Predicting Toxicity against the fathead Minnow by Adaptive Fuzzy Partition

Marco Pintore[a], Nadège Piclin[b,c], Emilio Benfenati[c], Giuseppina Gini[d], Jacques R. Chrétien[a,b]*

[a] BioChemics Consulting, Innovation Center, 16 rue Leonard de Vinci, 45074 Orléans, France
[b] Laboratory of Chemometrics & BioInformatics, University of Orléans, BP 6759, 45067 Orléans Cedex 2, France
[c] Istituto di Ricerche Farmacologiche Mario Negri, via Eritrea 62, 20157 Milano, Italy
[d] DEI, Politecnico di Milano, piazza Leonardo da Vinci 32, 20133 Milano, Italy

Abstract

Recent progress in the development of powerful tools suitable to design and to classify large chemical libraries can be fruitfully extended also to the ecotoxicity domain. Amidst these methods, Fuzzy Logic concepts, based on the possibility to handle the “concept of partial truth”, constitute interesting approaches to developing general predictive models.

In this work, a global strategy of Database Mining was applied on a data set of 568 chemicals, extracted from a toxicity database concerning the fathead minnow and divided into four classes, according to the toxicity ranges defined by the European Community legislation. Two large sets of molecular descriptors were tested on the 2D and 3D structures, and the best ones were selected with help of a procedure combining Genetic Algorithm concepts and stepwise method. After selecting the training set with a rational selection based on the Self Organizing Maps (SOM), structural-activity models were built by Adaptive Fuzzy Partition (AFP). This method consists in modeling relations between molecular descriptors and biological activities, by dynamically dividing the molecular descriptor hyperspace into a set of fuzzy subspaces. The best model was selected by a validation set, and its robustness was confirmed by predicting a test set of 80 chemicals never used to define the AFP models. An encouraging validation ratio of about 72% was obtained in the prediction of the experimental toxicity class. Furthermore, very similar results were obtained by using molecular descriptors computed on 2D or 3D structures.

1 Introduction

There are increasing needs to evaluate the effects of pollutants on the environment. This object requires also to analyze the high number of degradation and transformation compounds derived from the chemicals released into the environment, as well as the impurities present in the parent compounds [1]. Taking these products into account sensibly increases the already huge cost of experimentally assessing and developing, for example, a pesticide.

Furthermore, experimental tests are usually performed on animals and human toxicity is extrapolated from these results. In addition to ethical considerations associated to animal experimentation, most extrapolated models lack robustness and can be only applied to limited series of compounds.

An attractive alternative to laborious and expensive experimental studies consists in developing predictive tools based on computational methods [2–4]. These tools allow to evaluate a large number of compounds, for a range of toxicological end-points, automatically extracting new knowledge from all the information incorporated in the toxicity databases. The goal of the computational methods is to define relationships between biological activities and chemical structures, which can be represented in a numerical way by a large number of molecular descriptors including physicochemical, topological, quantum mechanical, constitutional and electronic parameters.

So far, up to 3000 descriptors have been exploited to build predictive models [5]. Efficient methods have to be used for selecting relevant parameters able to represent the ranges of toxicity levels. No general rule exists to define the best

* To receive all correspondence: Tel: (33)238417076, Fax: (33)238417221, E-mail: Jacques.Chretien@univ-orleans.

Key words: ecotoxicity, fathead minnow, fuzzy logic, SAR models

Abbreviations: GA, genetic algorithms; FL, fuzzy logic; AFP, adaptive fuzzy partition; SOM, self organizing map; SAR, structure – activity relationship; QSAR, quantitative structure – activity relationship; GA/SW, selection method combining genetic algorithms and the stepwise method

1611-020X/03/0204-0210 $ 17.50+.50/0

approach, the only essential condition consists in deriving a subset of descriptors from which it is possible to develop a generalist system. Several selection strategies are proposed in the literature, based, for example, on the stepwise method [6], principal component analysis (PCA) [7, 8], and Genetic Algorithms (GA) [9, 10]. The latter are probably the most powerful methods, as they are able to thoroughly explore the molecular descriptor hyperspace.

After selecting the most relevant descriptors to represent the biochemical activities in a toxicity database, the next step consists in defining the best computational method to develop robust predictive models. Standard Quantitative Structure – Activity Relationship (QSAR) approaches based, for example, on linear regression algorithms, such as Partial Least Squares [11] or non linear techniques, such as Back-Propagation Neural Networks [12], have proved their efficiency in several studies concerning ecotoxicity [8, 13 – 16]. But, although many international regulatory bodies recognize the potential benefits of QSAR techniques [17], they are scarcely used, as their real application in risk assessment problems is complicated by several factors [18]: i) the high variability in the experimental data, due to the wide range of biological answers; ii) the wide range of physiological and biochemical processes; iii) the lack of standardized protocols.

Besides methods based on regression algorithms that are able to predict exact values such as toxic doses, others techniques use discrimination algorithms to define the range of activities in which a given compound can be located. These classification methods allow to work on toxicity ranges that are directly linked to the requirements of the international regulations and, quite often, to derive more general models. In fact, the aim of the classification algorithms is not to fit all the activity values related to the training set compounds, but to find relations in the descriptor hyperspace able to separate different compound categories included in the data set. Then, the prediction results derived from classification methods, even if less precise, should be more general, as a category is more representative than isolated compounds [19].

Amidst these methods, Fuzzy Logic (FL) concepts [20] constitute interesting approaches to overcome the drawbacks related to a virtual screening of toxicity data sets. FL methods, based on the possibility to handle the "concept of partial truth", provide solutions to classification problems within the context of imprecise categories, in which toxicity can be included. The main ability of the fuzzy classification consists in representing the boundaries between neighboring activity classes as continuous, assigning to compounds a degree of membership of each class within a 0 to 1 range.

The aim of this work was then to apply a FL procedure, called Adaptive Fuzzy Partition (AFP) [21, 22], to a data set of 568 chemicals that are active against the fathead minnow (*Pimephales promelas*) and derived from the works of Brooke et al. [23 – 27]. This data set has already been studied by Russom et al. [28], which developed an expert system predicting the modes of action from the chemical structures. In this study, Russom et al. criticize the QSAR developed "on the assumption that compounds from the same chemical class should behave in a toxicology similar manner". In fact, the compounds included in different chemical classes can act through the same mode of action, and compounds within the same chemical class can show different modes of action.

In the proposed approach, a predictive model was derived by taking into account all chemical classes and modes of action. The data set of 568 compounds was divided into four intervals, according to toxicity ranges established by the European Community [29]. Two sets of molecular descriptors were analyzed, computed respectively on 2D and 3D structures, in order to test whether the 3D contribution sensibly increases the prediction power of the AFP models. The most relevant parameters were selected by a procedure derived from the Genetic Algorithm concepts combined with a stepwise technique [30]. Finally, the training set, on which the structure - activity models were built, was defined by an experimental design strategy based on the analysis of molecular diversity by Self Organizing Map (SOM) [31].

2 Materials and Methods

2.1 Compound selection

A data set of 568 compounds was derived from analyses of the chemicals in the fathead minnow acute toxicity database. A detailed description of the biological and chemical test protocols used in the study had been published [23 – 27]. Several chemical classes such as organophosphates, alkanes, ethers, alcohols, aldehydes, ketones, esters, amines and other nitrogen compounds, aromatic and sulfur compounds, and several modes of action, such as narcosis (I, II and III), oxidative phosphorylation uncoupling, respiratory inhibition, electrophile/proelectrophile reactivity, acetylcholinesterase (AChE) inhibition, and mechanisms of central nervous system (CNS) exposure are represented in this data set.

A ninety-six-hour lethal concentration killing 50% of the fathead minnow population (96h-LC50) was used to characterize toxicity. Four toxicity classes were generated according to the intervals established by the European Community legislation [29]: $LC50 < 1$ mg/l for class 1; 1 mg/l $< LC50 < 10$ mg/l for class 2; 10 mg/l $< LC50 < 100$ mg/l for class 3; $LC50 > 100$ mg/l for class 4.

The data set compounds were split in three sets: training, validation and test set. The test set includes molecules that were never used for developing the model. The validation set was used during the development of the model, based on the training set, to optimize the parameters and to validate the models.

2.2 Molecular descriptors

General molecular descriptors have proved a good compromise for data mining in large databases in terms of efficiency, as these parameters are able to take into account the main structural features of each molecule. Two sets of molecular descriptors were used to build the structure-activity models. The first one included 164 parameters computed on the 2D structures. Constitutional, information, topological, electrotopological, physicochemical, and electronic parameters were taken into account [32–35]. All the parameters were computed by ChemInter [36] and SciQSAR2D [37], excepting the lipophilicity descriptors that were calculated by Pallas [38]. The latter parameters were represented by the apparent octanol/water partition coefficient (logD), evaluated at various pH: 3, 5, 7.0, 7.4 and 9. More details about the different molecular descriptors used can be found in reference [39].

The second set of molecular descriptors, including 168 parameters, was computed on the 3D structures. Before calculating these parameters, all 3D molecular structures were optimized by a procedure exploiting tools included in the Hyperchem software [40] and subdivided in three steps consisting in: i) conformational analysis by using MM+ force field; ii) energy minimization of the lowest energy conformer by the steepest descent method; iii) after convergence of the minimization procedure, final optimization by the PM3 Hamiltonian. This preliminary steps were necessary to improve the generation of those descriptors depending on molecular geometry.

Most parameters were calculated by CODESSA 2.2.1 [41], particularly the constitutional, topological, geometrical, and electronic descriptors. Quantum-chemicals descriptors, i.e. total energy of the molecule, the energies of the lowest unoccupied and highest occupied molecular orbital (HOMO and LUMO), ionization potentials, heat of formation, etc. were computed by using the PM3 Hamiltonian. Finally, the same parameters derived in the 2D case by Pallas were used to represent the lipophilicity of the molecules at various pH.

2.3 Molecular descriptor selection

To select, amidst the two sets of molecular descriptors, the best parameters for classifying the data set compounds, a method based on GA concepts was used [42, 43]. GA, inspired by population genetics, is very effective for exploratory search, applicable to problems where little information is available, but it is not particularly suitable for local search. Then, a stepwise approach was combined with GA in order to reach local convergence [30–44], as it is quick and adapted to find solutions in "promising" areas already identified.

Finally, a specific index was derived from the fuzzy clustering method to evaluate the fitness function. This index has the advantage to be calculated quite quickly and to be able to estimate the descriptor relevance also by analyzing complex molecular distributions, in which finding separating edges between the different categories is difficult.

To prevent over-fitting and a poor generalization, a cross validation procedure was included in the algorithm during the selection procedure, randomly dividing the database into training and validation sets. The fitness score of each chromosome is derived from the combination of the scores of the training and validation sets.

The speed of convergence of this approach is sensibly dependent on the discriminant power of the initial set of descriptors. In fact, it has to be underlined that 95% of the processing time is devoted to compute the fitness scores whereas only 5% is devoted to the steps involving the genetic and stepwise procedures.

After completing the selection procedure, the 10 best fitness scores were evaluated to isolate the most relevant descriptors. The higher values were obviously favored, but when several chromosomes showed similar scores, i.e. with a variation within 2%, the descriptor set which included the lowest number of parameters was selected, in order to increase the possibilities to obtain a general classification model.

Finally, it has to be underlined that this selection method represents a pre-classification tool. A global method combining selection and classification steps could be also envisaged, by computing the fitness function, e.g., by AFP. But this approach would be very time consuming as regard to fitness score computation by fuzzy clustering and, therefore, it would not be suitable in data mining of large databases. Moreover, the evaluation of the best descriptor subset by the proposed procedure is "objective" and does not influence the efficiency of the successive data set classification by AFP.

More details about the strategy of molecular descriptor selection proposed and the proprietary software used can be found in reference [30].

The following parameters were used in the data processing of the data set of 568 chemicals:

i) fuzzy parameters: weighting coefficient = 1.5, tolerance convergence = 0.001, number of iterations = 0, cluster number = 6;
ii) genetic parameters: chromosome number = 10, chromosome size = 153 or 168 (number of descriptors used), initial active descriptors in each chromosome = 8, crossover point number = 1, percentage of rejections = 0.1, percentage of crossover = 0.8, percentage of mutation = 0.05, number of generations = 10;
iii) stepwise parameters: ascending coefficient = 0.02, descending coefficient = − 0.02.

2.4 Self Organizing Map

SOM [31] is a non-linear mapping technique which gives a 2D space representation of a given set of points from a multidimensional space derived from a large series of molecular descriptors. Each point of this set is related to a SOM node, which is characterized by N weighted connections varying between 0 and 1.

Training SOM consists in rearranging the layer nodes by gradually adjusting their weights. After selecting a first hyperspace point, the distances between its coordinates and each node of the SOM layer are calculated. The node having the shortest distance is called "winner" and the hyperspace point is "projected" on this node of the map. Then, the weights of the winning node and its neighbors are modified according to the equation:

$$w_{ij}(t+1) = w_{ji}(t) + \alpha(t)\gamma(t, r)(x_j - w_{ij}(t)) \quad (1)$$

where x_j is the component j of input vector x; w_{ij} represent the weight vector of the node i for the descriptor j; t and $\alpha(t)$ are respectively the iteration number and the learning rate; $\gamma(t, r)$ is the triangular neighborhood function depending on the iteration number and the distance r between the node i and the winning unit.

The learning rate $\alpha(t)$ is linearly decreased during the training process from $\alpha(0)$ to zero. The triangular function $\gamma(t, r)$ works on the whole map and it is discretely decreased with increasing the distance and the number of iterations.

The same procedure is successively repeated for all the hyperspace vectors and each point is associated with a node in the SOM layer. The points which are close in the descriptor hyperspace remain close in the SOM layer, occupying the same nodes or the neighboring ones. When SOM is applied on a chemical data set, the maps can then reveal similar compounds, if the Euclidean distance is accepted as a similarity measure.

The data set compounds were distributed in a map defined by 10 columns and 10 rows. The calculations were performed using proprietary software.

2.5 *Adaptive Fuzzy Partition (AFP)*

AFP is a supervised classification method implementing a fuzzy partition algorithm [45] and it was already presented and validated elsewhere [21, 22]. It models relations between molecular descriptors and chemical activities by dynamically dividing the descriptor space into a set of fuzzy partitioned subspaces. The aim of the algorithm is to select the descriptor and the cut position which allow to get the maximal difference between the two fuzzy rule scores generated by the new subspaces. The score is determined by the weighted average of the chemical activity values in an active subspace A and in its neighboring subspaces.

All the rules created during the fuzzy procedure are considered to establish the model between descriptor hyperspace and biochemical activities. Indicating with $P(x_{1,} ... x_n)$ a molecular vector in a n-dimensional descriptor hyperspace, a *rule* for a subspace S_k is defined by [46]:

if x_1 is associated with $\mu_{1k}(x_1)$ **and** x_2 is associated
with $\mu_{2k}(x_2)$ . . . **and** x_N is associated with $\mu_{Nk}(x_N)$ $\Rightarrow$
the score of the activity O for P is $O_{kP,}$ (2)

where x_i represents the value of the ith descriptor for the molecule P, μ_{ik} is the membership function related to the descriptor i for the subspace k, and O_{kP} is the biochemical activity value related to the subspace S_k. The "and" of the fuzzy rule is represented by the *Min operator* [47], which selects the minimal value amidst all the μ_{ik} components.

The membership functions are defined by trapezoidal shapes. The latter functions are based on the boundaries of the subspaces. If the width of a subspace S_k on the i^th^ dimension, after each cut, is represented by w_i, the p and q parameters defining the shape of the trapezoid are calculated by

$$p = \lambda_i w_i \text{ and } q = \upsilon_i w_i \quad (3)$$

where the parameters λ_i and υ_i vary so that $p \geq 1$ and $q \leq 1$. If $p = 1$ and $q = 1$, the membership function becomes a rectangle.

All the rules created during the fuzzy procedure are considered to establish the model between descriptor hyperspace and biochemical activities. The global score in the subspace S_k can be represented by

$$O_k = \frac{\sum_{j=1}^{M} \left(\text{Min}_i^N \mu_{ik}(x_i)_{Pj}\right) \cdot (A_{Pj})}{\sum_{j=1}^{M} \left(\text{Min}_i^N \mu_{ik}(x_i)_{P_j}\right)} \quad (4)$$

M is the number of molecular vectors in a given subspace, N is the total number of descriptors, $\mu_{ik}(x_i)_{Pj}$ is the fuzzy membership function related to the descriptor i for the molecular vector P_j, and A_{Pj} is the experimental activity of the compound P_j. A classic centroid defuzzification procedure [48] is implemented to determine the chemical activity of a new test molecule. All the subspaces k are considered and the general formula to compute the score of the activity O for a generic molecule Pj is

$$O(P_j) = \frac{\sum_{k=1}^{N\_subsp} \left(\text{Min}_i^N \mu_{ik}(x_i)_{Pj}\right) \cdot (O_k)}{\sum_{k=1}^{N\_subsp} \left(\text{Min}_i^N \mu_{ik}(x_i)_{P_j}\right)} \quad (5)$$

where N subsp represents the total number of subspaces.

The following parameters were used to process the data set of 568 chemicals:

maximal number of rules for each chemical activity = 35; minimal number of compounds for a given rule = 4; maximal number of cuts for each axis = 4; $p = 1.05 \div 1.55$ and $q = 0.55 \div 0.95$

3 Results and Discussion

3.1 *3D structures*

Before starting the database mining procedure, a test set was generated by randomly extracting 80 molecules from the data set of 568 chemicals. Then, the technique combining

Table 1. Most relevant descriptors selected for the 2D and 3D data sets of 568 pesticides.

Symbol	Definition	Descriptor family
	3D structures	
NrO	Relative number of oxygen atoms	Constitutional
NCRS	Relative negative charged surface area	Electronic
LogD-pH5	Lipophilicity at pH = 5	Physicochemical
LogP	Lipophilicity at pH = 7	Physicochemical
	2D structures	
Xvch10	Valence 10th order chain chi index	Topological
SdCH2	Sum of all H E-state values for (= CH_2)	Electro-topological
SHBint4	E-state of internal H bonds (5 path length)	Electro-topological
LogP	Lipophilicity at pH = 7	Physicochemical

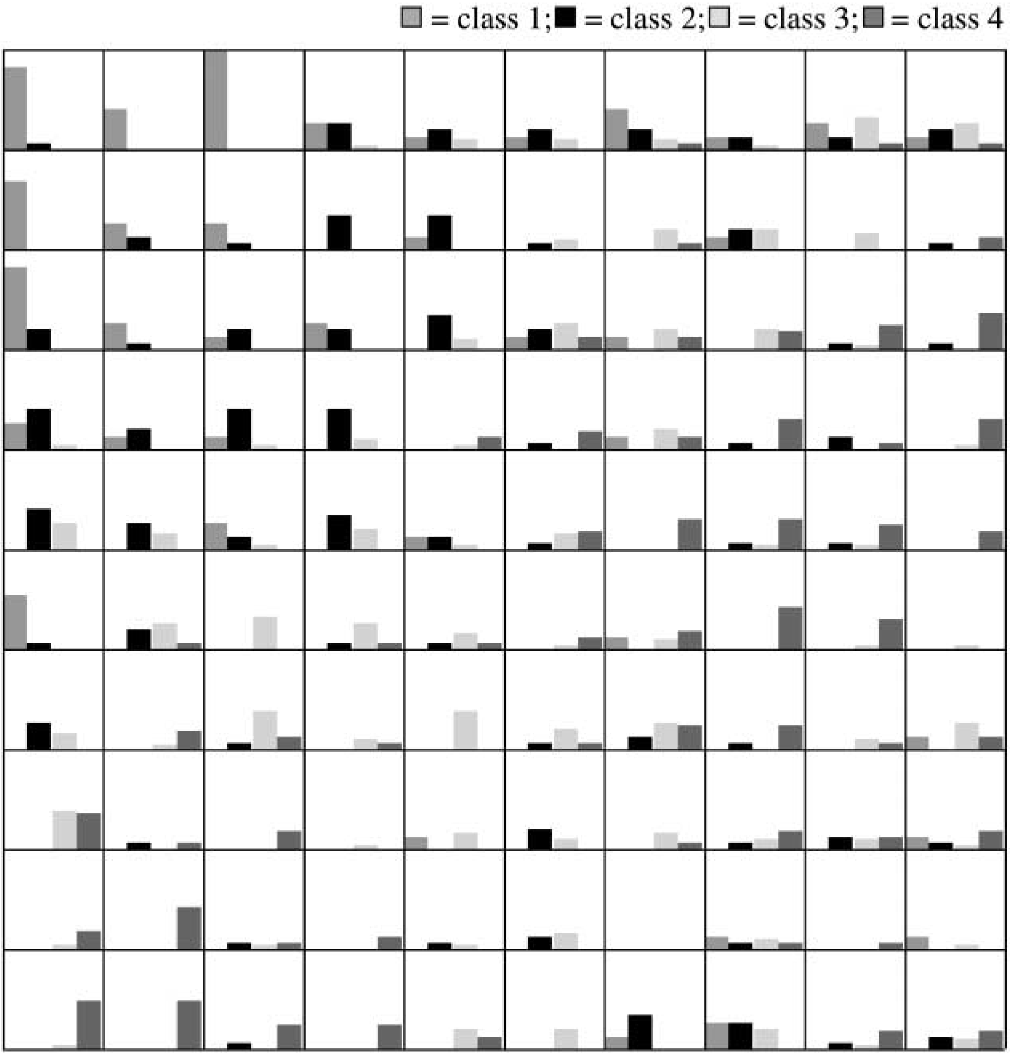


Figure 1. Molecular diversity analysis by SOM on the data set from which the AFP models were derived; the 4 descriptors computed on the 3D structures and selected by the GA/SW procedure were taken into account. The histogram heights represent the number of compounds of each class in any cell of the map.

GA and stepwise method (GA/SW) was applied on the molecular descriptors computed on the remaining 3D structures, in order to isolate the most relevant parameters. Table 1 shows the four descriptors selected and, amidst them, there are two parameters concerning lipophilicity (Log P and LogD-pH5), a very important property involved in the mechanism of molecular accumulation into the fish body [49]. The other parameters represent the relative

Table 2. Compound repartition in the training, validation, and test sets.

Classes	Training set	Validation set	Test set
1	43	15	15
2	105	21	21
3	137	23	23
4	123	21	21
All classes	408	80	80

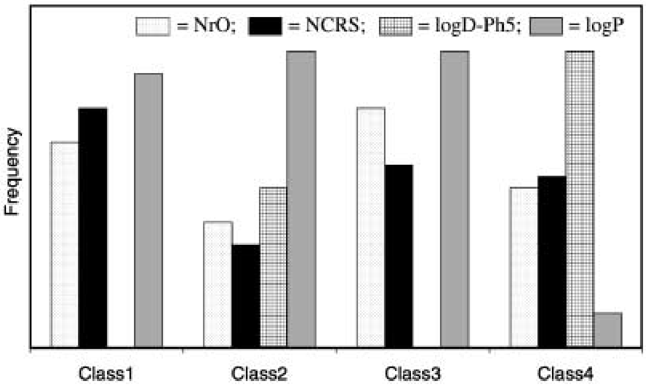


Figure 2. Representation of the descriptor contribution for each toxicity class, evaluated by computing descriptor frequency in the AFP rules.

number of oxygen atoms (RnO) and the relative negative charged surface area (RNCS).

The four descriptors selected were used to derive the SOM chart represented in Figure 1, which was employed to define the training and validation sets, maximizing, for each class, the molecular diversity; the compounds were selected in each cell of the map, according to the molecular frequency. The molecular distribution in the 4 classes is reported in Table 2, where the test set is also represented.

The map shows several regions characterized by a main toxicity class and the AFP method, working directly on the hyperspace, should be then able to define robust structure-activity relationships (SAR) for this data set. The AFP model was established on the 408 training set compounds distributed in the 4D descriptor space. Twenty-six rules were implemented to define each relationship between the molecular structures and the toxicity activities, and an example of rule defining an AFP subspace is represented by the following definition:

if $0 < x(\mathrm{RnO}) < 2$ and $2.3 < x(\log P) < 4.0 \Rightarrow$ the score (class 4) for a given compound is 1.

This relation evidences a subspace specially devoted to defining the least toxic class.

Figure 2 shows the descriptor contribution for each class, evaluated by computing the frequency of the molecular parameters in the rules. Even if all descriptors are taken into account to discriminate each class, their relative importance

Table 3. Comparison between experimental (Exp.) and predicted toxic classes for the 80 validation set compounds predicted by the AFP model established by molecular descriptors computed on 2D and 3D structures.

Name	Exp.	Predicted classes							
		3D structures				2D structures			
		1	2	3	4	1	2	3	4
Flucythrinate	1	0.79	0.17	0.02	0.06	0.91	0.09	0.00	0.04
Rotenone	1	0.06	0.52	0.02	0.06	0.00	0.85	0.05	0.04
Terbufos	1	1.00	0.17	0.02	0.06	0.84	0.00	0.05	0.04
2,2′-methylene bis(3,4,6-trichlorophenol)	1	0.82	0.00	0.02	0.06	0.94	0.00	0.00	0.04
1,3-dichloro-4,6-dinitrobenzene	1	0.55	0.25	0.21	0.06	0.32	0.55	0.23	0.04
Diphenyl phthalate	1	0.71	0.00	0.02	0.06	0.98	0.00	0.00	0.04
Manool	1	0.82	0.64	0.02	0.06	0.94	0.00	0.00	0.04
Nonylphenol	1	0.81	0.64	0.02	0.06	0.94	0.57	0.00	0.04
Pentachlorophenol	1	1.00	0.28	0.02	0.25	0.35	0.57	0.05	0.04
3-(3,4-dichlorophenoxy)benzaldehyde	1	1.00	0.34	0.02	0.06	1.00	0.00	0.05	0.04
n-undecyl cyanide	1	0.66	0.00	0.02	0.06	0.80	0.57	0.00	0.04
t-butylstyrene	1	0.48	0.34	0.02	0.06	0.84	0.43	0.05	0.04
Di-n-butylterephthalate	1	0.86	0.00	0.02	0.06	0.98	0.00	0.00	0.04
Di-n-hexylamine	1	1.00	0.14	0.02	0.05	0.05	0.00	0.35	0.00
3,5-dibromosalicylaldehyde	1	0.28	0.09	0.18	0.16	0.41	0.61	0.54	0.04
p-phenylazophenol	2	0.02	0.81	0.24	0.06	0.00	1.00	0.05	0.04
Chloroacetonitrile	2	0.00	0.08	0.00	0.68	0.00	0.00	0.00	1.00
2-undecanone	2	0.20	0.92	0.02	0.06	0.00	0.71	0.05	0.04
Amylbenzene	2	1.00	0.27	0.02	0.06	0.84	0.43	0.05	0.04
Dehydroabietic acid	2	0.83	0.16	0.02	0.06	0.57	0.71	0.00	0.04
Nonylamine	2	0.01	0.13	0.00	0.00	0.02	0.33	0.00	0.00
Hexane	2	0.08	0.97	0.00	0.06	0.02	0.56	0.00	0.04
4,6-dimethoxy-2-hydroxybenzaldehyde	2	0.11	0.14	0.41	0.03	0.00	0.13	0.67	0.16
4,9-dithiadodecane	2	0.06	1.00	0.02	0.06	0.00	0.85	0.05	0.04

Table 3. (cont.)

Name	Exp.	Predicted classes							
		3D structures				2D structures			
		1	2	3	4	1	2	3	4
2,5-dinitrophenol	2	0.13	0.16	0.77	0.07	0.55	0.13	0.23	0.77
2-chloro-5-nitrobenzaldehyde	2	0.00	0.04	0.44	0.25	0.39	0.54	0.13	0.04
Dibutyl succinate	2	0.32	0.98	0.11	0.06	0.41	0.73	0.19	0.04
Cyclohexane	2	0.01	0.16	1.00	0.06	0.02	0.16	0.54	0.04
1-naphthol	2	0.02	0.16	0.48	0.06	0.02	0.20	0.54	0.04
p-tert-butylphenol	2	0.02	0.22	0.84	0.06	0.02	0.20	0.71	0.04
3,8-dithiadecane	2	0.01	0.16	0.33	0.06	0.02	0.23	0.54	0.04
2,4,6-tribromophenol	2	0.49	0.55	0.02	0.06	0.34	0.57	0.05	0.04
Pentachloroethane	2	0.03	0.34	0.00	0.06	0.02	0.73	0.30	0.04
1,2,4-trimethylbenzene	2	0.04	0.75	0.00	0.06	0.02	0.49	0.19	0.04
Oxamyl	2	0.01	0.13	0.24	0.78	0.00	0.43	0.14	0.07
Propoxur (baygon)	2	0.19	0.23	0.44	0.06	0.00	0.13	0.54	0.04
4-butylaniline	3	0.00	0.16	0.28	0.06	0.02	0.23	0.54	0.04
Hexanal	3	0.01	0.13	0.42	0.41	0.02	0.13	0.54	0.04
2-allylphenol	3	0.02	0.16	0.49	0.06	0.02	0.19	0.54	0.04
3-pyridinecarboxaldehyde	3	0.01	0.13	0.24	0.56	0.02	0.13	0.75	0.07
Diethyl adipate	3	0.00	0.24	0.44	0.04	0.67	0.13	0.54	0.04
4-fluoroaniline	3	0.01	0.12	0.03	0.90	0.02	0.13	0.21	1.00
1-chloro-3-nitrobenzene	3	0.03	0.38	0.44	0.06	0.37	0.15	0.54	0.04
2,4-dimethoxybenzaldehyde	3	0.15	0.13	0.44	0.19	0.00	0.13	0.54	0.04
Methyl p-nitrobenzoate	3	0.00	0.15	0.44	0.03	0.49	0.11	0.64	0.04
Tert-octylamine	3	0.01	0.13	0.00	0.00	0.02	0.13	0.27	0.00
p-ethoxybenzaldehyde	3	0.43	0.24	0.44	0.04	0.48	0.13	0.54	0.04
1-fluoro-4-nitrobenzene	3	0.28	0.34	0.44	0.06	0.00	0.13	0.54	0.04
5-nonanone	3	0.02	0.32	0.06	0.06	0.01	0.20	0.73	0.04
Butyl ether	3	0.02	0.21	0.84	0.06	0.02	0.67	0.65	0.04
2-chloro-4-methylaniline	3	0.00	0.16	0.26	0.06	0.02	0.13	0.54	0.04
N,n-dimethylbenzylamine	3	0.01	0.13	1.00	0.31	0.02	0.13	0.98	0.67
2,6-dinitrophenol	3	0.06	0.13	0.91	0.04	0.00	0.13	0.52	0.00
A,a, a-trifluoro-o-tolunitrile	3	0.00	0.16	0.84	0.06	0.02	0.20	0.54	0.04
Trichloroethylene	3	0.27	0.13	0.38	0.41	0.02	0.13	0.54	0.04
3′-chloro-o-formotoluidide	3	0.02	0.16	0.56	0.07	0.02	0.13	0.54	0.04
2,4,5-trimethoxybenzaldehyde	3	0.00	0.15	0.44	0.10	0.00	0.13	0.54	0.04
2′-hydroxy-4′-methoxyacetophenone	3	0.05	0.15	0.24	0.16	0.00	0.13	0.66	0.16
2,5-dimethylfuran	3	0.01	0.12	0.24	0.79	0.02	0.13	0.54	0.94
4-methoxyphenol	4	0.01	0.16	0.29	0.57	0.00	0.13	0.44	0.73
4-toluidine	4	0.01	0.15	0.34	0.88	0.02	0.13	0.54	0.91
3,4-dimethyl-1-pentyn-3-ol	4	0.01	0.12	0.24	0.37	0.02	0.13	0.21	0.97
Pyrrole	4	0.01	0.14	0.03	0.93	0.02	0.13	0.12	0.97
Ethyl acetate	4	0.01	0.13	0.21	0.87	0.00	0.13	0.12	0.13
Butylamine	4	0.01	0.13	0.03	0.92	0.02	0.13	0.04	0.63
Methyl acetate	4	0.01	0.13	0.17	0.61	0.00	0.13	0.26	0.45
Tert-butyl acetate	4	0.04	0.15	0.24	0.68	0.00	0.13	0.49	0.69
5-diethylamino-2-pentanone	4	0.01	0.13	0.13	0.33	0.02	0.13	0.04	0.93
4-picoline	4	0.01	0.13	0.03	1.00	0.02	0.13	0.25	0.97
2-hexanone	4	0.01	0.13	0.42	0.58	0.02	0.13	0.54	0.04
1,6-dicyanohexane	4	0.00	0.13	0.76	0.39	0.02	0.13	0.54	0.04
N,n-bis(2,2-diethoxyethyl)methylamine	4	0.05	0.11	0.24	0.01	0.02	0.60	0.54	0.04
4-acetamidophenol	4	0.01	0.13	0.24	0.87	0.00	0.13	0.12	0.69
2-picoline	4	0.01	0.13	0.03	0.95	0.02	0.13	0.12	0.97
1,3-diaminopropane	4	0.01	0.00	0.00	1.00	0.02	0.13	0.00	1.00
3-pentanone	4	0.01	0.12	0.24	0.83	0.02	0.13	0.31	0.73
2-butanone	4	0.20	0.13	0.24	0.56	0.02	0.13	0.12	0.12
1,3-diethyl-2-thiobarbituric acid	4	0.01	0.01	0.00	0.99	0.02	0.13	0.00	1.00
1-(2-hydroxyethyl)piperazine	4	0.01	0.00	0.00	1.00	0.02	0.13	0.00	1.00
5,5-dimethylhydantoin	4	0.01	0.13	0.22	0.87	0.00	0.13	0.00	0.75

Table 4. Statistical values defining the robustness of the AFP model developed on the 2D and 3D data sets.

Class	Training set (%)	Validation set (%)	Test set (%)
		3D structures	
1	53	87	67
2	66	43	67
3	77	78	74
4	80	90	76
All classes	72	74	72
		2D structures	
1	80	67	53
2	57	57	56
3	65	83	74
4	90	76	81
All classes	71	72	72

is quite different. For example, logP is used to characterize classes 1, 2, and 3, whereas its contribution is negligible in class 4. In an analogous way, the descriptor logD-pH5 is specially devoted to defining classes 2 and 4, whereas the parameters RnO and RNCS, related to the molecular reactivity, play a major role to represent all classes.

The AFP models were validated by predicting the toxicity range of the 80 validation set compounds. The method allows to get the degrees of membership of the different classes for each compound, within a 0 to 1 range. The comparison between predicted and experimental values for all the validation set compounds is reported in Table 3. The validation results for the best AFP model are shown in Table 4. The experimental toxicity class was predicted correctly for 72% of the validation set compounds and, moreover, a similar score was obtained by testing the training set compounds, showing the model developed is general. But the robustness of this AFP model is chiefly confirmed by the validation statistics derived from the test set: in this case too about 72% of the molecules were predicted correctly. In fact, the main object in developing prediction models should not consist in getting impressive scores by predicting training and validation sets, but in developing robust models able to predict correctly also test sets never involved in the model building procedures.

3.2 2D structures

The above proposed global database mining procedure was applied on the set of descriptors computed only on the 2D structures, in order to evaluate whether it lost in prediction power compared with the previous set computed on 3D structures. Four relevant parameters were isolated by the GA/SW method (see Table 1) and, like in the 3D structure case, a lipophilicity parameter is present. The other descriptors represent a topological index (xvch10) and two electrotopological parameters related to H-bonds (SHBint4) and terminal double-bonded CH_2 groups (SdCH2). After selecting a new training set by SOM, but keeping the same proportions as indicated in Table 2, a new AFP model was built by using 36 rules. In this case too very similar results were obtained by predicting training, validation, and test sets, confirming the robustness of the model developed by AFP (Tables 3 and 4). But even more important is the fact that all these results indicate that prediction power is very similar in 2D and 3D models.

4 Conclusions

The need to better understand and predict the impact of the chemicals on human health and wildlife requires to develop ever more efficient SAR models. Fuzzy Logic concepts constitute an interesting solution to derive general classification models. In this work, an Adaptive Fuzzy Partition algorithm was applied on a data set of 568 chemicals, divided into four classes, defined by the European Community legislation, according to their different toxicity against fathead minnow. The AFP method consists in modeling molecular descriptor – activity relationships by dynamically dividing the descriptor hyperspace into a set of fuzzy subspaces. Two sets of molecular descriptors, respectively computed on 2D and 3D structures, were tested and the most relevant parameters were selected with help of a procedure based on genetic algorithm concepts and a stepwise method. The experimental toxicity class was predicted correctly for about 72% of the validation set compounds. Due to the high variability affecting the experimental procedures in the area of ecotoxicity [18] and the complexity of the phenomena related, these preliminary results are very encouraging. Furthermore, similar validation scores were obtained by using molecular descriptors computed on 2D or 3D structures. This underlines that probably, in ecotoxicity, where a large number of complex and interactive mechanisms define the biological phenomena, the role of the 3D descriptors in generating SAR models could be less important regarding other fields.

It has to be also underlined that the AFP method needs only few minutes to test several thousand molecules. Then, at present, work is underway to use these global SAR models to screen large databases, in accordance with the criteria of efficient and rapid structural alert.

Acknowledgement

Our gratitude goes to the European Commission for its support and the grant to Nadège Piclin (project IMAGETOX, HPRN-1999-00015). We thank Prof. A. Katritzky (University of Florida) and Prof. M. Karelson (University of Tartu) for the use of CODESSA.

References

[1] Jamet, P. (Ed.), *European Commission. Directorate-Generale XII. COST 66. Fate of pesticides in the soil and the environment*, ECSP-EEC-EAEC, Brussels 1994.

[2] Karcher, W., and Devillers, J. (Eds.), *Practical Applications of Quantitative Structure-Activity Relationships (QSAR) in Environmental Chemistry and Toxicology*, Kluwer Academic Publishers, Dordrecht (The Netherlands) 1990.

[3] Dearden, J. C., Barratt, M. D., Benigni, R., Bristol, D. W., Combes, R. D., Cronin, M. T. D., Judson, P. N., Payne, M. P., Richard, A. M., Tichy, M., Worth, A. P., and Yourick, J. J., The Development and Validation of Expert Systems for Predicting Toxicity, *ATLA 25*, 223 – 252 (1997).

[4] Gini, G., and Katrizky A., *Predictive Toxicology of Chemicals: Experiences and Impact of AI Tools*, AAAI Press, Menlo Park (USA) 1999.

[5] Todeschini, R., and Consonni, V., *Handbook of Molecular Descriptors*, Wiley-VCH, Mannheim (Germany) 2000.

[6] Ros, F., Guillaume, S., Rabatel, G., and Sevila, F., Recognition of overlapping particles in granular product images using statistics and neural networks, *Food Control 6*, 37 – 43 (1995).

[7] Ventura, S., Silva, M., Pérez-Bendito, D., and Hervás C., Computational Neural Networks in Conjunction with Principal Component Analysis for Resolving Highly Nonlinear Kinetics, *J. Chem. Inf. Comp. Sci. 37*, 287 – 291 (1997).

[8] Gini, G., Lorenzini, M., Benfenati, E., Grasso, P., and Bruschi, M., Predictive Carcinogenicity: a Model for Aromatic Compounds, with Nitrogen-containing Substituents, Based on Molecular Descriptors Using an Artificial Neural Network, *J. Chem. Inf. Comp. Sci. 39*, 1076 – 1080 (1999).

[9] Goldberg, D. E., *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley, New York (USA) 1989.

[10] Collins, R. J., and Jefferson, D. R., Selection in massively parallel genetic algorithms, in: Belew, R. K., and Booker, L. B. (Eds.), *Proceedings of the Fourth International Conference on Genetic Algorithms*, Morgan Kaufmann Publishers, San Mateo (USA) 1991, pp. 244 – 248.

[11] Wold, S., PLS for Multivariate linear Modeling, in: van de Waterbeemd, H. (Ed.), *Chemometric Methods in Molecular Design*, VCH, Weinheim (Germany) 1995, pp. 195 – 218.

[12] Hecht-Nielsen, R., Theory of the backpropagation neural network, in: *Proceedings of the International Joint Conference on Neural Networks*, IEEE Press, New York (USA) 1989, pp. 593 – 605.

[13] Rorije, E., and Peijnenburg, W. J. G. M., QSARs for oxidation of phenols in the aqueous environment, suitable for risk assessment, *J. Chemom. 10*, 79 – 93 (1996).

[14] Verhaar, H. J. M., Urrestarazu Ramos, E., and Hermens, J. L. M., Classifying environmental pollutants. 2: Separation of class 1 (baseline toxicity) and class 2 ('polar narcosis') type compounds based on chemical descriptors, *J. Chemom. 10*, 149 – 162 (1996).

[15] Devillers, J., A general QSAR model for predicting the acute toxicity of pesticides to Lepomis macrochirus, *SAR QSAR Environ. Res. 11*, 397 – 417 (2001).

[16] Devillers, J., QSAR modeling of large heterogeneous sets of molecules, *SAR QSAR Environ. Res. 12*, 515 – 528 (2001).

[17] 95/365/CE: Commission Decision of 25 July 1995 establishing the ecological criteria for the award of the Community ecolabel to laundry detergents. *Official Journal NO. L217, 13/09/1995*, pp. 0014 – 0030.

[18] Benfenati, E., Piclin, N., Roncaglioni, A., and Varì M. R., Factors influencing predictive models for toxicology, *SAR QSAR Environ. Res. 12*, 593 – 603 (2001).

[19] Schalkoff, R., *Pattern recognition statistical, structural and neural approaches*, John Wiley & Sons, New York (USA) 1992.

[20] Zadeh, L. A., 1977. Fuzzy sets and their applications to classification and clustering, in: Van Ryzin, J. (Ed.), *Classification and Clustering*, Academic Press, New York (USA) 1977, pp. 251 – 299.

[21] Pintore, M., Audouze, K., Ros, F., and Chrétien, J. R., Adaptive fuzzy partition in data base mining: application to olfaction, *Data Science Journal 1*, 99 – 110 (2002).

[22] Pintore, M., Piclin, N., Benfenati, E., Gini, G., and Chrétien, J. R., Database mining with adaptive fuzzy partition (AFP): application to the prediction of pesticide toxicity on rats, *Environ. Toxicol. Chem.*, in press (2002).

[23] Brooke, L. T., Call, D. J., Geiger, D. L., and Northcott, C. E. (Eds.), *Acute Toxicities of Organic Chemicals to Fathead Minnows (Pimephales promelas)*, Vol. 1, Center for Lake Superior Environmental Studies, University of Wisconsin, Superior (USA) 1984.

[24] Geiger, D. L., Northcott, C. E. Call, D. J., and Brooke, L. T. (Eds.), *Acute Toxicities of Organic Chemicals to Fathead Minnows (Pimephales promelas)*, Vol. 2, Center for Lake Superior Environmental Studies, University of Wisconsin, Superior (USA) 1985.

[25] Geiger, D. L., Poirier, S. H., Brooke, L. T., and Call, D. J. (Eds.), *Acute Toxicities of Organic Chemicals to Fathead Minnows (Pimephales promelas)*, Vol. 3, Center for Lake Superior Environmental Studies, University of Wisconsin, Superior (USA) 1986.

[26] Geiger, D. L., Call, D. J., and Brooke, L. T. (Eds.), *Acute Toxicities of Organic Chemicals to Fathead Minnows (Pimephales promelas)*, Vol. 4, Center for Lake Superior Environmental Studies, University of Wisconsin, Superior (USA) 1988.

[27] Geiger, D. L., Brooke, L. T. and Call, D. J. (Eds.), *Acute Toxicities of Organic Chemicals to Fathead Minnows (Pimephales promelas)*, Vol. 5. Center for Lake Superior Environmental Studies, University of Wisconsin, Superior (USA) 1990.

[28] Russom, C. L., Bradbury, S. P., Broderius, S. J., Hammermeister D. E., and Drummond R. A., Predicting modes of action from chemical structure: Acute toxicity in the fathead minnow (Pimephales promelas), *Environ. Toxicol. Chem. 16*, 948 – 967 (1997).

[29] Directive 92/32/ECC (1992), the seventh amendment to Directive 67/548/ECC, OJL 154 of 5.VI.92, p1.

[30] Ros, F., Pintore, M., and Chrétien, J. R., Molecular descriptor selection combining genetic algorithms and fuzzy logic: application to database mining procedures, *Chemometr. Intell. Lab. 63*, 15 – 26 (2002).

[31] Kohonen, T., *Self-Organizing Maps*, Springer-Verlag, Berlin (Germany) 2001.

[32] Kier, L. B., and Hall, L. H., *Molecular connectivity in structure analysis*, Wiley, New York (USA) 1986.

[33] Dearden, J. C., Physico-chemical descriptors, in: Karcher, W., and Devillers, J. (Eds.), *Practical Applications of Quantitative Structure-Activity Relationships (QSAR) in Environmental Chemistry and Toxicology*, Kluwer Academic, Dordrecht (The Netherlands) 1990, pp. 25 – 29.

[34] Sabljic, A., Topological indices and environmental chemistry, in: Karcher, W., and Devillers, J. (Eds.), *Practical Applications of Quantitative Structure-Activity Relationships (QSAR) in*

Environmental Chemistry and Toxicology, Kluwer Academic, Dordrecht (The Netherlands) 1990, pp. 61 – 82.

[35] Kier, L. B., and Hall, L. H., *Molecular Structure Description – The Electrotopological State*, Academic Press, San Diego (USA) 1999.

[36] ChemInter© 1.0, ChemInter.

[37] SciQSAR 2D®, SciVision, Burlington (USA) 1999.

[38] PALLAS© 2.0, CompuDrug Chemistry Ltd., Budapest (Hungary) 1994 – 1995.

[39] Pintore, M., Taboureau, O., Ros, F., and Chrétien J. R., Data Base Mining (DBM) applied to CNS activity, *Eur. J. Med. Chem. 36*, 349 – 359 (2001).

[40] HyperChem 5.1, Hypercube Inc., Gainesville (USA) 1997.

[41] CODESSA 2.20, Semichem Inc., Shawnee Mission (USA) 1994 – 1996.

[42] Kinnear, K. E., *Advances in Genetic Programming*, MIT Press, Cambridge (USA) 1994.

[43] Haupt, R. L., and Haupt S. E., *Practical Genetic Algorithms*, Wiley, New York (USA) 1999.

[44] Leardi, R., and Gonzales A. L., Genetic algorithms applied to feature selection in PLS regression: how and when to use them, *Chemometr. Intell. Lab. 41*, 195 – 207 (1998).

[45] Lin, Y., and Cunninghan, G. J., Building a Fuzzy System from Input-Output Data, *J. Intell. Fuzzy Syst. 2*, 243 – 250 (1994).

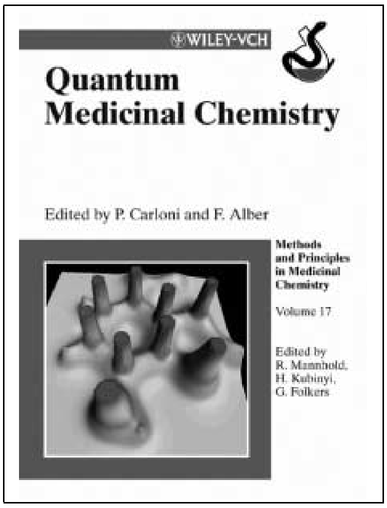
[46] Sugeno, M., and Yasakawa, T., A fuzzy-logic-based approach to qualitative modeling, *IEEE T. Fuzzy Syst. 1*, 7 – 31 (1993).

[47] Dubois, D., and Prade, H., An introduction to possibilistic and fuzzy logics, in: Shafer, G., and Pearl, J. (Eds.), *Readings in Uncertain Reasoning*, Morgan Kaufmann, San Francisco (USA) 1990, pp. 742 – 761.

[48] Gupta, M. M., and Qi, J., Theory of T-norms and fuzzy inference methods, *Fuzzy Set Syst. 40*, 431 – 450 (1991).

[49] Hermens, J. L. P., Quantitative structure-activity relationships for predicting fish toxicity, in: Karcher, W., and Devillers, J. (Eds.), *Practical Applications of Quantitative Structure-Activity Relationships (QSAR) in Environmental Chemistry and Toxicology*, Kluwer Academic, Dordrecht (The Netherlands) 1990, pp. 263 – 280.

Received on June 24, 2002; Accepted on October 15, 2002


WILEY-VCH
Quantum Medicinal Chemistry
Edited by P. Carloni and F. Alber
Methods and Principles in Medicinal Chemistry
Volume 17
Edited by R. Mannhold, H. Kubinyi, G. Folkers