

Theoretical Descriptor for the Correlation of the Fish Embryo Toxicity test (FET) By QSAR Model of LC_{50}

Jining L, Deling F, Lei W, Linjun Z and Lili S*

Nanjing Institute of Environmental Sciences, Ministry of Environmental Protection, Nanjing, 210042, China

Abstract

Quantitative structure–toxicity relationships were developed for the prediction of the fish embryo toxicity test using the CODESSA treatment, based on the linear heuristic method (HM) and support vector machines (SVM). Each kind of compound was represented by several calculated structural descriptors, derived for a diverse set of 97 compounds using DFT- B3LYP/6-31+G (d) level of the theory. A six-parameter correlation was found for the 97 compounds. In the HM method, the value of square of the correlation coefficient r^2 is 0.8142, s_2 is 0.0380, in the SVM method, the value of the r^2 is 0.7105 and s_2 is 0.0604. The HM model may be used for prediction of toxicity, safety and risk assessment of chemicals to achieve better ecotoxicological management.

Keywords: Quantitative Structure Activity Relationship (QSAR); Support Vector Machine (SVM); Heuristic method; Fish Embryo Toxicity Test (FET)

Introduction

The fish acute toxicity is a mandatory component in the base set of data requirements for eco-toxicity testing but fish suffer distress and perhaps pain [1]. Animal alternative considerations have also been incorporated into new REACH regulations through strong advocacy for the reduction of testing with live animals. REACH, as a regulatory tool, was envisaged to revise the existing chemical policy and to harmonize chemical legislations in Europe and the major aim of REACH is to systematically evaluate the risk to human health and the environment of approximately 30,000 chemical substances [2]. The fish embryo toxicity (FET) testan alternative approach to classical acute fish toxicity testing with live fish, has been a mandatory component in routine whole effluent testing in Germany since 2005 and has already been standardized at the international level. A modified version has been submitted by the German Federal Environment Agency (UBA) as a draft guideline for an alternative to chemical testing with intact fish [1]. The fish embryo test is a surrogate for the OECD 203 acute fish toxicity test (or other guideline equivalent acute fish assays). Recently, the regulations of the People's Republic of China on Control over Safety of Hazard Chemicals revised and passed, therefore, FET is useful for the acute fish toxicity testing. In order to analyze the applicability of the FET also in chemical testing, a comparative re-evaluation was carried out for a total of 97 substances, and QSAR was developed to evaluate fish embryo toxicity data. QSAR to evaluate fish embryo toxicity data is described to be useful for future validation exercises.

The quantitative structure-property/activity relationship (QSPR/QSAR) method is based on the assumption that the variation of the behavior of the compounds, as expressed by many measured physicochemical properties, can be correlated with changes in molecular features of the compounds termed descriptors [3]. This method can be used for the prediction of the properties of new compounds. It can also be applied to identify and describe important structural features of the molecules that are relevant to variations in molecular properties. Computational models are useful because they rationalize a large number of experimental observations.

The QSAR model is widely used for the prediction of physicochemical properties and biological activities in chemical,

environmental, and pharmaceutical areas [4,5], a potential or useful technique to estimate the toxicity especially for the compounds whose toxicity are not easy to test. The advantage of this approach over the experimental method lies in the fact that the descriptors used can be calculated from the structures alone and are not dependent on any experimental properties. Once the structure of a compound is known, a limited set of molecular properties, said descriptors, which can be obtained observing the molecular structure or semi-empirical quantum-mechanical calculations [6]. Therefore, once a reliable model is established, the model can be applied to help in identifying and describe important structural features of the molecules that are relevant to variations in molecular properties.

In this work we proposed a QSAR model to evaluate fish embryo toxicity data. 97 structures are calculated by the CODESSA software package, developed by Katritzky et al. [7,8] based on descriptors constitutional, topological, geometrical, electrostatic and quantum chemical.

Materials and Method

In order to develop the QSAR, a database of LC_{50} , reported in terms of micromolar (μM) concentration for correlation purposes, reported LC_{50} values were converted to their molar units and subsequently to free energy related negative logarithmic state, i.e., $\log(1/LC_{50})$. These LC_{50} values have been obtained from experiments and literature [9]. The selected series consists of total 97 numbers of compounds, divided in to two data sets, one was training set having 83 compounds and other was test set with the remaining 14 compounds based on random selection criteria. The biological activities were expressed in terms of median lethal concentration (EC_{50}).

*Corresponding author: Lili S, Nanjing Institute of Environmental Sciences, Ministry of Environmental Protection, Nanjing, 210042, China, Tel: +86 25 8528 7074; Fax: +86 25 8547 6477; E-mail: sll@nies.org

Received January 13, 2014; Accepted March 29, 2014; Published April 07, 2014

Citation: Jining L, Deling F, Lei W, Linjun Z, Lili S (2014) Theoretical Descriptor for the Correlation of the Fish Embryo Toxicity test (FET) By QSAR Model of LC_{50} . Organic Chem Curr Res 3: 125. doi:10.4172/2161-0401.1000125

Copyright: © 2014 Jining L, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Calculation of the descriptors

The chemicals in the data set (Table 1) have been optimized and their

electronic, geometric, and energetic parameters have been calculated with Gaussian 03 package [11]. The structures were optimized and

NO.	Compounds	status	MW	Experimental (µg/L)	HM (µg/L)	SVM (µg/L)
1	2-Aminoethanol	Training	61.08	6.5662	4.5114	4.8806
2	4-Aminophenol	Training	109.13	2.6720	1.8662	2.2334
3	Benzoicacid	Training	122.12	4.3031	2.1385	2.3162
4	4-Bromindole	Training	195.95	3.6720	1.678	1.4798
5	5-Bromindole	Training	196.05	3.7250	1.7252	1.5358
6	6-Bromindole	Training	196.05	3.96426	2.3408	2.2578
7	2-Bromophenol	Training	173.01	4.6543	2.0823	2.2993
8	3-Bromophenol	Training	173.01	4.7829	3.2246	2.7912
9	4-Bromophenol	Training	173.01	4.6657	3.3539	2.9999
10	n-Butylamine	Training	73.14	4.5552	3.743	2.408
11	sec-Butylamine	Training	73.14	4.9784	2.0591	2.2457
12	Butyldiglycol	Training	162.23	6.1077	1.8154	2.3225
13	2-Chloroaniline	Training	127.6	4.4517	3.8754	2.8491
14	3-Chloroaniline	Training	127.6	4.3222	3.0148	2.7056
15	4-Chloroaniline	Training	127.6	4.3280	1.7793	1.5822
16	4-Chlorophenol	Training	128.56	4.5800	1.6495	2.0511
17	Cyclohexanol	Training	100.16	6.1405	1.9123	2.32
18	Cyclohexylamine	Training	99.18	4.8019	2.7298	2.5952
19	n-Decylamine	Training	157.3	3.4977	1.4561	2.0633
20	2,4-Dibromophenol	Training	251.9	3.9003	1.4761	2.0776
21	2,6-Dibromophenol	Training	251.9	4.6208	3.8754	2.8491
22	Dibutylamine	Training	129.25	4.6069	3.3737	2.9902
23	2,4-Dichloroaniline	Training	162.02	4.3324	5.1731	2.8854
24	3,4-Dichloroaniline	Training	162.02	3.3031	3.3448	3.0048
25	Dicyclohexylamine	Training	181.32	4.4939	2.2432	2.8373
26	Diethylamine	Training	73.14	4.9696	2.0318	2.6624
27	Diethyleneglycol	Training	106.14	7.7059	3.6017	1.9352
28	N,N-Diethylmethylamine	Training	87.16	4.8450	3.3448	3.0048
29	N,N-Diisopropylethylamine	Training	129.25	5.0193	2.9621	2.8025
30	Diisobutylamine	Training	129.3	4.6738	2.3816	2.6403
31	Diisopropylamine	Training	101.2	4.9613	3.7035	3.1539
32	N,N-Dimethylamine	Training	45.09	5.5985	5.4871	4.7149
33	N,N-Dimethylanilin	Training	121.2	4.7342	1.1533	1.7563
34	N,N-Dimethylbutylamine	Training	101.19	4.7075	0.7109	1.3753
35	N,N-Dimethylcyclohexylamine	Training	127.23	4.7247	1.5988	1.7464
36	N,N-Dimethylethylamine	Training	73.14	4.9183	5.1469	5.2986
37	N,N-Dimethylformamide	Training	73.09	6.9746	4.8383	4.789
38	4,6-Dinitro-o-cresol	Training	198.14	2.5910	4.5685	3.9195
39	2,4-Dinitrophenol	Training	184.11	2.9542	2.4241	2.8188
40	Dipentylamine	Training	157.3	4.6313	2.7028	2.7623
41	Dipropylamine	Training	101.19	4.4936	2.4946	2.2544
42	Ethanol	Training	46.1	7.0625	3.2158	4.2721
43	Ethylacetate	Training	88.106	6.2692	2.9198	2.9311
44	Ethylenediamine	Training	60.1	5.5984	3.5851	4.5713
45	1-Ethylpiperidine	Training	113.2	4.8531	6.14	5.7386
46	2-Ethylpiperidine	Training	113.2	4.9729	3.7661	2.8756
47	n-Heptylamine	Training	115.22	4.4541	4.5401	4.1437
48	n-Hexylamine	Training	101.19	4.6261	5.746	5.3312
49	Hydroquinone	Training	110.11	3.8976	4.0228	4.1965
50	Hydroxyurea	Training	76.05	6.2530	5.1664	5.3695
51	Isobutylamine	Training	73.1	4.966694	4.201	4.4596
52	Isopentylamine	Training	87.16	4.7715	2.315	2.1711
53	Isopropylamine	Training	59.11	6.4425	2.6561	2.616
54	Methanol	Training	32.04	7.3443	2.668	2.6086
55	Methoxyaceticacid	Training	90.08	4.7298	4.6482	3.7387
56	2-Methoxyethanol	Training	76.09	7.3125	1.4622	1.8345

57	1-Methoxy-2-propanol	Training	90.12	7.1434	1.9533	2.2345
58	3-Methyl-1-butanol	Training	88.17	6.0097	1.07	2.513
59	N-Methylamine	Training	31.1	5.8525	2.0118	1.8028
60	N-Methylanilin	Training	107.2	4.5339	3.1082	2.2898
61	N-Methylformamide	Training	59.07	7.0408	1.2202	2.3192
62	1-Methylpiperidine	Training	99.18	4.8346	2.956	2.9168
63	2-Methylpiperidine	Training	99.18	5.0100	1.5135	1.2284
64	4-Methylpiperidine	Training	99.18	4.9681	3.0818	2.47
65	Morpholine	Training	87.12	5.7790	4.7142	5.0932
66	2-Nitroaniline	Training	138.1	4.3324	2.3356	2.3419
67	4-Nitrobenzoicacid	Training	167.12	4.4756	2.0638	2.3089
68	4-Nitrophenol	Training	139.11	4.7573	1.2928	1.818
69	n-Nonylamine	Training	143.27	4.0592	4.6492	2.7816
70	1-Octanol	Training	130.23	4.18752	4.3523	5.4814
71	n-Octylamine	Training	129.25	4.4058	0.5177	0.9196
72	Pentachlorophenol	Training	266.34	2.7403	2.8802	2.3573
73	n-Pentylamine	Training	87.16	4.4893	1.7798	1.2701
74	4-tert-Pentylphenol	Training	164.24	3.5440	2.5337	1.7304
75	Phenol	Training	94.11	5.3222	2.7298	2.5952
76	Piperidine	Training	85.2	5.0433	3.9809	2.4011
77	2-Propanol	Training	60.1	6.9721	5.3213	5.0889
78	n-Propylamine	Training	59.11	4.8984	3.5419	2.5303
79	Salicylicacid	Training	138.12	4.3577	4.8308	5.2073
80	Tetrachloroethylene	Training	165.83	4.4281	2.8316	2.9567
81	2,4,6-Tribromophenol	Training	330.8	3.6454	2.0737	2.352
82	Triethyleneglycol	Training	150.2	7.7315	0.9536	2.5847
83	Urea	Training	60.06	7.3598	2.354	2.2464
84	2,5-Hexanedion	Test	114.14	6.6691	6.0191	5.6191
85	Acetone	Test	58.08	4.0211	3.9211	3.0111
86	Acrolein	Test	56	2.5682	1.9682	1.5682
87	Chloroacetaldehyde	Test	78	3.5251	3.9551	2.9351
88	Diazinon	Test	304.34	4.4961	4.1161	4.0161
89	Diethyleneglycoldimethylether	Test	134.18	7.0436	6.5236	7.9436
90	Dimethylsulfoxide	Test	78.13	7.4638	6.4638	6.1238
91	D-Mannitol	Test	182.17	4.8908	5.0108	3.3508
92	Formamide	Test	45.041	6.9610	7.0610	2.9210
93	p-tert-Butylphenol	Test	150.22	3.2380	2.2380	1.2380
94	Quinone	Test	108.09	2.6739	1.9739	0.6739
95	Saccharinsodiumsalthydrate	Test	241.2	7.3211	6.2111	8.11211
96	Tridecylmono-octylether	Test	244.4	3.1568	2.1268	1.1128
97	Valproicacid	Test	144.21	4.3051	3.9151	2.3751

Table 1: The experimental values of all compounds and the predicted LC₅₀ by HM and SVM.

their vibrational frequencies calculated at the DFT level relying on the B3LYP functional. A comparison of all the methods, namely, AM1, MP3, HF, and DFT/B3LYP, indicates that the DFT/B3LYP method is more reliable than others and has a high predictive power [12]. After optimization, the CODESSA program was used to calculate five types of molecular descriptors constitutional (number of various types of atoms and bonds, number of rings, molecular weight, etc.), topological (Wiener index, Randic indices, Kier -Hall shape indices, etc.), geometrical (moments of inertia, molecular volume, molecular surface area, etc.), electrostatic (minimum and maximum partial charges, polarity parameter, charged partial surface area descriptors, etc.), and quantum chemical (reactivity indices, dipole moment, HOMO and LUMO energies, etc.). In the framework of the CODESSA program, the statistical structure-property correlation techniques can be used for the analysis of user-submitted experimental data in combination with the calculated molecular descriptors.

Development of linear model by HM [13]

The heuristic procedure was used to obtain quantitative structure property relationship. Once the molecular descriptors are generated, the HM in CODESSA is used to pre-select the descriptors and build the near model. The advantages of the HM are the high speed, usually produces correlations 2–5 times faster than other methods with comparable quality and no software restrictions on the size of the data set [14]. The HM can either quickly give a good estimation about what quality of correlation to expect from the data, or derive several best regression models. HM of the descriptors selection proceeds with a pre-selection of the descriptors to eliminate: (1) those descriptors that are not available for each structure; (2) descriptors having a small variation in magnitude for all structures; (3) descriptors that provide a F-test value below 1.0 on the one-parameter correlation and (4) the descriptors whose t-values are less than the user-specified value, etc. New descriptors were added one by one until the suitable number

descriptors without losing any important information in the model is achieved [15]. The final result is a list of the 10 best models according to the values of the F-test and correlation coefficient. The statistical quality of the generated models was gauged by the F-test (F), squared correlation co-efficient (R^2), square of cross-validate coefficient regression (R_{cv}^2), t-test and the standard deviation (s^2). Fischer's value (F), which represents F-ratio between the variance of calculated and observed activity, and chance statistics assuring that the results are not merely based on chance correlations and t-test reflects significance of the parameter within the model. From the above processes, seven descriptors are selected from descriptors pool and the linear model is produced by the HM [16]. The experimental values of all compounds are summarized in Table 1.

Development of linear model by SVM

The support vector machine algorithm was developed by Vapnik [17]. Support vector machine (SVM), as a novel type of learning machine is gaining rapid popularity due to its remarkable generalization performance [18], has gained much interest in pattern recognition and function approximation applications recently. In bioinformatics, SVMs have been successfully used to solve classification and correlation problems, with the basic idea of SVM is to map the original data into a higher dimensional feature space by a kernel function and then to do classification in this space by constructing an optimal separating hyperplane. Compared with traditional regression and neural networks methods, SVMs have some advantages, including the absence of local minima, good generalization ability, simple implementation, few free parameters, dimensional independence and its ability to condense information contained in the training set [19]. The flexibility in classification and ability to approximate continuous function make SVMs very suitable for QSAR and QSPR studies. Because it is difficult to predict in advance which descriptors are most relevant to the problem at hand, however, SVM is well known to tolerate irrelevant features [20-22]. In some cases, the best average performance was achieved when all the features were given to SVM [23].

Results and Discussion

Results of HM

The heuristic method is not made a prior choice of descriptors, but it performs an analysis on several intrinsic molecular properties available from the Gaussian output. We performed correlations for a growing number of descriptors in the range from 1 to 10.

Figure 1 shows the plots of R^2 , cross-validated coefficient (R_{cv}^2) and the standard deviation (s^2) for the training set as a function of the number of descriptors models. R^2 and R_{cv}^2 are increased with the increasing number of descriptors. However, the values of s^2 decreased with the increasing number of descriptors [24,25]. When adding another descriptor did not improve significantly the statistics of a model, it was determined that the optimum subset size had been achieved. From Figure 1, it can be seen that six descriptors appear to be sufficient for a successful regression model. Then the corresponding descriptors are applied as inputs for the non-linear model. Best models were selected on the basis of their statistical significance. High r^2 and high F and low s^2 values in each model reflect their good predictive potential. Pair wise correlations between the descriptors used in the QSAR models obtained for series is summarized in Table 2. The molecular descriptors used in the selected QSAR models are defined in Table 3. The predict results are shown in Figure 2.

Results of SVM

SVM is a machine-learning approach that allows one to learn from experimental data within a given set, and build a computer model to make predictions on new data sets. The associated parameters are c, p and g. $-t$ kernel function is the RBF kernel function: $\exp(-\text{gamma} * |\text{uv}|^2)$, the calculated optimal correlation parameters are c:3.9192, g:168.9664, p: 0.9892. The predict results are shown in Figure 3.

Compare with the HM and SVM

In the HM method, the R^2 and s^2 are 0.8142 and 0.0380 in training set, the R^2 and s^2 are 0.9238 and 1.8503 in test set. At the same time, in the SVM method, the R^2 and s^2 are 0.7105 and 0.0604 in training set,

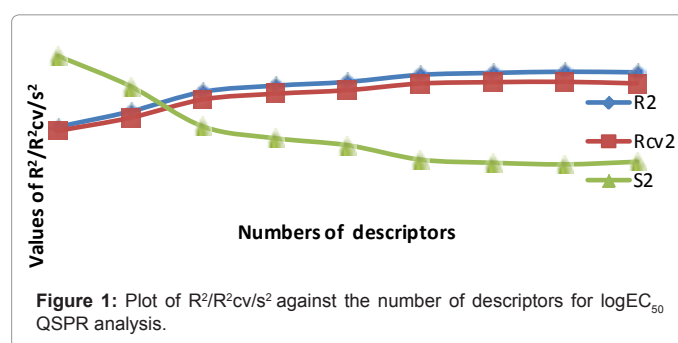


Figure 1: Plot of $R^2/R^2_{cv}/s^2$ against the number of descriptors for $\log EC_{50}$ QSPR analysis.

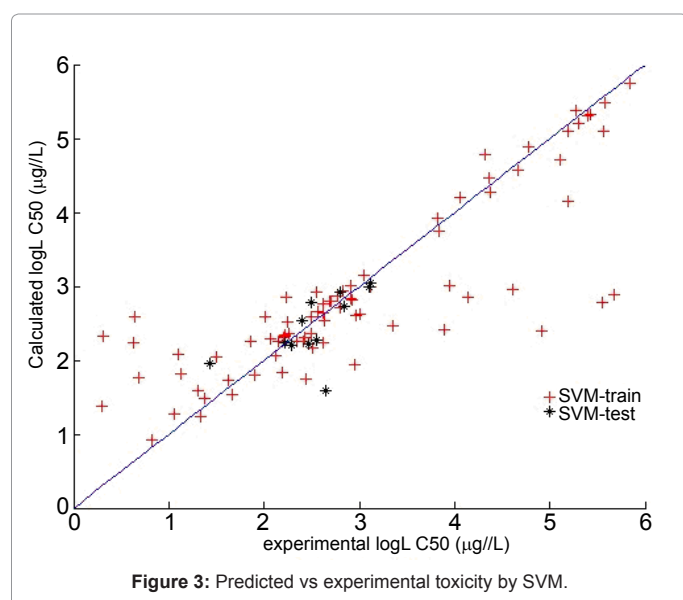
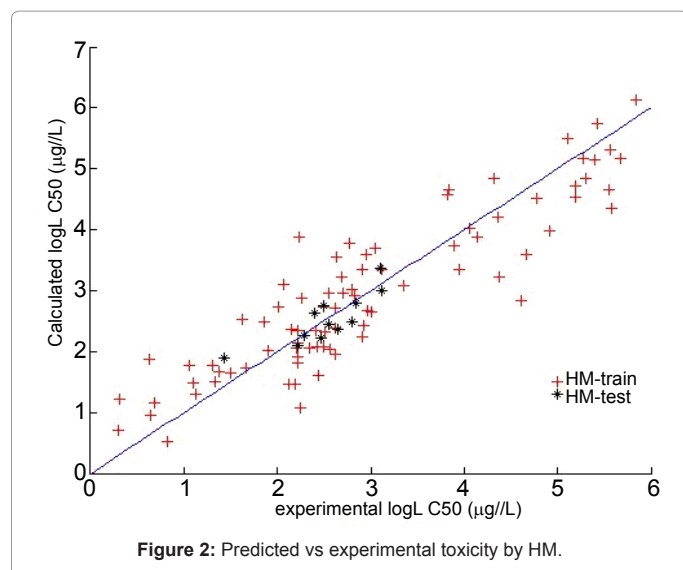
	A	B	C	D	E	F
A	1.0000	-0.0725	0.2135	0.3806	0.4239	0.2161
B	-0.0725	1.0000	-0.3540	-0.1499	-0.2106	0.1555
C	0.2135	-0.3540	1.0000	0.2619	0.1297	-0.2949
D	0.3806	-0.1499	0.2619	1.0000	0.2398	0.2922
E	0.4239	-0.2106	0.1297	0.2398	1.0000	-0.2729
F	0.2161	0.1555	-0.2949	0.2922	-0.2729	1.0000

Table 2: Correlation matrix of the six descriptors used in this work.

R^2	R^2_{cv}	F	X+DX	t-Test	Descriptors
0.8142	0.7733	55.50	6.0388e+01 6.3153e+00	9.5622	Intercept(A)
			-1.3840e+01 1.5885e+00	-8.7129	Avg valency of a C atom(B)
			-1.2775e+01 1.7259e+00	-7.4017	Relative number of double bonds(C)
			-1.9631e-02 2.5005e-03	-7.8508	Molecular volume(D)
			4.9342e-01 1.0389e-01	4.7492	HOMO-1 energy(E)
			-5.9297e-01 1.2477e-01	-4.7492	HOMO energy(F)
			-2.2411e+00 6.0338e-01	-3.7143	Average structural information content(order 1)(G)

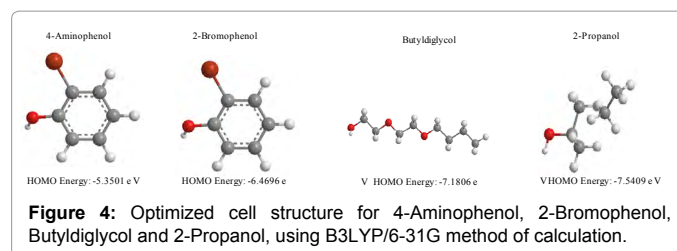
1, Avg valency of a C atom;2, Relative number of double bonds;3, PPSA-3 Molecular volume;4, HOMO-1 energy;5, HOMO energy;6, Average Structural Information content (order 1);
The fitted regression model is: $\log EC_{50} = (6.0388e+01 \ 6.3153e+00)*A + (-1.3840e+01 \ 1.5885e+00)*B + (-1.2775e+01 \ 1.7259e+00)*C + (-1.9631e-02 \ 2.5005e-03)*D + (4.9342e-01 \ 1.0389e-01)*E + (-5.9295e-01 \ 1.2477e-01)*F + (-2.2411e+00 \ 6.0338e-01)*G$

Table 3: Correlations of toxicity by the heuristic method.



the R^2 and s^2 are 0.7527 and 2.4351 in test set. The six descriptors are the most important physical chemistry properties for the construction of QSAR model for FET and the prediction of LC₅₀ with satisfied results. The predicted results of either training set or test set of the HM are better than those of SVM, both in the R^2 and s^2 . Therefore, the HM has a good generalized performance (Table 4) as a non-linear method.

Unsaturated double bond, based on positive and negative regression coefficients, can be judged according to their contribution to the toxic compounds. The coefficients of number of double bonds is positive, the LC₅₀ increase with the double bond value increases. As for the electron donation of the elide, it is noted that the HOMO energy can be used a new method for showing the basicity of elides. The descriptors that matter most is on the Highest Occupied Molecular Orbital (HOMO) calculated from quantum chemical calculation, which describes the effect of HOMO energies on the toxicity. The results also show that the contribution from the HOMO-1 becomes less important than HOMO. Molecular volume is important to toxicity, the information contained



	Training		Test	
	R^2	s^2	R^2	s^2
HM	0.8142	0.0380	0.9238	1.8503
SVM	0.7105	0.0604	0.7527	2.4351

Table 4: Performance comparison between HM and SVM.

in high-resolution structural determination of macromolecules. The HOMO-1, which is lower in energy than the HOMO, features a large contribution for toxicity.

The HOMO of 4-Aminophenol, 2-Bromophenol, Butyldiglycol and 2-Propanol, were calculated by B3LYP/6-31G method. 2-Propanol was the highest toxicity with the highest HOMO, as well as 4-Aminophenol was the lowest toxicity with the lowest HOMO (Figure 4). The structure of 4-Aminophenol was similar to 2-Bromophenol. The toxicity 2-Bromophenol, with higher HOMO, was higher than 4-Aminophenol. The Highest Occupied Molecular Orbital (HOMO) describes the effect of HOMO energies on the toxicity.

Conclusions

A QSAR study for fish embryo toxicity test (FET) was performed using HM and SVM, based on electrostatic, CPSA, constitutional, and quantum-chemical descriptors and satisfactory results were obtained. Through analyzing the obtained results, the present study gives rise to QSAR with good statistical significance and predictive capacity. The 6-parameter model can be considered as the best linear model the correlations are satisfying. Additionally, HM produced good model with vigorous predictive ability. Furthermore, the proposed approach can also be extended in other QSAR investigations. Fish embryo testing of chemicals has matured to the point that international standardization, method validation, and broadening of chemical coverage is rapidly occurring. Fish embryo tests offer a reasonable alternative to increased use of fish in the future.

Acknowledgment

Financial support from 2013 Commonweal and Environmental Protection Project of Ministry of Environmental Protection of the People's Republic of China (No.: 2013467028) and the National High Technology Research and Development Program of China (863 Program, No.: 2013AA06A308).

References

1. EC (European Commission), 2007. Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorization and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending Directive 1999/45/EC and repealing Council Regulation (EEC) No 793/93 and Commission Regulation (EC) No 1488/94 as well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC. *Off. J. Eur. Union* L 136, 3–280.
2. Braunbeck T, Boettcher M, Hollert H, Kosmehl T, Lammer E, et al. (2005) Towards an alternative for the acute fish LC(50) test in chemical assessment: the fish embryo toxicity test goes multi-species -- an update. *ALTEX* 22: 87-102.

3. Yao XJ, Liu MC, Zhang XY (2002) Radial basis function network-based quantitative structure–property relationship for the prediction of Henry's law constant. *Anal aChim Acta* 462: 101–117.
4. Katritzky AR, Fara DC, Petrukhin RO, Tatham DB, Maran U, et al. (2002) The present utility and future potential for medicinal chemistry of QSAR/QSPR with whole molecule descriptors. *Curr Top Med Chem* 2: 1333-1356.
5. Katritzky AR, Maran U, Lobanov VS, Karelson M (2000) Structurally Diverse Quantitative Structure-Property Relationship Correlations of Technologically Relevant Physical Properties. *J Chem Inf Comput Sci* 4:1-18.
6. Arning J, Stolte S, Baschen A, Stock F, Pitner WB, et al. (2008) Qualitative and quantitative structure activity relationships for the inhibitory effects of cationic head groups, functionalised side chains and anions of ionic liquids on acetylcholinesterase[J]. *Green Chem*10: 47–58.
7. Katritzky AR, Lobanov VS, Karelson M (1995) CODESSA: Training Manual, University of Florida, Gainesville, FL.
8. Katritzky AR, Lobanov VS, Karelson M (1994) CODESSA: Reference Manual, University of Florida, Gainesville, FL.
9. Lammer E, Carr GJ, Wendler K, Rawlings JM, Belanger SE, et al. (2009) Is the fish embryo toxicity test (FET) with the zebrafish (*Danio rerio*) a potential alternative for the fish acute toxicity test? *CompBiochemPhysiol C Toxicol Pharmacol* 149: 196-209.
10. Kammann U, Vobach M, Wosniok W (2006) Toxic effects of brominated indoles and phenols on zebrafish embryos. *Arch Environ Contam Toxicol* 51: 97-102.
11. Frisch MJ, Trucks GW, Schlegel HB (2003) Gaussian 03, revision A.1 [K]. Gaussian, Inc: Pittsburgh, PA.
12. Eroglu E, Palaz S, Oltulu O, Turkmen H, Ozaydin C (2007) Comparative QSTR Study Using Semi-Empirical and First Principle Methods Based Descriptors for Acute Toxicity of Diverse Organic Compounds to the Fathead Minnow[J]. *Int J Mol Sci* 8:1265-1283.
13. Katritzky AR, Lobanov VS, Karelson M (1994) CODESSA Reference Manual (version 2.0), Reference Manual, Version 2.0.
14. Katritzky AR, Lobanov VS, Karelson M CODESSA: Reference Manual, University of Florida, Gainesville, FL.
15. Devillers J, Chambon P (1986) Acute toxicity and QSAR of chlorophenols on *Daphnia magna*. *Bull Environ Contam Toxicol* 37: 599-605.
16. Katritzky AR, Petrukhin R, Jain R, Karelson M (2001) QSPR analysis of flash points. *J Chem Inf Comput Sci* 41: 1521-1530.
17. Vapnik VN (1998) *Statistical Learning Theory*. New York: John Wiley.
18. Liu HX, Zhang RS, Yao XJ, Liu MC, Hu ZD, et al. (2004) Prediction of the isoelectric point of an amino acid based on GA-PLS and SVMs. *J Chem Inf Comput Sci* 44: 161-167.
19. Chen C, Zhou X, Tian Y, Zou X, Cai P (2006) Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network. *Anal Biochem* 357: 116-121.
20. Yang Y, Pederson JO (1995) A comparative study on feature selection in text categorization. In Fisher DH (ed): *Proceedings of the ICML-95: 14th International Conference on Machine Learning*. San Francisco: Morgan Kaufmann: 412-420.
21. Rogati M, Yang (2002) High-performing feature selection for text classification. In *Proceedings of the 11th International Conference on Information and Knowledge Management*. New York: ACM Press: 659-661.
22. Brank J, Grobelnik M, Milic-Frayling N, Mladenic D (2002) Interaction of feature selection methods and linear classification models. *Proceedings of the ICML-02 Workshop on Text Learning*, Sydney, Australia.
23. Taira H, Haruno M (1999) Feature selection in SVM text categorization. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence and the Eleventh Innovative Applications of Artificial Intelligence Conference*. Menlo Park, CA: American Association for Artificial Intelligence: 480-486.
24. Katritzky AR, Tatham DB, Maran U (2001) Theoretical descriptors for the correlation of aquatic toxicity of environmental pollutants by quantitative structure-toxicity relationships. *J Chem Inf Comput Sci* 41: 1162-1176.
25. Bruzzone S, Chiappe C, Focardi SE (2011) Theoretical descriptor for the correlation of aquatic toxicity of ionic liquids by quantitative structure–toxicity relationships, *Chem Eng J* 175:17–23.