Inter*net* Electronic Journal of Molecular Design

April 2002, Volume 1, Number 4, Pages 203–218

Editor: Ovidiu Ivanciuc

Special issue dedicated to Professor Alexandru T. Balaban on the occasion of the 70th birthday
Part 4

Guest Editor: Mircea V. Diudea

Support Vector Machine Classification of the Carcinogenic Activity of Polycyclic Aromatic Hydrocarbons

Ovidiu Ivanciuc

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

Received: December 10, 2001; Accepted: March 21, 2002; Published: April 30, 2002

Citation of the article:

O. Ivanciuc, Support Vector Machine Classification of the Carcinogenic Activity of Polycyclic Aromatic Hydrocarbons, *Internet Electron. J. Mol. Des.* **2002**, *1*, 203–218, http://www.biochempress.com.

Inter*net* Electronic Journal of Molecular Design

BioChem Press
http://www.biochempress.com

Support Vector Machine Classification of the Carcinogenic Activity of Polycyclic Aromatic Hydrocarbons[#]

Ovidiu Ivanciuc*

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

Received: December 10, 2001; Accepted: March 21, 2002; Published: April 30, 2002

Internet Electron. J. Mol. Des. **2002**, *1* (4), 203–218

Abstract

Motivation. Structure–activity relationships (SAR) can be efficiently used to predict the carcinogenic hazard of new chemicals, before producing them on a large scale or even before synthesizing them. SAR models that detect potential carcinogens can also supplement short–term tests of genotoxicity, long–term tests of carcinogenicity in rodents, or epidemiological evidence in humans.

Method. Support vector machine (SVM) is an efficient classification algorithm that can provide highly predictive SAR models for the carcinogenic hazard. We have applied the SVM model to identify the carcinogenic activity of 46 methylated and 32 non–methylated polycyclic aromatic hydrocarbons (PAH). The PAH chemical structure was encoded by four theoretical descriptors computed with PM3, namely the energy of the highest occupied molecular orbital E_{HOMO}, the energy of the lowest unoccupied molecular orbital E_{LUMO}, the hardness HD, and the difference between E_{HOMO} and $E_{HOMO–1}$.

Results. A wide range of SVM experiments were performed using the dot, polynomial, radial basis function, neural, and anova kernels. The results obtained for the classification of PAH carcinogenicity demonstrate that the performances of SVM depend strongly on the kernel type and various parameters that control the kernel shape. The best prediction results were obtained with the radial basis function kernel with $\gamma = 0.5$, the anova kernel with $\gamma = 0.5$ and $d = 1$, and the anova kernel with $\gamma = 0.5$ and $d = 2$. In the first case, from 34 carcinogenic compounds, 28 were correctly classified, while from 44 non–carcinogenic compounds, 40 were correctly classified.

Conclusions. SAR models for predicting the carcinogenic hazard can benefit from the use of support vector machines, which determine a maximum separating hyperplane between carcinogenic and non–carcinogenic compounds. The solution of the SVM model is a unique hyperplane which can be computed very fast, but the classification results heavily depend on the kernel type and structural descriptors. Extensive cross–validation tests should be made to find the kernel with the optimum predictive power.

Keywords. Polycyclic aromatic hydrocarbons; structure–activity relationships; carcinogenicity; support vector machines; machine learning; kernel algorithm.

Abbreviations and notations

PAH, polycyclic aromatic hydrocarbons	E_{LUMO}, energy of the lowest unoccupied molecular orbital
SVM, support vector machines	HD, hardness, HD = $(E_{LUMO} – E_{HOMO})/2$
E_{HOMO}, energy of the highest occupied molecular orbital	ΔH, difference between E_{HOMO} and $E_{HOMO–1}$

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

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

1 INTRODUCTION

Extensive experimental tests and epidemiological studies demonstrated the causal relationship between cancer incidence and exposure to chemical compounds. The first report to link cancer and chemical exposure was a study on chimney sweeps by Pott [1], revealing soot as a carcinogenic agent. Effective cancer prevention can be obtained by stopping cigarette smoking, changing dietary habits, by restricting the number of carcinogens and exposure levels [2]. Due to the increasing exposure to industrial chemicals, food additives, drugs, cosmetics, pesticides, herbicides, and pollution agents, preventive measures must be taken in order to minimize the exposure to carcinogenic compounds. An effective prevention of cancer can be obtained by restricting the production and environmental emission of carcinogens, resulting in lower levels of exposure and fewer carcinogens produced by the chemical industry. This approach depends on reliable methods for the identification of carcinogenic compounds.

Structure–activity relationships (SAR) and quantitative structure–activity relationships (QSAR) are valuable statistical models for predicting the carcinogenic potential of new chemicals, not yet tested, and sometimes not yet synthesized [3]. Also, the interpretation of the short–term tests of genotoxicity, long–term tests of carcinogenicity in rodents, and epidemiological data can benefit from the use of sound statistical SAR models. The reach literature on SAR and QSAR models for the carcinogenic activity demonstrates the importance of this approach [3–28] in the molecular design of safer chemical compounds. A wide variety of structural descriptors were investigated as reliable indicators of the carcinogenic activity, from simple groups of atoms to topological indices and quantum indices. The relationship between these structural descriptors and the chemical carcinogenicity is explored with various statistical models, such as machine learning algorithms, clustering methods, discriminant analysis, linear regression, and artificial neural networks.

Support vector machines (SVM) represent a new class of machine learning algorithms for classification and regression with numerous applications in medicine and bioinformatics. In this study we present an application of SVM for the identification of the carcinogenic activity for a group of methylated and non–methylated polycyclic aromatic hydrocarbons (PAH) previously investigated in Refs. [23–28].

2 MATERIALS AND METHODS

2.1 Chemical Data

We have applied the SVM model to identify the carcinogenic activity of 32 PAH (presented in Figure 1) and 46 methylated PAH (presented in Figure 2) taken from literature [23–27]. From this set of 78 PAH, 34 are carcinogenic and 44 are non–carcinogenic. In Table 1 we present the

carcinogenic activity for all 78 PAH, together with their four theoretical descriptors computed with the PM3 semiempirical method [27], namely the energy of the highest occupied molecular orbital E_{HOMO}, energy of the lowest unoccupied molecular orbital E_{LUMO}, hardness HD, HD = (E_{LUMO} − E_{HOMO})/2, and difference between E_{HOMO} and E_{HOMO-1} denoted ΔH.

2.2 Support Vector Machines

Support vector machines were developed by Vapnik [29–31] as an effective algorithm for determining an optimal hyperplane to separate two classes of patterns [32–40]. 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 by maximizing the distance to the hyperplane of the closest patterns from the two classes.

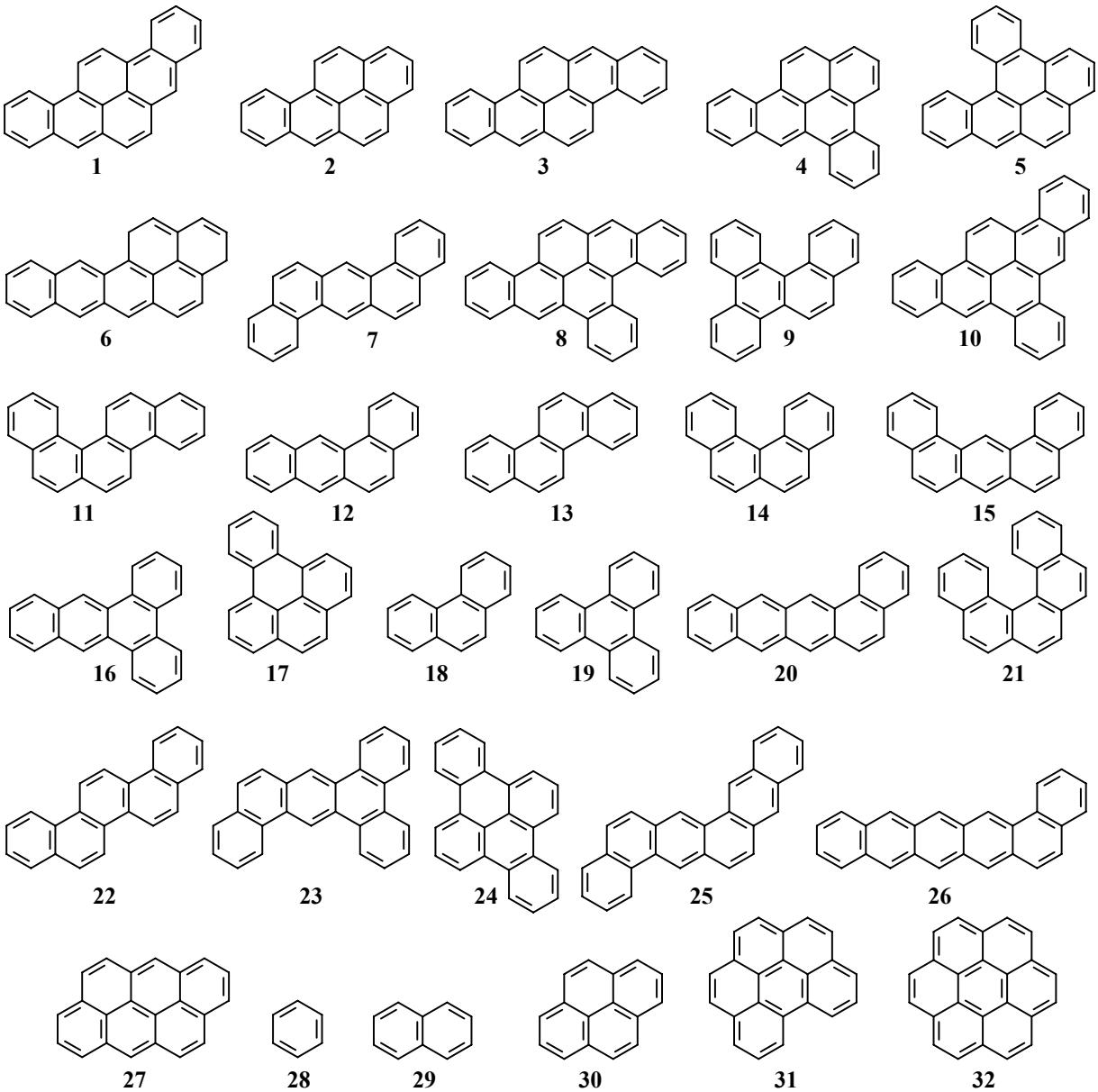

Figure 1. Molecular structure of the 32 benzenoid PAH [27].

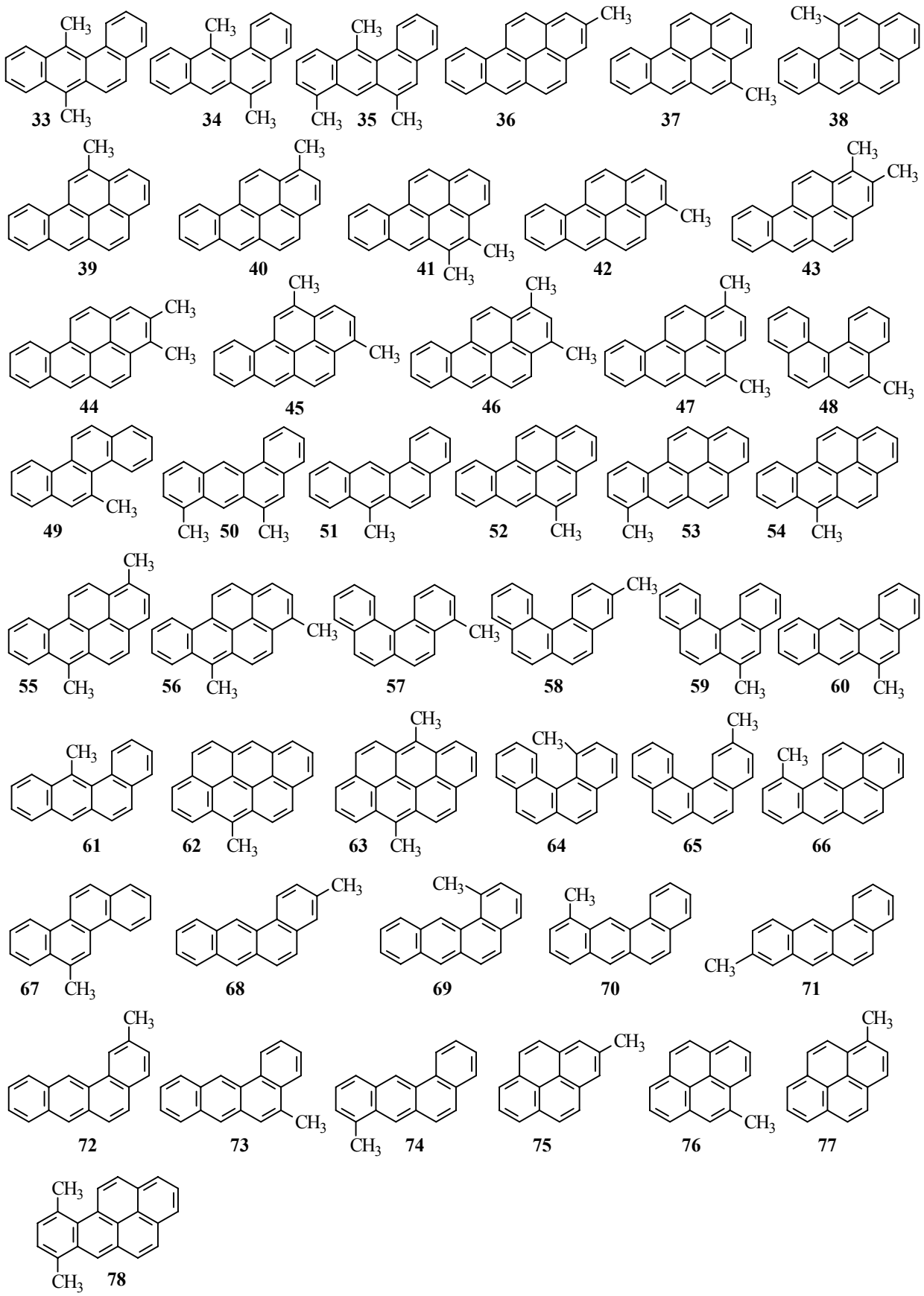

Figure 2. Molecular structure of the 46 methylated PAH [27].

Table 1. Structural Descriptors and Carcinogenic Activity for the 78 Polycyclic Aromatic Hydrocarbons [27]

PAH	Compound	E_{HOMO}[a]	E_{LUMO}[b]	HD[c]	ΔH[d]	*CA*[e]
1	dibenzo[3,4;9,10]pyrene	7.987	1.288	3.350	0.699	A
2	benzo[3,4]pyrene	8.042	1.221	3.411	0.860	A
3	dibenzo[3,4;8,9]pyrene	7.808	1.461	3.174	1.072	A
4	dibenzo[3,4;6,7]pyrene	8.140	1.163	3.489	0.676	A
5	dibenzo[1,2;3,4]pyrene	8.087	1.214	3.437	0.609	A
6	naphto[2,3;3,4]pyrene	7.848	1.421	3.214	0.913	A
7	dibenz[1,2;5,6]anthracene	8.377	0.918	3.730	0.326	A
8	tribenzo[3,4;6,7;8,9]pyrene	7.888	1.421	3.234	0.842	A
9	dibenzo[1,2;3,4]phenantrene	8.458	0.885	3.787	0.325	A
10	tribenzo[3,4;6,7;9,10]pyrene	8.087	1.228	3.430	0.607	A
11	dibenzo[1,2;5,6]phenantrene	8.519	0.782	3.869	0.098	I
12	benz[1,2]anthracene	8.328	0.934	3.697	0.563	I
13	chrysene	8.496	0.783	3.857	0.420	I
14	benzo[3,4]phenantrene	8.567	0.752	3.908	0.216	I
15	dibenz[1,2;7,8]anthracene	8.400	0.902	3.749	0.254	I
16	dibenz[1,2;3,4]anthracene	8.400	0.907	3.747	0.402	I
17	benzo[1,2]pyrene	8.335	0.967	3.684	0.489	I
18	phenantrene	8.740	0.535	4.103	0.236	I
19	triphenylene	8.773	0.556	4.109	0.000	I
20	benzo[1,2]naphthacene	7.962	1.297	3.333	0.918	I
21	dibenzo[3,4;5,6]phenantrene	8.481	0.813	3.834	0.246	I
22	picene	8.477	0.824	3.827	0.210	I
23	tribenz[1,2;3,4;5,6]anthracene	8.448	0.889	3.780	0.154	I
24	dibenzo[1,2;5,6]pyrene	8.409	0.930	3.740	0.304	I
25	phenanthra[2,3;1,2]anthracene	8.274	1.026	3.624	0.248	I
26	benzo[1,2]pentacene	7.693	1.573	3.060	1.088	I
27	anthanthrene	7.762	1.508	3.127	1.170	I
28	benzene	9.751	0.396	4.678	0.000	I
29	naphthalene	8.836	0.408	4.214	0.600	I
30	pyrene	8.249	1.010	3.619	0.793	I
31	benzo[*ghi*]perylene	8.139	1.167	3.486	0.569	I
32	coronene	8.290	1.063	3.614	0.000	I
33	7,12–dimethylbenz[*a*]anthracene	8.125	0.921	3.602	0.733	A
34	6,12–dimethylbenz[*a*]anthracene	8.158	0.911	3.624	0.673	A
35	6,8,12–trimethylbenz[*a*]anthracene	8.109	0.898	3.605	0.718	A
36	2–methylbenzo[*a*]pyrene	8.018	1.199	3.410	0.787	A
37	4–methylbenzo[*a*]pyrene	7.983	1.205	3.389	0.882	A
38	11–methylbenzo[*a*]pyrene	7.985	1.208	3.389	0.873	A
39	12–methylbenzo[*a*]pyrene	7.972	1.205	3.383	0.907	A
40	1–methylbenzo[*a*]pyrene	7.971	1.209	3.381	0.913	A
41	4,5–dimethylbenzo[*a*]pyrene	7.936	1.182	3.377	0.909	A
42	3–methylbenzo[*a*]pyrene	7.973	1.194	3.390	0.862	A
43	1,2–dimethylbenzo[*a*]pyrene	7.951	1.182	3.385	0.836	A
44	2,3–dimethylbenzo[*a*]pyrene	7.948	1.178	3.385	0.789	A
45	3,12–dimethylbenzo[*a*]pyrene	7.908	1.179	3.365	0.904	A
46	1,3–dimethylbenzo[*a*]pyrene	7.906	1.183	3.362	0.913	A
47	1,4–dimethylbenzo[*a*]pyrene	7.915	1.194	3.361	0.932	A
48	5–methylbenzo[*c*]phenanthrene	8.487	0.744	3.872	0.254	A
49	5–methylchrysene	8.410	0.787	3.812	0.436	A
50	6,8–dimethylbenz[*a*]anthracene	8.198	0.912	3.643	0.638	A
51	7–methylbenz[*a*]anthracene	8.218	0.931	3.644	0.647	A
52	5–methylbenzo[*a*]pyrene	7.986	1.207	3.390	0.897	A

[a] E_{HOMO}, energy of the highest occupied molecular orbital
[b] E_{LUMO}, energy of the lowest unoccupied molecular orbital
[c] HD, hardness
[d] ΔH, difference between E_{HOMO} and E_{HOMO-1}
[e] *CA*, carcinogenic activity (A, active; I, inactive)

Table 1. (Continued)

PAH	Compound	E_{HOMO}	E_{LUMO}	HD	ΔH	*CA*
53	7–methylbenzo[*a*]pyrene	7.993	1.206	3.394	0.875	A
54	6–methylbenzo[*a*]pyrene	7.942	1.215	3.364	0.933	A
55	1,6–dimethylbenzo[*a*]pyrene	7.876	1.204	3.336	0.983	A
56	3,6–dimethylbenzo[*a*]pyrene	7.878	1.189	3.345	0.929	A
57	4–methylbenzo[*c*]phenanthrene	8.510	0.740	3.885	0.237	I
58	3–methylbenzo[*c*]phenanthrene	8.505	0.728	3.889	0.242	I
59	6–methylbenzo[*c*]phenanthrene	8.527	0.733	3.897	0.140	I
60	6–methylbenz[*a*]anthracene	8.261	0.924	3.669	0.595	I
61	12–methylbenz[*a*]anthracene	8.219	0.920	3.650	0.650	I
62	6–methylanthanthrene	7.683	1.495	3.094	1.220	I
63	6,12–dimethylanthanthrene	7.606	1.483	3.062	1.280	I
64	1–methylbenzo[*c*]phenanthrene	8.460	0.722	3.869	0.284	I
65	2–methylbenzo[*c*]phenanthrene	8.528	0.731	3.899	0.175	I
66	10–methylbenzo[*a*]pyrene	7.985	1.206	3.390	0.881	I
67	6–methylchrysene	8.403	0.775	3.814	0.494	I
68	3–methylbenz[*a*]anthracene	8.291	0.915	3.688	0.522	I
69	1–methylbenz[*a*]anthracene	8.301	0.901	3.700	0.491	I
70	11–methylbenz[*a*]anthracene	8.277	0.923	3.677	0.593	I
71	9–methylbenz[*a*]anthracene	8.287	0.904	3.692	0.525	I
72	2–methylbenz[*a*]anthracene	8.278	0.915	3.682	0.579	I
73	5–methylbenz[*a*]anthracene	8.258	0.918	3.670	0.559	I
74	8–methylbenz[*a*]anthracene	8.256	0.922	3.667	0.617	I
75	2–methylpyrene	8.230	0.982	3.624	0.666	I
76	4–methylpyrene	8.178	0.994	3.592	0.830	I
77	1–methylpyrene	8.151	0.990	3.581	0.847	I
78	7,10–dimethylbenzo[*a*]pyrene	7.928	1.193	3.368	0.898	I

This powerful classification technique was applied with success in medicine, computational biology, bioinformatics, and structure–activity relationships, for the classification of: microarray gene expression data [41], translation initiation sites [42], genes [43], cancer type [44–47], pigmented skin lesions [48], HIV protease cleavage sites [49], GPCR type [50], protein class [51], membrane protein type [52], protein–protein interactions [53], protein subcellular localization [54–56], protein fold [57], protein secondary structure [58], specificity of GalNAc–transferase [59], DNA hairpins [60], organisms [61], aquatic toxicity mechanism of action [62].

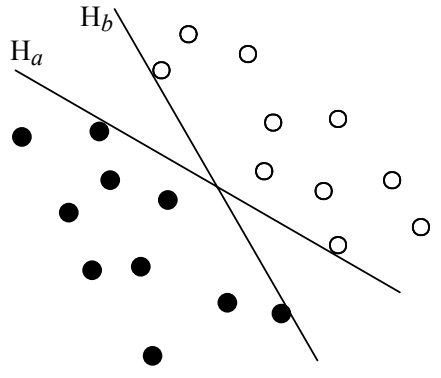

Figure 3. Two possible hyperplanes H_a and H_b that discriminate between patterns from the class +1 (black circles) and –1 (white circles).

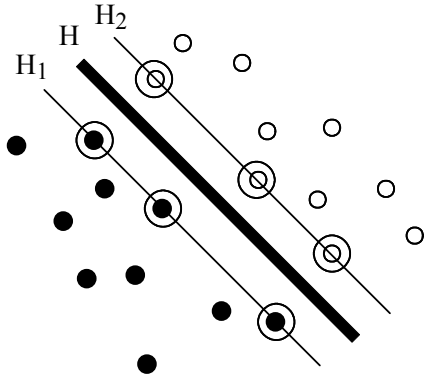

Figure 4. Example of patterns from the class +1 (black circles) and –1 (white circles) linearly separable by the maximal margin hyperplane H. The support vectors from the class +1 define the hyperplane H_1 while those from the class –1 define the hyperplane H_2.

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 3, 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 4. 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 4 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 MMH cannot be determined whenever, due to the partial overlapping of the +1 and –1 classes, a separating hypersurface does not exist. For this type of problems, the condition of 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 $\xi = (\xi_1, \xi_2, \ldots, \xi_l)$. Computing this soft margin separating hyperplane (SMSH) is equivalent to solving the following optimization problem:

$$\begin{cases} \text{minimize } \dfrac{1}{2}\|w\|^2 + C\sum_i \xi_i \\ \text{with } y_i(w \cdot x_i + b) \geq 1 - \xi_i \quad (i = 1, 2, \ldots, l) \\ \qquad \xi_i > 0 \end{cases} \tag{1}$$

Instead of solving the above problem directly, it is easier to solve the following Wolfe dual:

$$\begin{cases} \text{minimize } -\sum_i \alpha_i + \dfrac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \\ \text{with } \sum_i \alpha_i y_i = 0 \quad \text{and} \quad 0 \leq \alpha_i \leq C \end{cases} \tag{2}$$

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 4), 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. All calibration (training) patterns with $\alpha_i >$

0 in the solution are called support vectors. Patterns with $0< \alpha_i < C$ are called unbounded support vectors, while those with $\alpha_i = C$ are called bounded support vectors. SVM can be easily generalized to non–linear decision surfaces by replacing the inner product $(x_i \cdot x_j)$ with a kernel function $K(x_i,x_j)$.

All SVM models from the present paper for the classification of the PAH carcinogenic activity were obtained with mySVM [63], which is freely available for download. Links to Web resources related to SVM, namely tutorials, papers and software, can be found in BioChem Links [64] at http://www.biochempress.com. 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:

$$K(x, y) = x \cdot y \tag{3}$$

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

$$K(x, y) = (x \cdot y + 1)^d \tag{4}$$

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):

$$K(x, y) = \exp(-\gamma \parallel x - y \parallel^2) \tag{5}$$

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):

$$K(x, y) = \tanh(ax \cdot y + b) \tag{6}$$

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):

$$K(x, y) = \left(\sum_i \exp(-\gamma(x_i - y_i)) \right)^d \tag{7}$$

3 RESULTS AND DISCUSSION

The quality of the SVM models depends on the kernel type, various parameters that control the kernel shape, and the set of theoretical descriptors that describe the molecular structure. Using a quadratic programming algorithm, SVM offers a unique maximal separation hyperplane. However, similarly with over multivariate statistical models used in chemometrics and structure–activity studies, for a given set of molecules, there are no clear guidelines on selecting the optimum set of theoretical descriptors and decision function (kernel type and associated parameters). Therefore, the only practical way of finding an optimally predictive SVM model is through extensive experiments.

Table 2. Results for SVM Modeling of the PAH Carcinogenic Activity Using E_{HOMO}, E_{LUMO}, HD, and ΔH.[a]

Exp	*C*	*K*		SV	BSV	A/A	A/I	I/I	I/A	CAa	CAp	ASV	ABSV	TRa	TRp	TEa	TEp
1	10	D		48	44	28	6	36	8	0.82	0.78	43.3	39.3	0.80	0.76	0.76	0.69
2	100			49	43	28	6	36	8	0.82	0.78	42.6	38.6	0.80	0.76	0.74	0.68
3	1000			47	43	28	6	36	8	0.82	0.78	42.5	38.4	0.80	0.76	0.74	0.68
			d														
4	10	P	2	37	29	29	5	40	4	0.88	0.88	34.2	25.7	0.88	0.87	0.82	0.81
5	100		2	35	26	29	5	40	4	0.88	0.88	31.6	22.5	0.88	0.87	0.83	0.81
6	1000		2	36	26	29	5	40	4	0.88	0.88	31.9	21.2	0.88	0.88	0.83	0.80
7	10		3	33	19	28	6	40	4	0.87	0.88	31.2	16.5	0.88	0.88	0.82	0.81
8	100		3	34	16	29	5	40	4	0.88	0.88	31.2	13.6	0.88	0.88	0.78	0.73
9	1000		3	48	11	29	5	34	10	0.81	0.74	35.4	11.0	0.85	0.82	0.79	0.70
10	10		4	39	13	29	5	40	4	0.88	0.88	33.0	12.1	0.89	0.90	0.79	0.76
11	100		4	36	14	30	4	41	3	0.91	0.91	31.1	9.5	0.91	0.90	0.72	0.66
12	1000		4	37	9	32	2	40	4	0.92	0.89	32.1	4.0	0.87	0.84	0.69	0.58
13	10		5	35	11	31	3	41	3	0.92	0.91	30.5	7.7	0.93	0.91	0.68	0.60
14	100		5	36	4	28	6	38	6	0.85	0.82	30.7	2.9	0.96	0.94	0.68	0.62
15	1000		5	33	2	33	1	40	4	0.94	0.89	27.3	1.5	0.96	0.93	0.66	0.62
			γ														
16	10	R	0.5	36	19	28	6	40	4	0.87	0.88	34.2	17.3	0.87	0.88	0.86	0.88
17	100		0.5	31	14	29	5	42	2	0.91	0.94	29.7	11.9	0.91	0.93	0.83	0.77
18	1000		0.5	33	7	32	2	42	2	0.95	0.94	29.9	5.3	0.95	0.94	0.78	0.69
19	10		1.0	37	13	29	5	41	3	0.90	0.91	35.5	12.3	0.90	0.90	0.84	0.85
20	100		1.0	32	11	31	3	42	2	0.94	0.94	30.2	7.4	0.94	0.93	0.76	0.68
21	1000		1.0	29	5	32	2	42	2	0.95	0.94	27.1	3.8	0.96	0.94	0.79	0.74
22	10		2.0	41	10	29	5	41	3	0.90	0.91	37.1	9.2	0.91	0.92	0.81	0.79
23	100		2.0	34	6	32	2	42	2	0.95	0.94	31.4	4.5	0.95	0.94	0.74	0.69
24	1000		2.0	27	2	34	0	43	1	0.99	0.97	27.2	1.5	0.99	0.97	0.77	0.73
			a														
25	10	N	0.5	38	36	19	15	27	17	0.59	0.53	34.2	31.6	0.61	0.55	0.56	0.48
26	100		0.5	36	34	18	16	27	17	0.58	0.51	33.3	30.5	0.58	0.51	0.53	0.46
27	1000		0.5	36	34	17	17	27	17	0.56	0.50	33.0	30.1	0.57	0.51	0.53	0.46
28	10		1.0	32	30	20	14	29	15	0.63	0.57	28.0	26.4	0.64	0.58	0.59	0.51
29	100		1.0	30	30	26	8	23	21	0.63	0.55	27.9	26.0	0.64	0.58	0.58	0.52
30	1000		1.0	30	30	26	8	23	21	0.63	0.55	27.8	26.1	0.63	0.57	0.61	0.52
31	10		2.0	26	24	23	11	34	10	0.73	0.70	25.6	23.3	0.68	0.63	0.66	0.56
32	100		2.0	25	23	23	11	32	12	0.71	0.66	25.3	23.1	0.67	0.62	0.66	0.56
33	1000		2.0	24	22	23	11	33	11	0.72	0.68	25.2	23.0	0.67	0.62	0.66	0.56
			γ	*d*													
34	10	A	0.5	1	32	23	27	7	40	4	0.86	0.87	29.5	20.1	0.86	0.87	0.84	0.88
35	100		0.5	1	32	16	27	7	41	3	0.87	0.90	28.6	15.0	0.88	0.90	0.84	0.85
36	1000		0.5	1	31	14	29	5	40	4	0.88	0.88	28.3	11.7	0.89	0.89	0.78	0.73
37	10		1.0	1	33	20	29	5	40	4	0.88	0.88	29.9	17.8	0.88	0.89	0.84	0.85
38	100		1.0	1	33	13	30	4	39	5	0.88	0.86	29.7	11.8	0.89	0.88	0.82	0.76
39	1000		1.0	1	27	5	33	1	41	3	0.95	0.92	25.2	5.0	0.94	0.92	0.81	0.73
40	10		2.0	1	32	14	29	5	40	4	0.88	0.88	29.1	12.7	0.89	0.89	0.85	0.81
41	100		2.0	1	30	7	32	2	41	3	0.94	0.91	26.7	6.2	0.94	0.92	0.81	0.74
42	1000		2.0	1	24	6	33	1	41	3	0.95	0.92	25.4	4.7	0.95	0.93	0.76	0.68
43	10		0.5	2	33	15	29	5	40	4	0.88	0.88	29.9	13.2	0.89	0.89	0.84	0.84
44	100		0.5	2	32	8	32	2	42	2	0.95	0.94	28.9	7.1	0.95	0.94	0.80	0.73
45	1000		0.5	2	27	5	32	2	42	2	0.95	0.94	26.9	4.3	0.95	0.94	0.76	0.68

[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; A/A, number of active PAH (carcinogenic) classified as active; A/I, number of active PAH classified as inactive (non–carcinogenic); I/I, number of inactive PAH classified as inactive; I/A, number of inactive PAH classified as active; 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	A/A	A/I	I/I	I/A	CAa	CAp	ASV	ABSV	TRa	TRp	TEa	TEp
46	10	A	1.0	2	34	10	31	3	41	3	0.92	0.91	29.9	8.3	0.93	0.92	0.80	0.74
47	100		1.0	2	26	6	32	2	42	2	0.95	0.94	27.9	4.8	0.95	0.94	0.76	0.68
48	1000		1.0	2	27	3	34	0	43	1	0.99	0.97	24.9	1.8	0.99	0.97	0.73	0.68
49	10		2.0	2	32	6	33	1	41	3	0.95	0.92	29.3	5.4	0.95	0.92	0.80	0.74
50	100		2.0	2	26	4	34	0	42	2	0.97	0.94	26.2	3.2	0.98	0.96	0.76	0.71
51	1000		2.0	2	26	1	34	0	43	1	0.99	0.97	25.7	1.0	0.99	0.97	0.72	0.69
52	10		0.5	3	31	8	32	2	42	2	0.95	0.94	28.7	7.2	0.95	0.94	0.78	0.71
53	100		0.5	3	27	5	32	2	42	2	0.95	0.94	26.9	3.9	0.96	0.94	0.76	0.68
54	1000		0.5	3	25	1	34	0	43	1	0.99	0.97	22.9	0.9	0.99	0.97	0.78	0.74
55	10		1.0	3	30	6	32	2	42	2	0.95	0.94	29.0	4.9	0.95	0.94	0.77	0.69
56	100		1.0	3	27	2	34	0	43	1	0.99	0.97	25.6	1.3	0.99	0.97	0.74	0.71
57	1000		1.0	3	26	1	34	0	43	1	0.99	0.97	24.8	1.0	0.99	0.97	0.76	0.71
58	10		2.0	3	29	4	34	0	42	2	0.97	0.94	27.3	3.2	0.98	0.95	0.78	0.73
59	100		2.0	3	26	1	34	0	43	1	0.99	0.97	26.2	1.0	0.99	0.97	0.72	0.70
60	1000		2.0	3	26	1	34	0	43	1	0.99	0.97	28.0	0.9	0.99	0.97	0.72	0.69

The prediction performance of SVM models in structure–activity relationships strongly depends on the theoretical descriptors that numerically encode the molecular structure of all chemical compounds in the calibration (training) set. Although identifying the optimum set of structural descriptors is an important part of any SAR or QSAR study, procedures for descriptor selection are not currently available for SVM applications in SAR or QSAR. Therefore, in this study we have used the four quantum indices recently tested in a neural network model for the PAH carcinogenicity [27]: E_{HOMO}, the energy of the highest occupied molecular orbital; E_{LUMO}, energy of the lowest unoccupied molecular orbital; HD, hardness computed as HD = ($E_{LUMO} - E_{HOMO}$)/2; ΔH, difference between E_{HOMO} and E_{HOMO-1}.

A total of 60 SVM experiments were performed with the above four descriptors, with three values for the capacity parameter *C*, namely 10, 100, and 1000, and five kernels, namely dot, polynomial, radial basis function, neural, and anova. Each experiment consisted of a calibration phase that considered all 78 PAH, and a leave–10%–out cross–validation phase. As implemented in mySVM, *C* is scaled by 1/number of training examples. The calibration results reported in Table 2 are: SV, number of support vectors; BSV, number of bounded support vectors; A/A, number active PAH (carcinogenic) classified as active; A/I, number of active PAH classified as inactive (non–carcinogenic); I/I, number of inactive PAH classified as inactive; I/A, number of inactive PAH classified as active; CAa, accuracy; CAp, precision. For each SVM model we present in Table 2 the following leave–10%–out 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.

The first set of SVM experiments were obtained with the dot kernel (Table 2, experiments 1–3), but this simple kernel is not able to discriminate the carcinogenic and non–carcinogenic PAH. The number of support vectors is too large (almost 50), the prediction statistics are low, and the results in calibration are not very good, with classification errors for 6 active PAH and 8 inactive PAH.

The results obtained with the polynomial kernel, presented in Table 2 experiments 4–15, show an interesting trend: while calibration results improve when the polynomial degree increases from 2 to 5, the L10%O cross–validation prediction decreases with the increase of the polynomial degree. This is a clear demonstration of the fact that SVM models can be overfitted when too complex kernels are used. The L10%O cross–validation test is a reliable method for locating the SVM model with the best prediction power, although other cross–validation partitioning of the PAH set can offer equally good guiding.

The next group of models, presented in Table 2 experiments 16–24, was obtained with the radial basis function kernel, with $\gamma = 0.5$, 1.0, and 2.0. The same inverse relationship is identified between calibration statistics and L10%O cross–validation prediction: for example, the experiment 16 has the best prediction statistics (with CAa = 0.87, CAp = 0.88, TEa = 0.86, and TEp = 0.88) while the experiment 24 has the best calibration statistics (with CAa = 0.99, CAp = 0.97, TEa = 0.77, and TEp = 0.73). It is interesting to mention that the SVM from the experiment 24 has only one error in calibration, with one inactive compound classified as active; however, the prediction statistics are low compared with those from experiment 16.

The results obtained with the neural kernel, presented in Table 2 experiments 25–33, have the worst prediction statistics, with TEa between 0.53 and 0.61, and TEp between 0.46 and 0.56. While for the polynomial and radial kernels low prediction results are associated with high calibration performances, the neural kernel offers also low calibration statistics, with CAa between 0.56 and 0.72, CAp between 0.50 and 0.70. Our results show that the neural kernel is not suitable to separate the carcinogenic and non–carcinogenic PAH, but this finding should not be generalized. Additional SAR models must be investigated before a definite conclusion can be obtained regarding the utility of the hyperbolic tangent as a decision surface in SVM.

The last group of SVM models was obtained with the anova kernel (see Table 2, experiments 34–60), with overall good calibration (CAa between 0.86 and 0.99, CAp between 0.86 and 0.97) and L10%O prediction (TEa between 0.72 and 0.84, and TEp between 0.68 and 0.88) statistics. The experiments 48, 51, 54, 56, 57, 59, and 60 give the best calibration results, with only one inactive compound classified as active and all active compounds correctly classified. On the other hand, these seven SVM models have low prediction statistics, compared with those from experiments 34 and 43, both giving the best cross–validation results for anova kernels.

Experiment 34 (anova kernel, $\gamma = 0.5$, $d = 1$, 32 SV, A/I = 7, I/A = 4, CAa = 0.86, CAp = 0.87, TEa = 0.84, TEp = 0.88) and experiment 43 (anova kernel, $\gamma = 0.5$, $d = 2$, 33 SV, A/I = 5, I/A = 4, CAa = 0.88, CAp = 0.88, TEa =0.84, TEp = 0.84) have close statistics with experiment 16 (radial kernel, $\gamma = 0.5$, 36 SV, A/I = 6, I/A = 4, CAa = 0.87, CAp = 0.88, TEa =0.86, TEp = 0.88). These three SVM models have the best prediction results from the whole set of 60 experiments. Because their statistical indices are very close, all these three SAR models are equivalent from a statistical

point of view. Finding several SAR models with similar statistics is common when from a set of data one generates several structure–activity models. A comparison of the classification errors in the SVM calibration reveals that a group of PAH is responsible for the majority of these errors. In the experiment 16 the group A/I is formed by the PAH **7**, **9**, **48**, **49**, **50**, and **51**, while the I/A group is formed by the PAH **20**, **31**, **66**, and **78**. In the experiment 34 the group A/I is formed by the PAH **4**, **7**, **9**, **48**, **49**, **50**, and **51**, while the I/A group is formed by the PAH **20**, **31**, **66**, and **78**. In the experiment 43 the group A/I is formed by the PAH **7**, **9**, **48**, **49**, and **51**, while the I/A group is formed by the PAH **20**, **31**, **66**, and **78**. The consensus for these three SVM models is that six active PAH are classified as inactive, namely dibenz[1,2;5,6]anthracene, dibenzo[1,2;3,4]phenantrene, 5–methylbenzo[*c*]phenanthrene, 5–methylchrysene, 6,8–dimethylbenz[*a*]anthracene, and 7–methyl–benz[*a*]anthracene, while four inactive PAH are classified as active, namely benzo[1,2]naphthacene, benzo[*ghi*]perylene, 10–methylbenzo[*a*]pyrene, and 7,10–dimethylbenzo[*a*]pyrene. These SVM classification errors can be obtained due to errors in assigning their experimental carcinogenic activity, or because the four structural descriptors used in the SVM model are not appropriate for these PAH. Further investigations are needed to improve the classification of these PAH.

4 CONCLUSIONS

Support vector machines represent an efficient machine learning algorithm that separate two classes of patterns by determining a unique hyperplane that maximizes the separation between the two classes. In this study we have investigated the application of SVM for the classification of the carcinogenic activity for 32 PAH and 46 methylated PAH taken from literature [23–27]. From this set of 78 PAH, 34 are carcinogenic and 44 are non–carcinogenic. All SVM models were obtained with four theoretical descriptors computed with the PM3 semiempirical method, previously used to classify the same set of PAH with an artificial neural network [27], namely the energy of the highest occupied molecular orbital E_{HOMO}, energy of the lowest unoccupied molecular orbital E_{LUMO}, hardness HD, and difference between E_{HOMO} and E_{HOMO-1} denoted ΔH. In any SAR model, the selection of the best structural descriptors is equally important and difficult. Because there is no simple algorithm for descriptor selection in SVM models, we have used the theoretical indices from [27].

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. The prediction power of each SVM model was evaluated with a leave–10%–out cross–validation procedure. Our experiments with various kernels clearly demonstrate that the performance of the SVM classifier is strongly dependent on the kernel shape. Overall, the dot and neural kernels give low quality SAR models, with no practical use for the classification of carcinogenic PAH. The polynomial kernel gives fairly good results, while the best classification of carcinogenic PAH is obtained with the

radial and anova kernels. It is interesting to mention that for the polynomial, radial and anova kernels the SVM models with good calibration results have low leave–10%–out cross–validation results, while SVM models with good prediction results have low calibration results. This result clearly demonstrates that too complex kernels give overfitted SVM models, with low prediction power. Using complex kernels, SVM can be calibrated to maximize the separation of two classes of patterns, but the best calibration models are usually associate with poor predictions and only a cross–validation test can demonstrate the potential utility of an SVM model.

The best prediction results are obtained with the radial kernel ($\gamma = 0.5$) and anova kernel ($\gamma = 0.5$, $d = 1$; $\gamma = 0.5$, $d = 2$), with close calibration and cross–validation statistics. The general conclusion for these three SVM models is that six active PAH are classified as inactive, namely dibenz[1,2;5,6]anthracene, dibenzo[1,2;3,4]phenantrene, 5–methylbenzo[*c*]phenanthrene, 5–methylchrysene, 6,8–dimethylbenz[*a*]anthracene, and 7–methylbenz[*a*]anthracene, while four inactive PAH are classified as active, namely benzo[1,2]naphthacene, benzo[*ghi*]perylene, 10–methylbenzo[*a*]pyrene, and 7,10–dimethylbenzo[*a*]pyrene. These classification errors obtained in the SVM models can be explained by errors in assigning their experimental carcinogenic activity, or because other structural descriptors should be used for the classification of carcinogenic PAH.

This study demonstrates that SVM models can be used with success to discriminate between carcinogenic and non–carcinogenic PAH, providing reliable predictions. Further studies regarding the use of SVM in structure–activity relationships should explore the important problem of descriptor selection. Considerable effort should be directed also towards the investigation of various kernel functions, with the aim to develop reliable methods for selecting the best kernel for a particular classification problem.

Supplementary Material

The mySVM model files for experiments 16, 34, and 43 are available as supplementary material.

5 REFERENCES

[1] P. Pott, Chirurgical Observations Relative to the Cataracts, the Polypus of the Nose, the Cancer of the Scrotum, the Different Kinds of Ruptures and the Mortifications of the Toes and Feet, Hawes, Clarke and Collins, London, 1775.
[2] I. M. M. van Leeuwen and C. Zonneveld, From Exposure to Effect: A Comparison of Modeling Approaches to Chemical Carcinogenesis, *Mutat. Res*. **2001**, *489*, 17–45.
[3] N. Voiculetz, A. T. Balaban, I. Niculescu–Duvăz, and Z. Simon, *Modeling of Cancer Genesis and Prevention*, CRC Press, Boca Raton, Florida, 1990.
[4] A. Pullman and B. Pullman, Cancerisation par les Substances Chimiques et Structure Moleculaire, Masson, Paris, 1955.
[5] D. Malacarne, M. Taningher, R. Pesenti, M. Paolucci, A. Perrotta, and S. Parodi, Molecular Fragments Associated with Non–Genotoxic Carcinogens, as Detected Using a Software Program Based on Graph Theory: Their Usefulness to Predict Carcinogenicity, *Chem.–Biol. Interact*. **1995**, *97*, 75–100.
[6] A. Long and R. D. Combes, Using DEREK to Predict the Activity of Some Carcinogens/Mutagens Found in Foods, *Toxicol. Vitro* **1995**, *9*, 563–569.
[7] M. Sjögren, L. Ehrenberg, and U. Rannug, Relevance of Different Biological Assays in Assessing Initiating and

Promoting Properties of Polycyclic Aromatic Hydrocarbons with Respect to Carcinogenic Potency, *Mutat. Res*. **1996**, 358, 97–112.
[8] M. Taningher, D. Malacarne, T. Mancuso, M. Peluso, M. P. Pescarolo, and S. Parodi, Methods for Predicting Carcinogenic Hazards: New Opportunities Coming From Recent Developments in Molecular Oncology and SAR Studies, *Mutat. Res*. **1997**, *391*, 3–32.
[9] H. S. Rosenkranz, Y. P. Zhang, and G. Klopman, Studies on the Potential for Genotoxic Carcinogenicity of Fragrances and Other Chemicals, *Food Chem. Tox*. **1998**, *36*, 687–696.
[10] R. Benigni, The First US National Toxicology Program Exercise on the Prediction of Rodent Carcinogenicity: Definitive Results, *Mutat. Res*. **1998**, *387*, 35–45.
[11] A. R. Cunningham, H. S. Rosenkranz, Y. P. Zhang, and G. Klopman, Identification of 'Genotoxic' and 'Non–Genotoxic' Alerts for Cancer in Mice: The Carcinogenic Potency Database, *Mutat. Res*. **1998**, *398*, 1–17.
[12] H. S. Rosenkranz and M. H. Karol, Chemical Carcinogenicity: Can it be Predicted from Knowledge of Mutagenicity and Allergic Contact Dermatitis? *Mutat. Res*. **1999**, *431*, 81–91.
[13] H. S. Rosenkranz, Allergic Contact Dermatitis and its Relationship to Carcinogenicity, *Mutat. Res*. **2001**, *483*, 51–55.
[14] R. G. Harvey, Polycyclic Aromatic Hydrocarbons: Chemistry and Carcinogenesis, Cambridge University Press, Cabmridge, England, 1991.
[15] Y. Miyashita, T. Seki, Y. Takahashi, S.–I. Daiba, Y. Tanaka, Y. Yotsui, H. Abe, and S.–I. Sasaki, Computer–Assisted Structure–Carcinogenicity Studies on Polycyclic Aromatic Hydrocarbons by Pattern Recognition Methods, *Anal. Chim. Acta* **1981**, *133*, 603–613.
[16] Y. Miyashita, Y. Takahashi, S.–I. Daiba, H. Abe, and S.–I. Sasaki, Computer–Assisted Structure–Carcinogenicity Studies on Polynuclear Aromatic Hydrocarbons by Pattern Recognition Methods; The Role of the Bay and L–Regions, *Anal. Chim. Acta* **1982**, *143*, 35–44.
[17] W. C. Herndon and L. V. Szentpaly, Theoretical Model of Activation of Carcinogenic Polycyclic Benzenoid Aromatic Hydrocarbons. Possible New Classes of Carcinogenic Aromatic Hydrocarbons, *J. Mol. Struct.* (*Theochem*) **1986**, *148*, 141–152.
[18] Y. Miyashita, T. Okuyama, K. Yamaura, K. Jinno, and S.–I. Sasaki, Prediction of Carcinogenicity of Polynuclear Aromatic Hydrocarbons on the Basis of Their Chemical Structures, *Anal. Chim. Acta* **1987**, *202*, 237–240.
[19] J. Gayoso and S. Kimri, Sur une Tentative d'Unification des Théories Quantiques de la Cancérisation par les Polyacènes: I. Théorie des Régions M, L, et B, *Int. J. Quantum Chem*. **1990**, *38*, 461–486.
[20] J. Gayoso and S. Kimri, Sur une Tentative d'Unification des Théories Quantiques de la Cancérisation par les Polyacènes. II: Le Rôle de la Région K dans le Processus d'Activation Metabolique Conduisant au Cancérogène Ultime. Théorie des Régions M, L, et BK, *Int. J. Quantum Chem*. **1990**, *38*, 487–495.
[21] J. W. Flesher, J. Horn, and A. F. Lehner, Molecular Modeling of Carcinogenic Potential in Polycyclic Hydrocarbons, *J. Mol. Struct.* (*Theochem*) **1996**, *362*, 29–49.
[22] A. F. Lehner, J. Horn, and J. W. Flesher, Benzylic Carbonium Ions as Ultimate Carcinogens of Polynuclear Aromatic Hydrocarbons, *J. Mol. Struct.* (*Theochem*) **1996**, *366*, 203–217.
[23] P. M. V. B. Barone, A. Camilo Jr., and D. S. Galvão, Theoretical Approach to Identify Carcinogenic Activity of Polycyclic Aromatic Hydrocarbons, *Phys. Rev. Lett*. **1996**, *77*, 1186–1189.
[24] R. Vendrame, R. S. Braga, Y. Takahata, and D. S. Galvão, Structure–Activity Relationship Studies of Carcinogenic Activity of Polycyclic Aromatic Hydrocarbons Using Calculated Molecular Descriptors with Principal Component Analysis and Neural Network Methods, *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 1094–1104.
[25] R. S. Braga, P. M. V. B. Barone, and D. S. Galvão, Identifying Carcinogenic Activity of Methylated Polycyclic Aromatic Hydrocarbons (PAHs), *J. Mol. Struct.* (*Theochem*) **1999**, *464*, 257–266.
[26] P. M. V. B. Barone, R. S. Braga, A. Camilo Jr., and D. S. Galvão, Electronic Indices from Semi–Empirical Calculations to Identify Carcinogenic Activity of Polycyclic Aromatic Hydrocarbons, *J. Mol. Struct.* (*Theochem*) **2000**, *505*, 55–66.
[27] R. Vendrame, R. S. Braga, Y. Takahata, and D. S. Galvão, Structure–Carcinogenic Activity Relationship Studies of Polycyclic Aromatic Hydrocarbons (PAHs) with Pattern–Recognition Methods, *J. Mol. Struct.* (*Theochem*) **2001**, *539*, 253–265.
[28] D. J. G. Marino, P. J. Peruzzo, E. A. Castro, and A. A. Toropov, QSAR Carcinogenic Study of Methylated Polycyclic Aromatic Hydrocarbons Based on Topological Descriptors Derived from Distance Matrices and Correlation Weights of Local Graph Invariants, *Internet Electron. J. Mol. Des.* **2002**, *1*, 115–133, http://www.biochempress.com.
[29] V. Vapnik, *Estimation of Dependencies Based on Empirical Data*, Nauka, Moscow, 1979.
[30] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, 1995.
[31] V. Vapnik, *Statistical Learning Theory*, Wiley–Interscience, New York, 1998.
[32] C. J. C. Burges, A Tutorial on Support Vector Machines for Pattern Recognition, *Data Mining Knowledge Discov.* **1998**, *2*, 121–167.

[33] B. Schölkopf, C. J. C. Burges, and A. J. Smola, *Advances in Kernel Methods*: *Support Vector Learning*, MIT Press, Cambridge, MA, 1999.
[34] N. Cristianini and J. Shawe–Taylor, *An Introduction to Support Vector Machines*, Cambridge University Press, Cambridge, 2000.
[35] K.–R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf, An Introduction to Kernel–Based Learning Algorithms, *IEEE Trans. Neural Networks* **2001**, *12*, 181–201.
[36] C.–C. Chang and C.–J. Lin, Training ν–Support Vector Classifiers: Theory and Algorithms, *Neural Comput*. **2001**, *12*, 2119–2147.
[37] I. Steinwart, On the Influence of the Kernel on the Consistency of Support Vector Machines, *J. Machine Learning Res.* **2001**, *2*, 67–93, http://www.jmlr.org.
[38] A. Ben–Hur, D. Horn, H. T. Siegelmann, and V. Vapnik, Support Vector Clustering, *J. Machine Learning Res.* **2001**, *2*, 125–137, http://www.jmlr.org.
[39] R. Collobert and S. Bengio, SVMTorch: Support Vector Machines for Large–Scale Regression Problems, *J. Machine Learning Res*. **2001**, *1*, 143–160, http://www.jmlr.org.
[40] O. L. Mangasarian and D. R. Musicant, Lagrangian Support Vector Machines, *J. Machine Learning Res.* **2001**, *1*, 161–177, http://www.jmlr.org.
[41] M. P. S. Brown, W. Noble Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. Ares Jr., and D. Haussler, Knowledge–Based Analysis of Microarray Gene Expression Data by Using Support Vector Machines, *Proc. Natl. Acad. Sci. U. S. A*. **2000**, *97*, 262–267.
[42] A. Zien, G. Ratsch, S. Mika, B. Schölkopf, T. Lengauer, and K. R. Muller, Engineering Support Vector Machine Kernels that Recognize Translation Initiation Sites, *Bioinformatics* **2000**, *16*, 799–807.
[43] R. J. Carter, I. Dubchak, and S. R. Holbrook, A Computational Approach to Identify Genes for Functional RNAs in Genomic Sequences, *Nucleic Acids Res*. **2001**, *29*, 3928–3938.
[44] 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.
[45] 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.
[46] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, Gene Selection for Cancer Classification Using Support Vector Machines, *Machine Learning* **2002**, *46*, 389–422.
[47] C.–H. Yeang, S. Ramaswamy, P. Tamayo, S. Mukherjee, R. M. Rifkin, M. Angelo, M. Reich, E. Lander, J. Mesirov, and T. Golub, Molecular Classification of Multiple Tumor Types, *Bioinformatics* **2001**, *17*, S316–S322.
[48] S. Dreiseitl, L. Ohno–Machado, H. Kittler, S. Vinterbo, H. Billhardt, and M. Binder, A Comparison of Machine Learning Methods for the Diagnosis of Pigmented Skin Lesions, *J. Biomed. Informat*. **2001**, *34*, 28–36.
[49] 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.
[50] R. Karchin, K. Karplus, and D. Haussler, Classifying G–Protein Coupled Receptors with Support Vector Machines, *Bioinformatics* **2002**, *18*, 147–159.
[51] 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.
[52] Y.–D. Cai, X.–J. Liu, X. Xu, and K.–C. Chou, Support Vector Machines for Predicting Membrane Protein Types by Incorporating Quasi–Sequence–Order Effect, *Internet Electron. J. Mol. Des.* **2002**, *1*, 219–226, http://www.biochempress.com.
[53] J. R. Bock and D. A. Gough, Predicting Protein–Protein Interactions from Primary Structure, *Bioinformatics* **2001**, *17*, 455–460.
[54] S. J. Hua and Z. R. Sun, Support Vector Machine Approach for Protein Subcellular Localization Prediction, *Bioinformatics* **2001**, *17*, 721–728.
[55] Y.–D. Cai, X.–J. Liu, X.–B. Xu, and K.–C. Chou, Support Vector Machines for Prediction of Protein Subcellular Location, *Mol. Cell Biol. Res. Commun*. **2000**, *4*, 230–233.
[56] 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.
[57] C. H. Q. Ding and I. Dubchak, Multi–Class Protein Fold Recognition Using Support Vector Machines and Neural Networks, *Bioinformatics* **2001**, *17*, 349–358.
[58] 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.
[59] Y.–D. Cai, X.–J. Liu, X.–B. Xu, and K.–C. Chou, Support Vector Machines for Predicting the Specificity of GalNAc–Transferase, *Peptides* **2002**, *23*, 205–208.
[60] 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.
[61] C. W. Morris, A. Autret, and L. Boddy, Support Vector Machines for Identifying Organisms – A Comparison with Strongly Partitioned Radial Basis Function Networks, *Ecological Model*. **2001**, *146*, 57–67.
[62] 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.
[63] S. Rüping, mySVM, University of Dortmund, http://www–ai.cs.uni–dortmund.de/SOFTWARE/MYSVM/.
[64] BioChem Links, http://www.biochempress.com.