Inter*net* Electronic Journal of Molecular Design

March 2003, Volume 2, Number 3, Pages 195–208

Editor: Ovidiu Ivanciuc

Special issue dedicated to Professor Haruo Hosoya on the occasion of the 65th birthday
Part 7

Guest Editor: Jun–ichi Aihara

Aquatic Toxicity Prediction for Polar and Nonpolar Narcotic Pollutants with Support Vector Machines

Ovidiu Ivanciuc

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

Received: December 1, 2002; Accepted: January 25, 2003; Published: March 31, 2003

Citation of the article:
O. Ivanciuc, Aquatic Toxicity Prediction for Polar and Nonpolar Narcotic Pollutants with Support Vector Machines, *Internet Electron. J. Mol. Des.* **2003**, *2*, 195–208, http://www.biochempress.com.

Inter*net* Electronic Journal of Molecular Design

BioChem Press
http://www.biochempress.com

Aquatic Toxicity Prediction for Polar and Nonpolar Narcotic Pollutants with Support Vector Machines[#]

Ovidiu Ivanciuc*

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

Received: December 1, 2002; Accepted: January 25, 2003; Published: March 31, 2003

Internet Electron. J. Mol. Des. **2003**, *2* (3), 195–208

Abstract

Motivation. Narcotic pollutants, that act by nonspecifically disrupting the functioning of cell membranes, are categorized as polar and nonpolar compounds. The toxicity prediction of narcotic pollutants with QSAR (quantitative structure-activity relationships) depends on the reliable determination of the mechanism of toxic action. The classification of the chemical compounds as polar and nonpolar narcotic pollutants based on structural characteristics is of utmost importance in predicting the aquatic toxicity for new chemicals.

Method. Support vector machine (SVM) is a new machine learning algorithm that proved to be reliable in the classification of organic and bioorganic compounds. In this study we have investigated the application of SVM for the classification of 190 narcotic pollutants (76 polar and 114 nonpolar). Using an efficient descriptor selection algorithm, the energy of the highest occupied molecular orbital, the energy of the lowest unoccupied molecular orbital, and the most negative partial charge on any non–hydrogen atom in the molecule, all computed with the AM1 method, were found to be necessary for the discrimination of the polar and nonpolar compounds. The prediction power of each SVM model was evaluated with a leave–20%–out cross–validation procedure.

Results. The classification performances of SVM models generated with the dot, polynomial, radial basis function, neural, and anova kernels, show that the statistical performances of SVM depend strongly on the kernel type and various parameters that control the kernel shape. An SVM model obtained with the anova kernel offered the best results, with three errors in calibration and four errors in prediction, all for nonpolar chemicals.

Conclusions. SVM is a powerful and flexible classification algorithm, with many potential applications in molecular design, optimization of chemical libraries, and QSAR. In the present study we have demonstrated such an application for the identification of the aquatic toxicity mechanism.

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

1 INTRODUCTION

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 [1–16]. Significant progress has been made to classify chemical compounds according to their

[#] Dedicated to Professor Haruo Hosoya on the occasion of the 65th birthday.

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

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 [7–16].

Organic compounds that act as narcotic pollutants are considered to disrupt the functioning of cell membranes. Narcotic pollutants are represented by two classes of compounds, namely nonpolar (class 1) and polar (class 2) compounds. The toxicity of both polar and nonpolar narcotic pollutants depends on the octanol-water partition coefficient, but the toxicity of polar compounds depends also on the propensity of forming hydrogen bonds. Recently, Ren [15] developed nonlinear discriminant functions to separate polar and nonpolar narcotic pollutants based on their octanol-water partition coefficients and hydrogen bonding quantum descriptors computed with the AM1 method. 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 present the application of SVM for the classification of polar and nonpolar narcotic pollutants using the dataset explored in Ref. [15]. 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. A new algorithm for selecting relevant structural descriptors in SVM models was tested with good results in reducing the input space.

2 MATERIALS AND METHODS

2.1 Chemical Data

Ren [15] used five structural descriptors to discriminate between 76 polar and 114 nonpolar pollutants, 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}, the most negative partial charge on any non–hydrogen atom in the molecule Q^-, and the most positive partial charge on a hydrogen atom Q^+. All quantum descriptors were computed with the AM1 method. The 190 compounds investigated in the present study, together with their classification into polar/nonpolar pollutants, were taken from two recent studies [14,15] and are presented in Table 1 together with the three quantum descriptors used in the best SVM model to discriminate between their toxicity mechanism (E_{HOMO}, E_{LUMO}, and Q^-) and the experimental, calibration and prediction classification (nonpolar, +1; polar, –1).

Among the 190 compounds, 114 are nonpolar and 76 are polar pollutants. The nonlinear discriminant analysis [15] was tested in a leave-one-out cross-validation test that gave eight classification errors, namely 2–phenoxyethanol, 2,3,4–trimethoxyacetophenone, acetophenone,

benzophenone, 2,4–dichloroacetophenone, 2–hydroxy–4–methoxyacetophenone, 1,4–dichlorobenzene (all nonpolar compounds predicted as polar), and pyridine (polar compound predicted as nonpolar).

Table 1. Structure of the chemical compounds, theoretical descriptors (E_{HOMO}, E_{LUMO} and Q^-) and mechanism of toxic action (nonpolar, class +1; polar, class –1; experimental, Exp; calibration, Cal; prediction, Pre)

No	Compound	E_{HOMO}	E_{LUMO}	Q^-	SVM Class		
					Exp	Cal	Pre
1	methanol	–11.135	3.7775	–0.5353	+1	+1	+1
2	ethanol	–11.050	3.6513	–0.5360	+1	+1	+1
3	1–propanol	–10.940	3.6324	–0.5317	+1	+1	+1
4	2–propanol	–10.895	3.4925	–0.5469	+1	+1	+1
5	1–butanol	–10.940	3.5041	–0.5422	+1	+1	+1
6	2–butanol	–10.952	3.5536	–0.5456	+1	+1	+1
7	isobutanol	–10.858	3.5052	–0.5476	+1	+1	+1
8	*tert*–butyl alcohol	–10.991	3.4384	–0.5517	+1	+1	+1
9	1–pentanol	–10.940	3.5041	–0.5422	+1	+1	+1
10	3–pentanol	–10.805	3.4884	–0.5394	+1	+1	+1
11	1–hexanol	–10.930	3.4642	–0.5506	+1	+1	+1
12	1–heptanol	–10.924	3.4300	–0.5517	+1	+1	+1
13	1–octanol	–10.917	3.4174	–0.5526	+1	+1	+1
14	1–nonanol	–10.912	3.4031	–0.5539	+1	+1	+1
15	1–decanol	–10.907	3.3928	–0.5539	+1	+1	+1
16	1–undecanol	–10.903	3.3851	–0.5524	+1	+1	+1
17	1–dodecanol	–10.900	3.3793	–0.5506	+1	+1	+1
18	1,2–ethanediol	–10.946	3.2671	–0.5293	+1	+1	+1
19	1,3–propenediol	–9.493	1.0283	–0.5567	+1	+1	+1
20	2–methyl–2,4–pentanediol	–10.677	3.1360	–0.5777	+1	+1	+1
21	3–furanmethanol	–9.176	0.7497	–0.5465	+1	–1	–1
22	cyclohexanol	–10.304	0.9217	–0.4832	+1	+1	+1
23	2,2,2–trichloroethanol	–11.578	–0.4003	–0.5113	+1	+1	–1
24	butyldigol	–10.523	2.4765	–0.5258	+1	+1	+1
25	diethyleneglycol	–10.982	2.4265	–0.5148	+1	+1	+1
26	triethyleneglycol	–10.281	2.3815	–0.5460	+1	+1	+1
27	2–methoxyethanol	–10.807	2.8028	–0.5114	+1	+1	+1
28	2–ethoxyethanol	–10.687	2.6958	–0.5150	+1	+1	+1
29	2–isopropoxyethanol	–10.670	2.6498	–0.5233	+1	+1	+1
30	2–butoxyethanol	–10.650	2.6755	–0.5209	+1	+1	+1
31	2–(2–ethoxyethoxy)ethanol	–10.584	2.3600	–0.5514	+1	+1	+1
32	2–phenoxyethanol	–8.973	0.5669	–0.5669	+1	–1	–1
33	acetone	–10.668	0.8443	–0.4700	+1	+1	+1
34	2–propanone	–10.646	0.8489	–0.4779	+1	+1	+1
35	2–butanone	–10.541	0.8772	–0.4659	+1	+1	+1
36	3–pentanone	–10.420	0.9096	–0.4578	+1	+1	+1
37	2–octanone	–10.512	0.8723	–0.4751	+1	+1	+1
38	5–nonanone	–10.392	0.9090	–0.4763	+1	+1	+1
39	2–decanone	–10.509	0.8715	–0.4726	+1	+1	+1
40	3–methyl–2–butanone	–10.409	0.9131	–0.4635	+1	+1	+1
41	6–methyl–5–hepten–2–one	–9.445	0.8556	–0.4760	+1	+1	+1
42	2,3,4–trimethoxyacetophenone	–9.581	–0.4590	–0.4887	+1	+1	+1
43	acetophenone	–9.936	–0.3606	–0.4591	+1	+1	+1
44	3,3–dimethyl–2–butanone	–10.337	0.9430	–0.4722	+1	+1	+1
45	4–methyl–2–pentanone	–10.493	0.8962	–0.4713	+1	+1	+1
46	benzophenone	–9.875	–0.4759	–0.4512	+1	+1	+1
47	2,4–dichloroacetophenone	–9.890	–0.5146	–0.4423	+1	+1	+1
48	cyclohexanone	–10.616	3.3960	–0.5584	+1	+1	+1

Table 1. (Continued)

No	Compound	E_{HOMO}	E_{LUMO}	Q^-	SVM Class		
					Exp	Cal	Pre
49	ethyl acetate	–11.006	1.1370	–0.5045	+1	+1	+1
50	diethyl ether	–10.393	2.9807	–0.4057	+1	+1	+1
51	diiso–propyl ether	–10.383	2.8648	–0.5014	+1	+1	+1
52	dibutyl ether	–10.388	2.8852	–0.4487	+1	+1	+1
53	dipentyl ether	–10.389	2.8700	–0.4523	+1	+1	+1
54	diphenyl ether	–8.955	0.1708	–0.4029	+1	+1	+1
55	*tert*–butylmethyl ether	–10.431	2.9892	–0.4234	+1	+1	+1
56	furan	–9.317	0.7228	–0.2135	+1	+1	+1
57	tetrahydrofuran	–10.180	3.1103	–0.3943	+1	+1	+1
58	2,6–dimethoxytoluene	–9.424	0.2306	–0.3773	+1	+1	+1
59	1,4–dimethoxybenzene	–8.568	0.3924	–0.3696	+1	+1	+1
60	2–hydroxy–4–methoxyacetophenone	–9.119	–0.0249	–0.4636	+1	–1	–1
61	dichloromethane	–11.390	0.5946	–0.1854	+1	+1	+1
62	chloroform	–11.771	–0.3035	–0.2708	+1	+1	+1
63	tetrachloromethane	–12.379	–1.1170	–0.2974	+1	+1	+1
64	1,1–dichloroethane	–11.422	0.5822	–0.1724	+1	+1	+1
65	1,2–dichloroethane	–11.417	0.6838	–0.1151	+1	+1	+1
66	1,1,1–trichloroethane	–11.992	–0.2658	–0.1807	+1	+1	+1
67	1,1,2–trichloroethane	–11.513	0.3239	–0.1659	+1	+1	+1
68	1,1,2,2–tetrachloroethane	–11.655	–0.0738	–0.2785	+1	+1	+1
69	pentachloroethane	–11.870	–0.6817	–0.2966	+1	+1	+1
70	hexachloroethane	–12.182	–0.9677	–0.2913	+1	+1	+1
71	1,2–dichloropropane	–11.290	1.1169	–0.2122	+1	+1	+1
72	1,3–dichloropropane	–11.372	1.0193	–0.1625	+1	+1	+1
73	1,2,3–trichloropropane	–11.442	0.7594	–0.2074	+1	+1	+1
74	1–chlorobutane	–11.133	1.5109	–0.1880	+1	+1	+1
75	trichloroethene	–9.956	–0.0608	–0.0901	+1	+1	+1
76	tetrachloroethene	–9.902	–0.4367	–0.0372	+1	+1	+1
77	hexachlorobutadiene	–9.542	–1.3444	–0.1091	+1	+1	+1
78	lindane	–11.475	0.2284	–0.1923	+1	+1	+1
79	chlorobenzene	–9.561	0.1545	–0.1262	+1	+1	+1
80	1,2–dichlorobenzene	–9.602	–0.1425	–0.1028	+1	+1	+1
81	1,3–dichlorobenzene	–9.682	–0.1580	–0.1298	+1	+1	+1
82	1,4–dichlorobenzene	–9.523	–0.2162	–0.7993	+1	+1	+1
83	1,2,3–trichlorobenzene	–9.784	–0.3646	–0.1345	+1	+1	+1
84	1,2,4–trichlorobenzene	–9.623	–0.4691	–0.1004	+1	+1	+1
85	1,3,5–trichlorobenzene	–9.921	–0.4022	–0.1888	+1	+1	+1
86	1,2,3,4–tetrachlorobenzene	–9.735	–0.6518	–0.0587	+1	+1	+1
87	1,2,3,5–tetrachlorobenzene	–9.763	–0.6841	–0.1772	+1	+1	+1
88	1,2,4,5–tetrachlorobenzene	–9.655	–0.7308	–0.0512	+1	+1	+1
89	3–chlorotoluene	–9.444	0.1844	–0.2176	+1	+1	+1
90	4–chlorotoluene	–9.299	0.1351	–0.2161	+1	+1	+1
91	2,4–dichlorotoluene	–9.447	–0.1489	–0.2153	+1	+1	+1
92	2,4,5–trichlorotoluene	–9.475	–0.4355	–0.2593	+1	+1	+1
93	3,4–dichlorotoluene	–9.407	–0.1363	–0.2519	+1	+1	+1
94	pentachlorobenzene	–9.786	–0.8904	–0.0571	+1	+1	+1
95	2–chloronaphthalene	–8.868	–0.5063	–0.1939	+1	+1	+1
96	hexane	–11.084	3.7357	–0.1641	+1	+1	+1
97	octane	–11.066	3.6386	–0.1330	+1	+1	+1
98	decane	–11.063	3.5774	–0.1293	+1	+1	+1
99	benzene	–9.653	0.5551	–0.0921	+1	+1	+1
100	toluene	–9.330	0.5204	–0.1922	+1	+1	+1
101	*o*–xylene	–9.183	0.5231	–0.1838	+1	+1	+1
102	*m*–xylene	–9.186	0.5250	–0.1782	+1	+1	+1
103	*p*–xylene	–9.062	0.4871	–0.1846	+1	+1	+1

Table 1. (Continued)

No	Compound	E_{HOMO}	E_{LUMO}	Q^-	SVM Class		
					Exp	Cal	Pre
104	1,2,4–trimethylbenzene	–8.972	0.5030	–0.2105	+1	+1	+1
105	1,3,5–trimethylbenzene	–9.165	0.5756	–0.2229	+1	+1	+1
106	1,2,4,5–tetramethylbenzene	–8.832	0.4947	–0.2022	+1	+1	+1
107	ethylbenzene	–9.381	0.5281	–0.1464	+1	+1	+1
108	cumene	–9.383	0.5417	–0.1710	+1	+1	+1
109	1–methylnaphthalene	–8.584	–0.2668	–0.1574	+1	+1	+1
110	2–methylnaphthalene	–8.620	–0.2459	–0.1959	+1	+1	+1
111	biphenyl	–8.952	–0.0680	–0.1264	+1	+1	+1
112	cyclopentane	–10.970	3.6228	–0.1258	+1	+1	+1
113	cyclohexane	–10.937	3.6562	–0.0753	+1	+1	+1
114	methylcyclohexane	–10.822	3.6095	–0.2031	+1	+1	+1
115	nitrobenzene	–10.562	–1.0679	–0.4939	–1	–1	–1
116	2–nitrotoluene	–10.171	–1.0109	–0.5043	–1	–1	–1
117	3–nitrotoluene	–10.197	–1.0138	–0.4984	–1	–1	–1
118	4–nitrotoluene	–10.305	–1.0449	–0.5017	–1	–1	–1
119	2,3–dimethylnitrobenzene	–9.941	–0.9491	–0.5097	–1	–1	–1
120	3,4–dimethylnitrobenzene	–10.077	–0.9975	–0.5050	–1	–1	–1
121	2–chloronitrobenzene	–10.332	–1.0722	–0.4984	–1	–1	–1
122	3–chloronitrobenzene	–10.367	–1.2855	–0.4842	–1	–1	–1
123	4–chloronitrobenzene	–10.475	–1.3436	–0.4911	–1	–1	–1
124	2,3–dichloronitrobenzene	–10.283	–1.2297	–0.4900	–1	–1	–1
125	2,4–dichloronitrobenzene	–10.470	–1.3555	–0.4938	–1	–1	–1
126	2,5–dichloronitrobenzene	–10.218	–1.2921	–0.4879	–1	–1	–1
127	3,5–dichloronitrobenzene	–10.416	–1.4880	–0.4772	–1	–1	–1
128	2–chloro–6–nitrotoluene	–10.146	–0.8587	–0.4966	–1	–1	–1
129	4–chloro–2–nitrotoluene	–10.324	–1.2798	–0.4952	–1	–1	–1
130	4–chloro–3–nitrotoluene	–10.036	–1.0159	–0.5006	–1	–1	–1
131	phenol	–9.114	0.3976	–0.4958	–1	–1	–1
132	2–methylphenol	–8.960	0.4093	–0.4813	–1	–1	–1
133	3–methylphenol	–9.052	0.3732	–0.4963	–1	–1	–1
134	4–methylphenol	–8.880	0.4317	–0.4927	–1	–1	–1
135	2,4–dimethylphenol	–8.784	0.3979	–0.4980	–1	–1	–1
136	2,6–dimethylphenol	–8.885	0.3940	–0.4751	–1	–1	–1
137	3,4–dimethylphenol	–8.803	0.4360	–0.4982	–1	–1	–1
138	2,3,6–trimethylphenol	–8.833	0.3648	–0.4751	–1	–1	–1
139	2,4,6–trimethylphenol	–8.691	0.4322	–0.4750	–1	–1	–1
140	4–ethylphenol	–8.912	0.4334	–0.4931	–1	–1	–1
141	4–propylphenol	–8.903	0.4383	–0.4964	–1	–1	–1
142	4–*n*–butylphenol	–8.903	0.4362	–0.4930	–1	–1	–1
143	4–*tert*–butylphenol	–8.894	0.4709	–0.4990	–1	–1	–1
144	2–*tert*–butyl–4–methylphenol	–8.761	0.4780	–0.4381	–1	–1	–1
145	4–*n*–pentylphenol	–8.902	0.4370	–0.4951	–1	–1	–1
146	4–*tert*–pentylphenol	–8.885	0.4722	–0.4992	–1	–1	–1
147	2–allylphenol	–9.016	0.3597	–0.4818	–1	–1	–1
148	2–phenylphenol	–8.731	–0.0489	–0.4813	–1	–1	–1
149	1–naphthol	–8.455	–0.2472	–0.4810	–1	–1	–1
150	4–chlorophenol	–9.125	0.0946	–0.4928	–1	–1	–1
151	4–chloro–3–methylphenol	–9.051	0.0930	–0.4894	–1	–1	–1
152	4–chloro–3,5–dimethylphenol	–8.977	0.1466	–0.4982	–1	–1	–1
153	3–methoxyphenol	–8.941	0.4134	–0.4939	–1	–1	–1
154	4–methoxyphenol	–8.636	0.3034	–0.4790	–1	–1	–1
155	4–phenoxyphenol	–8.806	0.1133	–0.4904	–1	–1	–1
156	pyridine	–9.932	0.1385	–0.6610	–1	–1	–1
157	quinoline	–9.181	–0.4666	–0.6538	–1	–1	–1
158	aniline	–8.522	0.6392	–0.8545	–1	–1	–1

Table 1. (Continued)

No	Compound	E_{HOMO}	E_{LUMO}	Q^-	SVM Class		
					Exp	Cal	Pre
159	2–methylaniline	–8.435	0.6007	–0.9317	–1	–1	–1
160	3–methylaniline	–8.478	0.6051	–0.9380	–1	–1	–1
161	4–methylaniline	–8.356	0.6156	–0.9429	–1	–1	–1
162	2,3–dimethylaniline	–8.399	0.5917	–0.9301	–1	–1	–1
163	3,4–dimethylaniline	–8.314	0.6089	–0.9480	–1	–1	–1
164	*N*,*N*–dimethylaniline	–9.332	0.4336	–0.6200	–1	–1	–1
165	2–ethylaniline	–8.431	0.6081	–0.9294	–1	–1	–1
166	3–ethylaniline	–8.482	0.6107	–0.9510	–1	–1	–1
167	4–ethylaniline	–8.379	0.6219	–0.9589	–1	–1	–1
168	4–butylaniline	–8.376	0.6182	–0.9518	–1	–1	–1
169	2,6–diisopropylaniline	–8.338	0.6459	–0.8995	–1	–1	–1
170	2–chloroaniline	–8.376	0.3928	–0.6743	–1	–1	–1
171	3–chloroaniline	–8.458	0.3781	–0.6965	–1	–1	–1
172	4–chloroaniline	–8.577	0.2920	–0.9487	–1	–1	–1
173	2,4–dichloroaniline	–8.466	0.1239	–0.6755	–1	–1	–1
174	2,5–dichloroaniline	–8.589	0.0302	–0.6638	–1	–1	–1
175	3,4–dichloroaniline	–8.499	0.1307	–0.6796	–1	–1	–1
176	3,5–dichloroaniline	–8.687	0.0543	–0.6550	–1	–1	–1
177	2,3,4–trichloroaniline	–8.607	–0.1427	–0.6808	–1	–1	–1
178	2,3,6–trichloroaniline	–8.702	–0.2406	–0.6761	–1	–1	–1
179	2,4,5–trichloroaniline	–8.630	–0.1974	–0.6849	–1	–1	–1
180	4–bromoaniline	–8.393	0.4109	–0.6621	–1	–1	–1
181	α,α,α,4–tetrafluoro–3–methylaniline	–8.759	–0.3958	–0.6372	–1	–1	–1
182	α,α,α,4–tetrafluoro–2–methylaniline	–8.934	–0.4233	–0.8982	–1	–1	–1
183	pentafluoroaniline	–9.272	–1.0127	–0.8360	–1	–1	–1
184	3–benzyloxyaniline	–8.540	0.3454	–0.9448	–1	–1	–1
185	4–hexyloxyaniline	–8.371	0.4853	–0.9489	–1	–1	–1
186	2–nitroaniline	–9.068	–0.7937	–0.6488	–1	–1	–1
187	3–nitroaniline	–9.254	–0.9503	–0.9468	–1	–1	–1
188	4–nitroaniline	–9.160	0.7050	–0.6493	–1	–1	–1
189	2–chloro–4–nitroaniline	–9.256	–0.9066	–0.6434	–1	–1	–1
190	4–ethoxy–2–nitroaniline	–8.994	–0.8747	–0.8070	–1	–1	–1

2.2 Structure–Toxicity Models with Support Vector Machines

Support vector machines were developed by Vapnik [17–19] as an effective algorithm for determining an optimal hyperplane to separate two classes of patterns [20–30]. 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. The patterns that determine the separating hyperplane are called support vectors.

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 [31], translation initiation sites [32], genes [33], cancer type [34–37], pigmented skin lesions [38], HIV protease cleavage sites [39], GPCR type [40], protein class [41], membrane protein type [42], protein–protein interactions [43], protein subcellular localization [44–46], protein fold [47], protein secondary structure [48], specificity of GalNAc–transferase [49],

DNA hairpins [50], organisms [51], aquatic toxicity mechanism of action [16], carcinogenic activity of polycyclic aromatic hydrocarbons [52], structure–odor relationships for pyrazines [53], cancer diagnosis from the blood concentration of Zn, Ba, Mg, Ca, Cu, and Se [54].

In this study we have investigated the application of SVM for the classification of polar and nonpolar pollutants using structural descriptors. The 190 compounds presented in Table 1 were taken from the literature [14,15], and consist of 114 nonpolar data compounds (SVM class +1) and 76 polar compounds (SVM class –1). All SVM models from the present paper for the classification of polar and nonpolar pollutants were obtained with mySVM [55], which is freely available for download. Links to Web resources related to SVM, namely tutorials, papers and software, can be found in BioChem Links [56] 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–20%–out (L20%O) 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:

$$K(x, y) = x \cdot y \tag{1}$$

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

$$K(x, y) = (x \cdot y + 1)^d \tag{2}$$

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):

$$K(x, y) = \exp(-\gamma \parallel x - y \parallel^2) \tag{3}$$

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 (values 0, 1, and 2):

$$K(x, y) = \tanh(ax \cdot y + b) \tag{4}$$

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):

$$K(x, y) = \left(\sum_i \exp(-\gamma(x_i - y_i)) \right)^d \tag{5}$$

2.3 Descriptor Selection in Support Vector Machines

All studies that develop QSAR models from a large set of structural descriptors use a wide range of algorithms for selecting significant descriptors. Currently, there is no widely accepted algorithm

for selecting the best group of descriptors for an SVM model. Because an exhaustive test of all combinations of descriptors requires too large computational resources, we have used a heuristic method for descriptor selection. This heuristic algorithm starts from the set of 5 structural descriptors used by Ren [15] (namely, log K_{ow}, E_{HOMO}, E_{LUMO}, Q^-, and Q^+) and develops SVM models by applying the following steps:

(1) Starting from the complete group of N descriptors, all SVM models with one descriptor each are computed. For each descriptor or group of descriptors, 78 experiments were performed using the dot, polynomial, radial basis function, neural, and anova kernels, with various parameters (see Eqs. (1)–(5) and Table 2). The prediction performances of each SVM experiment are evaluated with the L20%O cross–validation procedure, and the accuracy index AC is computed for each experiment, namely AC = (TP + TN)/(TP + FP + TN + FN), where TP is the true positive number, FP is the false positive number, TN is the true negative number, and FN is the false negative number. The descriptor that gives the maximum prediction AC is selected for further experiments.

(2) Using the descriptor selected in step (1) and each of the remaining $N - 1$ descriptors, pairs of descriptors are tested in SVM models. The pair of descriptors with the maximum prediction AC is selected for further experiments.

(3) In each step, a new descriptor is selected, namely the one that, together with the descriptors selected in previous steps, gives the maximum prediction AC. The process stops when prediction AC does not increase by adding a new descriptor, or when a certain maximum number of descriptors are selected.

3 RESULTS AND DISCUSSION

The results of the descriptor selection algorithm show that SVM models obtained with E_{HOMO} (the energy of the highest occupied molecular orbital), E_{LUMO} (the energy of the lowest unoccupied molecular orbital), and Q^- (the most negative partial charge on any non–hydrogen atom in the molecule) give the maximum prediction $AC_p = 0.98$. Because adding a fourth descriptor does not increase the prediction AC, we will discuss only SVM models obtained with these three quantum descriptors. The SVM results obtained with E_{HOMO}, E_{LUMO}, and Q^- are presented in Table 2. The calibration of the SVM models was performed with the whole set of 190 compounds (114 nonpolar, SVM class +1; 76 polar, SVM class –1). The calibration results reported in Table 2 are: TP_c, true positive in calibration, the number of +1 patterns (nonpolar compounds) computed in class +1; FN_c, false negative in calibration, the number of +1 patterns computed in class –1; TN_c, true negative in calibration, the number of –1 patterns (polar compounds) computed in class –1; FP_c, false positive in calibration, the number of –1 patterns computed in class +1; SV_c, number of support vectors in calibration; BSV_c, number of bounded support vectors in calibration; AC_c, calibration accuracy.

Using sophisticated kernels, SVM can be calibrated to perfectly discriminate two populations of patterns, but only a cross–validation prediction test can demonstrate the potential utility of an SVM model. For each SVM model we present in Table 2 the following leave–20%–out cross-validation statistics: TP_p, true positive in prediction; FN_p, false negative in prediction; TN_p, true negative in prediction; FP_p, false positive in prediction; SV_p, average number of support vectors in prediction; BSV_p, average number of bounded support vectors in prediction; AC_p, prediction accuracy.

Table 2. Results for SVM classification of polar and nonpolar pollutants using E_{HOMO}, E_{LUMO} and Q^-. [a]

Exp	*C*	*K*			TP_c	FN_c	TN_c	FP_c	SV_c	BSV_c	AC_c	TP_p	FN_p	TN_p	FP_p	SV_p	BSV_p	AC_p
1	10	D			105	9	76	0	27	23	0.95	104	10	76	0	22.2	18.4	0.95
2	100				106	8	76	0	25	21	0.96	104	10	76	0	20.2	16.2	0.95
3	1000				106	8	76	0	25	21	0.96	108	6	76	0	19.6	15.6	0.97
			d															
4	10	P	2		109	5	75	1	21	12	0.97	108	6	75	1	18.0	9.2	0.96
5	100		2		109	5	76	0	20	10	0.97	108	6	74	2	15.2	6.0	0.96
6	1000		2		109	5	76	0	19	9	0.97	108	6	72	4	14.8	5.4	0.95
7	10		3		112	2	76	0	21	7	0.99	108	6	73	3	15.2	4.8	0.95
8	100		3		113	1	76	0	19	2	0.99	107	7	73	3	15.2	1.2	0.95
9	1000		3		114	0	76	0	18	0	1.00	106	8	73	3	14.4	0.0	0.94
10	10		4		112	2	76	0	22	5	0.99	106	8	73	3	17.0	2.4	0.94
11	100		4		114	0	76	0	20	0	1.00	106	8	72	4	15.8	0.0	0.94
12	1000		4		114	0	76	0	20	0	1.00	106	8	72	4	15.8	0.0	0.94
13	10		5		114	0	76	0	19	1	1.00	107	7	70	6	15.0	0.4	0.93
14	100		5		114	0	76	0	20	0	1.00	107	7	70	6	15.0	0.0	0.93
15	1000		5		114	0	76	0	20	0	1.00	107	7	70	6	15.0	0.0	0.93
			γ															
16	10	R	0.5		109	5	76	0	26	14	0.97	107	7	75	1	23.6	11.0	0.96
17	100		0.5		112	2	76	0	20	4	0.99	108	6	74	2	17.0	4.2	0.96
18	1000		0.5		113	1	76	0	19	2	0.99	108	6	74	2	15.8	0.6	0.96
19	10		1.0		112	2	76	0	35	7	0.99	109	5	75	1	34.0	5.4	0.97
20	100		1.0		113	1	76	0	28	2	0.99	109	5	75	1	26.4	1.4	0.97
21	1000		1.0		114	0	76	0	21	0	1.00	109	5	75	1	21.8	0.0	0.97
22	10		2.0		113	1	76	0	45	5	0.99	109	5	74	2	44.8	3.0	0.96
23	100		2.0		114	0	76	0	43	0	1.00	109	5	75	1	40.8	0.0	0.97
24	1000		2.0		114	0	76	0	43	0	1.00	109	5	75	1	40.8	0.0	0.97
			a	*b*														
25	10	N	0.5	0.0	102	12	68	8	26	24	0.89	102	12	66	10	24.2	21.4	0.88
26	100		0.5	0.0	102	12	64	12	28	25	0.87	104	10	63	13	23.4	20.0	0.88
27	1000		0.5	0.0	102	12	64	12	28	24	0.87	103	11	62	14	22.0	18.6	0.87
28	10		1.0	0.0	98	16	60	16	34	32	0.83	95	19	61	15	30.6	28.0	0.82
29	100		1.0	0.0	98	16	60	16	34	32	0.83	100	14	56	20	31.4	29.0	0.82
30	1000		1.0	0.0	98	16	60	16	34	32	0.83	95	19	60	16	29.6	27.4	0.82
31	10		2.0	0.0	85	29	48	28	60	58	0.70	80	34	55	21	45.2	43.8	0.71
32	100		2.0	0.0	87	27	48	28	58	55	0.71	80	34	55	21	45.2	43.2	0.71
33	1000		2.0	0.0	85	29	47	29	60	58	0.69	86	28	48	28	47.6	45.0	0.71
34	10		0.5	1.0	95	19	53	23	53	50	0.78	92	22	52	24	41.4	38.6	0.76
35	100		0.5	1.0	92	22	53	23	49	46	0.76	89	25	51	25	39.4	36.4	0.74
36	1000		0.5	1.0	92	22	53	23	49	45	0.76	89	25	50	26	39.2	36.4	0.73
37	10		1.0	1.0	85	29	47	29	61	58	0.69	87	27	50	26	44.6	42.8	0.72
38	100		1.0	1.0	98	16	59	17	35	33	0.83	83	31	52	24	43.8	41.2	0.71
39	1000		1.0	1.0	98	16	59	17	35	33	0.83	84	30	46	30	48.0	45.4	0.68
40	10		2.0	1.0	86	28	43	33	64	64	0.68	86	28	50	26	35.6	34.0	0.72
41	100		2.0	1.0	86	28	43	33	64	64	0.68	94	20	55	21	26.6	24.8	0.78
42	1000		2.0	1.0	86	28	43	33	64	64	0.68	97	17	46	30	34.0	32.6	0.75

Table 2. (Continued)

Exp	*C*	*K*	*a*	*b*	TP_c	FN_c	TN_c	FP_c	SV_c	BSV_c	AC_c	TP_p	FN_p	TN_p	FP_p	SV_p	BSV_p	AC_p
43	10	N	0.5	2.0	87	27	46	30	67	65	0.70	90	24	44	32	54.2	52.0	0.71
44	100		0.5	2.0	84	30	46	30	63	60	0.68	85	29	44	32	51.0	48.4	0.68
45	1000		0.5	2.0	84	30	46	30	62	60	0.68	84	30	44	32	50.2	47.8	0.67
46	10		1.0	2.0	83	31	45	31	64	62	0.67	71	43	50	26	52.0	50.4	0.64
47	100		1.0	2.0	83	31	45	31	64	62	0.67	82	32	45	31	51.6	49.4	0.67
48	1000		1.0	2.0	83	31	45	31	64	62	0.67	82	32	45	31	51.6	49.4	0.67
49	10		2.0	2.0	85	29	46	30	63	60	0.69	75	39	65	11	46.0	44.6	0.74
50	100		2.0	2.0	97	17	58	18	37	35	0.82	79	35	68	8	42.0	40.0	0.77
51	1000		2.0	2.0	97	17	58	18	37	35	0.82	82	32	65	11	38.2	35.8	0.77
			γ	*d*														
52	10	A	0.5	1	110	4	76	0	26	16	0.98	106	8	75	1	22.0	12.6	0.95
53	100		0.5	1	111	3	76	0	17	9	0.98	108	6	74	2	15.4	5.6	0.96
54	1000		0.5	1	112	2	76	0	14	4	0.99	109	5	73	3	13.2	2.8	0.96
55	10		1.0	1	111	3	76	0	26	11	0.98	109	5	75	1	20.4	8.6	0.97
56	100		1.0	1	111	3	76	0	18	5	0.98	110	4	74	2	16.0	3.6	0.97
57	1000		1.0	1	113	1	76	0	17	3	0.99	110	4	72	4	14.6	1.6	0.96
58	10		2.0	1	111	3	76	0	24	6	0.98	110	4	76	0	20.6	4.6	0.98
59	100		2.0	1	113	1	76	0	18	3	0.99	109	5	73	3	17.8	1.6	0.96
60	1000		2.0	1	114	0	76	0	14	0	1.00	109	5	70	6	15.2	0.0	0.94
61	10		0.5	2	112	2	76	0	24	7	0.99	107	7	75	1	18.4	4.8	0.96
62	100		0.5	2	112	2	76	0	20	3	0.99	108	6	74	2	16.8	1.6	0.96
63	1000		0.5	2	114	0	76	0	15	0	1.00	107	7	74	2	14.2	0.0	0.95
64	10		1.0	2	112	2	76	0	21	4	0.99	108	6	75	1	18.8	2.4	0.96
65	100		1.0	2	114	0	76	0	20	0	1.00	107	7	73	3	16.6	0.0	0.95
66	1000		1.0	2	114	0	76	0	20	0	1.00	107	7	73	3	16.6	0.0	0.95
67	10		2.0	2	114	0	76	0	24	2	1.00	108	6	73	3	24.6	1.0	0.95
68	100		2.0	2	114	0	76	0	22	0	1.00	108	6	73	3	23.0	0.0	0.95
69	1000		2.0	2	114	0	76	0	22	0	1.00	108	6	73	3	23.0	0.0	0.95
70	10		0.5	3	112	2	76	0	21	4	0.99	108	6	74	2	17.0	1.8	0.96
71	100		0.5	3	114	0	76	0	17	0	1.00	107	7	73	3	15.4	0.0	0.95
72	1000		0.5	3	114	0	76	0	17	0	1.00	107	7	73	3	15.4	0.0	0.95
73	10		1.0	3	114	0	76	0	20	0	1.00	107	7	74	2	20.4	0.0	0.95
74	100		1.0	3	114	0	76	0	20	0	1.00	107	7	74	2	20.4	0.0	0.95
75	1000		1.0	3	114	0	76	0	20	0	1.00	107	7	74	2	20.4	0.0	0.95
76	10		2.0	3	114	0	76	0	38	0	1.00	108	6	74	2	37.2	0.0	0.96
77	100		2.0	3	114	0	76	0	38	0	1.00	108	6	74	2	37.2	0.0	0.96
78	1000		2.0	3	114	0	76	0	38	0	1.00	108	6	74	2	37.2	0.0	0.96

[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 (TP_c, true positive in calibration; FN_c, false negative in calibration; TN_c, true negative in calibration; FP_c, false positive in calibration; SV_c, number of support vectors in calibration; BSV_c, number of bounded support vectors in calibration; AC_c, calibration accuracy) and L20%O prediction results (TP_p, true positive in prediction; FN_p, false negative in prediction; TN_p, true negative in prediction; FP_p, false positive in prediction; SV_p, average number of support vectors in prediction; BSV_p, average number of bounded support vectors in prediction; AC_p, prediction accuracy).

The results from Table 2 show that the classification results depend on the kernel type and parameters: dot kernel, with AC_c between 0.95 and 0.96 and AC_p between 0.95 and 0.97; polynomial kernel, with AC_c between 0.97 and 1 and with AC_p between 0.93 and 0.96; radial basis function kernel, with AC_c between 0.97 and 1 and with AC_p between 0.96 and 0.97; neural kernel, with AC_c between 0.68 and 0.89 and with AC_p between 0.64 and 0.88; anova kernel, with AC_c between 0.98 and 1 and with AC_p between 0.94 and 0.98. The overfitting of SVM models is clearly detected in several cases. For example, as the degree of the polynomial kernel increases from 2 to 5,

AC_c increases from 0.97 to 1, while AC_p decreases from 0.96 to 0.93. These results show that SVM models are capable of overfitting, and the only sound method to identify the optimum model is by comparing prediction statistics.

The maximum prediction $AC_p = 0.98$ was obtained in experiment 58, with the anova kernel, $SV_c = 24$, $SV_p = 20.6$, and $AC_c = 0.98$ (see Table 2). The SVM model from experiment 58 has three classification errors in calibration, all nonpolar compounds (class +1) situated on the polar (class –1) region of the SVM hyperplane: **21**, 3–furanmethanol; **32**, 2–phenoxyethanol; **60**, 2–hydroxy–4–methoxyacetophenone. The leave-20%-out cross-validation has four errors in prediction, all nonpolar compounds predicted to be polar, *i.e.*, the three compounds from calibration (**21**, **32**, and **60**) and **23**, 2,2,2–trichloroethanol. These results show that SVM models obtained with E_{HOMO}, E_{LUMO}, and Q^- are capable of discriminating between polar and nonpolar pollutants. Good prediction results are obtained also with a group of SVM models that have $AC_p = 0.97$, namely experiments 3 (polynomial kernel); 19–24 (radial kernel); 55 and 56 (anova kernel). The results from Table 2 show that several experiments have $AC_c = 1$: experiments 9 and 11–15 (polynomial kernel); 21, 23, and 24 (radial kernel); 60, 63, 65–69, and 71–78 (anova kernel). However, the corresponding prediction values for AC_p are between 0.93 and 0.96 for the polynomial and anova kernels, and only the experiments with the radial kernel, having $AC_p = 0.97$, can be regarded as interesting alternatives to experiment 58.

4 CONCLUSIONS

Narcotic pollutants, that act by nonspecifically disrupting the functioning of cell membranes, are categorized as polar and nonpolar compounds. The toxicity prediction of narcotic pollutants with QSAR (quantitative structure-activity relationships) depends on the reliable determination of the mechanism of toxic action. The classification of the chemical compounds as polar and nonpolar narcotic pollutants based on structural characteristics is of utmost importance in predicting the aquatic toxicity for new chemicals. 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 190 narcotic pollutants (76 polar and 114 nonpolar) using literature data [14,15]. Using an efficient descriptor selection algorithm, the energy of the highest occupied molecular orbital E_{HOMO}, the energy of the lowest unoccupied molecular orbital E_{LUMO}, and the most negative partial charge on any non–hydrogen atom in the molecule Q^-, all computed with the AM1 method, were found to be necessary for the discrimination of the polar and nonpolar compounds.

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–20%–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. The best prediction results were obtained with the anova kernel, followed by the radial basis function kernel.

This study demonstrates that SVM models can be used with success to discriminate between polar and nonpolar pollutants, providing reliable predictions. The heuristic algorithm proposed here for the efficient selection of structural descriptors for SVM models was able of significantly reducing the dimensionality of the input space. Further studies regarding the use of SVM in structure–activity relationships should compare this heuristic algorithm with other descriptor selection methods, such as the genetic algorithm. 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 experiment 58 is available as supplementary material.

5 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] E. Urrestarazu Ramos, W. H. J. Vaes, H. J. M. Verhaar, and J. L. M. Hermens, Quantitative Structure–Activity Relationships for the Aquatic Toxicity of Polar and Nonpolar Narcotic Pollutants, *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 845–852.
[15] S. Ren, Classifying Class I and Class II Compounds by Hydrophobicity and Hydrogen Bonding Descriptors, *Environ. Toxicol*. **2002**, *17*, 415–423.
[16] 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.
[17] V. Vapnik, *Estimation of Dependencies Based on Empirical Data*, Nauka, Moscow, 1979.
[18] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, 1995.
[19] V. Vapnik, *Statistical Learning Theory*, Wiley–Interscience, New York, 1998.
[20] C. J. C. Burges, A Tutorial on Support Vector Machines for Pattern Recognition, *Data Mining Knowledge Discov.* **1998**, *2*, 121–167.
[21] B. Schölkopf, K. –K. Sung, C. J. C. Burges, F. Girosi, P. Niyogi, T. Poggio, and V. Vapnik, Comparing Support Vector Machines with Gaussian Kernels to Radial Basis Function Classifiers, *IEEE Trans. Signal Process.* **1997**, *45*, 2758–2765.
[22] V. N. Vapnik, An Overview of Statistical Learning Theory, *IEEE Trans. Neural Networks* **1999**, *10*, 988–999.
[23] B. Schölkopf, C. J. C. Burges, and A. J. Smola, *Advances in Kernel Methods*: *Support Vector Learning*, MIT Press, Cambridge, MA, 1999.
[24] N. Cristianini and J. Shawe–Taylor, *An Introduction to Support Vector Machines*, Cambridge University Press, Cambridge, 2000.
[25] 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.
[26] C.–C. Chang and C.–J. Lin, Training ν–Support Vector Classifiers: Theory and Algorithms, *Neural Comput*. **2001**, *12*, 2119–2147.
[27] 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.
[28] 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.
[29] 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.
[30] O. L. Mangasarian and D. R. Musicant, Lagrangian Support Vector Machines, *J. Machine Learning Res.* **2001**, *1*, 161–177, http://www.jmlr.org.
[31] 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.
[32] 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.
[33] 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.
[34] 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.
[35] 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.
[36] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, Gene Selection for Cancer Classification Using Support Vector Machines, *Machine Learning* **2002**, *46*, 389–422.
[37] 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.
[38] 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.
[39] 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.
[40] R. Karchin, K. Karplus, and D. Haussler, Classifying G–Protein Coupled Receptors with Support Vector Machines, *Bioinformatics* **2002**, *18*, 147–159.
[41] 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.
[42] 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.
[43] J. R. Bock and D. A. Gough, Predicting Protein–Protein Interactions from Primary Structure, *Bioinformatics* **2001**, *17*, 455–460.
[44] S. J. Hua and Z. R. Sun, Support Vector Machine Approach for Protein Subcellular Localization Prediction, *Bioinformatics* **2001**, *17*, 721–728.
[45] 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.
[46] 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.
[47] C. H. Q. Ding and I. Dubchak, Multi–Class Protein Fold Recognition Using Support Vector Machines and Neural Networks, *Bioinformatics* **2001**, *17*, 349–358.
[48] 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.
[49] 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.
[50] 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.
[51] 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.
[52] 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.
[53] O. Ivanciuc, Structure–Odor Relationships for Pyrazines with Support Vector Machines, *Internet Electron. J. Mol. Des.* **2002**, *1*, 269–284, http://www.biochempress.com.
[54] O. Ivanciuc, Support Vector Machines for Cancer Diagnosis from the Blood Concentration of Zn, Ba, Mg, Ca, Cu, and Se, *Internet Electron. J. Mol. Des*. **2002**, *1*, 418–427, http://www.biochempress.com.
[55] S. Rüping, mySVM, University of Dortmund, http://www–ai.cs.uni–dortmund.de/SOFTWARE/MYSVM/.
[56] BioChem Links, http://www.biochempress.com.