CODEN IEJMAT *Internet Electronic Journal of Molecular Design* **2005**, *4*, 835–849 ISSN 1538–6414
BioChem Press http://www.biochempress.com

Inter*net* Electronic Journal of
Molecular Design

December 2005, Volume 4, Number 12, Pages 835–849

Editor: Ovidiu Ivanciuc

Special issue dedicated to Professor Danail Bonchev on the occasion of the 65th birthday

Support Vector Machines for Prediction of Mechanism of Toxic Action from Multivariate Classification of Phenols Based on MEDV Descriptors

Zhong–Sheng Yi [1] and Shu–Shen Liu [1,2]

[1] Department of Material and Chemistry Engineering, Guilin University of Technology, Guilin 541004, P. R. China
[2] State Key Laboratory of Pollution Control and Resources Reuse, School of Environment, Nanjing University, Nanjing 210093, P. R. China

Received: May 8, 2005; Revised: November 13, 2005; Accepted: December 1, 2005; Published: December 31, 2005

Citation of the article:
Z.–S. Yi and S.–S. Liu, Support Vector Machines for Prediction of Mechanism of Toxic Action from Multivariate Classification of Phenols Based on MEDV Descriptors, *Internet Electron. J. Mol. Des.* **2005**, *4*, 835–849, http://www.biochempress.com.

Inter*net* Electronic Journal of Molecular Design

BioChem Press
http://www.biochempress.com

Support Vector Machines for Prediction of Mechanism of Toxic Action from Multivariate Classification of Phenols Based on MEDV Descriptors

Zhong–Sheng Yi [1,*] and Shu–Shen Liu [1,2]

[1] Department of Material and Chemistry Engineering, Guilin University of Technology, Guilin 541004, P. R. China

[2] State Key Laboratory of Pollution Control and Resources Reuse, School of Environment, Nanjing University, Nanjing 210093, P. R. China

Received: May 8, 2005; Revised: November 13, 2005; Accepted: December 1, 2005; Published: December 31, 2005

Internet Electron. J. Mol. Des. **2005**, *4* (12), 835–849

Abstract

Motivation. Phenols are widely used in agriculture as biocides and disinfectants and in various industries. Most synthetic phenolic compounds are toxic and are classified as hazardous pollutants. Their mechanism of toxic action (MOA) classes are usually predicted by quantitative structure–activity relationships (QSAR) models. In this study, we report the support vector machine (SVM) model for identifying four MOA of phenols.

Method. The structures of 221 phenols were described by the molecular electronegativity distance vector (MEDV). The SVM algorithm with one–against–one multi–class classification method was used to construct the QSAR models for four MOA classes (polar narcotics, weak acid respiratory uncouplers, precursors to soft electrophiles, and soft electrophiles). The predictive power of each model was estimated by leave–one–out (LOO) cross validation method.

Results. In order to find MOA classifiers with high predictive power, we have investigated 345 SVM models generated from two SVM methods and two kernels including linear and radial basis function (RBF). The key factors affecting the quality of SVM models are kernel type, its corresponding parameters that control the kernel shape, and the capacity parameter C. We used a RBF kernel with $\gamma = 0.0004$ and a capacity parameter $C = 128$, which has the highest accuracy index for leave–one–out cross–validation. The accuracy index for all 221 compounds (with 13 compounds misclassified) is 94.1%. To test the stability of this SVM model, we have uniformly chosen 155 from all 221 compounds for training, and the remaining compounds were included in the test set. The training set was used to construct a new SVM model with the parameters of $\gamma = 0.0004$ and $C = 128$. It has been shown that 16 compounds (8 in the training set and 8 in testing set) were misclassified, which gives an accuracy index of 92.8%. These results show that the SVM model has a high quality for predicting the aquatic toxicity mechanism for new chemical compounds, when appropriate SVM parameters and molecular descriptors are used.

Conclusions. The SVM method based on MEDV descriptors allows satisfactory classification of phenols with respect to four MOA which are based on experimental toxicity to the ciliate *Tetrahymena pyriformis*. This approach can be used to predict the aquatic toxicity mechanism and to select the appropriate QSAR model based on MEDV descriptors for new phenolic compounds.

Keywords. Support vector machines; structure–toxicity relationships; quantitative structure–activity

Dedicated on the occasion of the 65th birthday to Danail Bonchev.

* Correspondence author; E–mail: yzs@glite.edu.cn or samtyty@hotmail.com.

relationships; QSAR; aquatic toxicity; mechanism of action; ciliate *Tetrahymena pyriformis*.

Abbreviations and notations

MOA, mechanism of toxic action	ν–SVMC, ν support vector classification
SVM, support vector machines	QSAR, quantitative structure–activity relationships
MEDV, molecular electronegativity distance vector	LDA, stepwise linear discriminant analysis
C–SVMC, *C* support vector classification	LOO, leave–one–out

1 INTRODUCTION

Today, more and more chemical compounds are used widely in our society because of rapid increase of industry and economy. Most of these chemical compounds can be environmental pollutants. Because of this, during the past decade a great deal of effort has been put into the study of the relationships between a compound's structure and its toxicity. Significant progress has been made to classify chemical compounds according to their mechanism of toxicity and to screen them for their environmental risk assessment [1–5]. The toxicological data are necessary to assess the impact of such compounds on the environment. Because of the time and financial resources required, toxicological data are not available for all chemical compounds. Quantitative structure–activity relationships (QSAR) are used as scientifically credible tools to predict the acute toxicity of chemicals when few empirical data are available. QSAR has been widely used in modeling and predicting toxicities of organic compounds [2].

Phenols are versatile and important industrial organic chemicals. They are widely used in agriculture as biocides and disinfectants and in various industries such as coal conversion, metal casting, paper manufacturing, and resin production [6]. Most synthetic phenolic compounds are toxic and are classified as hazardous pollutants. Phenols, especially chlorosubstitued phenols, have been of interest to environmental toxicologists, and their toxic potencies have been assessed in several test systems.

The toxicity of phenols involves a numbers of different mechanisms and modes of action. Phenols exhibit toxicity *via* several mechanisms. Most substituted phenols act by the polar narcosis mechanism. Some phenols act by other mechanisms which include respiratory uncoupling, pro–electropilic and soft electrophilic reactivity [7]. A number of QSAR investigations have been performed to predict the mechanism of toxic action (MOA) of phenols [8–11]. Existing methods for classifying compounds according to MOAs can be grouped into two types of approaches. One is a qualitative approach based on simple structural characteristics. The other is based on statistical analysis of physico–chemical properties [5]. The first approach is simple and relatively successful for phenols with only few substituents, but when substituents associated with different MOAs are present in a molecule, it is limited because of the limitation to the type of the substituents in training set. The second classification (based on physico–chemical properties) has some disadvantages, which include the availability and use of the descriptors, the difficulty of mechanistic interpretation with some types of descriptors, and the fact that the property profile of the initial compounds may

differ significantly from the metabolically activated toxicant. Recently, several classification models based on various statistics methods have been derived for a set of phenols using quantum mechanical based descriptors, additional whole molecule descriptors and empirical physico–chemical descriptors [5,7,12,13].

Support vector machines (SVM) represent a new class of machine learning algorithms developed by Vapnik [14,15]. SVM found numerous applications in various classification and regression models such as MOA prediction [12,16–18], classification of microarray gene expression data [19], estimation of aqueous solubility [20], classification of organophosphate nerve agent simulants [21]. In this study we have investigated the application of SVM, based on the MEDV [22–28], for the recognition of the aquatic toxicity mechanism for the compounds previously explored by Aptula [5] and Ren [7].

2 MATERIALS AND METHODS

2.1 Chemical Data

In this study we have investigated the application of SVM for 221 phenols from four MOA classes [5], which included 153 polar narcosis (marked as label 1), 18 weak acid respiratory uncouplers(marked as label 2), 27 precursors to soft electrophiles (marked as label 3) and 23 soft electrophiles (marked as label 4). The classification was based on 37 MEDV descriptors calculated from the method of the MEDV descriptors [22–28]. The MOA of phenols are presented in table 1. For testing the stability of model, we uniformly chose 2/3 of all 221 compounds (155 compounds) as training set, the rest 66 compounds as testing set.

2.2 Multi–class classification for Support Vector Machines

A detailed description of the theory of SVM can be seen in several excellent books and tutorials [15,29,30]. In SVM, the input space is transformed into a higher dimensional feature space by using different kernels that perform a nonlinear mapping, and then, find out the maximal margin hyperplane between two classes in that higher dimensional space. Furthermore, SVM solved the classification problem by support vector that determined the separating hyperplane.

Although SVM were originally designed for binary classification, it can also be extend to solve multi–class classification. How to effectively extend SVM for multi–class classification is still an on–going research issue. Currently there are two types of approaches for multi–class SVM. One of the frequently used methods is to decompose the multi–class problem into a set of binary classification problems and then combine these binary classifiers. There are two approaches that can be used for this purpose [31], *i.e.*, “one–against–all” and “one–against–one”.

Table 1. Structure of phenols and mechanisms of toxic action (1, polar narcosis; 2, weak acid respiratory uncouplers; 3, precursors to soft electrophiles; 4, soft electrophiles).

No	Name	MOA_{EXP}	MOA_{All}	MOA_{TRN}
	Training set			
1	1,3,5–trihydroxybenzene	***1***	***3***	***1***
3	2,3,5–trichlorophenol	1	1	1
4	2,3,5–trimethylphenol	1	1	1
5	2,3,6–trimethylphenol	1	1	1
7	2,3–dimethylphenol	1	1	1
8	2,4,5–trichlorophenol	1	1	1
10	2,4,6–tribromoresorcinol	1	1	1
11	2,4,6–trichlorophenol	1	1	1
13	2,4,6–tris (dimethylaminomethyl) phenol	1	1	1
14	2,4–dibromophenol	1	1	1
15	2,4–dichlorophenol	1	1	1
17	2,4–dimethylphenol	1	1	1
18	2,5–dichlorophenol	1	1	1
20	2,6–di–*tert*–butyl–4–methylphenol	1	1	1
21	2,6–dichloro–4–fluorophenol	1	1	1
23	2,6–difluorophenol	1	1	1
24	2,6–dimethoxyphenol	1	1	1
25	2–allylphenol	1	1	1
27	2–bromophenol	1	1	1
28	2–chloro–4,5–dimethylphenol	1	1	1
30	2–chlorophenol	1	1	1
31	2–cyanophenol	1	1	1
33	2–ethylphenol	1	1	1
34	2–fluorophenol	1	1	1
35	2–hydroxy–4,5–dimethylacetophenone	1	1	1
37	2–hydroxy–4–methoxybenzophenone	1	1	1
38	2–hydroxy–5–methylacetophenone	1	1	1
40	2–hydroxybenzylalcohol	1	1	1
41	2–hydroxyethylsalicylate	1	1	1
43	2–methoxy–4–propenylphenol	1	1	1
44	2–methoxyphenol	1	1	1
45	2–phenylphenol	1	1	1
47	3,4,5–trimethylphenol	1	1	1
48	3,4–dichlorophenol	1	1	1
50	3,5–dibromosalicylaldehyde	1	1	1
51	3,5–dichlorophenol	1	1	1
53	3,5–diiodosalicylaldehyde	1	1	1
54	3,5–dimethoxyphenol	1	1	1
55	3,5–dimethylphenol	1	1	1
57	3–acetamidophenol	1	1	1
58	3–bromophenol	1	1	1
60	3–chloro–5–methoxyphenol	1	1	1
61	3–chlorophenol	1	1	1
63	3–ethoxy–4–hydroxybenzaldehyde	1	1	1
64	3–ethoxy–4–methoxyphenol	1	1	1
65	3–ethylphenol	1	1	1
67	3–hydroxy–4–methoxybenzylalcohol	1	1	1
68	3–hydroxyacetophenone	1	1	1
70	3–hydroxybenzoic acid	1	1	1
71	3–hydroxybenzyl alcohol	1	1	1
73	3–isopropylphenol	1	1	1
74	3–methoxyphenol	1	1	1
75	3–phenylphenol	1	1	1

Table 1. (Continued)

No	Name	MOA_{EXP}	MOA_{All}	MOA_{TRN}
77	4–*tert*–octylphenol	1	1	1
78	4–*tert*–butylphenol	1	1	1
80	4–allyl–2–methoxyphenol	1	1	1
81	4–benzyloxyphenol	1	1	1
82	4–bromo–2,6–dichlorophenol	1	1	1
84	4–bromo–3,5–dimethylphenol	1	1	1
85	4–bromo–6–chloro–2–cresol	1	1	1
87	4–butoxyphenol	1	1	1
88	4–chloro–2–isopropyl–5–methylphenol	1	1	1
90	4–chloro–3,5–dimethylphenol	1	1	1
91	4–chloro–3–ethylphenol	1	1	1
92	4–chloro–3–methylphenol	1	1	1
94	4–chlororesorcinol	***1***	***3***	***1***
95	4–cyanophenol	1	1	1
97	4–ethylphenol	1	1	1
98	4–fluorophenol	1	1	1
100	4–hexyloxyphenol	1	1	1
101	4–hexylresorcinol	1	1	1
102	4–hydroxy–2–methylacetophenone	1	1	1
104	4–hydroxy–3–methoxybenzonitrile	1	1	1
105	4–hydroxy–3–methoxybenzylalcohol	1	1	1
107	4–hydroxy–3–methoxyphenethylalcohol	1	1	1
108	4–hydroxyacetophenone	1	1	1
110	4–hydroxybenzamide	1	1	1
111	4–hydroxybenzoic acid	1	1	1
112	4–hydroxybenzophenone	1	1	1
114	4–hydroxyphenethylalcohol	1	1	1
115	4–hydroxyphenylacetic acid	1	1	1
117	4–iodophenol	1	1	1
118	4–isopropylphenol	1	1	1
120	4–phenylphenol	1	1	1
121	4–propylphenol	1	1	1
122	4–*sec*–butylphenol	1	1	1
124	5–bromo–2–hydroxybenzylalcohol	1	1	1
125	5–bromovanillin	1	1	1
127	5–methylresorcinol	***1***	***3***	***1***
128	5–pentylresorcinol	1	1	1
130	α,α,α–trifluoro–4–cresol	1	1	1
131	ethyl–3–hydroxybenzoate	1	1	1
132	ethyl–4–hydroxy–3–methoxyphenylacetate	1	1	1
134	isovanillin	1	1	1
135	3–cresol	1	1	1
137	methyl–4–hydroxybenzoate	1	1	1
138	methyl–4–methoxysalicylate	1	1	1
140	2–cresol	1	1	1
141	2–vanillin	1	1	1
142	4–cresol	1	1	1
144	phenol	1	1	1
145	resorcinol	***1***	***3***	***1***
147	salicylaldoxime	1	1	1
148	salicylamide	1	1	1
149	salicylhydrazide	1	1	1
151	salicylic acid	1	1	1
152	syringaldehyde	1	1	1
154	2,3,4,5–tetrachlorophenol	2	2	2
155	2,3,5,6–tetrachlorophenol	2	2	2

Table 1. (Continued)

No	Name	MOA_{EXP}	MOA_{All}	MOA_{TRN}
157	2,3–dinitrophenol	2	2	2
158	2,4,6–trinitrophenol	2	2	2
159	2,4–dichloro–6–nitrophenol	2	2	2
161	2,5–dinitrophenol	2	2	2
162	2,6–dichloro–4–nitrophenol	***2***	***4***	***4***
164	2,6–dinitro–4–cresol	2	2	2
165	2,6–dinitrophenol	2	2	2
167	3,4–dinitrophenol	2	2	2
168	4,6–dinitro–2–cresol	2	2	2
169	pentabromophenol	2	2	2
171	pentafluorophenol	2	2	2
172	1,2,3–trihydroxybenzene	3	3	3
174	2,3–dimethylhydroquinone	3	3	3
175	2,4–diaminophenol	3	3	3
177	2–aminophenol	3	3	3
178	3,5–di–*tert*–butylcatechol	3	3	3
179	3–aminophenol	3	3	3
181	4–acetamidophenol	***3***	***1***	***1***
182	4–amino–2,3–dimethylphenol	3	3	3
184	4–aminophenol	3	3	3
185	4–chlorocatechol	***3***	***1***	***1***
187	5–amino–2–methoxyphenol	3	3	3
188	5–chloro–2–hydroxyaniline	3	3	3
189	6–amino–2,4–dimethylphenol	3	3	3
191	catechol	***3***	***1***	***1***
192	chlorohydroquinone	***3***	***1***	***1***
194	methoxyhydroquinone	***3***	***3***	***1***
195	methylhydroquinone	***3***	***3***	***1***
197	tetrachlorocatechol	3	3	3
198	trimethylhydroquinone	3	3	3
199	2,6–dibromo–4–nitrophenol	***4***	***2***	***4***
201	2–amino–4–nitrophenol	4	4	4
202	2–chloro–4–nitrophenol	4	4	4
204	2–nitrophenol	4	4	4
205	2–nitroresorcinol	4	4	4
207	3–hydroxy–4–nitrobenzaldehyde	4	4	4
208	3–methyl–4–nitrophenol	4	4	4
209	3–nitrophenol	4	4	4
211	4–chloro–2–nitrophenol	4	4	4
212	4–chloro–6–nitro–3–cresol	4	4	4
214	4–methyl–2–nitrophenol	4	4	4
215	4–methyl–3–nitrophenol	4	4	4
217	4–nitrocatechol	4	4	4
218	4–nitrophenol	4	4	4
219	4–nitrosophenol	4	4	4
221	5–hydroxy–2–nitrobenzaldehyde	***4***	***1***	***1***
	Testing set			
2	2–*tert*–butyl–4–methylphenol	1	1	1
6	2,3–dichlorophenol	1	1	1
9	2,4,6–tribromophenol	1	1	1
12	2,4,6–trimethylphenol	1	1	1
16	2,4–difluorophenol	1	1	1
19	2,5–dimethylphenol	1	1	1
22	2,6–dichlorophenol	1	1	1
26	2–bromo–4–methylphenol	1	1	1
29	2–chloro–5–methylphenol	1	1	1

Table 1. (Continued)

No	Name	MOA_{EXP}	MOA_{All}	MOA_{TRN}
32	2–ethoxyphenol	1	1	1
36	2–hydroxy–4–methoxyacetophenone	1	1	1
39	2–hydroxyacetophenone	1	1	1
42	2–isopropylphenol	1	1	1
46	2–*tert*–butylphenol	1	1	1
49	3,4–dimethylphenol	1	1	1
52	3,5–dichlorosalicylaldehyde	1	1	1
56	3,5–di–*tert*–butylphenol	1	1	1
59	3–chloro–4–fluorophenol	1	1	1
62	3–cyanophenol	1	1	1
66	3–fluorophenol	1	1	1
69	3–hydroxybenzaldehyde	1	1	1
72	3–iodophenol	1	1	1
76	3–*tert*–butylphenol	1	1	1
79	4,6–dichlororesorcinol	1	1	1
83	4–bromo–2,6–dimethylphenol	1	1	1
86	4–bromophenol	1	1	1
89	4–chloro–2–methylphenol	1	1	1
93	4–chlorophenol	1	1	1
96	4–ethoxyphenol	1	1	1
99	4–heptyloxyphenol	1	1	1
103	4–hydroxy–3–methoxyacetophenone	1	1	1
106	4–hydroxy–3–methoxybenzylamine	1	1	1
109	4–hydroxybenzaldehyde	1	1	1
113	4–hydroxybenzylcyanide	1	1	1
116	4–hydroxypropiophenone	1	1	1
119	4–methoxyphenol	1	1	1
123	4–*tert*–pentylphenol	1	1	1
126	5–fluoro–2–hydroxyacetophenone	1	1	1
129	6–*tert*–butyl–2,4–dimethylphenol	1	1	1
133	ethyl–4–hydroxybenzoate	1	1	1
136	methyl–3–hydroxybenzoate	1	1	1
139	nonylphenol	1	1	1
143	4–cyclopentylphenol	1	1	1
146	salicylaldehyde	1	1	1
150	salicylhydroxamic acid	1	1	1
153	vanillin	1	1	1
156	2,3,5,6–tetrafluorophenol	***2***	***2***	***1***
160	2,4–dinitrophenol	2	2	2
163	2,6–diiodo–4–nitrophenol	***2***	***2***	***4***
166	3,4,5,6–tetrabromo–2–cresol	2	2	2
170	pentachlorophenol	2	2	2
173	1,2,4–trihydroxybenzene	***3***	***3***	***1***
176	2–amino–4–*tert*–butylphenol	3	3	3
180	3–methylcatechol	3	3	3
183	4–amino–2–cresol	3	3	3
186	4–methylcatechol	***3***	***3***	***1***
190	bromohydroquinone	***3***	***1***	***1***
193	hydroquinone	***3***	***3***	***1***
196	phenylhydroquinone	***3***	***1***	***1***
200	2–amino–4–chloro–5–nitrophenol	4	4	4
203	2–chloromethyl–4–nitrophenol	4	4	4
206	3–fluoro–4–nitrophenol	4	4	4
210	4–amino–2–nitrophenol	4	4	4
213	4–hydroxy–3–nitrobenzaldehyde	4	4	4
216	4–nitro–3–(trifluoromethyl)–phenol	***4***	***4***	***2***
220	5–fluoro–2–nitrophenol	4	4	4

In the "one–against–all" approach, *k* SVM models are constructed where *k* is the number of classes and the *i*–th SVM is trained with all of the examples in the *i*–th class with positive labels, and all the other examples with negative labels. In the case of "one–against–one" approach, $k(k-1)/2$ classifiers are constructed, with each classifier trained to discriminate between a class pair *i*–th and *j*–th [30]. Several published works have shown that the "one–against–one" is more suitable for practical use than the "one–against–all" [31–33]. In our present study, the "one–against–one" approach implemented in LIBSVM [30] is used for the multi–class classification problem.

All SVM models from our study for the classification of 4 classes were obtained with LIBSVM [30], which can be downloaded freely. This tool provided the SVM with two classification methods, *C*–SVMC and ν–SVMC, and four kernels, linear, polynomial, radial basic function (RBF), and sigmoid kernel. However, in our experiment only the linear and RBF kernel were tested because the SVM based on the polynomial and sigmoid kernel required a too long time for training. The two kernels and their parameter used in this study are $K(x,x_i)=(x^T x_i)$, $K(x,x_i)=e^{-\gamma\|x-x_i\|^2}$ with γ = 0, 0.00025×2^N (N = 0, 1, 2… 9) respectively. The capacity parameter *C* of *C*–SVMC or ν of ν–SVMC took the value $C = 1\times 2^N$ (N = 0, 1, 2…12), $\nu = 0.001\times 2^N$ (N = 0, 1, 2…9). The predictive ability of each SVM model was tested against LOO cross–validation method.

2.3 Calculation of Molecular Descriptors

From the literature [24], the original MEDV descriptor x_v $(v = 1, 2, 3, \cdots, 91)$ can be calculated. First, the relative electronegativity (*e*) of a non–hydrogen atom is calculated using the atomic type, atomic attributes, and intrinsic state (*I*) of the atom defined in Table 2:

$$e_i = I_i + \sum_{j\neq i}^{all\ j} (I_i - I_j)/d_{ij}^2 \tag{1}$$

where d_{ij} is the shortest graph distance between two atoms, atom *i* and *j*. Then, the MEDV descriptor x_v is calculated from the following formula:

$$x_v = x_{kl} = \sum_{i\in k,\ j\in l} \frac{e_i e_j}{d_{ij}^2} \qquad (k,l = 1, 2, \cdots, 13;\ l \geq k,\ v = 1, 2, \cdots, 91) \tag{2}$$

where *k* or *l* is the atomic type of the atom *i* or *j* in the molecule.

From the MEDV descriptors, only 48 MEDV have one or more nonzero elements where 4 descriptors (x_{37}, x_{38}, x_{47} and x_{48}) contain 1 nonzero element, 3 descriptors (x_6, x_{46} and x_{55}) contain 2 nonzero elements, and 4 descriptors (x_{37}, x_{49}, x_{52} and x_{85}) have 3 nonzero elements. The 11 descriptors with too few nonzero elements should be eliminated from the 48 descriptors with nonzero elements. So, there are in fact 37 nonzero MEDV descriptors to be employed in our study.

Table 2. The atomic types, atomic attributes and intrinsic state (*I*) for various non–hydrogen atoms

atom	type	attribute	*I*	atom	Type	attribute	*I*	Atom	type	Attribute	*I*
$-CH_3$	1	1	2.0000	~C≈	3	16	1.8333	≥N=	7	30	2.2361
$-CH_2-$	2	2	1.5000	–OH	9	17	2.4495	–SH	9	31	1.7691
–CH<	3	3	1.3333	–O–	10	18	1.8371	–S–	10	32	1.1567
>C<	4	4	1.2500	=O	9	19	3.6742	=S	9	33	2.3134
$=CH_2$	1	5	3.0000	~O	9	20	3.0619	>S=	11	34	1.1340
$=CH_2-$	2	6	2.0000	$-NH_2$	5	21	2.2361	≥S≤	12	35	1.1227
=C<	3	7	1.6667	–NH–	6	22	1.6771	–F	13	36	2.6458
=C=	2	8	2.5000	>N–	7	23	1.0882	–Cl	13	37	1.9108
≡CH	1	9	4.0000	=NH	5	24	3.3541	–Br	13	38	1.6536
≡C–	2	10	2.5000	=N–	6	25	2.2361	–I	13	39	1.5345
$\sim CH_2$	1	11	2.5000	≡N	5	26	4.4721	$-PH_3$	5	40	1.6149
~CH–	2	12	1.7500	~NH	5	27	2.7951	–PH–	6	41	1.0559
~CH<	3	13	1.5000	~N–	6	28	1.9566	>P–	7	42	0.8696
~CH~	2	14	2.0000	~N~	6	29	2.2361	≥P<	8	43	0.9006
–C≈	3	15	1.6667								

[a] The symbols "~" and "≈" represent one and two conjugated double bonds, respectively

3 RESULTS AND DISCUSSION

In structure–activity studies, the performance of SVM depends on the combination of several parameters. Those parameters included SVM methods, kernel type and the various parameters that control kernel shape. Up to now most SVM software cannot work effectively for model selection because there is no general criterion to select SVM approach, kernel type and parameters of kernel. In this study, using all 211 compounds as training set, we tested two SVM methods with grid–search methods for classification, *C*–SVMC and ν–SVMC with various parameters mentioned above, for a total 345 SVM model for each MOA experiment.

Table 3. The highest accuracy index classification rate of two SVM methods

SVM methods	Kernel	The highest rate (%)
C–SVMC	Linear	86.36
	RBF	89.54
ν–SVMC	Linear	88.18
	RBF	89.09

The calculated result shows that when employing *C*–SVMC with RBF kernel, the highest accuracy index (the percent of the number of compounds correctly classified to the number of overall compounds in certain class) of LOO is 89.55% (see Table 3), namely 23 compounds are misclassified, which indicates that the classification power of *C*–SVMC is higher than that of ν–SVMC. However, three *C*–SVMC models with the parameters of $C = 128$ and $\gamma = 0.0004$, $C = 4$ and $\gamma = 0.0064$, and $C = 2$ and $\gamma = 0.0064$, respectively, have the same rate values of 89.55%. The accuracy index for LOO of the SVM models with other parameters are shown in Figure 1. From Figure 1, it is obvious that when $0.0002<\gamma<0.128$, the rates are generally the highest and the rates remain almost unchanged with the parameter *C*.

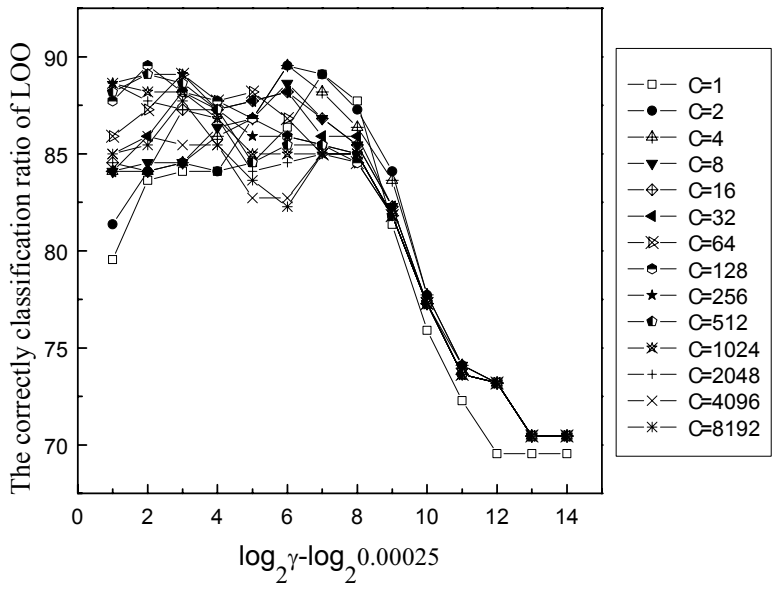

Figure 1. The relation between the accuracy index of LOO and parameter γ with different parameter C.

As mentioned above, there are three models with the highest LOO accuracy index. How to choose the best model which has the highest predicted power is a key problem. According to the theory of SVM, the number of support vector is very important. The accuracy index always increases with the number of support vector, but the generalization ability of SVM model will decrease. In other words, the smaller the number of support vector is, the larger the generalization power of SVM model is [15]. The number of support vector of three models which have the same rate values of 89.55%, is 84, 118 and 121, respectively. So we chose the SVM model with small number of support vector. Namely, the best SVM model is *C*–SVMC with a BRF kernel and the corresponding parameters of $C = 128$ and $\gamma = 0.0004$.

After choosing SVM parameters, we generated the following 12 discriminant functions to classify the 4 MOA classes according to the one–against–one procedure.

Discriminant for the 1st class of MOA:

$$1\ vs\ 2\quad f(x) = sign\left(\sum_{i=1}^{43}\alpha_i y_i e^{-0.0004\|x-x_i\|^2} - 0.9915\right) \tag{3}$$

$$1\ vs\ 3\quad f(x) = sign\left(\sum_{i=1}^{43}\alpha_i y_i e^{-0.0004\|x-x_i\|^2} + 6.3487\right) \tag{4}$$

$$1\ vs\ 4\quad f(x) = sign\left(\sum_{i=1}^{43}\alpha_i y_i e^{-0.0004\|x-x_i\|^2} + 0.1347\right) \tag{5}$$

Discriminant for the 2nd class of MOA:

$$1\ vs\ 2\quad f(x) = sign\left(\sum_{i=1}^{8}\alpha_i y_i e^{-0.0004\|x-x_i\|^2} - 0.9915\right) \tag{6}$$

$$2\ vs\ 3\ \ f(x) = sign\left(\sum_{i=1}^{8} \alpha_i y_i e^{-0.0004\|x-x_i\|^2} + 0.5923\right) \tag{7}$$

$$2\ vs\ 4\ \ f(x) = sign\left(\sum_{i=1}^{8} \alpha_i y_i e^{-0.0004\|x-x_i\|^2} + 3.6532\right) \tag{8}$$

Discriminant for the 3rd class of MOA

$$1\ vs\ 3\ \ f(x) = sign\left(\sum_{i=1}^{24} \alpha_i y_i e^{-0.0004\|x-x_i\|^2} + 6.3487\right) \tag{9}$$

$$2\ vs\ 3\ \ f(x) = sign\left(\sum_{i=1}^{24} \alpha_i y_i e^{-0.0004\|x-x_i\|^2} + 0.5923\right) \tag{10}$$

$$3\ vs\ 4\ \ f(x) = sign\left(\sum_{i=1}^{24} \alpha_i y_i e^{-0.0004\|x-x_i\|^2} - 0.9222\right) \tag{11}$$

Discriminant for the 4th class of MOA

$$1\ vs\ 4\ \ f(x) = sign\left(\sum_{i=1}^{9} \alpha_i y_i e^{-0.0004\|x-x_i\|^2} + 0.1347\right) \tag{12}$$

$$2\ vs\ 4\ \ f(x) = sign\left(\sum_{i=1}^{9} \alpha_i y_i e^{-0.0004\|x-x_i\|^2} + 3.6532\right) \tag{13}$$

$$3\ vs\ 4\ \ f(x) = sign\left(\sum_{i=1}^{9} \alpha_i y_i e^{-0.0004\|x-x_i\|^2} - 0.9222\right) \tag{14}$$

LIBSVM uses the following voting strategy: if discriminant says the unknown sample *x* is in the *i*–th class, then the vote for the *i*–th class is added by one. Otherwise, the *j*–th is increased by one. Then we predict the unknown sample *x* is in the class with the largest number of votes. According to the result of voting strategy, 13 compounds are misclassified (see Table 4). These compounds are: **1**, 1,3,5–trihydroxybenzene; **94**, 4–chlororesorcinol; **127**, 5–methylresorcinol; **145**, resorscinol; **162**, 2,6–dichloro–4–nitrophenol; **181**, 4–acetamidophenol; **185**, 4–chlorocatechol; **190**, bromohydroquinone; **191**, catechol; **192**, chlorohydroquinone; **196**, phenylhydroquinone; **199**, 2,6–dibromo–4–nitrophenol; **221**, 5–hydroxy–2–nitrobenzaldehyde (see the column MOA_{ALL} in Table 1). The numbers of compounds misclassified are 4, 1, 6 and 2 for four classes of MOAs of phenols, respectively. The accuracy index is 208/221 = 94.12% for all classes, 149/153 = 97.39 for class 1, 17/18 = 94.44% for class 2, 21/27 = 77.77% for class 3, and 21/23 = 91.30% for class 4.

In order to test stability of the SVM model, first, we used 155 compounds in the training set to construct the SVM classifier with chose parameters: BRF kernel, parameter C = 128 and $\gamma = 0.0004$. There are 8 compounds in the training set and 8 ones in the testing set to be misclassified (see Table 4). These compounds are: **156**, 2,3,5,6–tetrafluorophenol; **162**, 2,6–

dichloro–4–nitrophenol; **163**, 2,6–diiodo–4–nitrophenol; **173**, 1,2,4–trihydroxybenzene; **181**, 4–acetamidophenol; **185**, 4–chlorocatechol; **186**, 4–methylcatechol; **190**, bromohydroquinone; **191**, catechol; **192**, chlorohydroquinone; **193**, hydroquinone; **194**, methoxyhydroquinone; **195**, methylhydroquinone; **196**, phenylhydroquinone; **216**, 4–nitro–3–(trifluoromethyl)–phenol; **221**, 5–hydroxy–2–nitrobenzaldehyde respectively (see Table 1).

Table 4. The number of misclassifications of two models which are 221 compounds model and 155 compounds model

Type of MOA	The number of misclassification for 221 compounds model	The number of misclassification for 155 compounds model	
		Training set	Testing set
1, polar narcosis	4/153	0/107	0/46
2, weak acid respiratory uncouplers	1/18	1/13	2/5
3, precursors to soft electrophiles	6/27	6/19	5/8
4, soft electrophiles	2/23	1/16	1/7

And then, *k*–fold cross–validation was used for model validation. All compounds were randomly divided into *k* (*k* = 2, 5, 10) subsets with equal percentages of each MOA present in each subset. Then a model was fitted taking *k*–1 of these subsets as the training set to construct the SVM classifier with the following parameters: BRF kernel, parameter C = 128 and γ = 0.0004, and the remaining one as a test set. The cross–validation accuracy is 88.64%, 88.18%, and 90.91%, respectively.

Aptula *et al*. [5] studied this data set employing 3–6 molecular descriptors with LDA, and those LDA models achieved 86–89% overall accuracy index for the four mechanisms. In those models, two equalized complementary subsets subdivided from all compounds were taken as external prediction and training set respectively, and used the leave–one–out cross validation method to avoid pitfalls accidentally. From this result, we concluded that several compounds belonging to polar narcotics (class 1) are predicted as pro–electrophiles (class 3) and that some others belonging to class 3 are predicted as class l. There are also several compounds that are misclassified between respiratory uncouplers (class 2) and soft electrophiles (class 4). In the first step of the Ren's [7] two–step method, all compounds are classified into two groups, one including class 1 and class 3 and the other including class 2 and mode 4. In the second step, class 1 and class 3, class 2 and class 4 were discriminated respectively. The quality of this two–step is better than LDA (see Table 5). Yao [12] also used the same data set and divided all compounds into two groups similar to Aptula. A model constructed by SVM achieved 93.6% overall accuracy. In Spycher's [13] study, the data set has some small changes and additional descriptors. It contains 220 compounds (155 class 1, 19 class 2, 24 class 3, and 22 class 4), and two compounds of the training set were assigned to a different MOA than Aptula, Yao, Ren and this study. No. **179** and **187** in Table 1 were reassigned as class 1, **199** was classified as class 2, and **198** was omitted. A 21–dimensional model that successfully discriminated between the four MOAs was developed. Its overall predictive power was estimated to 92% using 5–fold cross–validation.

Compared with the results of Ren, Aptula and Yao, the present study also gives a better result of classification of MOAs, but the accuracy index for precursors to soft electrophiles (class 3) is lower than that of Ren, and Aptula in some cases. This may indicate that the MEDV descriptors are not efficient for this mechanism. It is also possible that some phenols act by more than one mechanism, or mixed mechanisms. Spycher's overall predictive power using 5–fold cross–validation was better than that of this study, but the data set is different (see Table 5).

Table 5. The accuracy index of three methods (%).

Type of MOAs	Ren [7]	Aptula [5]	Spycher [13] [a]	Spycher [13] [b]	Yao [12]	This study
1, polar narcosis	95.7	84.3–96.7	98.1	96.8	–	97.39
2, weak acid respiratory uncouplers	73.2	55.6–77.8	68.4	94.7	–	94.44
3, precursors to soft electrophiles	83.5	37.0–92.6	79.2	79.2	–	77.77
4, soft electrophiles	86.4	69.6–87.0	86.4	77.3	–	91.30
Overall prediction	90.9	–	92.3	92.7	93.6 [c]	94.12

[a] Result in CPG NN; [b] Result in multinomial logistic regression; [c] Average of two experiment.

Table 6. Analysis of result of this study, each row indicates how the compounds of each class (class 1– to class 4) are assigned by SVM to different modes of action (columns)

	Class 1	Class 2	Class 3	Class 4
Class 1	149	0	6	1
Class 2	0	17	0	1
Class 3	4	0	21	0
Class 4	0	1	0	21

Aptula suggested that weak acid respiratory uncoupling phenols (class 2) and precursors to soft electrophile phenols (class 3) were similar in the sense of mechanisms classification. However, after analyzing the distance between mechanisms, Ren suggested that when using 4–mechanism model and when the prediction accuracy for class 2 was high, the low prediction accuracy for class 3 does not necessarily imply that class 2 and class 3 are similar in the molecular descriptor space. Yao summarized the study of Aptula, showing that several samples belonging to class 1 are predicted as class 3 and that some others belonging to class 3 are predicted as class 1. There are also several compounds misclassified between class 2 and class 4. Obviously, our results are similar to ones of the literature [12] (also see Table 6).

4 CONCLUSIONS

In this study we have investigated the application of SVM with one–against–one multi–class classification method for the classification of 221 phenols compounds from four MOA classes (polar narcosis, weak acid respiratory uncouplers, precursors to soft electrophiles and soft electrophiles). The MOA classification was based on MEDV descriptors. The prediction power of each SVM model was evaluated with a leave–one–out cross–validation procedure. Because there is no general criterion for SVM model selection, we have investigated two SVM with grid–search

method methods (*C*–SVMC with parameter $C = 1\times2^N$, $N = 0, 1, 2\ldots12$ and ν–SVMC with $\nu = 0.001\times2^N$, $N = 0, 1, 2\ldots9$) with two kernels (linear and RBF kernels with $\gamma=0, 0.00025\times2^N$, $N=0, 1, 2\ldots 9$). There are total 345 SVM models have been built. The result showed that the quality of SVM models classifiers for MOA depends strongly on SVM methods, the kernel type and various parameters that control the kernel shape. We took a RBF kernel with $\gamma = 0.0004$ and capacity parameter of $C = 128$ of *C*–SVMC to construct the final SVM model which has the highest accuracy index of leave–one–out cross validation. The accuracy index of all 221 compounds is 94.1%, 13 compounds are misclassified. To test the stability of this SVM model, we have uniformly chosen 2/3 from all 221 compounds as a training set which is used to obtain a new SVM model, the rest compounds being used as a testing set. A total of 16 compounds (8 in the training set and 8 in testing set) were misclassified, which gives an accuracy index of 92.8%. The result indicates that SVM model has a high ability for predicting the aquatic toxicity mechanism, if we employ appropriate parameters of SVM and molecular descriptors.

Acknowledgment

We are especially grateful to the Guangxi Natural Science Fund (No. 0236063) and Guilin University of Technology youth fund for their financial supports.

5 REFERENCES

[1] A. R. Katritzky and D. B. Tatham, 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] T. W. Schultz, M. T. D. Cronin, and T. I. Netzeva, The present status of QSAR in toxicology, *J. Mol. Struct. (Theochem)*. **2003**, *622*, 23–38.

[3] S. Ren, P. D. Frymier, and T. W. Schultz, An Exploratory Study of the Use of Multivariate Techniques to Determine Mechanisms of Toxic Action, *Ecotox. Environ. Safety* **2003**, *55*, 86–97.

[4] S. Ren, Ecotoxicity Prediction Using Mechanism– and Non–mechanism–based QSARs: a Preliminary Study, *Chemosphere*. **2003**, *53*, 1053–1065.

[5] A. O. Aptula, T. I. Netzeva, I. V. Valkova, M. T. D. Cronin, T. W. Schultz, and R. Kuhne, G. Schuurmanna, Multivariate Discrimination between Modes of Toxic Action of Phenols, *Quant. Struct.–Act. Relat.* **2002**, *21*, 12–22.

[6] Z. Rappoport, *The chemistry of phenols*. John Wiley & Sons Ltd: Chichester, 2003.

[7] S. Ren, Two–step Multivariate Classification of the Mechanisms of Toxic Action of Phenols, *QSAR Comb. Sci.* **2003**, *22*, 596–603.

[8] M. T. Cronin and T. W. Schultz, Validation of Vibrio Fisheri Acute Toxicity Data: Mechanism of Action–based QSARs for Non–polar Narcotics and Polar Narcotic Phenols, *Sci. Total Environ.* **1997**, *204*, 75–88.

[9] S. Ren, Determining the Mechanisms of Toxic Action of Phenols to *Tetrahymena pyriformis*, *Environ. Toxicol.* **2002**, *17*, 119–127.

[10] S. Ren, Use of Molecular Descriptors in Separating Phenols by Three Mechanisms of Toxic Action, *Quant. Struct.–Act. Relat.* **2002**, *21*, 486–492.

[11] S. Ren, Classifying Class I and Class II Compounds By Hydrophobicity and Hydrogen Bonding Descriptors, *Environ. Toxicol.* **2002**, *17*, 415–423.

[12] X. J. Yao, A. Panaye, J. P. Doucet, H. F. Chen, R. S. Zhang, B. T. Fan, M. C. Liu, Z. D. Hu, Comparative Classification Study of Toxicity Mechanisms Using Support Vector Machines and Radial Basis Function Neural Networks, *Anal. Chim. Acta*. **2005**, *535*, 259–273.

[13] S. Spycher, E. Pellegrini, and J. Gasteiger, Use of Structure Descriptors To Discriminate between Modes of Toxic Action of Phenols, *J. Chem. Inf. Model*. **2005**, *45*, 200–208.

[14] V. N. Vapnik, *Statistical Learning Theory*, Wiley, New York, 1998.

[15] N. Cristianini and J. Shawe–Taylor, *An Introduction to Support Vector Machines and Other Kernel–based Learning Methods*, Cambridge University Press, Cambridge, 2000.
[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] 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.
[18] O. Ivanciuc, Support Vector Machines Prediction of the Mechanism of Toxic Action from Hydrophobicity and Experimental Toxicity Against *Pimephales promelas* and *Tetrahymena pyriformis*, *Internet Electron. J. Mol. Des.* **2004**, *3*, 802–821, http://www.biochempress.com.
[19] M. P. S. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. Sugnet, J. Manuel Ares, and D. Haussler, *Support Vector Machine Classification of Microarray Gene Expression Data*; UCSC–CRL–99–09; University of California: Santa Cruz, 9, 1999.
[20] P. Lind and T. Maltseva, Support Vector Machines for the Estimation of Aqueous Solubility, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1855–1859.
[21] O. Sadik, J. Walker H. Land, A. K. Wanekaya, M. Uematsu, M. J. Embrechts, L. Wong, D. Leibensperger, and A. Volykin, Detection and Classification of Organophosphate Nerve Agent Simulants Using Support Vector Machines with Multiarray Sensors, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 499–507.
[22] S. S. Liu, C. S. Yin, Y. Y. Shi, S. X. Cai, and Z. L. Li, MEDV–13 for QSAR Studies on the COX–2 Inhibition by Indomethacin Amides and Esters, *Chin. J. Chem.* **2001**, *19*, 751–756.
[23] S. S. Liu, C. S. Yin, Z. L. Li, and S. X. Cai, QSAR Study of Steroid Benchmark and Dipeptides Based on MEDV–13, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 321–329.
[24] S.–S. Liu, H.–L. Liu, Y.–Y. Shi, and L.–S. Wang, QSAR of Cyclooxygenase–2 (COX–2) Inhibition by 2,3–Diarylcyclopentenones Based on MEDV–13, *Internet Electron. J. Mol. Des.* **2002**, *1*, 310–318, http://www.biochempress.com.
[25] S. S. Liu, S. H. Cui, D. Q. Yin, Y. Y. Shi, and L. S. Wang, QSAR Studies on the COX–2 Inhibition by 3,4–Diarylcycloxazolones Based on MEDV Descriptor, *Chin. J. Chem.* **2003**, *21*, 1510–1516.
[26] S. S. Liu, S. H. Cui, and L. S. Wang, Prediction of the Aqueous Solubilities of Polychlorinated Biphenyls, *Chin. Chem. Lett.* **2004**, *15*, 467–470.
[27] S. S. Liu, S. X. Cai, C. Z. Cao, and Z. L. Li, Molecular Electronegative Distance Vector (MEDV) Related to 15 Properties of Alkanes, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1337–1348.
[28] S.–S. Liu, S.–H. Cui, Y.–Y. Shi, and L.–S. Wang, A Novel Variable Selection and Modeling Method based on the Prediction for QSAR of Cyclooxygenase–2 Inhibition by Thiazolone and Oxazolone Series, *Internet Electron. J. Mol. Des.* **2002**, *1*, 610–619, http://www.biochempress.com.
[29] S. R. Gunn. *Support Vector Machines for Classification and Regression*; Image Speech and Intelligent Systems Research Group, University of Southampton: 1998.
[30] C. C. Chang and C. J. Lin *LIBSVM – A Library for Support Vector Machines*, 2.6; 2001. http://www.csie.ntu.edu.tw/~cjlin/libsvm.
[31] C. W. Hsu and C. J. Lin, A Comparison of Methods for Multi–class Support Vector Machines. *IEEE Trans. Neural Networks*. **2002**, *13*, 415–425.
[32] J. C. Platt, N. Cristianini, and J. Shawe–Taylor, Large Margin DAGs for Multiclass Classification, *Adv. Neural Inform. Process. Syst.* **2000**, *12*, 547–553.
[33] J. Weston and C. Watkins, *Multi–class Support Vector Machines*; Technical Report CSD–TR–98–04, Royal Holloway University of London: Egham, UK, May 20, 1998.