QSAR in Ecotoxicity: An Overview of Modern Classification Techniques

Paolo Mazzatorta,*,† Emilio Benfenati,† Paola Lorenzini,†,‡ and Marco Vighi‡

Istituto di Ricerche Farmacologiche "Mario Negri" Milano, Via Eritrea, 62, 20157 Milano, Italy, and Università degli Studi di Milano Bicocca, Dip. di Scienze dell'Ambiente e del Territorio (DISAT), Piazza della Scienza 1, 20126 Milano, Italy

Received September 2, 2003

This study deals with classification for toxicity prediction. Using a data set of 235 pesticides and 153 descriptors, we built several models using seven classification algorithms: nearest mean classifier, linear discriminant analysis, quadratic discriminant analysis, regularized discriminant analysis, soft independent modeling of class analogy, K nearest neighbors classification, classification, and regression tree. The performance of the models was then compared with the classifier, the end-points, the number of descriptor, and the diversity of the data set. Finally, we made a critical analysis of the models and descriptors.

1. INTRODUCTION

The quantitative structure−activity relationship (QSAR) approach is based on the assumption that the structure of a molecule must contain the features responsible for its physical, chemical, and biological properties and on the possibility of representing a molecule by numerical descriptors. With increasing demand for reliable chemical data validated QSAR methods are acquiring new significance. For instance, the risk assessment of the vast range of chemicals that may pose a threat to the environment and human health requires such a wealth of chemical, physicochemical, and toxicological data that experimental capacities are and will remain largely insufficient to satisfy every need. The critical application of QSAR methods may fill many of the gaps.

Numerous different models are used in QSAR approaches. Here we will consider classification systems which are quite common in the case of carcinogenicity[1−3] because carcinogenicity classes are defined by regulatory bodies such as IARC, EPA. For ecotoxicity most of the QSAR models are regressions, referring to the dose that gives the toxic effect in 50% of the animals (for instance LD_{50}: lethal dose for 50% of the test animals). This dose is a continuous value so regression seems the most appropriate algorithm. However, classification offers some advantages in ecotoxicology too: (i) the regulatory values are indicated as toxicity classes and (ii) classification can allow better management of data which are often noisy. For this reason we have investigated classification in the past[2,4−6] and in the present study.

No general rule exists to define the best approach to a specific classification problem.[7,8] In several cases a selection of descriptors is the only essential condition for developing a general system. The next step involves defining the best computational method to develop robust structure−activity models.

The present paper deals with some of the most common classification techniques and how they behave for toxicology modeling. The rest of the paper is then organized as follows: section 2 describes the data set and the algorithms and gives some details about the descriptors and the validation procedure. General and in-depth analyses of the models are grouped in section 3. The paper ends with conclusions, discussion, acknowledgments, and references.

2. MATERIALS AND METHODS

Data Set. We investigated a data set of 235 common agrochemical compounds, developed within the EC funded project COMET (Computerized Molecular Evaluation of Toxicity).[9,10] The selection of these compounds was based first of all on the existence of reliable information about their toxicity toward different end-points. The most frequently studied animals were trout, rat, daphnia, quail, and duck. The toxicity values are the result of a wide bibliographic search. The literature sources are *the Pesticide Manual*,[11] the *ECOTOX database system*,[12] and *Micromedex*,[13] which include several databases. If different values were observed, the lowest one was kept, unless it was an outlier, by considering the 95% interval of confidence. An evaluation of the variability is given elsewhere (ref 10). The data set used is available on request.

The toxicity value was expressed using the following formula

$$y_i = \log_{10}\left(\frac{1}{LC_{50,i}}\right) \quad (1)$$

where LC_{50} is the water concentration which kills 50% of the aquatic animals (trout and daphnia), expressed as mmol/L. In the case of rat and birds, we used instead LD_{50}, defined above, expressed as mmol/kg of body weight.

Then the values were scaled in the interval [−1; 1]. Four classes were defined as follows:

- Class 1: [−1; −0.5)
- Class 2: [−0.5; 0)
- Class 3: [0; 0.5)
- Class 4: [0.5; 1]

* Corresponding author phone: +39-02-39014499; fax: +39-02-39001916; e-mail: mazzatorta@marionegri.it.
† Istituto di Ricerche Farmacologiche "Mario Negri" Milano.
‡ Università degli Studi di Milano Bicocca.

10.1021/ci034193w CCC: $27.50
Published on Web 12/18/2003

Table 1. Data Sets Used in the Study

	end-point				
	trout (tr)	rat (rat)	daphnia (dap)	quail (q)	duck (d)
Chemical Class					
anilines (A)	39		35		
ureas (U)	31	31	28		
carbamates (C)	26	26	26		
halogenated aromatic (Ha)	23		23		
organophosphoric (OP)	57	59	49	37	28
halogen (CA)	83		78	49	41
heterocyclic (Het)	119	120	115		
Target					
herbicides (Her)	95	96	87		
fungicides (F)	47	47	47		
insecticides (I)	73	74	66		

These 235 pesticides were grouped by chemical class, toxic activity, and end-point (Table 1). We also split the pesticides according to their pattern of use: insecticides, herbicides, and fungicides. In some cases there were too few molecules to allow acceptable models. Overall, we tested about 200 models.

Descriptors. We examined a pool of about 150 descriptors calculated with different software: Hyperchem 5.0 (Hypercube Inc., Gainsville, FL, U.S.A.), CODESSA 2.2.1 (SemiChem Inc., Shawnee, KS, U.S.A.), and Pallas 2.1 (CompuDrug; Budapest, Hungary). They are split into six categories according to the classification in the software CODESSA: constitutional (34 descriptors), geometrical (14), topological (38), electrostatic (57), quantum-chemicals (6), and physicochemical descriptors (4). For a better description consult.[14−16]

To obtain a good model, the variables that best describe the molecules must be selected. Some of these descriptors may not add information, just increasing the noise, making analysis more complex. Also with a relatively low number of variables the risk of overfitting is reduced. The descriptors were selected by principal components analysis (PCA).

Classification Algorithms. Classification is the process of dividing a data set into mutually exclusive groups so that the members of each group are as "close" as possible to one another, and different groups are as "far" as possible from each other, where distance is measured with respect to specific variable(s) involved in the prediction. The classification algorithms form a model able to define to which particular class each object belongs. This procedure is also known in the literature as Supervised Learning in order to distinguish it from the Unsupervised Learning or Clustering, in which the existence of classes or clusters is deduced from the data.[17] Classes were defined classifying the compound's toxicity as before split.

Seven classification algorithms were used:

- NMC (nearest mean classifier)
- LDA (linear discriminant analysis)
- QDA (quadratic discriminant analysis)
- RDA (regularized discriminant analysis)
- SIMCA (soft independent modeling of class analogy)
- KNN (K nearest neighbors classification)
- CART (classification and regression tree)

The first four are parametric statistical systems based on Fisher's discriminant analysis, the fifth and sixth are nonparametric statistical methods, and the last is a classification tree.

The software SCAN (Software for Chemometric ANalysis) version 1.1 for Windows[18] was used.

NMC. This algorithm uses the Euclidean distance (d_e) between the object *s* and *t* with respect to the *j*th coordinate. Classifying, it assigns the objects to the class with the nearest centroid:

$$d_e = \sqrt{\sum_{j=1}^{p}(x_{sj} - x_{tj})^2} \tag{2}$$

where x_{sj} = coordinates of the *j*th coordinate of the *s* object, x_{tj} = coordinates of the *j*th coordinate of the *t* object, and p = number of variables.

LDA. Fisher's linear discrimination[7] is an empirical method based on p-dimensional vectors of attributes. Thus the classes are separated by a hyperplane which divides the p-dimensional space of attributes to distinguish the classes as well as possible.

QDA. The quadratic discrimination is similar to the linear one, but in this case the hypersurfaces, which divide the classes, are quadratic.

RDA. The variations introduced in this model are intended to overcome the main problems that afflict both linear and quadratic discrimination. The most efficient regulation was by Friedman,[19] who proposed a compromise between the two previous techniques using a biparametric method for the estimation (λ and γ).

SIMCA. This model, proposed by Wold and Sjöström,[20] is one of the first used in chemometrics for modeling classes and, unlike the other techniques, is not parametric. The idea is to consider each class separately and to look for a representation using the principal components. An object is put in a class on the basis of the residual distance, rsd^2, from the model representing the class itself

$$r_{igj}{}^2 = (\hat{x}_{igj} - x_{igj})^2 \tag{3}$$

$$\mathrm{rsd}^2 = \frac{\sum_j r_{igj}{}^2}{(p - M_j)} l \sqrt{\sum_{j=1}^{p}(x_{sj} - x_{tj})^2} \tag{4}$$

where $\hat{x}_{igj}$ = coordinates of the object's projections on the inner space of the mathematical model for the class, x_{igj} = object's coordinates, p = number of variables, and M_j = number of the principal components significant for the *j* class.

KNN. This technique classifies each record in a data set based on a combination of the classes of the *k* record(s) most similar to it in a historical data set (where $k = 1$).

CART. CART is a tree-shaped structure that represents sets of decisions. These decisions generate rules for the classification of a data set. CART provides a set of rules that can be applied to a new (unclassified) data set to predict which records will have a given outcome. It segments a data set by creating two-way splits.

Validation. The most common validation methods are as follows: (i) leave-one-out (LOO); (ii) leave-more-out (LMO); (iii) train & test; and (iv) bootstrap. Here, we only used the first one systematically because in the literature it is considered the best one when working on a small data set such as the ones we used.[21]

Table 2. Overview of the Results Obtained with Different Algorithms on the Data Sets[a]

	best	worst[b]	mean	max	high gap	low gap
A-tr[c]	RDA	SIMCA	73.5	79.5	SIMCA	CART
A-dap	CART, RDA	SIMCA	78.3	80.0		RDA
U-tr	LDA, RDA	KNN, SIMCA, NMC	72.8	80.7		RDA, LDA
U-dap	SIMCA	NMC	85.2	92.9		
U-rat	RDA	SIMCA	77.1	83.9	SIMCA	NMC
C-tr	RDA	SIMCA	75.6	89.5	SIMCA	RDA
C-dap	RDA	CART, LDA, NMC	77.0	84.6	SIMCA	LDA
C-rat	RDA	NMC	73.1	80.8	SIMCA	LDA
Ha-tr	RDA	SIMCA	82.6	91.3		RDA
Ha-dap	LDA, RDA	SIMCA	79.1	82.6		RDA, LDA, NMC
OP-tr	CART	NMC	69.8	77.2	SIMCA	LDA
OP-dap	CART	LDA, SIMCA	66.3	77.6		RDA, NMC
OP-rat	CART	SIMCA, NMC	67.0	74.6	SIMCA	RDA
OP-q	CART	SIMCA	77.1	86.5	SIMCA	NMC
OP-d	CART	others	67.9	67.9		
CA-tr	CART	NMC	66.1	70.7	SIMCA	RDA
CA-dap	CART	SIMCA, NMC	66.5	71.4	CART	LDA
CA-q	CART	SIMCA	74.2	79.2	CART	RDA
CA-d	CART	SIMCA	78.1	80.5	CART	RDA, NMC
Het-tr	CART	others	63.0	63.0		
Het-dap	RDA	SIMCA	69.0	70.4		
Het-rat	RDA	others	66.7	66.7		
Her-tr	CART	SIMCA	66.7	71.6		CART
Her-dap	CART	QDA, SIMCA	66.1	70.0		
Her-rat	CART	others	59.4	59.4		
F-tr	CART	others	74.5	74.5		
F-dap	CART	KNN, SIMCA, NMC	64.7	67.5	QDA	LDA
F-rat	CART	SIMCA	65.1	70.2	CART	NMC
I-tr	RDA	others	71.2	71.2		
I-dap	RDA	others	63.6	63.6		
I-rat	RDA	others	60.8	60.8		

[a] "best" lists the algorithm(s) with the best nonerror rate (NER%); "worst" lists the algorithm(s) with the worst NER%; the average error of all the algorithms is reported in "mean"; the best NER% is in "max"; where an algorithm has a significant difference between the NER% in fitting and the NER% in cross-validation the algorithm(s) is reported in "high gap"; the most stable algorithm (minimum difference between NER% in fitting and NER% in cross-validation) is reported in "low gap". [b] In cases where only one algorithm has an acceptable NER% value, *others* is written in the "worst" column. [c] Legend as in Table 1.

According to LOO, given n objects you compute n models. For each model the training set consists of $n - 1$ objects, and the evaluation set consists of the object left. To estimate the predictive ability we considered the gap between the experimental value (fitting) and the predicted one (cross-validation) for all the n objects which, one by one, are left out of the model. The final model is built using all the data in order to exploit as much information as possible. This validation method allows minimum perturbation, because each model is computed using, in practice, all the available examples. This is a good way to compare prediction for different techniques, but it can give an optimistic evaluation of the predictive ability, especially for data sets with a large number of samples.[21]

The results are expressed in NER% (nonerror rate), in fitting, on the objects present in the training set, and in prediction, on the objects present in the test set

$$\mathrm{NER\%} = \frac{\sum_g c_g}{n} \times 100 \qquad (5)$$

where $g = 1,..., G$ number of classes and $n =$ number of objects.

3. RESULTS AND DISCUSSION

Here we present the best classification models and analyze their performances. We carried out a preliminary study on the whole data set. Descriptors were selected using PCA, and an external test set was randomly extracted from the original pool for validation of the models.

The complete set of pesticides was randomly split into a training set of 165 and a test set of 70. We tried the classifiers discussed above. The results were best with LDA and RDA, which gave 62.8%.

To better the prediction ability of the models, we split the data set as in Table 1, reducing its heterogeneity but also the universality of the model. We studied different end-points to see how they influenced the models.

Classification Algorithms. About 200 models were computed, using, where possible, all the algorithms present in SCAN software and dividing the original data set into 31 daughter sets obtained considering the different chemical classes of the compounds (anilines, ureas, carbamates, halogen aromatic compounds, organophosphorus compounds, heterocyclic compounds, halogen compounds), the different pattern of use (herbicides, fungicides, insecticides), and the different end-points (trout, daphnia, rat, quail, duck) (Table 1).

Table 2 shows the results obtained for each of the 31 daughter sets.

CART and RDA appear to be the classification algorithms which best describe the pesticides considered, with the selected descriptors. NMC, and especially SIMCA, are the worst (Figure 1).

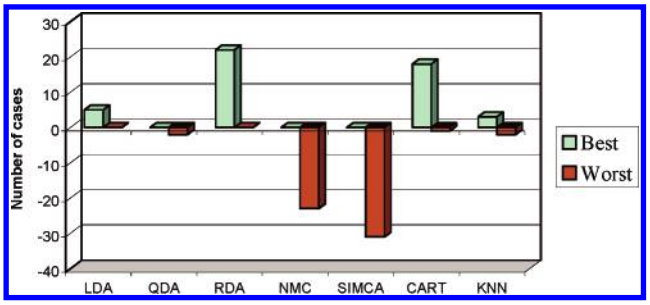

Figure 1. Evaluation of the algorithms' performances.

Table 3. Misclassification Matrix for Toxicity Prediction against Definition of Ureas Using SIMCA

		assigned class				
U-dap		1	2	3	4	no. of objects
real class	1	17				17
	2		5			5
	3			4		4
	4			2		2

Table 4. Misclassification Matrix for Toxicity Prediction against Definition of Ureas Using RDA and LDA

		assigned class				
U-dap		1	2	3	4	no. of objects
real class	1	17				17
	2	1	4			5
	3	1		2	1	4
	4				2	2

RDA has a further advantage. Because of the definition of RDA, the classification is easier to understand. The regularization was used to overcome the problems of linear and quadratic discriminant analysis, and by varying the parameters λ and γ, RDA became equal to LDA or QDA or NMC.

SIMCA is the algorithm most often afflicted with overfitting: the error on the training set is driven to a very small value, but the model is not able to predict new compounds. Indeed, it is the algorithm that most often has the biggest difference between NER% in fitting and NER% in cross-validation. CART also sometimes shows a large difference between NER% in fitting and NER% in cross-validation of chemicals (high gap in Table 2), and this is a limitation of this classification system.

Individual Models. We cannot discuss all the numerous individual models on chemical classes; below we just discuss a few examples.

The best result was on the urea data set toward *Daphnia magna*. It was obtained using SIMCA and the following five descriptors: relative number of C atoms, minimal partial charge for a C atom [Zefirov's PC], molecular volume/XYZ box, average bonding information content (order 1), and FNSA-3 fractional PNSA (PNSA-3/TMSA) [Zefirov's PC]. This model gave a NER% in fitting of 100.0 and a NER% in validation of 92.9. Table 3 shows the misclassification matrix for this model.

Although this model has a very high NER%, it does not correctly classify the molecules belonging to the most toxic class (class 4). Thus the models obtained by RDA and LDA are preferable, because they correctly assign the object in the most toxic class (Table 4).

This model has a worse NER% (96.4 in fitting and 89.3% in validation), but it is preferable because it errs on the side of the less toxic class. The most serious error is that a molecule belonging to the third class, *pencycuron*, is assigned in the first class. This is because *pencycuron* is the only molecule belonging to the third class with the carbonyl group typical of the first class.

Table 5. Misclassification Matrix for Toxicity Prediction against Definition of Halogen Aromatic Using RDA

		assigned class				
Ha-tr		1	2	3	4	no. of objects
real class	1	3		1		4
	2		6	1		7
	3			12		12
	4					0

On the halogenated aromatic compounds, using the data set toward trout, the best model was obtained using RDA ($\lambda = 0.50$; $\gamma = 0.25$). The descriptors selected were as follows: LogD pH5, molecular volume/XYZ box, average information content (order 0), min partial charge for a C atom [Zefirov's PC], average bonding information content (order 0), and XY shadow/XY rectangle. For this model the NER% was 91.3 in fitting and 91.3 in validation. Table 5 shows the misclassification matrix.

The result is good not only because of the few mistakes in classifying the object, but particularly because the wrong assignments are overestimates which makes them more acceptable. However, such a model may be less robust than other ones, because there are no compounds in the most toxic class, due to the limited number of halogenated aromatic compounds.

Good results were also obtained for the carbamates data set toward trout. We used RDA ($\lambda = 0.11$; $\gamma = 0.00$) and the following descriptors: relative number of H atoms, TMSA total molecular surface area [Zefirov's PC], RNCS relative negative charged SA (SAMNEG*RNCG), DPSA-1 difference in CPSAs (PPSA1-PNSA1) [Zefirov's PC], and information content (order 0). The NER% is 96.1 in fitting and 89.5 in validation.

The classification is good except for one molecule, *thiram*, which belongs in the fourth class but is assigned to the first. This serious error was made by all the classification techniques, so it is probably due to some peculiarity of this molecule that makes it an outlier.

CART was the best algorithm on the organophosphorus compounds. It used molecular surface area, radix index (order 0), HA dependent HDCA-2/TMSA [Zerfirov's PC], ZX shadow, and ZX shadow/ZX rectangle as molecular descriptors, giving a NER% of 83.7 in fitting and of 77.5 in validation.

Analyzing the erroneously classified molecules, we found two wrong toxicity values. *Pirimiphos methyl* and *pirimiphos ethyl* are two organophosphorus insecticides which differ only in the presence of an ethyl group instead of a methyl group. This difference does not seem enough to strongly modify their toxic action, and in fact they both are classified in the fourth class in our model. However, the toxic data that we used[11] placed *pirimiphos methyl* in the second class and *pirimiphos ethyl* in the fourth. A closer study of the literature confirmed the toxic similarity of both the compounds.[22]

Anilifos, belonging to the first class, was also classified in the fourth, but after a better bibliographic research its toxic

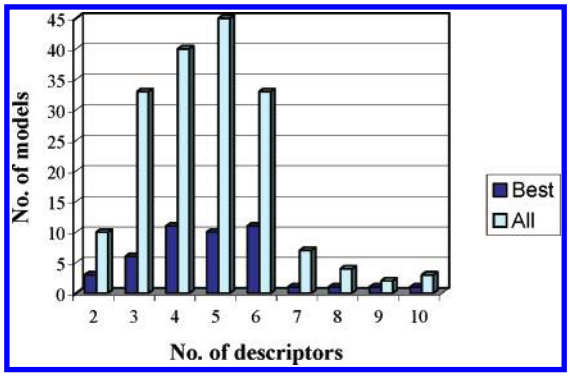

Figure 2. Number of descriptors used by the best models and by all models.

value proved wrong. The one we selected refers to a 3-h test, while for all the other compounds we used 48-h test. So in this case too the mistake in classification is due to a human error in the compilation of the data set and not to a bad performance of the model.

This shows an interesting ability of the QSAR models: they can identify outliers according to the model, and it may happen, as we have seen here, that the anomalous behavior of these compounds is actually due to a wrong literature value.

Number and Characteristics of the Descriptors. A major consideration regards the number of variables used. Figure 2 shows the number of models developed and the numbers of descriptors selected. Most of the models were obtained using just three, four, five, or six variables out of the over 150 descriptors, even if we consider only the models with the best performance.

PCA was done on the data set. In data sets with many variables, groups of variables often behave similarly. One reason is that more than one variable may be measuring the same driving principle governing the behavior of the system. Many systems have only a few such driving forces, but in the attempt to describe the chemical structures properly, an abundance of system variables are generated.

We can simplify our problem by replacing a group of variables with a single new variable. PCA is a quantitatively rigorous method for achieving this. The method generates a new set of variables, called principal components. Each principal component is a linear combination of the original variables. All the principal components are orthogonal to each other so there is no redundant information.

Table 6. Misclassification Matrix for Toxicity Prediction against Definition of Carbamates Using RDA

		assigned class				
C-tr		1	2	3	4	no. of objects
real class	1	1	1	1		3
	2		9			9
	3			11		11
	4	1			2	3

Figure 3 shows some of these descriptors are, in reality, closely related or not decisive in explaining the data. These loading plots allow an analysis of the influence of each variable on the principal components and their direct and inverse correlations. The coordinates of each variable are defined by the couple of loadings (autovectors of the covariance matrix) that each variable has in the two components considered. Big negative or positive loadings indicate that these variables are significantly represented in the component. Variables which are close in the loading plot bring common or similar information. Variables in the opposite position from the origin are in inverse relation.

LogD has special weight in these models, mostly for the organochlorinated pesticides, because it indicates how the compound can penetrate biological membranes. Many QSAR studies have used LogP in the past. This parameter is frequently used in models for aquatic toxicity and has been directly related to the phenomenon of narcosis. LogP is the logarithm of the partition coefficient of the un-ionized species between *n*-octanol and water, irrespective of pH, while LogD is referred to defined pH and in thus takes account of the pH effect. LogP and LogD are strongly related in our models.

The models were also very sensitive for the descriptors representing the molecular shape, especially for modeling anilines and organophosphorus compounds.

On the whole, only 40 descriptors were used frequently in all the models.

Table 8 lists the 40 most frequent descriptors and their nature: constitutional (C), geometrical (G), topological (T), electrostatic (E), quantum-chemical (Q), and physicochemical (PC).

Despite the unequal original distribution of the descriptors (i.e. there are many more electrostatic and topological descriptors than physicochemical ones), the most frequent descriptors in the best models are well distributed in each category (Figure 4).

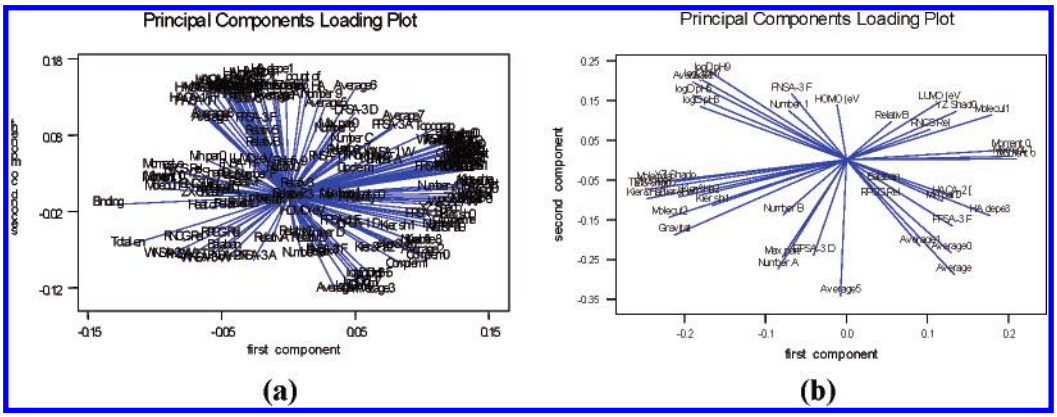

Figure 3. Loading plots on the plane of the first two principal components of all the 153 descriptors (a) and of the 40 most frequent descriptors in the models (b).

Table 7. Misclassification Matrix for Toxicity Prediction against Definition of Organophosphorus Ureas Using RDA

		assigned class				
OP-dap		1	2	3	4	no. of objects
real class	1	4			1	5
	2		4		2	6
	3		2	14	1	17
	4		3	2	16	21

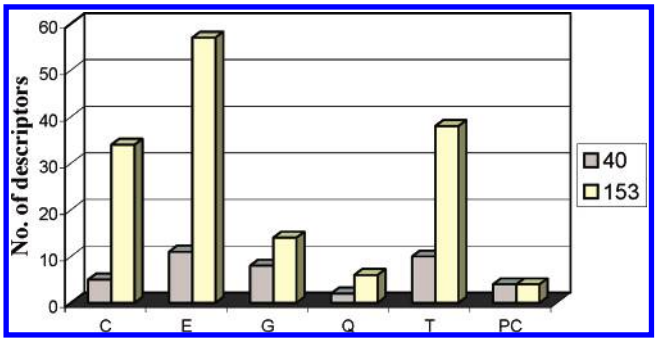

Figure 4. Distribution of the descriptors selected within the categories: constitutional (C), geometrical (G), topological (T), electrostatic (E), quantum-chemical (Q), physicochemical (PC).

This shows that the model is able to extract the correct information, even in the presence of high redundancy, as appears in Figure 3a. Nevertheless, we still do not know how much the system is affected by redundancy, but luckily redundancy can force the model toward some similar descriptors by chance.

This kind of observation indicates how important selection of the variables is. Using all the descriptors not only makes the computational procedure dull but also risks hiding the information on some relevant descriptors in the noise of all the others.

Often what determines the quality of the performances is not the algorithm used but selection of the variables. Other publications too[23−25] reach this conclusion.

End-point. Comparing the different end-points (Figure 5) we notice that results for daphnia and trout are similar: in some cases daphnia was better modeled, while in others trout gave better results. This is to be expected, since both kinds of toxicity refer to similar mechanisms. Rat toxicity seems harder to model: in five cases out of seven daphnia toxicity was modeled better. This can be explained by the greater complexity of toxicity mechanisms in the rat, while for aquatic toxicity LogP can explain a higher percentage.[26]

Congeneric Data Sets. The comparison of results on the whole data set of pesticides and on chemical classes confirms the difficulty of modeling compounds that differ widely in their structure and chemical activity. Historically, the first QSAR models were elaborated on similar substances. Models for different compounds have been developed,[5,6,27] but it is generally recognized that QSARs on molecules belonging to the same chemical class or with the same action mechanism have the best results.[28−30]

From Figure 6 it is clear that separation of the data set based on the target of the pesticide (herbicide, fungicide, insecticide) generally results in poor local models. In fact, this *classification* has no chemical or biological basis and merely relates to the pattern of use.

4. CONCLUSIONS

Many different QSAR models have been published, but they are very difficult to compare. Each model uses different data sets, descriptors, algorithms, and end-points. We considered the best out of hundreds of classifier models, with the same descriptors as starting point and data sets of pesticides for some end-points. This allowed us to compare results in relation to different aspects.

Descriptors. Analyzing the most widely used descriptors we see that all the different types were selected with high frequency and that LogP is one of the fundamental variables. Generally the parameters that represent hydrophobicity are important, since they indicates how the molecule can penetrate biological membranes.

In the test on the anilines data set HOMO and the Balaban index also have a fundamental role. They are best able to distinguish the agrochemicals from the other aniline compounds, while LogD and HOMO describe anilines belonging

Table 8. The Most Frequent Descriptors

	name		name
T	average bonding information content (order 0)	G	momentum of inertia B
T	average bonding information content (order 1)	G	momentum of inertia C
T	average bonding information content (order 2)	C	number of benzene rings
T	average complementary information content (order 0)	C	number of O atoms
T	average information content	C	number of P atoms
T	Balaban index	C	relative number of rings
E	DPSA-3 difference in CPSAs (PPSA3-PNSA3) [Zefirov's PC]	E	RNCS relative negative charged SA (SAMNEG*RNCG) [Zefirov's PC]
E	FNSA-3 fractional PNSA (PNSA-3/TMSA) [Zefirov's PC]	E	RPCG relative positive charge (QMPOS/QTPLUS) [Zefirov's PC]
E	FPSA-3 fractional PPSA (PPSA-3/TMSA) [Zefirov's PC]	E	TMSA total molecular surface area [Zefirov's PC]
G	gravitation index (all bonds)	G	YZ shadow
E	HA dependent HDCA-2/TMSA [Zefirov's PC]	G	YZ shadow/YZ rectangle
E	HACA-2 [Zefirov's PC]	G	ZX Shadow
T	Kier shape index (order 3)	E	max partial charge (Qmax) [Zefirov's PC]
T	Kier&Hall index (order1)	E	min partial charge for a C atom [Zefirov's PC]
T	Kier&Hall index (order2)	PC	LogD pH3
T	Kier&Hall index (order3)	PC	LogD pH5
G	molecular surface area	PC	LogD pH7.4
G	molecular volume/XYZ box	PC	LogD pH9
C	molecular weight	Q	HOMO (eV)
G	momentum of inertia A	Q	LUMO (eV)

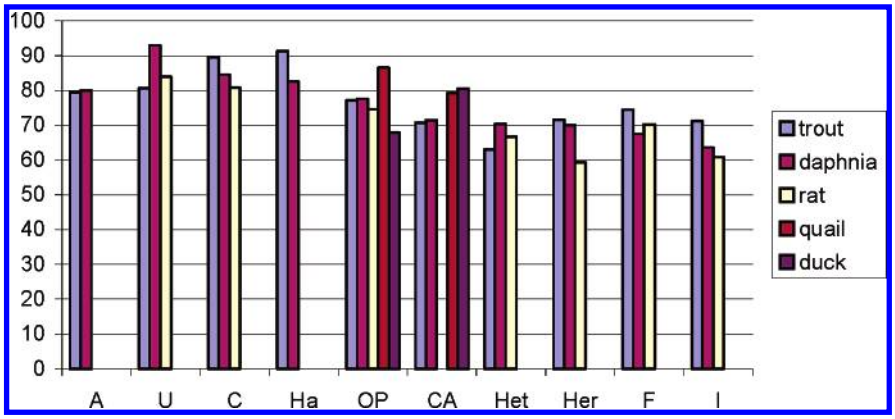

Figure 5. Results for different end-points.

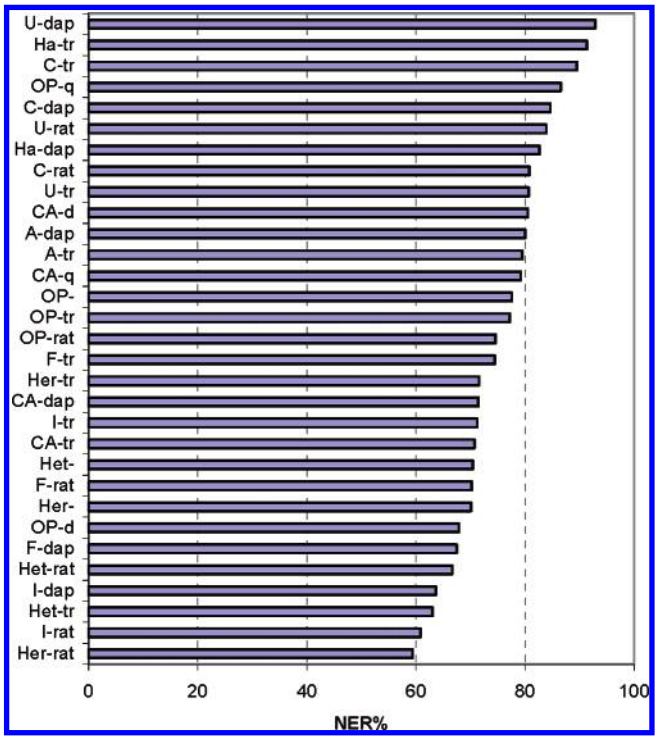

Figure 6. Best NER% for best models, labeled as in Table 1.

to the class II of polar narcotics according to refs 27 and 30. This result is confirmed by Ramos et al.[31] We found that the toxicity of the compounds with an effect greater than narcotic is well modeled by the hydrophobicity and the descriptors HOMO and LUMO. They indicate the ability to form hydrogen bonds, a parameter that seems to explain the higher baseline toxicity for class II substances.

Classification Algorithms. Then, we compared different algorithms used for the QSAR approach. The best were discriminant analysis (particularly RDA) and the classification tree (CART), but each model presents its own positive and negative aspects.

For discriminant analysis, which was particularly powerful and gives precise mathematical rules, we found important limitations: (i) if the data are not distributed homogeneously over the whole interval, discriminant analysis is not able to model the objects correctly; (ii) it involves the removal of variables with a constant value within the class, but if this attribute changes value between classes, it could be a good instrument for differentiation but discriminant analysis cannot recognize it, though other techniques do.

For LDA linearity could be a further weakness of the system because, even if it represents data with a linear correlation well, it is not able to model more complex relations. QDA was able to model nonlinear relations and was more robust to small deviations from normality, but attributes must not have nil variance within the class, and a large number of parameters must be calculated.

RDA is a combination of the two previous algorithms and has the advantage of improved performance for small data sets and with a high number of correlations and descriptors.

NMC does not consider the scale differences, so if the class centroids are not very spaced—as in our case—it often gives a poor performances.

SIMCA has the advantage of better performance, especially when there are a lot of variables and/or correlations without being enough degrees of freedom for the computation of QDA, but this often causes overfitting. A big disadvantage of this algorithm is due to its sensitivity to data scaling.

KNN gives good performance mainly when it can work with a large number of objects in the training set. This method is very effective when the surfaces of separation between classes are not linear and are complex. Limitations are related to the absence of a real mathematical model and the consequent empirical nature. Another defect is that it is sensitive to data scaling.

CART, was the best algorithm in our case, appeared insensitive to data scaling and robust to the presence of outliers, though it showed a tendency to overfit the data set.

Data Set. From this study it appeared to be easier to predict aquatic toxicity than mammalian or bird toxicity. This can be explained by the fact that aquatic toxicity is well modeled by LogP.

As expected, data sets of compounds with a homogeneous chemical basis are easier to model than heterogeneous ones. This apparently obvious statement calls for comment: (i) the homogeneity of the data set on a chemical basis can be misleading, because metabolism in the organism may produce a new chemical; (ii) on the other hand, this result provides support for those who prefer this approach rather than using data sets with homogeneous biological activity.

Discussion. The good performance of these models showed the high potential of the classification methods. Another application of the classification techniques and, more generally, of the QSAR approach involves analysis of the experimental data for a quick check for possible errors. This use, partially proposed by Hermens,[32] helps manage different laboratory values: the classification could highlight values

that need to be checked for mistakes on the basis of the model. The results should be carefully evaluated. For instance in the case of two very similar compounds, like two homogeneous substances in a series, with very different toxic values, we need to check the experimental data. However, toxicology is a very complex discipline. Two very similar compounds (from a human point of view) may show different toxicological behavior, but the model is not able to catch the complexity of the real world. This, for instance, is the case of p- and o-aminoanisole. The o-analogue is highly carcinogenic, while the p-analogue is not carcinogenic, because it undergoes a catabolic process that transforms it into a nontoxic metabolite.

Finally, in assessing performance of the model analysis of the error is as important as the model's ability to correctly assign the class. Overestimating the toxicity is much less dangerous than underestimating it. A critical observation of the outliers is fundamental to understanding the limits of the models or the presence of atypical molecules.

ACKNOWLEDGMENT

We acknowledge financial support from the European Commission for the DEMETRA project (Development of Environmental Modules for Evaluation of Toxicity of Pesticide Residues in Agriculture), contract QLK5-CT-2002-00691.

REFERENCES AND NOTES

(1) Franke, R.; Gruska, A.; Giuliani, A.; Benigni, R. Prediction of rodent carcinogenicity of aromatic amines: a quantitative structure−activity relationships model. *Carcinogenesis* **2001**, *22*, 1561−1571.
(2) Gini, G.; Lorenzini, M.; Benfenati, E.; Grasso, P.; 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. Comput. Sci.* **1999**, *39*, 1076−1080.
(3) Kabankin, A. S. B.; Kurlyandskii, A. Discriminant Analysis of the Relationship Between Topological Molecular Structure and Carcinogenicity of Aromatic Amines. *Pharm. Chem. J.* **2001**, *35*, 257−259.
(4) Gini, G.; Balestri, M.; Benfenati, E.; Pelagatti, S. Prediction of ecotoxicity of pesticides: comparison of multivariate analysis, neural network and classifiers. *Artificial Intelligence and Mathematics International Symposium*. 5−7 Jan. 2000, Ft. Lauderdale, U.S.A.
(5) Pintore, M.; Piclin, N.; Benfenati, E.; Giuseppina, G.; Chrétien, J. R. Database mining with adaptive fuzzy partition: application to the prediction of pesticide toxicity on rats. *Environ. Toxicol. Chem.* **2003**, *22*, 983−991.
(6) Pintore, M.; Piclin, N.; Benfenati, E.; Giuseppina, G.; Chrétien, J. R. Predicting toxicity against the fathead minnow by adaptive fuzzy partition. *QSAR Comb. Sci.* **2003**, *22*, 210−219.
(7) Fisher, R. A. The use of multiple measurements in taxonomic problems. *Ann. Eugenic*. **1936**, *7*, 179−188.
(8) Marengo, E.; Todeschini, R. Linear Discriminant Hierarchical Clustering: A modeling and Cross-Validate Divisive Clustering Method. *Chemom. Intell. Lab. Sys.* **1993**, *19*, 43−51.
(9) COMET, http://www.marionegri.it/ambsal/chem-toxi/comet.htm.
(10) Benfenati, E.; Piclin, N.; Roncaglioni, A.; Varì, M. R. Factors Influencing Predictive Models For Toxicology. *SAR QSAR Environ. Res.* **2001**, *12*, 593−603.
(11) Tomlin, C. D. S. *The Pesticide Manual*, 11th ed.; British Crop Protection Council: Berks, 1997.
(12) U.S. Environmental Protection Agency. ECOTOX User Guide: ECOTOXicology Database System. Version 3.0. 2002. Available: http:/www.epa.gov/ecotox/.
(13) *Micromedex, TOMES CPS System*. Copyright 1974−1998 Micromedex Inc. www.micromedex.com.
(14) Dearden, J. C. Physicochemical descriptors. In *Practical application of quantitative structure−activity relationships (QSAR) in environmental chemistry and toxicology*; Karcher, W., Devillers, Eds.; Kluwer: Dordrecht, 1990; pp 25−29.
(15) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. *CODESSA Comprehensive Descriptors for Structural and Statistical Analysis*; Reference Manual, version 2.0; Gainesville, FL, 1994.
(16) Sabljić, A. Topological indices and environmental chemistry. In *Practical application of quantitative structure−activity relationships (QSAR) in environmental chemistry and toxicology*; Karcher, W., Devillers, Eds.; Kluwer: Dordrecht, 1990; pp 61−82.
(17) Michie, D.; Spiegelhalter, D. J.; Taylor, C. C. *Machine learning, neural and statistical classification*; Hellis Horwood: 1994.
(18) *Scan: software for chemometric analysis. Reference Manual. Release 1 for Windows*; Minitab Inc.: State College, PA, 1995.
(19) Friedman, J. H. Regularized Discriminant Analysis. *J. Am. Statist. Assoc.* **1989**, *84*, 165.
(20) Wold, S.; Sjöström, M. "SIMCA: a method for analyzing chemical data in terms of similarity and analogy". In *Chemometrics Theory and Application*; Kowalski, B. R., Ed.; American Chemical Society Symposium Series 52; American Chemical Society: Washington, DC, 1977; pp 243−282.
(21) Helma, C.; Gottmann, E.; Kramer, S. Knowledge discovery and data mining in toxicology. *Statistical Methods Medical Res.* **2000**, *9*, 329−358.
(22) Vighi, M.; Masoero, M.; Calamari, D. QSARs for organophosphorus pesticides on Daphnia and honeybees. *Sci. Tot. Environ.* **1991**, *109−110*, 605−622.
(23) Basak, S. C.; Grunwald, G. D.; Host, G. E.; Niemi, G. J.; Bradbury, S. P. A comparative study of molecular similarity, statistical and neural network methods for predicting toxic modes of action. *Environ. Toxicol. Chem.* **1998**, *17*, 1056 1064.
(24) Drew, M. G. B.; Lumley, J. A.; Price, N. R. Predicting ecotoxicology or organophosphorous insecticides: Successful parameter selection with the genetic function algorithm. *Quant. Struct.-Act. Relat.* **1999**, *18*, 573−583.
(25) Mooney, R.; Shavlik, J.; Towell, G.; Gove, A. An experimental comparison of symbolic and connectionist learning algorithms. *IJCAI 89 Machine Learning* **1989**, 775−780.
(26) Cronin, M. T. D.; Dearden, J. C.; Duffy, J. C.; Edwards, R.; Manga, N.; Worth, A. P.; Worgan, A. D. P. The importance of hydrophobicity and electrophilicity descriptors in mechanistically based QSARs for toxicological endpoints. *SAR QSAR Environ. Sci.* **2002**, *13*, 167−176.
(27) ACS symposium series *Classical and Three-Dimensional QSAR in Agrochemistry*; Hermens, J. L. M., Verhaar, H. J. M.; American Chemical Society: Washington, DC, 1995.
(28) Ramos, E. U.; Vaes, W. H. J.; Verhaar, H. J. M.; Hermens, J. L. M. Polar narcosis: designing a suitable training set for QSAR studies. *Environ. Sci. Pollut. R.* **1997**, *4*, 83−90.
(29) Russom, C. L.; Bradbury, S. P.; Broderius, S. J.; Hammermeister, D. E.; Drummond, R. A. Predicting modes of toxic action from chemical structure: Acute toxicity in the Fathead Minnow (Pimephales Promelas). *Environ. Toxicol. Chem.* **1997**, *16*, 948−967.
(30) Verhaar, H. J. M.; Solbé, J.; Speksnijder, J.; van Leeuwen, C. J.; Hermens, J. L. M. Classifying environmental pollutants: Part 3. External validation of the classification system. *Chemosphere* **2000**, *40*, 875−883.
(31) Ramos, E. U.; Vaes, W. H. J.; Verhaar, H. J. M.; Hermens, J. L. M. Quantitative Structure−Activity Relationships for the Aquatic Toxicity of Polar and Nonpolar Narcotic Pollutants. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 845−852.
(32) Hermens, J. Prediction of environmental toxicity based on structure−activity relationships using mechanistic information. *Sci. Total Environ.* **1995**, *171*, 235−242.

CI034193W