CODEN IEJMAT *Internet Electronic Journal of Molecular Design* **2007**, *6*, 229–236 ISSN 1538–6414
BioChem Press http://www.biochempress.com

Inter*net* Electronic Journal of Molecular Design

August 2007, Volume 6, Number 8, Pages 229–236

Editor: Ovidiu Ivanciuc

Support Vector Machines QSAR for the Toxicity of Organic Chemicals to *Chlorella vulgaris* with SVM Parameters Optimized with Simplex

Zhong–Sheng Yi [1] and Li–Tang Qin [1]

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

Received: September 23, 2006; Accepted: December 5, 2006; Published: August 31, 2007

Citation of the article:
Z.–S. Yi and L.–T. Qin, Support Vector Machines QSAR for the Toxicity of Organic Chemicals to *Chlorella vulgaris* with SVM Parameters Optimized with Simplex, *Internet Electron. J. Mol. Des.* **2007**, *6*, 229–236, http://www.biochempress.com.

Copyright © 2007 BioChem Press

Inter*net* **Electronic** Journal of Molecular Design

BioChem Press
http://www.biochempress.com

Support Vector Machines QSAR for the Toxicity of Organic Chemicals to *Chlorella vulgaris* with SVM Parameters Optimized with Simplex

Zhong–Sheng Yi [1,*] and Li–Tang Qin [1]

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

Received: September 23, 2006; Accepted: December 5, 2006; Published: August 31, 2007

Internet Electron. J. Mol. Des. **2007**, *6* (8), 229–236

Abstract

Motivation. The key to a successful application of support vector machines (SVM) is to selecte proper parameters, but there is no general method for selecting the best set of SVM parameters. The predictive power of SVM models depends strongly on the set of parameters that control the model. In this paper we used the simplex optimization method to search for the optimum set of SVM parameters, namely the capacity parameter *C*, the insensitive loss parameter ε and the parameter γ that controls the shape of the RBF kernel.

Method. The leave–one–out cross–validation correlation coefficient q^2 is used as objective function for the simplex optimization of SVM parameters.

Results. SVM quantitative structure–activity relationships (QSAR) models were built for the toxicity of organic chemicals to *Chlorella vulgaris*. The SVM models with simplex optimized parameters are compared with multi–linear regression QSAR models obtained in the same conditions.

Conclusions. A series of QSAR models with one to three variables were obtained for the acute toxicity of 91 organic chemicals to *Chlorella vulgaris*. The SVM models with parameters optimized with simplex have better statistics than the multi–linear regression QSAR equations. The results from the present investigation demonstrate that the simplex algorithm is an efficient approach in finding the best set of SVM parameters for QSAR models.

Keywords. SVM; support vector machines; QSAR; quantitative structure–activity relationships; simplex optimization; *Chlorella vulgaris*; acute toxicity.

1 INTRODUCTION

In recent years, with the development of modern industrial technology, a large number of organical chemicals have entered into the natural environment, on which human life relies for existence, and some chemicals have seriously polluted the environment. Due to the high cost and time, it is difficult to determine experimentally the toxicity of a large number of chemicals. Due to their high predictive power, quantitative structure–activity relationships (QSAR) are powerful tools

* Correspondence author; E–mail: yzs@glite.edu.cn and yizhsh@gmail.com.

for predicting chemical toxicity of environmental pollutants [1,2]. Many interesting applications of QSAR models were developed in environmental toxicology, demonstrating their theoretical and practical importance.

Recently, by using multi–linear regression (MLR), Cronin and his co–workers [3] assessed and modeled the toxicity of selected 91 organic compounds to *Chlorella vulgaris* using three descriptors, which included hydrophobicity expressed by the 1–octanol/water partition coefficient ($\log K_{ow}$), electrophilicity expressed by the lowest unoccupied molecular orbital (LUMO) and a function of molecular size corrected for the presence of hereroatoms expressed by the first–order delta valence connectivity index ($\Delta^1\chi^v$), and concluded that method selection in QSAR is task–dependent and there should be clear indications supporting the need of more complied, nonlinear, methods that may deliver better QSAR. As a power classified and regression tool, the support vector machines (SVM) have wide application in QSAR, and we attempt to re–build this model using it.

SVM is a popular algorithm developed from the statistics learning theory by Vapnik [4,5]. Due to its remarkable generalization performance, the SVM has attracted attention and gained extensive application such as MOA prediction [6–8], classification of microarray gene expression data [9], estimation of aqueous solubility [10], classification of organophosphate nerve agent simulants [11]. The key of the SVM's models quality is the selection of SVM parameters, similarly with artificial neural networks (ANN), and the molecular structure descriptors. Many studies demonstrated the importance of the SVM parameters in developing a predictive QSAR [12–14]. In this study, a modified simplex optimization was used to optimize SVM parameters and various QSAR models were built with difference descriptors for the toxicity of selected chemicals to *Chlorella vulgaris*. Compared with the result of multi–linear regression (MLR), SVM QSAR models have higher prediction power.

2 MATERIALS AND METHODS

2.1 Data Set and Structural Descriptors

The toxicity data (pEC_{50}) of a total of 91 chemicals covering a wide range of physicochemical properties and structural features, were collected from literature [3] and are listed in Table 1. The data set was split into a training set of 73 chemicals and a test set of 18 chemicals. The training set and test set were used to test the reliability of SVM model using the parameters through simplex optimization. Cronin [3] obtained a successful MLR model by using logarithm of the octanol–water partition coefficient ($\log K_{ow}$), the energy of the lowest unoccupied molecular orbital (E_{LUMO}) and the first–order delta valence connectivity index ($\Delta^1\chi^v$). In order to compare the performance of the MLR QSAR with that of SVM models, the same descriptors were used in this study.

Table 1. The Acute *Chlorella vulgaris* Toxicities and Descriptors for 91 Compounds

No	Name	logK_{ow}	E_{LUMO}	$\Delta^1\chi^v$	pEC_{50} (Exp)	pEC_{50} Eq. (1)	pEC_{50} Eq. (2)
			Training Set				
1	methanol	–0.77	3.778	0.051	–4.06	–4.43	–4.16
2	ethanol	–0.31	3.565	–0.130	–3.32	–3.94	–3.74
4	butan–2–ol	0.61	3.554	–0.500	–2.98	–3.06	–2.92
5	2–hydroxyethyl methacrylate	0.47	–0.074	–2.275	–2.82	–1.71	–1.72
6	2–hydroxyethyl acrylate	–0.21	–0.102	–2.044	–2.79	–2.34	–2.36
7	methyl acrylate	0.80	0.001	–1.509	–2.75	–1.67	–1.69
9	butanone	0.29	0.882	–0.686	–2.51	–2.56	–2.57
10	methyl methacrylate	1.38	0.055	–1.736	–2.24	–1.14	–1.14
11	pentan–3–one	0.99	0.910	–0.717	–2.23	–1.97	–1.98
12	crotonaldehyde	0.52	–0.141	–0.972	–1.98	–2.02	–2.03
14	trans–2–pentenal	1.05	–0.115	–1.024	–1.88	–1.56	–1.58
15	phenol	1.47	0.398	–1.477	–1.46	–1.22	–1.23
16	allyl methacrylate	1.68	0.045	–2.240	–1.42	–0.74	–0.73
17	aniline	0.90	0.639	–1.260	–1.34	–1.83	–1.84
19	anisole	2.11	0.483	–1.644	–1.09	–0.66	–0.66
20	2–fluorophenol	1.71	0.013	–2.166	–1.08	–0.73	–0.72
21	2–fluoroaniline	1.26	0.266	–1.998	–1.05	–1.22	–1.23
22	3–cresol	1.96	0.396	–1.622	–1.01	–0.77	–0.77
24	2–hydroxyaniline	0.62	0.474	–1.806	–0.91	–1.86	–1.88
25	2–methoxyphenol	1.32	0.392	–2.194	–0.88	–1.15	–1.15
26	2,6–dimethylaniline	1.84	0.595	–1.476	–0.87	–0.97	–0.97
27	benzaldehyde	1.47	–0.435	–1.732	–0.81	–0.93	–0.93
29	2–hydroxybenzaldehyde	1.81	–0.434	–2.282	–0.80	–0.49	–0.47
30	nitrobenzene	1.85	–1.068	–2.293	–0.78	–0.29	–0.25
31	methidathion	2.42	–2.550	–1.436	–0.73	0.35	0.42
32	4–cresol	1.94	0.429	–1.622	–0.66	–0.80	–0.80
34	4–tolualdehyde	1.99	–0.430	–1.866	–0.65	–0.46	–0.44
35	2–ethoxyphenol	1.85	0.422	–2.184	–0.62	–0.72	–0.71
36	3–cyanobenzaldehyde	1.18	–0.917	–2.452	–0.57	–0.84	–0.82
37	3–nitrotoluene	2.42	–1.017	–2.481	–0.50	0.23	0.28
39	2,4–dinitroaniline	1.72	–1.475	–3.840	–0.36	0.15	0.26
40	4–bromophenol	2.59	0.020	–1.578	–0.35	–0.16	–0.14
41	4–bromoaniline	2.26	0.218	–1.289	–0.33	–0.57	–0.56
42	3–chloroaniline	1.88	0.263	–1.350	–0.31	–0.88	–0.88
44	3,5–dinitroaniline	1.89	–1.780	–3.846	0.03	0.37	0.50
45	2–chlorobenzaldehyde	2.33	–0.683	–1.838	0.06	–0.11	–0.09
46	4–iodophenol	2.91	0.024	–1.590	0.16	0.11	0.13
47	4–ethylbenzaldehyde	2.52	–0.423	–1.842	0.16	–0.02	0.00
49	2–isopropylphenol	2.88	0.408	–1.754	0.17	0.03	0.05
50	3,5–dichloroaniline	2.90	–0.042	–1.428	0.24	0.08	0.10
51	1,3,5–trimethyl–2–nitrobenzene	3.22	–0.857	–2.808	0.25	0.95	1.01
52	2,6–dichloroaniline	2.82	–0.006	–1.416	0.26	0.00	0.01
54	1,2–dichlorobenzene	3.43	–0.142	–1.011	0.37	0.43	0.46
55	1,3–dinitrobenzene	1.49	–1.911	–3.532	0.38	–0.02	0.09
56	2,4–dinitrophenol	1.67	–1.807	–3.976	0.40	0.23	0.36
57	1,4–dinitrobenzene	1.47	–2.208	–3.532	0.41	0.05	0.16
59	phosmet	2.78	–2.349	–2.322	0.47	0.84	0.93
60	methylparathion	2.86	–2.068	–3.077	0.60	1.05	1.15

Table 1. (Continued)

No	Name	logK_{ow}	E_{LUMO}	$\Delta^1\chi^v$	pEC_{50} (Exp)	pEC_{50} Eq. (1)	pEC_{50} Eq. (2)
61	malathion	2.36	–2.658	–2.457	0.64	0.61	0.71
62	2,6–dichloro–4–nitroaniline	2.80	–1.096	–2.970	0.64	0.71	0.78
64	2,4–dinitrotoluene	1.98	–1.841	–3.760	0.70	0.44	0.56
65	cyanophos	2.75	–1.832	–2.349	0.79	0.69	0.76
66	6–chloro–2,4–dinitroaniline	2.46	–1.667	–4.080	0.80	0.88	1.02
67	2,6–dibromo–4–nitrophenol	3.57	–1.452	–3.315	0.81	1.54	1.63
69	2,5–dichloronitrobenzene	3.03	–1.296	–2.633	0.97	0.86	0.92
70	piperine	2.70	–0.767	–3.985	0.97	0.82	0.92
71	4–tert–butylbenzaldehyde	3.32	–0.391	–2.210	1.00	0.74	0.78
72	2,4,6–trichloroaniline	3.69	–0.240	–1.485	1.11	0.81	0.83
74	4–chloro–2,6–dinitroaniline	2.46	–1.895	–4.080	1.19	0.94	1.09
75	1,2–dinitrobenzene	1.69	–1.840	–3.526	1.23	0.13	0.24
76	1–chloro–4–nitrobenzene	2.39	–1.344	–2.476	1.25	0.29	0.35
77	dicapthon	3.58	–2.124	–3.256	1.36	1.71	1.82
79	2,3,5,6–tetrachloroaniline	4.47	–0.560	–1.535	1.48	1.56	1.58
80	2,4–dichloro–6–nitrophenol	3.07	–1.431	–3.174	1.50	1.08	1.17
81	2,6–dichlorobenzaldehyde	3.08	–0.473	–1.931	1.50	0.48	0.52
82	fenthion	4.09	–1.628	–1.486	1.56	1.52	1.56
84	pentachlorophenol	5.12	–0.978	–1.931	1.69	2.33	2.33
85	fenitrothion	3.30	–2.027	–3.256	1.71	1.45	1.56
86	1,2,4–trichloro–5–nitrobenzene	3.47	–1.536	–2.774	1.88	1.33	1.41
87	1,3,5–trichloro–2,4–dinitrobenzene	2.97	–2.037	–4.179	1.89	1.44	1.59
89	4–(dibutylamino)benzaldehyde	5.06	–0.097	–2.392	2.18	2.17	2.16
90	2,3,5,6–tetrachloronitrobenzene	4.38	–1.419	–2.895	2.34	2.10	2.15
91	pentabromophenol	4.85	–1.193	–1.459	3.10	2.03	2.04
	Testing set						
3	2–methyl–propan–2–ol	0.35	3.438	–0.728	–3.16	–3.18	–3.06
8	butan–1–ol	0.88	3.425	–0.428	–2.73	–2.82	–2.69
13	trans–2–hexenal	1.58	–0.115	–1.094	–1.94	–1.10	–1.10
18	2–heptanone	1.98	0.879	–0.902	–1.18	–1.09	–1.08
23	4–methoxyphenol	1.34	0.313	–2.200	–0.97	–1.11	–1.11
28	2–cresol	1.95	0.396	–1.616	–0.81	–0.78	–0.78
33	3,4–dimethylphenol	2.23	0.436	–1.750	–0.65	–0.52	–0.52
38	4–chlorophenol	2.39	0.095	–1.603	–0.42	–0.34	–0.33
43	benzyl Methacrylate	2.53	0.079	–3.044	–0.21	0.19	0.23
48	2–methyl–1,4–naphthoquinone	2.20	–1.493	–3.045	0.16	0.33	0.41
53	2–tert–butyl phenol	3.27	0.407	–1.747	0.29	0.36	0.37
58	3–nitrobenzaldehyde	1.47	–1.404	–3.102	0.45	–0.29	–0.22
63	methyl azinphos	2.75	–2.494	–2.186	0.69	0.82	0.90
68	thiometon	3.15	–2.632	1.134	0.94	0.27	0.37
73	2–chloro–6–nitrotoluene	3.09	–1.219	–2.636	1.17	0.89	0.95
78	2,6–di–tert–butyl–4–methylphenol	5.89	0.464	–2.040	1.45	2.62	2.55
83	2,4–di–tert–butylphenol	4.36	0.431	–1.936	1.60	1.32	1.33
88	phenylazophenol	3.96	–0.768	–3.417	2.16	1.71	1.78

2.2 Support Vector Machines Regression

Initially, SVM is developed for pattern recognition problems, but now, with the introduction of insensitive loss function, SVM has been extended to solve nonlinear regression estimation with excellent performances [10]. Here, SVM can be called support vector regression (SVR). A detailed description of the regression theory of SVM can be seen in several excellent books and tutorials [15–18]. For this reason, we will only briefly describe the main ideas of SVM regression here.

A support vector machine is first trained on a group of objects having known target values. After training, the SVM model is used to predict or estimate target values for objects where these values are unknown. A kernel–induced feature space with function $k(x_i, x)$ is used for the mapping of objects onto target values. Thus a non–linear feature mapping will allow the treatment of non–linear problems in linear space. The prediction or approximation function used by a basis SVM is $f(x) = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) k(x_i, x) + b$, where α_i and α_i^* are Lagrange multipliers which are mostly zero and have $\alpha_i \times \alpha_i^* = 0$, where x_i is a feature vector corresponding to a training object, and $k(x_i, x)$ is a kernel function such as linear, polynomial, radical basis function and sigmoid. The components of vector α and the constant b represent the hypothesis and are optimized during training. It may be useful to think of kernel $k(x_i, x)$ as comparing patterns, or as evaluating the proximity of objects in their feature space. Thus a test point is evaluated by comparing it to all training points. Training points with non–zero weights α_i are called the support vectors.

For a given dataset and one of the two SVM methods, mu–SVR and nu–SVR, the kernel function with parameters respectively and the capacity parameter C of SVM must be selected to determine a specific SVM model. All SVM models used mu–SVR and the radical basis function (RBF) $k(x, x_i) = e^{-\gamma \|x - x_i\|^2}$, and all calculations were performed with LIBSVM software by Chang and Lin [17]. Therefore, there are 3 parameters to be optimized, namely the capacity parameter C, the insensitive loss parameter ε and the parameter γ that controls the shape of the RBF kernel. The target function of the simplex optimization was the correlation coefficient of leave–one–out cross validation, q^2.

3 RESULTS AND DISCUSSION

3.1 Multi–linear Regression

Three descriptors, logK_{ow}, E_{LUMO} and $^1\chi^v$, were chosen from 110 descriptors in Ref. [3] to characterize 91 compounds. The multi–linear models built by various combinations of those descriptors are list in Table 2. This result show that logK_{ow} is the most important descriptor, the next is E_{LUMO} and then $^1\chi^v$, because q^2 for these descriptors is 0.7410, 0.4648, 0.2437 and r^2 is 0.7561, 0.4819, 0.2881, respectively. From the model built with three descriptors (logK_{ow}, E_{LUMO}, and $^1\chi^v$), we can find that the importance of descriptors $^1\chi^v$ is very small. When $^1\chi^v$ is used in the three parameters model, the q^2 only increase about 0.02 and R^2 only 0.022 compared with the two parameters (logK_{ow} and E_{LUMO}) model. The three variables model of logK_{ow}, E_{LUMO} and $^1\chi^v$ is:

$$\begin{aligned} pEC_{50} &= -(2.7602 \pm 0.1735) + (0.8376 \pm 0.0474) \log K_{ow} \\ &\quad - (0.2678 \pm 0.0540) E_{LUMO} - (0.2786 \pm 0.0664) {}^1\chi^v \\ n &= 91,\ R^2 = 0.8903,\ q^2 = 0.8752,\ RMSE = 0.4826,\ F = 244 \end{aligned} \quad (1)$$

The predicted toxicities of 91 organic chemicals from Eq. (1) are list in Table 1.

Table 2. The Result of Building Linear Models for Different Combination between Three Descriptors

No	No. samples	q^2	R^2	S	F	Descriptors
1	91	0.7410	0.7561	0.7196	279	$\log K_{ow}$
2	91	0.4648	0.4819	1.0489	83	E_{LUMO}
3	91	0.2437	0.2881	1.2296	36	$\Delta^1\chi^v$
4	91	0.8577	0.8682	0.5291	296	$\log K_{ow}$ E_{LUMO}
5	91	0.8417	0.8593	0.5466	275	logKow $\Delta^1\chi^v$
6	91	0.4699	0.4966	1.0340	45	E_{LUMO} $\Delta^1\chi^v$
7	91	0.8754	0.8903	0.4826	244	$\log K_{ow}$ E_{LUMO} and $\Delta^1\chi^v$

Table 3. The Initial Simplex and the Best Parameters of SVM

No.	ε	C	γ	q^2
1	0.0078	40.7294	0.0108	0.8702
2	0.0655	40.7294	0.0108	0.8042
3	0.0226	402.2643	0.0108	0.7957
4	0.0226	128.0000	0.0472	0.8414
62	0.0021	62.2426	0.0039	0.8879

3.2 Simplex Optimization of the SVM Parameters

The QSAR models were obtained by considering $\log K_{ow}$, E_{LUMO} and $\Delta^1\chi^v$ as SVR input descriptors, the toxicity of organic chemicals as output, the parameters C, ε and γ as simplex factors, q^2 of SVR as simplex object function. The starting simplexes were generated by gold section. The optimization converges when the difference is lower than 10^{-4} among each peak of the last simplex. The initial simplex (No. 1 to 4) and the best parameters (No. 62) of SVR are listed in Table 3. The convergence parameters were $q^2 = 0.8879$, $C = 62.2426$, $\varepsilon = 0.0021$, $\gamma = 0.0039$. The other sets of optimum parameters are listed in Table 4 (No. 1 to 7).

Table 4. The Best Parameters of SVM for the Different Combination between Three Descriptors

No.	No. samples	ε	C	γ	q^2	R^2	S	Descriptors
1	91	0.0034	129.2170	0.0042	0.7811	0.7922	0.4492	$\log K_{ow}$
2	91	0.0456	15.6773	0.0079	0.5099	0.5139	1.0368	E_{LUMO}
3	91	0.3834	0.7121	0.0069	0.3876	0.3891	1.3095	$\Delta^1\chi^v$
4	91	0.0881	17.5386	0.0546	0.8855	0.8997	0.2141	$\log K_{ow}$, E_{LUMO}
5	91	0.0212	11.6073	0.0312	0.8673	0.8858	0.2434	$\log K_{ow}$, $\Delta^1\chi^v$
6	91	0.0066	26.1815	0.0266	0.4974	0.5081	1.0496	E_{LUMO}, $\Delta^1\chi^v$
7	91	0.0021	62.2426	0.0039	0.8879	0.8956	0.2238	$\log K_{ow}$, E_{LUMO}, $\Delta^1\chi^v$
8	73	0.0021	62.2426	0.0039	0.8851	0.8931	0.2317	$\log K_{ow}$, E_{LUMO}, $\Delta^1\chi^v$
9	73	0.0050	63.9971	0.0149	0.8896	0.8964	0.2229	$\log K_{ow}$, E_{LUMO}, $\Delta^1\chi^v$

After choosing the SVR parameters, we generated the following prediction or approximation function of three descriptors

$$pEC_{50} = \sum_{i=1}^{72}((\alpha_i - \alpha_i^*)\exp(-\gamma^*\|x_i, x\|^2)) - 0.9778 \quad (2)$$

where support vector samples is 72, γ* is the optimum parameter that controls the RBF shape. The

estimated toxicities of organic chemicals by Eq. (2) are listed in Table 1, and the plot of estimated by Eq. (1) versus observed toxicity of organic chemicals is shown in Figure 1. A comparison of the results from Tables 2 and 4 show that q^2 and R^2 of SVR models are larger than that of the corresponding linear models.

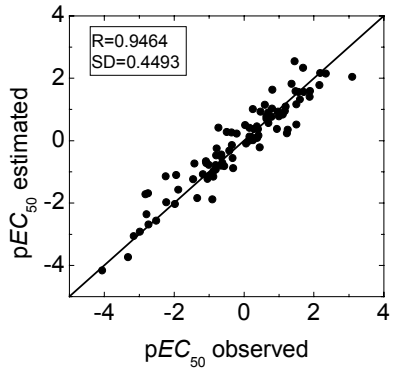

Figure 1. Plot of the observed algal toxicity against that predicted in Eq. (2)

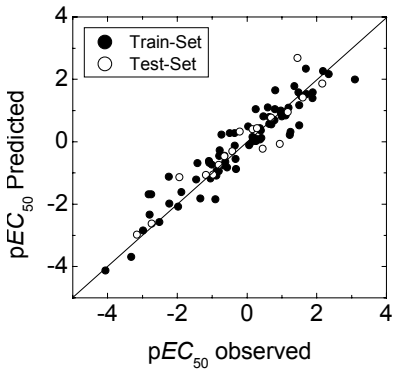

Figure 2. Plot of the observed algal toxicity against that predicted in Eq. (3)

A good QSAR model should have not only excellent estimation ability for the training set but also a better predictive power for the external samples. In order to validate the predictive ability of the model, 73 samples are selected from all 91 organic chemicals to construct a training set and the remaining 18 formed a test set. The 73 samples in the training set are employed to develop a QSAR model and then the SVR model is used to predict the toxicity of organic chemicals in the test set. The statistical parameters of the training set, optimized with simplex, are presented in Table 4: No. 7 is obtained from 91 chemicals, No. 8 is obtained from 73 chemicals using the same parameters with No. 7, and No. 9 is from 73 chemicals. The statistics of the models with 91 and 73 compounds are similar.

$$pEC_{50} = \sum_{i=1}^{52}((\alpha_i - \alpha_i^*)\exp(-\gamma^*\|x_i, x\|^2)) - 0.0868 \quad (3)$$

4 CONCLUSIONS

The predictive power of SVM models depends strongly on the set of parameters that control the model. In this paper we used the simplex optimization method to search for the optimum set of SVM parameters, namely the capacity parameter *C*, the insensitive loss parameter ε and the parameter γ that controls the shape of the RBF kernel. The leave–one–out cross–validation correlation coefficient q^2 is used as objective function for the simplex optimization of SVM parameters. SVM quantitative structure–activity relationships models were built for the toxicity of organic chemicals to *Chlorella vulgaris*. A series of QSAR models with one to three variables were obtained for the acute toxicity of 91 organic chemicals to *Chlorella vulgaris*. The SVM models with

parameters optimized with simplex have better statistics than the multi–linear regression QSAR equations. The results from the present investigation demonstrate that the simplex algorithm is an efficient approach in finding the best set of SVM parameters for QSAR models.

5 REFERENCES

[1] L. S. Wang, *Chemistry of Organic Pollution*. Higher Education Press, Beijing, 2004.
[2] T. W. Schultz, M. T. D. Cronin, T. I. Netzeva, The present status of QSAR in toxicology. *J Mol. Struct.* (*Theochem*) **2003**, *622*, 23–38.
[3] M. T. D. Cronin, T. I. Netzeva, J. C. Dearden, R. Edwards, A. D. P. Worgan, Assessment and Modeling of the Toxicity of Organic Chemicals to *Chlorella vulgaris*: Development of a Novel Database. *Chem. Res. Toxicol.* **2004**, *17*, 545–554.
[4] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer Verlag: New York, 1995.
[5] V. N. Vapnik, *Statistical Learning Theory*. Wiley: New York, 1998.
[6] 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.
[7] 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.
[8] 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.
[9] M. P. S. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. Sugnet, J. Manuel Ares, D. Haussler *Support Vector Machine Classification of Microarray Gene Expression Data*; UCSC–CRL–99–09; University of California: Santa Cruz, 9, 1999.
[10] P. Lind, T. Maltseva, Support Vector Machines for the Estimation of Aqueous Solubility. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1855–1859.
[11] O. Sadik, J. Walker H. Land, A. K. Wanekaya, M. Uematsu, M. J. Embrechts, L. Wong, D. Leibensperger, 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.
[12] C. X. Xue, R. S. Zhang, H. X. Liu, M. C. Liu, Z. D. Hu, B. T. Fan, Support Vector Machines–Based Quantitative Structure–Property Relationship for the Prediction of Heat Capacity. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1267–1274.
[13] 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.
[14] K. R. Muller, G. Ratsch, S. Sonnenburg, S. Mika, M. Grimm, N. Heinrich, Classifying 'Drug–likeness' with Kernel–based Learning Methods. *J. Chem. Inf. Model.* **2005**, *45*, 249–53.
[15] N. Cristianini, J. Shawe–Taylor, *An Introduction to Support Vector Machines and Other Kernel–based Learning Methods*. Cambridge University Press: Cambridge, 2000.
[16] S. R. Gunn. *Support Vector Machines for Classification and Regression*; Image Speech and Intelligent Systems Research Group, University of Southampton: 1998.
[17] C.–C. Chang, C.–J. Lin *LIBSVM – a library for support vector machines*, 2.6; 2001. http://www.csie.ntu.edu.tw/~cjlin/libsvm.
[18] O. Ivanciuc, Applications of support vector machines in chemistry. In: *Reviews in Computational Chemistry*, K. B. Lipkowitz and T. R. Cundari, (eds.), Wiley–VCH, Weinheim, 2007; Vol. 23, pp 291–400.