Inter*net* Electronic Journal of Molecular Design

March 2002, Volume 1, Number 3, Pages 157–172

Editor: Ovidiu Ivanciuc

Special issue dedicated to Professor Alexandru T. Balaban on the occasion of the 70th birthday
Part 3

Guest Editor: Mircea V. Diudea

Support Vector Machine Identification of the Aquatic Toxicity Mechanism of Organic Compounds

Ovidiu Ivanciuc

Sealy Center for Structural Biology, Department of Human Biological Chemistry & Genetics, University of Texas Medical Branch, Galveston, Texas 77555–1157

Received: January 17, 2002; Accepted: February 19, 2002; Published: March 31, 2002

Citation of the article:

O. Ivanciuc, Support Vector Machine Identification of the Aquatic Toxicity Mechanism of Organic Compounds, *Internet Electron. J. Mol. Des.* **2002**, *1*, 157–172, http://www.biochempress.com.

Inter*net* Electronic Journal of Molecular Design

BioChem Press
http://www.biochempress.com

Support Vector Machine Identification of the Aquatic Toxicity Mechanism of Organic Compounds[#]

Ovidiu Ivanciuc*

Sealy Center for Structural Biology, Department of Human Biological Chemistry & Genetics, University of Texas Medical Branch, Galveston, Texas 77555–1157

Received: January 17, 2002; Accepted: February 19, 2002; Published: March 31, 2002

Internet Electron. J. Mol. Des. **2002**, *1* (3), 157–172

Abstract

Motivation. Because numerous organic chemicals can be environmental pollutants, considerable efforts were directed towards the study of the relationships between a compound's structure and its toxicity. Significant progress has been made to classify chemical compounds according to their mechanism of toxicity and to screen them for their environmental risk assessment. The prediction of the mechanism of action using structural descriptors has major applications in selecting the appropriate quantitative structure–activity relationships (QSAR) model, to identify chemicals with similar toxicity mechanism, and in extrapolating toxic effects between different species and exposure regimes.

Method. Support vector machine (SVM) is a new machine learning algorithm that found numerous applications in various classification studies. In this study we have investigated the application of SVM for the recognition of the aquatic toxicity mechanism of 88 organic compounds. For each compound, the chemical structure was encoded by four structural descriptors, namely the octanol–water partition coefficient log K_{ow}, the energy of the highest occupied molecular orbital E_{HOMO}, the energy of the lowest unoccupied molecular orbital E_{LUMO}, and the average acceptor superdelocalizability S^{N}_{av}.

Results. Extensive simulations using the dot, polynomial, radial basis function, neural, and anova kernels demonstrate that the classification performances of SVM depend strongly on the kernel type and various parameters that control the kernel shape. The best prediction results were obtained with a polynomial kernel of degree 2.

Conclusions. Support vector machines represent a powerful and flexible classification algorithm, with many potential applications in QSAR and molecular design. The results reported in the present study demonstrate such an application in the identification of the aquatic toxicity mechanism.

Keywords. Support vector machines; structure–toxicity relationships; aquatic toxicity; mechanism of action.

Abbreviations and notations

MOA, mechanism of action	E_{LUMO}, energy of the lowest unoccupied molecular orbital
SVM, support vector machines	log K_{ow}, octanol–water partition coefficient
E_{HOMO}, energy of the highest occupied molecular orbital	S^{N}_{av}, average acceptor superdelocalizability

1 INTRODUCTION

The quantitative structure–activity relationships (QSAR) models consider that the physical, chemical, and biological properties of a chemical compound are related to, and can be modeled

[#] Dedicated on the occasion of the 70th birthday to Professor Alexandru T. Balaban.

* Correspondence author; E–mail: ivanciuc@netscape.net.

from from, the molecular structure of that compound. In the field of aquatic toxicology QSAR is a reliable scientific tool for modeling the toxic effect of organic compounds and for predicting the ecological risk associated with new, not yet tested compounds [1]. Using QSAR models and comprehensive investigation of the chemical reactivity of organic compounds, significant progress has been made to classify chemical compounds according to their mechanism of toxicity and to screen them for their environmental risk assessment. Using theoretical descriptors computed from the molecular structure and various classification algorithms, the prediction of the mechanism of action (MOA) has major applications in identifying chemicals with similar toxicity mechanism, in selecting the appropriate QSAR model, and in extrapolating toxic effects across species and exposure regimes when limited experimental data are available [2–13].

A structure–toxicity study compared the toxicity of various organic compounds with different MAOs for the fathead minnow (*Pimephales promelas*) and *Tetrahymena pyriformis* using hydrophobicity and electronic indices as numerical descriptors for the chemical structure [8]. In a recent study, discriminant analysis and logistic regression were used for the same set of compounds to discriminate between narcotic and reactive MOA [13]. Support vector machines (SVM) represent a new class of machine learning algorithms that found numerous applications in various classification and regression models. In this study we have investigated the application of SVM for the recognition of the aquatic toxicity mechanism for the compounds previously explored in Refs. [8] and [13]. The influence of the kernel type on the SVM performances was extensively explored using various kernels, namely the dot, polynomial, radial basis function, neural, and anova kernels.

2 SUPPORT VECTOR MACHINES

The support vector machines were developed by Vapnik [14–16] as a powerful tool for pattern classification in two classes by determining an optimal hyperplane that separates the classes [17,18]. The SVM algorithm generates a separating hypersurface in the input space that optimally separates two classes of patterns. In the first step, using various kernels that perform a nonlinear mapping, the input space is transformed into a higher dimensional feature space. Then, a maximal margin hyperplane (MMH) is computed in the feature space. MMH maximizes the distance to the hyperplane of the closest patterns from the two classes. This powerful classification technique found interesting applications in molecular modeling: recognition of translation initiation sites [19], cancer diagnosis [20–22], identification of HIV protease cleavage sites [23], protein class prediction [24], protein–protein interactions [25], protein subcellular localization [26,27], protein fold recognition [28], protein secondary structure prediction [29], and DNA hairpins recognition [30]. In this section the SVM algorithm is outlined first for the linearly separable case. When complete separation into two classes is not desirable due too significant errors in the data the SVM uses slack variables. Finally, kernel functions are introduced for patterns with a non–linear separation surface.

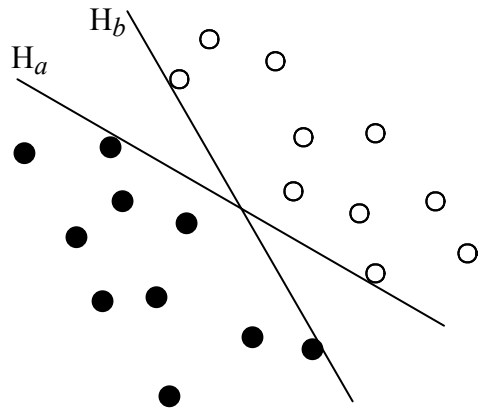

Figure 1. Two possible hyperplanes H_a and H_b that discriminate between patterns from the class +1 (black circles) and –1 (white circles).

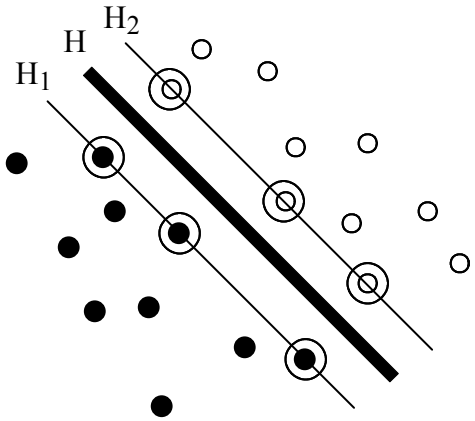

Figure 2. Example of patterns from the class +1 (black circles) and –1 (white circles) linearly separable by the maximal margin hyperplane H. The support vectors from the class +1 define the hyperplane H_1 while those from the class –1 define the hyperplane H_2.

Let S be a set of l vectors $x_i \in R^n$, $i = 1, 2, \ldots, l$, in an n–dimensional space. Each vector x_i belongs to either of two classes identified by the label $y_i \in \{-1, +1\}$. If the two classes are linearly separable, then there exists a hyperplane that divides the set S leaving all the vectors of the same class on the same side. However, as one can see from Figure 1, this hyperplane is not unique because both hyperplanes H_a and H_b discriminate between patterns from class +1 (black circles) and –1 (white circles), and between them one can find an infinite number of hyperplanes with the same property. This is a well–known problem in chemometrics, and various pattern recognition methods were devised to solve it. SVM is a new approach to find a unique hyperplane that maximizes the separation between the two classes of patterns, as depicted in Figure 2. The maximal margin hyperplane (MMH) H is defined by $w \cdot x + b = 0$, where w is the normal to the hyperplane, $b/\|w\|$ the perpendicular distance to the origin and $\|w\|$ the Euclidean norm of w. The +1 class of patterns is bordered by the hyperplane H_1 defined by $w \cdot x + b = +1$, while the –1 class of patterns is bordered by the hyperplane H_2 defined by $w \cdot x + b = -1$. Hyperplanes H, H_1, and H_2 are parallel and no patterns are situated between H_1 and H_2. The +1 patterns that are situated on H_1 and the –1 patterns that are situated on H_2 are the support vectors, depicted in Figure 2 within a larger circle. These support vectors are used to define the separating hyperplane. Let d_+ be the shortest distance from the separating hyperplane H to the closest positive pattern, and d_- be the shortest distance from the separating hyperplane H to the closest negative pattern. The distance between H_1 and H_2 defines the margin, equal to $d_+ + d_-$. Because $d_+ = d_- = 1/\|w\|$, the margin is $2/\|w\|$.

The maximal margin hyperplane H is computed as the solution to the following problem:

$$\begin{cases} \text{minimize } \dfrac{1}{2}\|w\|^2 \\ \text{with } y_i(w \cdot x_i + b) \geq 1 \quad (i = 1, 2, \ldots, l) \end{cases} \tag{1}$$

The above equation is a quadratic programming problem, solved by the Karush–Kuhn–Tucker

theorem. If we denote by $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_l)$ the l non–negative Lagrange multipliers associated with the constraints, the solution to the problem from Eq. (1) is equivalent to determining the solution of the following Wolfe dual problem:

$$\begin{cases} \text{maximize } \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j (x_i \cdot x_j) y_i y_j \\ \text{with } \sum_i \alpha_i y_i = 0 \qquad\qquad \alpha_i \geq 0 \end{cases} \tag{2}$$

The solution for w is:

$$w = \sum_i \alpha_i y_i x_i \tag{3}$$

The only α_i that can have a nonzero value in equation (3) are those for which the constraints of the first problem are satisfied with the equality sign. For patterns that can be easily separated with a linear decision plane most of the α_i are usually null, and the vector w is a linear combination of a small percentage of the vectors x_i. These vectors from both the +1 and –1 classes are termed support vectors (depicted in Figure 2 within a larger circle) and they are the only vectors of S needed to determine the MMH. In real applications, good SVM models are obtained by using a small fraction of the +1 and –1 patterns. The problem of classifying a new data vector x is now simply solved by looking at the sign of $w \cdot x + b$ with b obtained from the Karush–Kuhn–Tucker conditions.

In real applications, it may happen that a significant fraction or even almost all x vectors are used as support vectors. This situation can appear from various causes, such as poor descriptors selected in the x vector or experimental errors in determining the class y. By forcing the algorithm to find a perfect separation hypersurface for the +1 and –1 classes, too many support vectors are used and the solution obtained is overfitted, giving erroneous predictions for patterns not used in obtaining the SVM model. The above classification model cannot be applied whenever, due to the partial overlapping of the +1 and –1 classes, a separating hypersurface does not exist. To deal with these problems, the perfect separation of the +1 and –1 classes is relaxed, and the SVM is extended to deal with imperfect separation cases. By introducing l non–negative slack variables $\xi = (\xi_1, \xi_2, \ldots, \xi_l)$ a hyperplane is defined by minimizing the trade–off between margin and training error:

$$y_i (w \cdot x_i + b) \geq 1 - \xi_i \quad (i = 1, 2, \ldots, l) \tag{4}$$

The solution to the problem:

$$\begin{cases} \text{minimize } \frac{1}{2}\|w\|^2 + C\sum_i \xi_i \\ \text{with } y_i (w \cdot x_i + b) \geq 1 - \xi_i \quad (i = 1, 2, \ldots, l) \end{cases} \tag{5}$$

is called the soft margin separating hyperplane (SMSH). The vectors satisfying the above constraints with the equality sign are called support vectors and are the only vectors needed to determine the decision surface. Similarly to the linearly separable case, the dual formulation

requires the solution of a quadratic programming problem with linear constraints:

$$\begin{cases} \text{maximize } \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j (x_i \cdot x_j) y_i y_j \\ \text{with } \sum_i \alpha_i y_i = 0 \qquad\qquad 0 \le \alpha_i \le C \end{cases} \tag{6}$$

where C is a capacity parameter. The above formulation of the separation hypersurface allows for the presence of +1 or –1 patterns in the margin of the hyperplane (between hyperplanes H_1 and H_2 from Figure 2), or for the presence of +1 patterns in the –1 region bordered by H_2, or for the presence of –1 patterns in the +1 region bordered by H_1.

When each vector x in input space is mapped into a vector $z = \Phi(x)$ in a higher dimensional feature space, the above classification method can be extended to include nonlinear separating hypersurfaces. The Lagrangian function in the high dimensional feature space is:

$$L = \sum_i \alpha_i - 0.5 \sum_{i,j} \alpha_i \alpha_j y_i y_j \left\langle \Phi(x_i), \Phi(x_j) \right\rangle \tag{7}$$

The dot product $\langle \Phi(x), \Phi(y) \rangle$ in feature space is substituted with a kernel function $K(x,y)$. The kernel functions usually used in pattern clustering are the dot kernel, the polynomial kernel, and the radial basis function kernel. With a suitable kernel, SVM can separate in the feature space the data that in the original input space was non–separable. The non–negative slack variables $\xi_i \ge 0$ are used to control the imperfect separation for non–linear separation surfaces, by using a penalty constant C that sets the degree of penalty for patterns situated between H_1 and H_2 or misclassified:

$$\begin{cases} \text{maximize } \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \left\langle \Phi(x_i), \Phi(x_j) \right\rangle \\ \text{with } \sum_i \alpha_i y_i = 0 \qquad\qquad 0 \le \alpha_i \le C \end{cases} \tag{8}$$

More details on SVM can be found in references [14–18], while their various applications in molecular design are found in [19–30]. All SVM models from the present paper for the classification of the aquatic toxicity of organic compounds were obtained with mySVM [31], which is freely available for download. Links to Web resources related to SVM, namely tutorials, papers and software, can be found in BioChem Links [32] at http://www.biochempress.com.

3 MATERIALS AND METHODS

The 88 compounds investigated in the present study were taken from two recent studies [8,13] and are presented in Table 1 together with the four theoretical descriptors used to discriminate between their mechanism of action, namely the octanol–water partition coefficient log K_{ow}, the energy of the highest occupied molecular orbital E_{HOMO}, the energy of the lowest unoccupied molecular orbital E_{LUMO}, and the average acceptor superdelocalizability S^N_{av}. The compounds are

classified either as narcotics, which include non–polar and polar narcotics, or reactive compounds, which included respiratory uncouplers, soft electrophiles, and proelectrophiles. The data set consists of 48 narcotic compounds (class +1) and 40 reactive compounds (class –1).

Table 1. Structure of the Chemical Compounds, Theoretical Descriptors (log K_{ow}, S^{N}_{av}, E_{HOMO}, and E_{LUMO}) and Mechanism of Toxic Action (Narcotic +1, Reactive –1)

No	Chemical compound	log K_{ow}	S^{N}_{av}	E_{HOMO}	E_{LUMO}	MOA
1	Phenol	1.46	0.285	–9.17	0.29	+1
2	*o*–Cresol	2.12	0.285	–9.04	0.31	+1
3	*p*–Cresol	1.94	0.285	–8.95	0.33	+1
4	2,4–Dimethylphenol	2.30	0.285	–8.85	0.35	+1
5	2,4,6–Trimethylphenol	3.42	0.286	–8.90	0.28	+1
6	2,3,6–Trimethylphenol	3.42	0.286	–8.90	0.28	+1
7	4–Ethylphenol	2.58	0.285	–9.01	0.32	+1
8	4–Propylphenol	3.18	0.285	–8.99	0.33	+1
9	2–Methyl–3–butyn–2–ol	0.33	0.285	–10.98	1.80	–1
10	2–Allylphenol	2.64	0.285	–9.11	0.26	+1
11	4–*Tert*–butylphenol	3.31	0.286	–8.99	0.36	+1
12	4–*Tert*–pentylphenol	3.98	0.286	–8.98	0.36	+1
13	4–Phenylphenol	3.36	0.285	–8.87	–0.09	+1
14	Catechol	0.88	0.310	–8.92	0.24	–1
15	Resorcinol	0.80	0.287	–9.06	0.27	–1
16	3–Methoxyphenol	1.58	0.287	–9.25	0.13	+1
17	4–Methoxyphenol	1.34	0.288	–9.11	0.17	+1
18	4–Phenoxyphenol	3.75	0.292	–8.91	0.09	+1
19	2–Chlorophenol	2.15	0.299	–9.04	0.03	+1
20	4–Chlorophenol	2.48	0.299	–9.01	0.05	+1
21	4–Chloro–3–methylphenol	3.10	0.299	–8.95	0.05	+1
22	4–Chlorocatechol	1.97	0.325	–8.88	0.00	–1
23	2,4–Dichlorophenol	2.92	0.322	–9.01	–0.19	+1
24	2,4,6–Trichlorophenol	3.69	0.327	–9.13	–0.44	+1
25	2,3,4,6–Tetrachlorophenol	4.45	0.339	–9.19	–0.68	–1
26	2,3,4,5–Tetrachlorophenol	4.21	0.338	–9.05	–0.59	–1
27	Tetrachlorocatechol	4.29	0.340	–9.05	–0.62	–1
28	2,4,6–Tribromophenol	4.02	0.320	–9.56	–0.80	+1
29	2–Nitrophenol	1.85	0.311	–9.90	–1.22	–1
30	2,6–Dinitrophenol	1.91	0.339	–10.69	1.96	–1
31	2,5–Dinitrophenol	1.75	0.340	–10.54	–2.28	–1
32	2,4–Dinitrophenol	1.54	0.340	–10.79	–1.88	–1
33	*Tert*–butyldinitrophenol	3.36	0.339	–10.43	–1.84	–1
34	4,6–Dinitro–*o*–cresol	2.56	0.339	–10.53	–1.82	–1
35	Aniline	0.90	0.279	–8.61	0.42	+1
36	4–Toluidine	1.39	0.279	–8.46	0.40	+1
37	4–Ethylaniline	1.96	0.279	–8.50	0.39	+1
38	4–Butylaniline	3.15	0.279	–8.50	0.38	+1
39	3–Benzyloxyaniline	2.79	0.280	–9.12	0.11	+1
40	4–Hexyloxyaniline	3.66	0.283	–8.57	0.24	+1
41	2–Chloroaniline	1.90	0.292	–8.66	0.13	+1
42	2,3,4–Trichloroaniline	3.33	0.318	–8.71	–0.26	+1
43	2,3,5,6–Tetrachloroaniline	4.10	0.331	–8.91	–0.55	–1
44	2,3,4,5,6–Pentafluoroaniline	2.22	0.336	–9.52	–1.15	–1
45	α,α,α–4–Tetrafluoro–3–toluidine	2.62	0.308	–9.18	–0.72	+1
46	α,α,α–4–Tetrafluoro–2–toluidine	2.62	0.308	–10.42	–1.12	+1
47	4–Ethoxy–2–nitroaniline	2.47	0.308	–9.04	–1.06	–1
48	Isopropylbenzene	3.66	0.282	–9.53	0.38	+1
49	1,2,4–Trimethylbenzene	3.78	0.283	–9.09	0.38	+1
50	Butylbenzene	4.26	0.282	–9.50	0.36	+1

Table 1. (Continued)

No	Chemical compound	log K_{ow}	S^N_{av}	E_{HOMO}	E_{LUMO}	MOA
51	Amylbenzene	4.91	0.282	–9.53	0.36	+1
52	Biphenyl	4.09	0.283	–9.18	–0.10	+1
53	Chlorobenzene	2.86	0.296	–9.39	0.06	+1
54	1,2–Dichlorobenzene	3.38	0.309	–9.29	–0.17	+1
55	1,2,4–Trichlorobenzene	4.02	0.324	–9.23	–0.43	+1
56	3,4–Dichlorotoluene	4.22	0.310	–9.16	–0.16	+1
57	Pentachloropyridine	4.34	0.344	–9.37	–1.02	–1
58	Bromobenzene	2.99	0.294	–9.81	–0.05	+1
59	Nitrobenzene	1.85	0.309	–10.60	–1.13	–1
60	3–Nitrotoluene	2.45	0.309	–10.27	–1.09	–1
61	1–Fluoro–4–nitrobenzene	1.80	0.321	–10.85	–1.41	–1
62	1–Chloro–3–nitrobenzene	2.41	0.324	–10.06	–1.31	–1
63	1–Chloro–2–nitrobenzene	2.24	0.323	–9.99	–1.19	–1
64	1,4–Dinitrobenzene	1.46	0.339	–11.31	–2.25	–1
65	2,4–Dinitrotoluene	2.00	0.337	–11.18	–1.84	–1
66	1,3–Dichloro–4,6–dinitrobenzene	2.49	0.367	–10.63	–2.08	–1
67	2–Butyn–1–ol	0.16	0.304	–10.26	1.66	–1
68	*Cis*–3–hexen–1–ol	1.34	0.286	–9.89	1.02	–1
69	1–Hexen–3–ol	1.12	0.297	–10.50	0.94	–1
70	4–Pentyn–2–ol	–0.08	0.296	–10.73	1.95	–1
71	2–Phenyl–3–butyn–2–ol	1.68	0.285	–9.76	0.17	–1
72	2–Propyn–1–ol	–0.37	0.304	–10.66	1.73	–1
73	3–Butyn–2–ol	–0.06	0.302	–10.80	1.77	–1
74	2–Decyn–1–ol	3.33	0.304	–10.21	1.63	–1
75	3–Butyn–1–ol	–0.50	0.296	–10.86	1.79	–1
76	2–Butyn–1,4–diol	–1.83	0.306	–10.21	1.52	–1
77	4–Chloroaniline	1.83	0.292	–8.59	0.10	+1
78	4–Bromoaniline	2.26	0.290	–8.78	0.06	+1
79	4–Fluoroaniline	1.15	0.289	–8.74	0.07	+1
80	4–Nitroaniline	1.31	0.304	–9.42	–1.01	–1
81	Pentachlorophenol	5.12	0.351	–9.14	–0.79	–1
82	4–Nonylphenol	6.36	0.285	–9.02	0.31	–1
83	Pentabromophenol	5.74	0.340	–9.63	–1.14	–1
84	3,4–Dichloroaniline	2.69	0.305	–8.67	–0.11	+1
85	4–Octylaniline	5.27	0.279	–8.51	0.38	+1
86	2,4,6–Tri(*tert*)butylphenol	6.95	0.289	–9.04	–0.43	+1
87	2,6–Di(*tert*)butyl–4–methylphenol	6.07	0.288	–8.70	0.39	+1
88	4–Nitrophenol	1.91	0.312	–10.17	–1.08	–1

The discriminant analysis for the 88 compounds in Table 1 used the four theoretical descriptors, namely log K_{ow}, E_{HOMO}, E_{LUMO}, and S^N_{av}, to obtain a good classification in polar or reactive compounds [13]. Six out of the 48 narcotic compounds were erroneously classified as reactive compounds, namely: 2,4–dichlorophenol, 2,4,6–trichlorophenol, 2,4,6–tribromophenol, α,α,α–4–tetrafluoro–3–toluidine, α,α,α–4–tetrafluoro–2–toluidine, and 2,4,6–tri(*tert*)butylphenol. From the 40 reactive compounds, four were erroneously classified as narcotic compounds, namely: catechol, resorcinol, 2–phenyl–3–butyn–2–ol, and 4–nonylphenol.

The variable selection in the logistic regression showed that E_{LUMO} does not improve the classification [13]. Using three structural descriptors (log K_{ow}, E_{HOMO}, and S^N_{av}) the logistic regression gives classification results similar with those obtained with the discriminant analysis. From the 48 narcotic compounds, four were erroneously classified as reactive compounds, namely:

2,4–dichlorophenol, 2,4,6–trichlorophenol, 2,4,6–tribromophenol, α,α,α–4–tetrafluoro–2–toluidine, and 1,2,4–trichlorobenzene. Four out of the 40 reactive compounds were erroneously classified as narcotic compounds, namely: resorcinol, 4–ethoxy–2–nitroaniline, 2–phenyl–3–butyn–2–ol, and 4–nonylphenol. All SVM models for the classification of the 88 organic compounds from Table 1 into narcotic and reactive were obtained with mySVM [31] using the same four structural descriptors from ref. [13], namely log K_{ow}, E_{HOMO}, E_{LUMO}, and S^{N}_{av}. Before computing the SVM model, the input vectors were scaled to zero mean and unit variance. The prediction power of each SVM model was evaluated with a leave–10%–out cross–validation procedure, and the capacity parameter C took the values 10, 100, and 1000. We present below the kernels and their parameters used in this study.

The dot kernel. The inner product of x and y defines the dot kernel:

$$K(x, y) = x \cdot y \tag{9}$$

The polynomial kernel. The polynomial of degree d (values 2, 3, 4, and 5) in the variables x and y defines the polynomial kernel:

$$K(x, y) = (x \cdot y + 1)^d \tag{10}$$

The radial kernel. The following exponential function in the variables x and y defines the radial basis function kernel, with the shape controlled by the parameter γ (values 0.5, 1.0, and 2.0):

$$K(x, y) = \exp(-\gamma \parallel x - y \parallel^2) \tag{11}$$

The neural kernel. The hyperbolic tangent function in the variables x and y defines the neural kernel, with the shape controlled by the parameters a (values 0.5, 1.0, and 2.0) and b (considered 0):

$$K(x, y) = \tanh(ax \cdot y + b) \tag{12}$$

The anova kernel. The sum of exponential functions in x and y defines the anova kernel, with the shape controlled by the parameters γ (values 0.5, 1.0, and 2.0) and d (values 1, 2, and 3):

$$K(x, y) = \left(\sum_i \exp(-\gamma(x_i - y_i)) \right)^d \tag{13}$$

4 RESULTS AND DISCUSSION

Similarly with other multivariate statistical models, the performances of SVM classifiers in structure–activity studies depend on the combination of several parameters, and the kernel type is the most important one. Because the use of SVM models in chemometrics, structure–activity studies, and QSAR is only in the beginning, there are no clear guidelines on selecting the most effective kernel for a certain classification problem. Another important problem in SVM applications for structure–activity models is the selection of those structural descriptors that can discriminate the investigated set of compounds. For the moment, this is an unexplored problem, and in this study we have used four structural descriptors (log K_{ow}, S^{N}_{av}, E_{HOMO}, and E_{LUMO}) from [13].

Table 2. Results for SVM Modeling of the Mechanism of Aquatic Toxicity Using log K_{ow}, S^{N}_{av}, E_{HOMO}, and E_{LUMO}.[a]

Exp	*C*	*K*		SV	BSV	+/+	+/–	–/–	–/+	CAa	CAp	ASV	ABSV	TRa	TRp	TEa	TEp
1	10	D		25	20	47	1	27	13	0.84	0.78	23.1	17.7	0.85	0.80	0.84	0.76
2	100			29	19	47	1	31	9	0.89	0.84	22.6	17.6	0.85	0.79	0.84	0.76
3	1000			30	19	47	1	32	8	0.90	0.85	23.3	17.3	0.86	0.80	0.86	0.79
			d														
4	10	P	2	14	0	48	0	34	6	0.93	0.89	16.2	0.0	0.94	0.90	0.92	0.88
5	100		2	14	0	48	0	34	6	0.93	0.89	16.2	0.0	0.94	0.90	0.92	0.88
6	1000		2	14	0	48	0	34	6	0.93	0.89	16.2	0.0	0.94	0.90	0.92	0.88
7	10		3	24	0	48	0	40	0	1.00	1.00	20.4	0.0	0.98	0.97	0.85	0.80
8	100		3	24	0	48	0	40	0	1.00	1.00	20.4	0.0	0.98	0.97	0.85	0.80
9	1000		3	24	0	48	0	40	0	1.00	1.00	20.4	0.0	0.98	0.97	0.85	0.80
10	10		4	21	0	48	0	33	7	0.92	0.87	20.2	0.0	0.98	0.98	0.87	0.81
11	100		4	21	0	48	0	33	7	0.92	0.87	20.2	0.0	0.98	0.98	0.87	0.81
12	1000		4	21	0	48	0	33	7	0.92	0.87	20.2	0.0	0.98	0.98	0.87	0.81
13	10		5	24	0	48	0	32	8	0.91	0.86	21.8	0.0	0.95	0.93	0.87	0.81
14	100		5	24	0	48	0	32	8	0.91	0.86	21.8	0.0	0.95	0.93	0.87	0.81
15	1000		5	24	0	48	0	32	8	0.91	0.86	21.8	0.0	0.95	0.93	0.87	0.81
			γ														
16	10	R	0.5	23	2	48	0	29	11	0.88	0.81	25.4	1.5	0.91	0.86	0.82	0.75
17	100		0.5	24	0	48	0	32	8	0.91	0.86	26.4	0.0	0.90	0.84	0.83	0.76
18	1000		0.5	24	0	48	0	32	8	0.91	0.86	26.4	0.0	0.90	0.84	0.83	0.76
19	10		1.0	37	1	48	0	28	12	0.86	0.80	36.5	0.9	0.88	0.83	0.75	0.69
20	100		1.0	37	0	48	0	29	11	0.88	0.81	35.9	0.0	0.88	0.83	0.75	0.69
21	1000		1.0	37	0	48	0	29	11	0.88	0.81	35.9	0.0	0.88	0.83	0.75	0.69
22	10		2.0	53	0	48	0	28	12	0.86	0.80	51.0	0.0	0.91	0.87	0.66	0.61
23	100		2.0	53	0	48	0	28	12	0.86	0.80	51.0	0.0	0.91	0.87	0.66	0.61
24	1000		2.0	53	0	48	0	28	12	0.86	0.80	51.0	0.0	0.91	0.87	0.66	0.61
			a														
25	10	N	0.5	19	16	40	8	31	9	0.81	0.82	19.5	16.7	0.78	0.78	0.78	0.82
26	100		0.5	18	16	40	8	31	9	0.81	0.82	18.9	15.9	0.77	0.78	0.76	0.75
27	1000		0.5	18	16	40	8	31	9	0.81	0.82	17.8	15.0	0.79	0.79	0.79	0.75
28	10		1.0	24	22	38	10	28	12	0.75	0.76	21.0	18.6	0.75	0.77	0.80	0.83
29	100		1.0	24	21	38	10	28	12	0.75	0.76	20.7	18.2	0.75	0.77	0.76	0.82
30	1000		1.0	24	21	38	10	28	12	0.75	0.76	20.6	18.1	0.76	0.77	0.76	0.82
31	10		2.0	28	25	35	13	26	14	0.69	0.71	22.7	20.4	0.73	0.74	0.77	0.82
32	100		2.0	27	25	35	13	26	14	0.69	0.71	21.7	19.6	0.74	0.75	0.78	0.82
33	1000		2.0	27	25	35	13	26	14	0.69	0.71	21.7	19.5	0.74	0.75	0.78	0.82
			γ	*d*													
34	10	A	0.5	1	21	7	48	0	35	5	0.94	0.91	18.6	6.3	0.94	0.90	0.87	0.84
35	100		0.5	1	17	1	48	0	33	7	0.92	0.87	16.7	0.3	0.94	0.90	0.85	0.80
36	1000		0.5	1	17	0	48	0	33	7	0.92	0.87	16.6	0.0	0.94	0.90	0.85	0.80
37	10		1.0	1	22	5	48	0	31	9	0.90	0.84	20.6	3.1	0.94	0.90	0.84	0.78
38	100		1.0	1	19	0	48	0	40	0	1.00	1.00	18.2	0.0	0.97	0.96	0.87	0.81
39	1000		1.0	1	19	0	48	0	40	0	1.00	1.00	18.2	0.0	0.97	0.96	0.87	0.81
40	10		2.0	1	27	0	48	0	31	9	0.90	0.84	25.7	0.0	0.95	0.92	0.86	0.79
41	100		2.0	1	27	0	48	0	31	9	0.90	0.84	25.7	0.0	0.95	0.92	0.86	0.79
42	1000		2.0	1	27	0	48	0	31	9	0.90	0.84	25.7	0.0	0.95	0.92	0.86	0.79
43	10		0.5	2	23	0	48	0	32	8	0.91	0.86	20.4	0.0	0.97	0.95	0.85	0.78
44	100		0.5	2	23	0	48	0	32	8	0.91	0.86	20.4	0.0	0.97	0.95	0.85	0.78
45	1000		0.5	2	23	0	48	0	32	8	0.91	0.86	20.4	0.0	0.97	0.95	0.85	0.78

[a] The table reports the experiment number Exp, capacity parameter *C*, kernel type *K* (dot D; polynomial P; radial basis function R; neural N; anova A) and corresponding parameters, calibration results (SV, number of support vectors; BSV, number of bounded support vectors; +/+, number of +1 patterns (narcotic compounds) classified in class +1; +/–, number of +1 patterns classified in class –1; –/–, number of –1 patterns (reactive compounds) classified in class –1; –/+, number of –1 patterns classified in class +1; CAa, accuracy; CAp, precision), and cross-validation results (ASV, average number of support vectors; ABSV, average number of bounded support vectors; TRa, training accuracy; TRp, training precision; TEa, test accuracy; TEp, test precision).

Table 2. (Continued)

Exp	*C*	*K*	γ	*d*	SV	BSV	+/+	+/–	–/–	–/+	CAa	CAp	ASV	ABSV	TRa	TRp	TEa	TEp
46	10	A	1.0	2	25	0	48	0	40	0	1.00	1.00	25.3	0.0	0.89	0.84	0.84	0.77
47	100		1.0	2	25	0	48	0	40	0	1.00	1.00	25.3	0.0	0.89	0.84	0.84	0.77
48	1000		1.0	2	25	0	48	0	40	0	1.00	1.00	25.3	0.0	0.89	0.84	0.84	0.77
49	10		2.0	2	38	0	48	0	29	11	0.88	0.81	34.1	0.0	0.95	0.93	0.82	0.73
50	100		2.0	2	38	0	48	0	29	11	0.88	0.81	34.1	0.0	0.95	0.93	0.82	0.73
51	1000		2.0	2	38	0	48	0	29	11	0.88	0.81	34.1	0.0	0.95	0.93	0.82	0.73
52	10		0.5	3	24	0	48	0	31	9	0.90	0.84	22.2	0.0	0.93	0.90	0.84	0.78
53	100		0.5	3	24	0	48	0	31	9	0.90	0.84	22.2	0.0	0.93	0.90	0.84	0.78
54	1000		0.5	3	24	0	48	0	31	9	0.90	0.84	22.2	0.0	0.93	0.90	0.84	0.78
55	10		1.0	3	30	0	48	0	32	8	0.91	0.86	29.3	0.0	0.91	0.87	0.84	0.77
56	100		1.0	3	30	0	48	0	32	8	0.91	0.86	29.3	0.0	0.91	0.87	0.84	0.77
57	1000		1.0	3	30	0	48	0	32	8	0.91	0.86	29.3	0.0	0.91	0.87	0.84	0.77
58	10		2.0	3	49	0	48	0	27	13	0.85	0.79	41.1	0.0	0.94	0.90	0.76	0.68
59	100		2.0	3	49	0	48	0	27	13	0.85	0.79	41.1	0.0	0.94	0.90	0.76	0.68
60	1000		2.0	3	49	0	48	0	27	13	0.85	0.79	41.1	0.0	0.94	0.90	0.76	0.68

Table 3. Results for SVM Modeling of the Mechanism of Aquatic Toxicity Using Three Structural Descriptors: log K_{ow}, S^{N}_{av}, and E_{HOMO}. For Notations See the Footnote of Table 2.

Exp	*C*	*K*		SV	BSV	+/+	+/–	–/–	–/+	CAa	CAp	ASV	ABSV	TRa	TRp	TEa	TEp
61	10	D		26	22	47	1	29	11	0.86	0.81	23.1	19.1	0.85	0.80	0.84	0.76
62	100			26	22	47	1	26	14	0.83	0.77	23.3	19.2	0.85	0.79	0.83	0.75
63	1000			24	18	47	1	29	11	0.86	0.81	24.0	18.1	0.86	0.81	0.85	0.77
			d														
64	10	P	2	23	8	47	1	31	9	0.89	0.84	20.9	5.3	0.94	0.91	0.88	0.81
65	100		2	27	7	47	1	32	8	0.90	0.85	22.5	4.9	0.92	0.88	0.89	0.83
66	1000		2	26	5	47	1	38	2	0.97	0.96	22.7	4.1	0.93	0.90	0.91	0.86
67	10		3	22	2	47	1	34	6	0.92	0.89	20.7	1.6	0.97	0.96	0.80	0.74
68	100		3	18	0	48	0	40	0	1.00	1.00	18.3	0.0	0.92	0.88	0.81	0.75
69	1000		3	18	0	48	0	40	0	1.00	1.00	18.3	0.0	0.92	0.88	0.81	0.75
70	10		4	20	0	48	0	32	8	0.91	0.86	18.8	0.0	0.96	0.93	0.84	0.76
71	100		4	20	0	48	0	32	8	0.91	0.86	18.8	0.0	0.96	0.93	0.84	0.76
72	1000		4	20	0	48	0	32	8	0.91	0.86	18.8	0.0	0.96	0.93	0.84	0.76
73	10		5	21	0	48	0	32	8	0.91	0.86	19.9	0.0	0.97	0.95	0.85	0.77
74	100		5	21	0	48	0	32	8	0.91	0.86	19.9	0.0	0.97	0.95	0.85	0.77
75	1000		5	21	0	48	0	32	8	0.91	0.86	19.9	0.0	0.97	0.95	0.85	0.77
			γ														
76	10	R	0.5	29	4	47	1	32	8	0.90	0.85	26.4	3.9	0.93	0.90	0.88	0.81
77	100		0.5	25	1	48	0	39	1	0.99	0.98	23.6	0.7	0.92	0.87	0.80	0.74
78	1000		0.5	24	0	48	0	40	0	1.00	1.00	23.0	0.0	0.92	0.88	0.81	0.77
79	10		1.0	37	2	48	0	31	9	0.90	0.84	33.2	1.8	0.87	0.82	0.75	0.68
80	100		1.0	36	0	48	0	40	0	1.00	1.00	30.3	0.0	0.91	0.87	0.79	0.72
81	1000		1.0	36	0	48	0	40	0	1.00	1.00	30.3	0.0	0.91	0.87	0.79	0.72
82	10		2.0	49	1	48	0	39	1	0.99	0.98	46.1	0.6	0.95	0.93	0.69	0.63
83	100		2.0	50	0	48	0	40	0	1.00	1.00	46.1	0.0	0.97	0.95	0.69	0.63
84	1000		2.0	50	0	48	0	40	0	1.00	1.00	46.1	0.0	0.97	0.95	0.69	0.63
			a														
85	10	N	0.5	19	16	40	8	30	10	0.80	0.80	20.1	17.4	0.77	0.77	0.77	0.78
86	100		0.5	19	16	40	8	30	10	0.80	0.80	19.1	16.7	0.77	0.78	0.76	0.75
87	1000		0.5	18	15	41	7	31	9	0.82	0.82	19.0	16.7	0.77	0.78	0.76	0.75
88	10		1.0	20	20	38	10	31	9	0.78	0.81	19.6	17.1	0.77	0.78	0.78	0.75
89	100		1.0	20	18	39	9	30	10	0.78	0.80	18.8	16.4	0.77	0.78	0.75	0.71
90	1000		1.0	20	18	39	9	30	10	0.78	0.80	18.8	16.3	0.77	0.78	0.76	0.73
91	10		2.0	21	18	39	9	30	10	0.78	0.80	19.7	17.3	0.78	0.78	0.76	0.72
92	100		2.0	22	19	39	9	29	11	0.77	0.78	19.3	17.2	0.77	0.78	0.77	0.72
93	1000		2.0	22	19	39	9	29	11	0.77	0.78	19.3	17.5	0.76	0.78	0.75	0.73

Table 3. (Continued)

Exp	*C*	*K*	γ	*d*	SV	BSV	+/+	+/–	–/–	–/+	CAa	CAp	ASV	ABSV	TRa	TRp	TEa	TEp
94	10	A	0.5	1	28	10	48	0	34	6	0.93	0.89	20.4	8.2	0.92	0.88	0.87	0.81
95	100		0.5	1	21	6	47	1	38	2	0.97	0.96	19.3	2.8	0.95	0.93	0.89	0.83
96	1000		0.5	1	21	1	48	0	33	7	0.92	0.87	18.3	0.6	0.96	0.94	0.85	0.76
97	10		1.0	1	27	7	48	0	36	4	0.95	0.92	22.7	4.7	0.93	0.89	0.86	0.82
98	100		1.0	1	22	1	48	0	39	1	0.99	0.98	21.8	0.6	0.96	0.93	0.85	0.75
99	1000		1.0	1	22	0	48	0	40	0	1.00	1.00	20.7	0.0	0.97	0.95	0.84	0.74
100	10		2.0	1	28	2	48	0	33	7	0.92	0.87	26.2	1.5	0.94	0.90	0.83	0.76
101	100		2.0	1	24	0	48	0	33	7	0.92	0.87	25.0	0.0	0.94	0.91	0.82	0.77
102	1000		2.0	1	24	0	48	0	33	7	0.92	0.87	25.0	0.0	0.94	0.91	0.82	0.77
103	10		0.5	2	28	1	48	0	39	1	0.99	0.98	23.8	1.5	0.97	0.95	0.85	0.80
104	100		0.5	2	28	0	48	0	40	0	1.00	1.00	23.3	0.0	0.93	0.89	0.83	0.74
105	1000		0.5	2	28	0	48	0	40	0	1.00	1.00	23.3	0.0	0.93	0.89	0.83	0.74
106	10		1.0	2	33	0	48	0	40	0	1.00	1.00	30.3	0.0	0.93	0.89	0.86	0.79
107	100		1.0	2	33	0	48	0	40	0	1.00	1.00	30.3	0.0	0.93	0.89	0.86	0.79
108	1000		1.0	2	33	0	48	0	40	0	1.00	1.00	30.3	0.0	0.93	0.89	0.86	0.79
109	10		2.0	2	37	0	48	0	40	0	1.00	1.00	34.2	0.0	0.93	0.90	0.81	0.74
110	100		2.0	2	37	0	48	0	40	0	1.00	1.00	34.2	0.0	0.93	0.90	0.81	0.74
111	1000		2.0	2	37	0	48	0	40	0	1.00	1.00	34.2	0.0	0.93	0.90	0.81	0.74
112	10		0.5	3	31	0	48	0	40	0	1.00	1.00	25.7	0.0	0.94	0.90	0.84	0.78
113	100		0.5	3	31	0	48	0	40	0	1.00	1.00	25.7	0.0	0.94	0.90	0.84	0.78
114	1000		0.5	3	31	0	48	0	40	0	1.00	1.00	25.7	0.0	0.94	0.90	0.84	0.78
115	10		1.0	3	36	0	48	0	29	11	0.88	0.81	31.9	0.0	0.90	0.84	0.82	0.75
116	100		1.0	3	36	0	48	0	29	11	0.88	0.81	31.9	0.0	0.90	0.84	0.82	0.75
117	1000		1.0	3	36	0	48	0	29	11	0.88	0.81	31.9	0.0	0.90	0.84	0.82	0.75
118	10		2.0	3	48	0	48	0	31	9	0.90	0.84	42.2	0.0	0.90	0.85	0.73	0.66
119	100		2.0	3	48	0	48	0	31	9	0.90	0.84	42.2	0.0	0.90	0.85	0.73	0.66
120	1000		2.0	3	48	0	48	0	31	9	0.90	0.84	42.2	0.0	0.90	0.85	0.73	0.66

The statistical results obtained for the first set of SVM experiments are presented in Table 2. The SVM models were obtained with the above four descriptors, and with three values for the capacity parameter *C*, namely 10, 100, and 1000. The calibration of the SVM models was performed with the whole set of 88 compounds. The calibration results reported in Table 2 are: SV, number of support vectors; BSV, number of bounded support vectors; +/+, number of +1 patterns (narcotic compounds) predicted in class +1; +/–, number of +1 patterns predicted in class –1; –/–, number of –1 patterns (reactive compounds) predicted in class –1; –/+, number of –1 patterns predicted in class +1; CAa, accuracy; CAp, precision. The high flexibility of multivariate statistical models in approximating a wide range of mathematical functions comes with a significant danger: overfitting. Using sophisticated kernels, SVM can be calibrated to perfectly discriminate two populations of patterns, but only a cross–validation test can demonstrate the potential utility of an SVM model. For each SVM model we present in Table 2 the following leave–10%–out (L10%O) cross-validation statistics: ASV, average number of support vectors; ABSV, average number of bounded support vectors; TRa, training accuracy; TRp, training precision; TEa, test accuracy; TEp, test precision. As implemented in mySVM, *C* is scaled by 1/number of training examples.

The first group of SVM models computed with log K_{ow}, S^{N}_{av}, E_{HOMO}, and E_{LUMO} were obtained with the dot kernel, but the number of support vectors is too large, the prediction statistics are low, and a significant fraction of –1 patterns (reactive compounds) are classified as narcotic. On the

other hand, from the 48 narcotic compounds only one is classified as reactive, which is a first sign that with these four structural descriptors reactive compounds are more difficult to classify than narcotic compounds.

A significant improvement for the classification of narcotic and reactive compounds is obtained with the polynomial kernel, as presented in Table 2, experiments 4–15. Using 24 support vectors, a polynomial of degree 3 (experiments 7–9) can perfectly discriminate between narcotic and reactive compounds, with fairly good leave–10%–out (L10%O) cross–validation results, namely TEa = 0.85 and TEp = 0.80. These results are not sensitive to the value of the capacity parameter *C*. When a polynomial of degree 2 is used (experiments 4–6), all 48 narcotic compounds are correctly included in the +1 class, while six reactive compounds are misclassified as narcotic compounds, namely resorcinol, 4–chlorocatechol, 2,6–dinitrophenol, 2,3,5,6–tetrachloroaniline, 3–nitrotoluene, and 4–nonylphenol. Resorcinol and 4–nonylphenol were also outliers in the discriminant analysis and logistic regression [13]. Experiments 4–6 have the best overall prediction statistics, with TEa = 0.92 and TEp = 0.88, and a polynomial of degree 2 should be considered the best choice for discriminating the aquatic toxicity mechanism of action for organic compounds.

The next group of models, presented in Table 2 experiments 16–24, was obtained with the radial basis function kernel. The best results are obtained for $\gamma = 0.5$; for example, in experiment 18, with 24 support vectors all narcotic compounds are correctly classified, while eight reactive compounds are misclassified as narcotic compounds. Compared with the polynomial kernel, the radial basis function kernel performs worse with the data from this study, and the two classes of compounds cannot be separated. In our tests, the neural kernel gave the worst results, as can be seen from Table 2 experiments 25–33, having the largest number of compounds incorrectly classified. The neural kernel is not able to correctly classify even the narcotic compounds, which are an easy task for the polynomial, radial, and anova kernels.

The last group of models from in Table 2, experiments 34–60, was obtained with the anova kernel. Several combinations of the *C*, γ, and *d* provide a complete separation of the narcotic and reactive compounds: $\gamma = 1.0$ and $d = 1$ for $C = 100$ and 1000; $\gamma = 1.0$ and $d = 2$ for $C = 10$, 100 and 1000. Similarly with other kernels, the classification performances of the anova kernel vary significantly with the parameters that control the shape of the kernel. In calibration, the anova kernel needs only 19 support vectors to completely separate the narcotic and reactive compounds (experiments 38 and 39), while the polynomial kernel needs 24 support vectors for a complete separation (experiments 7–9). However, the best prediction results in the L10%O cross–validation test are those obtained in experiments 4–6 with a polynomial of degree 2; this SVM model has 14 support vectors in calibration and six reactive compounds are misclassified as narcotic compounds. Compared with the radial basis function, neural, or anova kernels, the polynomial of degree 2 defines a simple kernel, but with a better prediction power obtained with the smallest number of

support vectors. These results indicate that, similarly with other multivariate statistical models used in structure–activity studies, the best SVM model (kernel type and parameters) must be selected to maximize the predictive potential, not the calibration results.

A second group of SVM models, reported in Table 3, were obtained with log K_{ow}, S^{N}_{av}, and E_{HOMO}, because the logistic regression showed that E_{LUMO} does not improve the classification [13]. The general trends revealed by the data from Table 2 can be identified also in this set of experiments, but our intention was to determine if E_{LUMO} can be deleted without loosing the predictive power. The dot and neural kernels show a poor discrimination between narcotic and reactive compounds, while the polynomial, radial, and anova kernels computed with certain parameters provide a complete separation of the two classes of compounds. Among the experiments from Table 3, the best predictions are obtained in the experiment 66, with a polynomial of degree 2, C = 1000, and 26 support vectors. In calibration, one narcotic compound was computed as reactive and two reactive compounds were classified as narcotic, while for prediction the cross–validation statistics indicate a good prediction power for this model, *i.e.* TEa = 0.91 and TEp = 0.86.

While a comparison of the results reported in tables 2 and 3 clearly shows that SVM with a polynomial of degree 2 kernel provide the best predictions, it is not very clear if E_{LUMO} can be deleted without degrading the performance of the SVM model. The advantages of the SVM models from the experiments 4–6 are a lower number of support vectors (14 in calibration and an average of 16.2 in the prediction test) and slightly better prediction statistics (TEa = 0.92 and TEp = 0.88). The advantage of the experiment 66 is represented by a lower number of classification errors in calibration (one for narcotic compounds and two for reactive compounds), but the number of support vectors is significantly larger (26 in calibration and an average of 22.7 in the prediction test). The SVM model files for experiments 6 and 66 are available as supplementary material. These files can be used to make predictions for new organic compounds, using as input the four structural descriptors, namely log K_{ow}, S^{N}_{av}, E_{HOMO}, and E_{LUMO}.

5 CONCLUSIONS

Support vector machines represent an attractive new class of machine learning algorithms that can have significant applications in developing structure–activity models, chemometrics, and design of chemical libraries. In the SVM approach, two clusters of patterns are optimally separated with a hyperplane that maximizes the separation between the two classes. Using various kernels, a non–linear mapping transforms the input space into a higher dimensional feature space, and then a quadratic programming algorithm determines a unique maximal margin hyperplane. While many chemometrics algorithms and SAR models are currently used for the non–linear classification of patterns, their application is usually plagued by the existence of multiple minima. For example, artificial neural networks are very flexible and can easily model highly non–linear separating

surfaces, but the optimization process usually ends in the local minimum that is closest to the starting point. As a result, starting with different random sets of connection weights, typically the optimized parameters will be different from an experiment to another, making neural network models difficult to replicate. In contrast, for a given kernel SVM determines a unique maximal separation hyperplane, using a fast quadratic programming algorithm. The possibility to discriminate clusters separated by non–linear surfaces, the unique solution for the class separation, and the fast optimization are three important advantages of SVM.

In this study we have investigated the application of SVM for the recognition of the aquatic toxicity mechanism for 88 compounds previously explored with discriminant analysis and logistic regression [8,13]. Four theoretical descriptors were used to discriminate between narcotic compounds (including non–polar and polar narcotics) and reactive compounds (including respiratory uncouplers, soft electrophiles, and proelectrophiles), namely the octanol–water partition coefficient log K_{ow}, the energy of the highest occupied molecular orbital E_{HOMO}, the energy of the lowest unoccupied molecular orbital E_{LUMO}, and the average acceptor superdelocalizability S^{N}_{av}. Only calibration models were presented, and some compounds were misclassified both in discriminant analysis and logistic regression [13].

The SVM applications in structure–activity models, chemometrics, and chemical libraries clustering are only in the beginning and for the moment there are no clear rules on selecting the most efficient parameters that control the SVM performances, namely the kernel and the set of structural descriptors that are essential for the SVM model. We have explored the influence of the kernel type on the SVM performances by testing various kernels, namely the dot, polynomial, radial basis function, neural, and anova kernels. Because there is no simple algorithm for descriptor selection in SVM models, we have used the theoretical indices from [8,13].

The role of a classifier is to learn the classification rule from training patterns and then to apply the rule to new patterns in order to obtain reliable predictions. Therefore, for a classifier, one of the most important properties is its generalization ability or its ability to make correct predictions for patterns not used in the calibration phase. The prediction power of each SVM model was evaluated with a leave–10%–out cross–validation procedure. After experimenting with various kernels and associated parameters, our results clearly demonstrate that the performance of the SVM classifier is strongly dependent on the kernel shape. With four structural descriptors, the dot, radial, and neural kernels show a poor discrimination between narcotic and reactive compounds, while the polynomial and anova kernels computed with certain parameters give a complete separation of the two classes of compounds. Similar results obtained with three structural descriptors indicate a somewhat unexpected result: a low degree polynomial kernel offers superior separation compared with the radial and neural kernels that have a more complex shape.

Our results show that with the four structural descriptors utilized, reactive compounds are more

difficult to classify than narcotic compounds. This can indicate either that this class is not homogeneous, or that more adequate descriptors must be used in order to describe the characteristics of the reactive compounds. In calibration, several kernels were able to give a complete separation of the two classes of compounds, unlike the discriminant analysis and logistic regression [13]. However, caution must be exerted in appreciating statistical models only by their calibration (training) results, because the goal of developing structure–activity models is to obtain equations that have a high predictive power. In our experiments, the SVM models with the best calibration performances were surpassed in the L10%O cross–validation by the polynomial of degree 2 kernel. This result clearly demonstrates that too complex kernels give overfitted SVM models, with low prediction power. Using sophisticated kernels, SVM can be calibrated to perfectly discriminate two populations of patterns, but only a cross–validation test can demonstrate the potential utility of an SVM model. Sometimes, the complete separation of the two classes of patterns is not possible to achieve due to errors in the experimental data or because the theoretical descriptors that describe the molecular structure of each compound are not adequate for the investigated property. Another important parameter that must be monitored in an SVM study is the number of support vectors, and whenever possible, SVM models with a lower number of support vectors must be preferred. In this study we have not addressed the important problem of selecting significant descriptors in SVM models. In QSAR studies it is generally accepted that it is more important to screen a wide variety of structural descriptors instead of using too sophisticated mathematical models. The same is true for SVM models, and considerable effort should be directed towards the development of efficient algorithms for descriptor selection.

Supplementary Material

The mySVM model files for experiments 6 and 66 are available as supplementary material.

6 REFERENCES

[1] A. R. Katritzky, D. B. Tatham, and U. Maran, Theoretical Descriptors for the Correlation of Aquatic Toxicity of Environmental Pollutants by Quantitative Structure–Toxicity Relationships, *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 1162–1176.

[2] H. J. M. Verhaar and C. J. Van Leeuwen, and J. L. M. Hermens, Classifying Environmental Pollutants. 1: Structure–Activity Relationships for Prediction of Aquatic Toxicity, *Chemosphere* **1992**, *25*, 471–491.

[3] S. P. Bradbury, Predicting Modes of Toxic Action From Chemical Structure: An Overview, *SAR QSAR Environ. Res*. **1994**, *2*, 89–104.

[4] O. G. Mekenyan and G. D. Veith, The Electronic Factor in QSAR: MO–Parameters, Competing Interactions, Reactivity and Toxicity, *SAR QSAR Environ. Res*. **1994**, *2*, 129–143.

[5] S. P. Bradbury, Quantitative Structure–Activity Relationships and Ecological Risk Assessment: An Overview of Predictive Aquatic Toxicology Research, *Toxicol. Lett*. **1995**, *79*, 229–237.

[6] S. Karabunarliev, O. G. Mekenyan, W. Karcher, C. L. Russom, and S. P. Bradbury, Quantum–Chemical Descriptors for Estimating the Acute Toxicity of Electrophiles to the Fathead Minnow (*Pimephales promelas*): An Analysis Based on Molecular Mechanisms, *Quant. Struct.–Act. Relat*. **1996**, *15*, 302–310.

[7] C. L. Russom, S. P. Bradbury, S. J. Broderium, D. E. Hammermeister, and R. A. Drummond, Predicting Modes of Toxic Action From Chemical Structure: Acute Toxicity in the Fathead Minnow (*Pimephales promelas*), *Environ. Toxicol. Chem*. **1997**, *16*, 948–967.

[8] A. P. Bearden and T. W. Schultz, Structure–Activity Relationships for *Pimephales* and *Tetrahymena*: A

Mechanism of Action Approach, *Environ. Toxicol. Chem*. **1997**, *16*, 1311–1317.
[9] A. B. A. Boxall, C. D. Watts, J. C. Dearden, G. M. Bresnen, and R. Scoffin, Classification of Environmental Pollutants Into General Mode of Toxic Action Classes Based on Molecular Descriptors, in: *Quantitative Structure–Activity Relationships in Environmental Sciences VII*, Eds. F. C. Fredenslund and G. Schüürmann, SETAC Press, Pensacola, Florida, USA, 1997, pp. 315–327.
[10] A. P. Bearden and T. W. Schultz, Comparison of *Tetrahymena* and *Pimephales* Toxicity Based on Mechanism of Action, *SAR QSAR Environ. Res*. **1998**, *9*, 127–153.
[11] S. C. Basak, G. D. Grunwald, G. E. Host, G. J. Niemi, and S. P. Bradbury, A Comparative Study of Molecular Similarity, Statistical, and Neural Methods for Predicting Toxic Modes of Action, *Environ. Toxicol. Chem*. **1998**, *17*, 1056–1064.
[12] T. W. Schultz, Structure–Toxicity Relationships for Benzenes Evaluated with *Tetrahymena pyriformis*, *Chem. Res. Toxicol*. **1999**, *12*, 1262–1267.
[13] S. Ren and T. W. Schultz, Identifying the Mechanism of Aquatic Toxicity of Selected Compounds by Hydrophobicity and Electrophilicity Descriptors, *Toxicol. Lett*. **2002**, *129*, 151–160.
[14] V. Vapnik, *Estimation of Dependencies Based on Empirical Data*, Nauka, Moscow, 1979.
[15] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, 1995.
[16] V. Vapnik, *Statistical Learning Theory*, Wiley–Interscience, New York, 1998.
[17] C. J. C. Burges, A Tutorial on Support Vector Machines for Pattern Recognition, *Data Mining Knowled. Discov.* **1998**, *2*, 121–167.
[18] N. Cristianini and J. Shawe–Taylor, *An Introduction to Support Vector Machines*, Cambridge University Press, Cambridge, 2000.
[19] A. Zien, G. Ratsch, S. Mika, B. Scholkopf, T. Lengauer, and K. R. Muller, Engineering Support Vector Machine Kernels that Recognize Translation Initiation Sites, *Bioinformatics* **2000**, *16*, 799–807.
[20] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler, Support Vector Machine Classification and Validation of Cancer Tissue Samples Using Microarray Expression Data, *Bioinformatics* **2000**, *16*, 906–914.
[21] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C. H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J. P. Mesirov, T. Poggio, W. Gerald, M. Loda, E. S. Lander, and T. R. Golub, Multiclass Cancer Diagnosis Using Tumor Gene Expression Signatures, *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98*, 15149–15154.
[22] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, Gene Selection for Cancer Classification Using Support Vector Machines, *Mach. Learn.* **2002**, *46*, 389–422.
[23] Y. D. Cai, X. J. Liu, X. B. Xu, and K. C. Chou, Support Vector Machines for Predicting HIV Protease Cleavage Sites in Protein, *J. Comput. Chem.* **2002**, *23*, 267–274.
[24] Y. D. Cai, X. J. Liu, X. B. Xu, and K. C. Chou, Prediction of Protein Structural Classes by Support Vector Machines, *Comput. Chem.* **2002**, *26*, 293–296.
[25] J. R. Bock and D. A. Gough, Predicting Protein–Protein Interactions from Primary Structure, *Bioinformatics* **2001**, *17*, 455–460.
[26] S. J. Hua and Z. R. Sun, Support Vector Machine Approach for Protein Subcellular Localization Prediction, *Bioinformatics* **2001**, *17*, 721–728.
[27] Y. D. Cai, X. J. Liu, X. B. Xu, and K. C. Chou, Support Vector Machines for Prediction of Protein Subcellular Location by Incorporating Quasi–Sequence–Order Effect, *J. Cell. Biochem.* **2002**, *84*, 343–348.
[28] C. H. Q. Ding and I. Dubchak, Multi–Class Protein Fold Recognition Using Support Vector Machines and Neural Networks, *Bioinformatics* **2001**, *17*, 349–358.
[29] S. J. Hua and Z. R. Sun, A Novel Method of Protein Secondary Structure Prediction with High Segment Overlap Measure: Support Vector Machine Approach, *J. Mol. Biol.* **2001**, *308*, 397–407.
[30] W. Vercoutere, S. Winters–Hilt, H. Olsen, D. Deamer, D. Haussler, and M. Akeson, Rapid Discrimination Among Individual DNA Hairpin Molecules at Single–Nucleotide Resolution Using an Ion Channel, *Nat. Biotechnol.* **2001**, *19*, 248–252.
[31] S. Rüping, mySVM, University of Dortmund, http://www–ai.cs.uni–dortmund.de/SOFTWARE/MYSVM/.
[32] BioChem Links, http://www.biochempress.com.