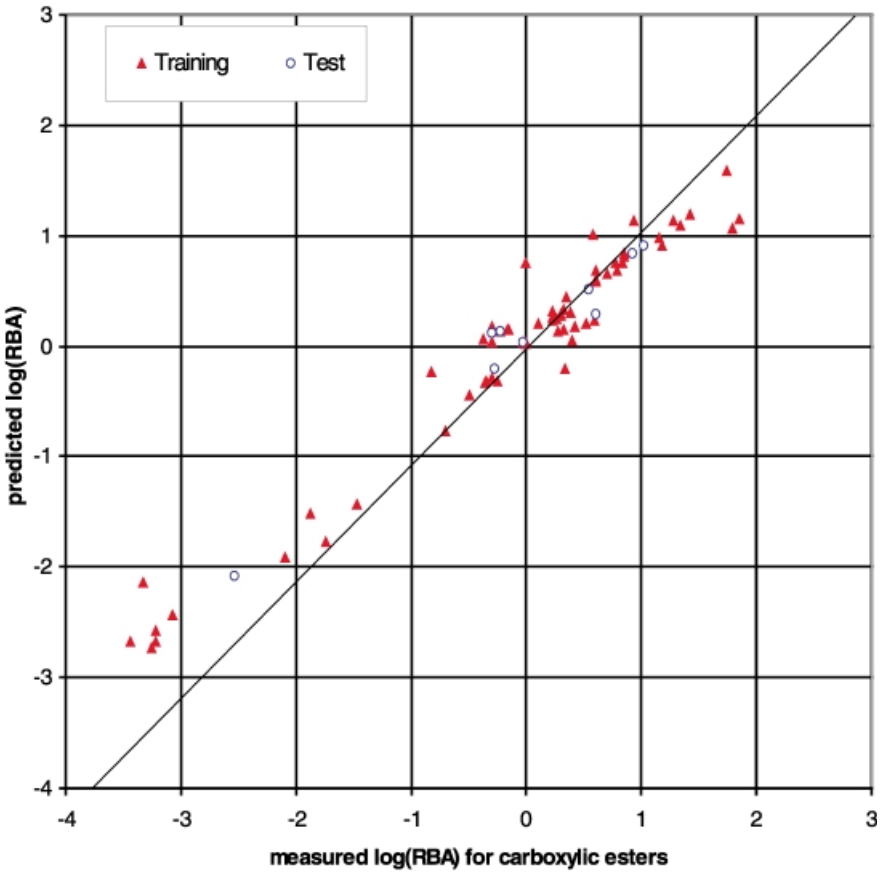

Figure 2. Measured vs. predicted estrogen receptor binding affinity (RBA) data, values in log(RBA), where RBA is the binding affinity relative to 17*beta*-estradiol (100%). Reproduced with permission from *Water Quality Research J. Canada 36*, 619–630 (2001), copyright CAWQ, Burlington, Ontario, Canada.

effects on organisms. The binding affinity of compounds to the hormone-receptor proteins is one possible endpoint related to that. Consequently, efficient models of the receptor binding of substances, and their agonistic and antagonistic effects are highly desirable. A comparatively small, but representative data set was developed and analyzed by Van Helden *et al.* [23] for binding to the progesterone receptor. The same group [24] also used a combination of genetic algorithms and neural networks. Later, Niculescu and Kaiser [25] used the same data with different neural network methodologies. Independent of the particular method applied, neural network models were found to perform better than any of the traditional QSAR methods.

Kaiser and Niculescu [26] also analyzed a data set of 1 000 compounds known to bind to the estrogen receptor complex. Specifically, the estrogen receptor binding affinity (RBA) values of compounds relative to 17*beta*-estradiol were used with a 20%-leave-out cross-validation and external test set of another 118 substances. Both sets comprise steroidal and non-steroidal structures with a large variety of substitutents and molecular sizes and cover approximately eight orders of magnitude in activity although, because of the underlying goal to create compounds with high binding affinity, the data available are somewhat skewed towards the log(RBA) range of 0 to +3, with relatively few values in the log(RBA) −5 to 0 range. While the overall model clearly shows promise, it is not yet refined enough to be of practical use. However, several sub-sets, defined by certain structural conditions, such as the presence of halogens, or carboxylic acid groups, or carboxylic acid ester groups, gave good correlations of predicted versus measured results, indicating that such sub-models maybe useful at this time in predicting estrogen receptor binding affinity for such chemicals. Figure 2 shows the model results filtered for the carboxylic ester moiety-containing compounds in this data set; this set comprises 75 and 9 compounds in the training and test sets, respectively, and this group has standard deviation errors of 0.28 (training) and 0.20 (test) and Pearson's r^2 of >0.94, respectively. While there are several studies in the literature on the prediction of steroid RBA values, they all use much smaller training sets and different assay selections than that used in [26]. Therefore, any comparison of the results between these studies is not possible.

3b Various effects and properties

There are numerous studies on the application of artificial intelligence methods, including neural networks to the modeling of a host of biological effects and physicochemical properties of compounds. They will not be reviewed here; suffice it to mention two important examples. Mammalian carcinogenicity and mutagenicity data sets of several hundred compounds and different routes of exposure and types of effect, have been analyzed by Bristol and coworkers [27]. Devillers *et al.* [28] used a BNN to investigate the biodegradability of organic chemicals.

4 Conclusions

Most studies with neural networks find that the versatility and modeling capabilities of neural networks lead to results that are at least at par, but frequently superior to those obtained with traditional QSAR methods. Particularly for data sets involving non-congeneric structures, such as many of the substances tested with acute toxicity bioassays in the studies mentioned here, non-linear neural networks usually perform much better by recognizing and adapting to non-linear relationships. Further advancements and, ultimately, reliable neural-network-based effect prediction programs can be expected to result from this research.

5 References

[1] U.S.Environmental Protection Agency, AQUIRE, http://www.epa.gov/ecotox, (2002).
[2] TerraBase Inc., http://www.terrabase-inc.com, (2002).
[3] U.S.Environmental Protection Agency, ECOSAR, http://www.epa.gov/oppt/newchems/21ecosar.htm, (2000).
[4] Kaiser, K. L. E., Dearden, J. C., Klein, W., and Schultz, T. W., A note of caution to users of ECOSAR, *Wat. Qual. Res. J. Canada 34,* 179–182 (1999).
[5] Russom, C. L., Bradbury, S. P., Broderius, S. J., and Hammermeister, D. E., Predicting modes of toxic action from chemical structure: acute toxicity in the fathead minnow (*Pimephales promelas*), *Environ. Toxicol. Chem. 16,* 948–967 (1997).
[6] Kaiser, K. L. E., Niculescu, S. P., and McKinnon, M. B., On simple linear regression, multiple linear regression, and elementary probabilistic neural network with Gaussian kernel's performance in modeling toxicity values to fathead minnow, in: Chen, F. and Schüürmann, G. (Ed.), *QSAR in Environmental Sciences -VII,* SETAC Press, Pensacola, FL. 1997, pp.285–297.
[7] Eldred, D. V., Weikel, C. L., Jurs, P. C., and Kaiser, K. L. E., Prediction of fathead minnow acute toxicity of organic compounds from molecular structure, *Chem. Res. Toxicol 12,* 670–678 (1999).
[8] Kaiser, K. L. E., and Niculescu, S. P., Using probabilistic neural networks to model the toxicity of chemicals to the fathead minnow (*Pimephales promelas*): a study based on 865 compounds, *Chemosphere 38,* 3237–3245 (1999).
[9] Moore, D. R. J., Breton, R. L., and MacDonald, D. B., A comparison of model performance for six QSAR packages that predict acute toxicity to fish, *Environ. Toxicol. Chem.*, in press.
[10] U.S.Environmental Protection Agency, ASTER, Assessment Tools for the Evaluation of Risk, http://www.epa.gov/med/databases/aster.htm (2002).
[11] Accelrys Inc., TOPKAT, http://www.accelrys.com (2002).
[12] Kaiser, K. L. E., and Niculescu, S. P., Modeling acute toxicity of chemicals to *Daphnia magna*: a probabilistic neural network approach, *Environ. Toxicol. Chem. 20,* 420–431 (2001).

[13] Schultz, T. W., and Cronin, M. T. D., Response-surface analyses for toxicity to *Tetrahymena pyriformis.* Reactive carbonyl-containing aliphatic chemicals, *J. Chem. Inf. Comput. Sci. 39,* 304 – 309 (1999).
[14] Serra, J. R., Jurs, P. C., and Kaiser, K. L. E., Linear regression and computational neural network prediction of *Tetrahymena* acute toxicity for aromatic compounds from molecular structure, *Chem. Res. Toxicol. 14,* 1535 – 1545 (2001).
[15] Niculescu, S. P., Kaiser, K. L. E., and Schultz, T. W., Modeling the toxicity of chemicals to *Tetrahymena pyriformis* using molecular fragment descriptors and probabilistic neural networks, *Arch. Environ. Contam. Toxicol. 39,* 289 – 298 (2000).
[16] Kaiser, K. L. E., Niculescu, S. P., and Schultz, T. W., Probabilistic neural network modeling of the toxicity of chemicals to *Tetrahymena pyriformis* with molecular fragment descriptors, *SAR QSAR Environ. Res. 13,* 57 – 67 (2002).
[17] Burden, F. R., Winkler, D. A., A QSAR model for the acute toxicity of substituted benzenes to *Tetrahymena pyriformis* using Bayesian-regularized neural networks, *Chem. Res. Toxicol. 13,* 436 – 440 (2000).
[18] Kaiser, K. L. E., and Devillers, J., 1994. *Ecotoxicity of Chemicals to Photobacterium phosphoreum*. Gordon and Breach Science Publishers, Reading, MA.
[19] Devillers, J., Bintein, S., Domine, D., and Karcher, W., A general QSAR model for predicting the toxicity of organic chemicals to luminescent bacteria (MicrotoxR test), *SAR QSAR Environ. Res. 4,* 29 – 38 (1995).
[20] Devillers, J., and Domine, D., A noncongeneric model for predicting toxicity of organic molecules to *Vibrio fischeri*, *SAR QSAR Environ. Res. 10,* 61 – 70 (1999).
[21] Kaiser, K. L. E., and Niculescu, S. P., Neural network modeling of *Vibrio fischeri* toxicity data with structural physicochemical parameters and molecular indicator variables, in: Walker, J. D. (Ed.), *Proc. 8th Intl. Workshop on QSAR in Environmental Sciences,* SETAC, Pensacola, FL. 2002, pp.16 – 20.
[22] Devillers, J., QSAR modeling of large heterogenous sets of molecules, *SAR QSAR Environ. Res. 12,* 515 – 528 (2001).
[23] Van Helden, S. P., Hamersma, H., and van Geerestein, V. J., Prediction of the progesterone receptor binding of steroids using a combination of genetic algorithms and neural networks, in: Devillers, J. (Ed.), *Genetic Algorithms in Molecular Modeling,* Academic Press, London. 1996, pp.159 – 192.
[24] So, S.-S., van Helden, S. P., van Geerestein, V. J., and Karplus, M., Quantitative structure-activity relationship studies of progesterone receptor binding steroids, *J. Chem. Inf. Comput. Sci. 40,* 762 – 772 (2000).
[25] Niculescu, S. P., and Kaiser, K. L. E., Modeling the relative binding affinity of steroids to the progesterone receptor with probabilistic neural networks, *Quant. Struct.-Act. Relat. 20,* 223 – 226 (2001).
[26] Kaiser, K. L. E., and Niculescu, S. P., On the PNN modelling of estrogen receptor binding data for carboxylic acid esters and organochlorine compounds, *Wat. Qual. Res. J. Canada 36,* 619 – 630 (2001).
[27] Bahler, D., Stone, B., Wellington, C., and Bristol, D. W., Symbolic, neural, and Bayesian machine learning models for predicting carcinogenicity of chemical compounds, *J. Chem. Inf. Comput. Sci. 40,* 906 – 914 (2000).
[28] Devillers, J., Domine, D., and Boethling, R. S., Use of a backpropagation neural network and autocorrelation descriptors for predicting the biodegradation of organic chemicals, in: Devillers, J. (Ed.), *Neural Networks in QSAR and Drug Design,* Academic Press, London. 1996, pp. 65 – 82.

Received on ; Accepted on November 7, 2002