

Study on aromatic hydrocarbons toxicity to *Chlorella vulgaris* based on QSAR model

Shenglong Yang & Cuihua Wang*

East China Sea Fisheries Research Institute, Chinese Academy of Fisheries Sciences, Shanghai, 200090, China

*[E-mail: swatipriyanka1a@gmail.com]

Received 24 September 2014; revised 11 November 2016

Predicting ability based on the quantitative structure–activity relationships (QSAR) model of unknown aromatic hydrocarbons toxicity is one of the tasks of security precaution. To establish the QSAR model between the physical and chemical properties of aromatic hydrocarbons and the inhibited activity of *Chlorella vulgaris* (C. Vulgaris), the optimized geometries, based on the 96 hr-EC50 of 25 aromatic hydrocarbons with C. Vulgaris were carried out at the B3LYP/6-311G** level by density functional theory (DFT) calculation. With matlab2010(a) software, multiply linear regression (MLR) and three types principal components regression (PCR) methods were used to develop the QSAR model. Three methods were introduced to select the PCs, namely k-values (K), eigenvalue ranking (EV) and correlation ranking (CR) procedures. The different results then were compared. After eliminated one of the collinear descriptor and constants, 6, 1 and 4 PCs were selected by K, EV, CR for PCR method respectively. LOO cross-validated coefficient (R^2_{CV}) of training set to MLR, K-PCR, EVPCR, CRPCR were 0.673, 0.817, 0.874, 0.907, the R^2 of prediction were 0.569, 0.678, 0.519, 0.79, respectively.

[Keywords: PCR; aromatic hydrocarbons; *Chlorella vulgaris*; Toxicity]

Introduction

The frequent occurrence of oil spills, such as Gulf of Mexico oil spill, tanker collided in Dalian bay and Bohai sea oil spill accidents, along with the rapid development of industry had attracted more and more attention. Aromatic hydrocarbons, one of the major pollutants in petroleum, poured into the ocean with disastrous effect on ocean ecosystem^{1, 2}. Due to their carcinogenic, mutagenic, strongly hepatotoxic and extremely resistant to enzymatic degradation for us³⁻⁵, more and more researchers had paid more attention to the damage of aromatic hydrocarbons.

Marine algae, one kind of the main foods for different marine organisms in food chain, such as fish and shellfish through ingestion of marine copepods, play an important role in keeping ocean ecosystem balance. For its high sensitivity to contaminants and the short life cycle, they are regarded as one of the ideal index biological indexes for polluted water^{6, 7}. Toxicity of aromatic hydrocarbons on aquatic organisms have been studied⁸, but those are both time

consumption and expensive. Quantitative structure-activity relationships (QSAR), Combining the chemical substructure and biological toxicity, was widely used in toxicology to explain toxigenicity mechanism and prediction⁹⁻¹³.

In QSAR studies, a regression model of the form $y = X \cdot b + e$ be used to describe a set of predictor variables (X) with a predicted variable (y) by means of a regression vector (b). However, the collinearity, which often existed between independent variables, creates a severe problem in certain types of mathematical problem such as matrix inversion¹⁴. A better predictive model can be obtained by orthogonalization of the variables using principal component analysis (PCA) and then the obtained consequent method is called principal component regression (PCR)¹⁵⁻¹⁷. In order to reduce the dimensionality of the independent variable space, a limited number of principal components (PCs) are used and therefore a major question of PCA is how

many and which PCs constitute a good subset of predictive purpose?

The factor with high eigenvalue is considered as the most significant one, so one simple way is choose by condition number k using the means of the PCs with k value less than 100¹⁸. Another common way is ranking the factors in the order of decreasing eigenvalue (Eigenvalue ranking PCR, EVPCR)¹⁹. The factors are introduced into the PCR model based on their decreasing eigenvalue until no further improvement of the PCR model is obtained. But the eigenvalue magnitude is not necessarily a measure of its significance for the calibration²⁰. The correlation ranking principal component regression (CRPCR) method that the factors are ranked by their correlation coefficient with the dependent variable, and the selection procedure same as EVPCR. Better results are often obtained by this model²¹.

In this study, a single-celled green algae, *Chlorella vulgaris* (*C. Vulgaris*), was selected to test the toxicity of 25 aromatic hydrocarbons and to obtain the median effective inhibition concentration (EC_{50}) at the 96 hr acute toxicity. Based on these toxicology data and the geometries of all 25 compounds by DFT calculation, the relationship between the molecule structural features and bioactive toxicity (pEC_{50}) were discussed by principal components regression (PCR) model. Three methods were introduced to select the PCs and compared to multiply linear regression (MLR) model.

Materials and Methods

C. vulgaris, obtained from institute of oceanology, Chinese Academy of Sciences, was cultivated in Erlenmeyer flasks with f/2 culture medium without Si (pH 8.0±0.1, salinity 25.0±0.5), the light was provided by a cool white fluorescent lamp at the density of 2500-3500lx (a light/dark (L/D) 12h/12h cycles), the ambient temperature was 24±1°C²².

For preliminary experiments, the grades for 5-7 poisonous groups were designed by equal space between logarithm in a 96 hr growth test. *C.vulgaris* cells (10^5 cell/mL) at the exponential growth stage were put into different concentrations of aromatic hydrocarbons solution dissolved by acetone (the concentration of acetone was under 0.1 mg/L). The blank sample without aromatic hydrocarbons and control sample with acetone but without aromatic hydrocarbons were also exposed to the same conditions as the test samples. All experiments were done with *C. vulgaris* cells in the exponential growth stage unless stated. The cell density of *C.vulgaris* was

recorded at 24 hr, 48 hr, 72 hr and 96 hr and the growth rate “ r ” of *C. vulgaris* is calculated by the following formula²³

$$r = (\ln N_t - \ln N_0) / (t - t_0) \quad (1)$$

Where N_0 is the initial cell density, N_t is the cell density after the cultivated for t hr, t_0 is the initial time and t is the terminal time.

The growth inhibition rate “ η ” of *C. vulgaris* is calculated by the following formula:

$$\eta = (r_{t_1} - r_{t_2}) / r_{t_1} \quad (2)$$

Where r_{t_1} is the growth rate of the blank sample and r_{t_2} is the growth rate of the test sample.

All the experiments were performed in triplicate and the average results were presented in the EC_{50} (mg/L). The EC_{50} and its 95% confidence intervals are calculated with the probit method using the software Origin pro 8.0. The bioactive toxicity used in the study is expressed as:

$$pEC_{50} = -\log EC_{50} \quad (3)$$

The bioactive toxicity of aromatic hydrocarbons is increased with the value of pEC_{50} .

The experimental and predicted bioactive toxicities of 25 aromatic hydrocarbons used in the study were listed in Table 1. The training set, randomly selected was used to develop QSAR model and the test set was used to evaluate the model's predictive ability.

The geometries of all 25 compounds, carried out at the B3LYP/6-311** level with Gaussian 03 program by DFT calculation²⁴ are performed to ensure they are minimal at potential energy surface. The calculated values of the thermodynamic parameters of all 25 compounds using DFT calculation were presented in table 2.

Similarity and diversity sampling are two fundamental research themes in chemical database analysis^{25,26}. A diverse subset of “representative” compounds is defined to scan only a subset of the huge database each time. In this study, diversity analysis was performed for the dataset to make sure the structures of the training cases can represent those of the whole ones.

The perfect multi-collinearity is not strictly correlated between the each independent variables and violated the independent, so the perfect multi-collinearity was tested by condition number (k) and correlation coefficient²⁷.

In order to remove the collinearity and decrease the redundancy existed between the descriptors, the

correlation between the descriptors with pEC_{50} were detected. Among the collinear descriptors, the one with the lowest correlation with pEC_{50} was deleted. Among the collinear descriptors, the one with the lowest correlation with pEC_{50} was deleted. The descriptors were constant or near constant were also removed. A new data matrix(D) were gathered with the remaining descriptors. In order to examine the generalization of the construction model, the original data set was divided into two set. One of them was training with 20 sets of data (80%), and the other was prediction with 5 sets of data.

Before the PCA performing or any other modeling, the descriptors were normalization by matlab 2010(a) function „*ascore*“. The PCA was also

carried out by matlab2010(a) function „*princomp*“^[28]. The first 30PCs were found to explain more than 96% of variances in the original descriptors data matrix. The extracted PCs were taken as the predictor variable for PCR models.

The MLR model was employed in this paper with different numbers input variables. One was obtained by taking all the original 14 independent variables to the toxicity variable pEC_{50} ²⁸, and the other was developed with the new data matrix (D). The determination coefficient (R^2), F-test(F), mean square error (MSE) and mean absolute error (MAE) were calculated to evaluate the goodness of fit. The stability of the model was tested by the LOO cross-validated coefficient (R_{cv}^2)^{28,29}.

Table 1 – The bioactive toxicities of 25 aromatic hydrocarbons on the *Chlorella vulgaris*

No.	CAS Name	Obs	MLR		PCR		EVPCR		CRPCR	
			Pre	Res	Pre	Res	Pre	Res	Pre	Res
Training set										
22	phenol	-2.399	-2.317	-0.081	-2.302	-0.097	-2.287	-0.112	-2.353	-0.045
15	o-chloroaniline	-2.242	-2.072	-0.169	-2.069	-0.173	-1.766	-0.476	-2.034	-0.208
17	p-Chloroaniline	-2.229	-2.135	-0.095	-2.180	-0.050	-1.858	-0.372	-2.133	-0.096
16	m-Chloroaniline	-2.202	-2.172	-0.030	-2.184	-0.018	-1.850	-0.352	-2.146	-0.056
11	p-nitroaniline	-1.894	-1.980	0.086	-1.993	0.098	-2.168	0.274	-1.918	0.024
14	m-nitroanilinev/3-Nitroaniline	-1.782	-1.612	-0.170	-1.641	-0.140	-1.981	0.200	-1.575	-0.206
12	nitrobenzene	-1.665	-1.240	-0.425	-1.213	-0.452	-1.655	-0.010	-1.335	-0.330
23	nitrophenol	-1.627	-1.791	0.164	-1.734	0.108	-2.049	0.423	-1.704	0.078
7	2,4-dinitrotoluene	-1.605	-1.374	-0.231	-1.408	-0.197	-0.940	-0.665	-1.387	-0.218
25	p-nitrophenol	-1.570	-2.075	0.505	-2.064	0.494	-2.158	0.588	-2.033	0.463
18	2-Chlorotoluene	-1.478	-1.429	-0.049	-1.377	-0.101	-1.132	-0.346	-1.454	-0.024
19	3-Nitrochlorobenzene	-1.202	-1.011	-0.191	-1.019	-0.183	-1.062	-0.140	-1.071	-0.131
10	o-dinitrobenzene	-0.940	-1.073	0.133	-1.063	0.123	-1.215	0.275	-1.106	0.167
9	2,4-Dichloronitrobenzene	-0.681	-0.903	0.222	-0.913	0.232	-0.473	-0.208	-0.912	0.231
3	fluorene	-0.193	-0.205	0.012	-0.192	-0.001	-0.277	0.084	-0.231	0.038
6	benzo(a)anthracene	0.658	1.011	-0.354	1.030	-0.372	1.017	-0.360	1.083	-0.425
20	anthracene	0.775	0.768	0.007	0.667	0.108	0.282	0.492	0.635	0.139
2	Pyrene	0.783	0.298	0.485	0.361	0.421	0.405	0.378	0.365	0.417
5	phenanthrene	0.845	0.426	0.419	0.359	0.486	0.273	0.571	0.294	0.551
21	benzo(a)pyrene	1.046	1.282	-0.236	1.331	-0.285	1.291	-0.245	1.413	-0.367
Test set										
13	nitroaniline	-1.815	-1.994	0.180	-2.039	0.224	-2.216	0.401	-1.967	0.152
24	m-nitrophenol	-1.418	-1.710	0.292	-1.662	0.244	-1.908	0.490	-1.633	0.215
8	2,4-nitrochlorobenzene	-1.053	-0.793	-0.261	-0.726	-0.328	-0.503	-0.551	-0.717	-0.337
1	naphthalene	-0.663	0.012	-0.675	-0.140	-0.523	-0.507	-0.156	-0.289	-0.374
4	acenaphthene	-0.233	-0.085	-0.148	-0.165	-0.068	-0.338	0.105	-0.221	-0.012

Three methods were introduced to select the PCs. One is selected by condition number k by the means of the PCs with k less than 100¹⁸. The factors are introduced into the PCR model based on their decreasing eigenvalue until no further improvement of the PCR model is obtained in the EVPCR procedure. The PRESS (the predicted residual sum of squares) and LOO cross-validated coefficient (R^2_{cv}) were calculated in each step to evaluate the performances of the model. The determination coefficient (R^2), mean square error (MSE) and mean absolute error (MAE) were also calculated. The final number of PCs was obtained by PRESSCV, PRESSP and R^2_{cv} . The CRPCR procedure was similar to the EVPCR method, the difference between the PCs were introduced into model based on their decreasing correlation with the pEC_{50} ¹⁹.

Results and Discussion

The averaged distance of one compound to the other compounds were computed. After that the distances were normalized into [0,1] by the matlab2010(a) function mapminmax. The Fig.1 and Table 2 were obtained to explore the diversity of the

molecules of all data set. The closer to 1 the distance is, the more diverse to each other the compound is. The Fig.1 and Table.2 showed that most of the diversity differences of the compounds followed in the interval [0,0.1], it's mean most of the diversity differences were not distinct. Only one compound's (9) diversity differences was bigger than 0.8. Fig.1 also suggested that the training data set could represent the whole compounds, test set not included extreme data. So the above selection could lead to develop a good model with high-accuracy.

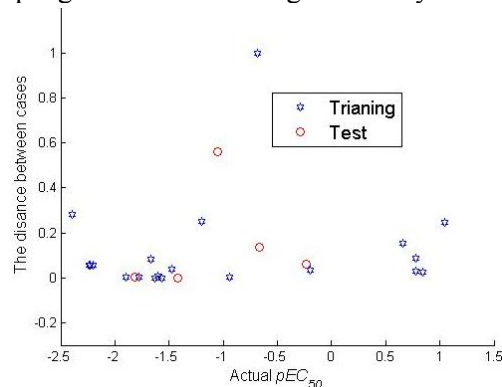


Fig.1—Scatter plot of the diversity distance VS. actual *Chlorella vulgaris* toxicity for the case in the whole data set

Table 2—the diversity analysis for the case in the whole data set

Range the distance

	0~0.1	0.1~0.2	0.2~0.3	0.3~0.4	0.4~0.5	0.5~0.6	0.6~0.7	0.7~0.8	0.8~0.9	0.9~1
Number of the case in the rang	18	2	3	0	0	1	0	0	0	1

The perfect multicollinearity between the independent variables was tested. From the correlation coefficient matrix (Table 4), there are many values bigger than 0.9, and the k values even bigger than 1000 (Table 3). It's could concluded that

there were perfect multicollinearity between the independent variables from the above results. The perfect multicollinearity could lead inaccuracy to MLR model.

Table 3—The values of the eigenvalue and condition number

eigenvalue	6.668	4.004	1.703	0.955	0.444	0.141	0.052
k	1	1.665	3.916	6.985	15.019	47.291	127.334
eigenvalue	0.019	0.013	0.002	0.0007	0	0	0
k	358.387	532.762	4610.03	9216.751	38390448	1.53E+11	1.15E+19

After eliminated one of the collinear descriptor and constants, 8 descriptors were remained. Using the new data matrix(D) as the input variables to develop MLR model. The linear equation as follows:

$$pEC_{50} = 0.2233 + 0.0004 \times E + 11.0059 \times E_{HOMO} - 9.5991 \times E_{LUMO} - 3.4254 \times qH^+ + 1.1453 \times q^- - 0.1265 \times C_v^0 + 0.0279 \times S^0 + 0.0032 \times V_i \quad (4)$$

$F = 27.3613$, $R^2 = 0.9522$, $R^2_{cv} = 0.6729$, $MSE = 0.0648$,

$N = 20$

Eq.(4) showed that the model gave R^2 of 0.9522 and MSE of 0.0648, but R^2_{cv} of 0.6729. It's suggested the training model with high accuracy. The model was used to predict the test set, the statistical parameters were given as follows: $R^2 = 0.5689$ and $MSE = 0.1325$ and the whole data set has $R^2 = 0.9316$ and $MSE = 0.0783$. The predicted values of pEC_{50} were

listed in the table. Fig.2 showed the predicted pEC50 values VS. observed pEC50 values for all of the 25 compounds studied. The prediction results showed that the MLR model could not use to describe the toxicity of these organic compounds. The results of the PCA on the new data matrix are given in the Table 5.

Taking training set to develop the PCR model. The table 5 showed the seventh $k > 100$, so took six principal component in this paper and the sum contribution of the six principal component bigger than 99.62%. The principal component linear equation as follows:

$$\text{pEC}_{50} = -0.9801 - 0.6206 \times \text{PC1} - 0.0025 \times \text{PC2} - 0.1655 \times \text{PC3} + 0.0729 \times \text{PC4} - 0.2547 \times \text{PC5} - 0.3337 \times \text{PC6} \quad (5)$$

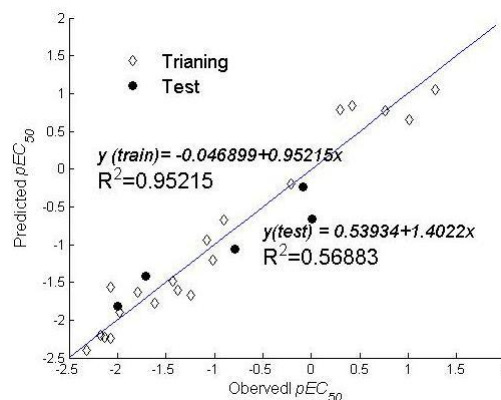


Fig.2—Scatter plot of the predicted VS. observed *Chlorella vulgaris* toxicity by MLR

Table 4—Correlation matrix of the independent variables

	E	E _{HOMO}	E _{LUMO}	qH ⁺	q ⁻	μ	a	ZPE	H ⁰	G ⁰	E _{th}	Cv ⁰	S ⁰	V _i
E	1	0.377	0.29	0.174	-0.134	-0.038	-0.083	0.195	1.000	1.000	0.178	-0.191	-0.435	-0.164
E _{HOMO}		1.000	0.744	-0.29	-0.2	-0.601	0.504	0.648	0.378	0.378	0.633	0.272	-0.125	0.409
E _{LUMO}			1.000	-0.15	-0.28	-0.529	0.05	0.243	0.290	0.290	0.223	-0.192	-0.509	-0.015
qH ⁺				1.000	-0.5	0.560	-0.582	-0.544	0.174	0.174	-0.542	-0.440	-0.301	-0.57
q ⁻					1.00	-0.469	0.516	0.468	-0.134	-0.134	0.476	0.547	0.574	0.579
μ						1.000	-0.625	-0.682	-0.039	-0.039	-0.674	-0.413	-0.126	-0.597
a							1.000	0.942	-0.083	-0.083	0.949	0.945	0.740	0.987
ZPE								1.000	0.195	0.195	1.000	0.850	0.561	0.919
H ⁰									1.000	1.000	0.178	-0.190	-0.435	-0.164
G ⁰										1.000	0.178	-0.19	-0.435	-0.164
E _{th}											1.000	0.863	0.582	0.928
Cv ⁰												1.000	0.907	0.971
S ⁰													1.000	0.810
V _i														1.000

Table 5—The results of PCA

	Eigenvalue	% of variance explained	Cumulative percent of variances	k value	Correlation coefficient
PC1	3.323	41.538	41.538	1.000	-0.948
PC2	2.485	31.063	72.602	1.337	-0.003
PC3	0.944	11.797	84.399	3.521	-0.135
PC4	0.608	7.604	92.003	5.462	0.048
PC5	0.476	5.952	97.955	6.978	-0.147
PC6	0.133	1.661	99.616	25.014	-0.102
PC7	0.026	0.329	99.945	126.401	0.033
PC8	0.004	0.055	100.000	750.526	-0.011

Taking the Eq(5) to predict the test set has $R^2=0.678$ and $\text{MSE}=0.099$. The whole data set gave R^2 of 0.8781 and MSE of 0.0729. Fig.3 showed that

the predicted pEC50 values VS. observed pEC50 values for all of the 25 compounds studied. The statistic parameters suggested that the PCR model

also can well describe the toxicity of all organic compounds easily and directly.

The results of EVPCR are showed in table 6. The coefficients of determination for the cross validation, training data set and prediction data set are represented by R^2_{cv} , R^2_t and R^2_p respectively. F is F-test value. MAE_p denote mean absolute error. In addition, the $PRESS_{CV}$ and $PRESS_T$ are all

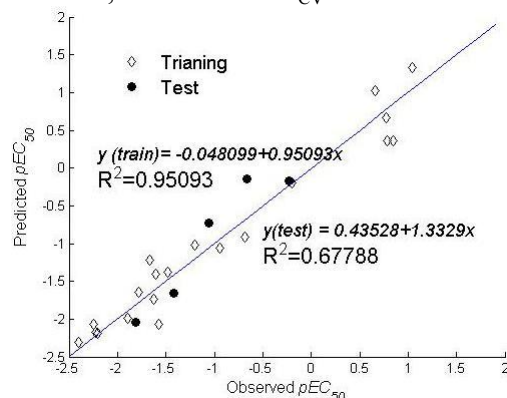


Fig.3—Scatter plot of the predictd VS. observed *Chlorella vulgaris* toxicity by PCR

include in Table. The plot of $PRESS_{CV}$ and $PRESS_P$ against the number of PCs are shown in Fig.4. The results showed that the first PC1 could be used to develop the PCR model. From the first row of Table 6, the R^2 of training, prediction data set and LOO

cross validation are 0.898, 0.519 and 0.874, respectively, which suggest at least 87.4% of variances of the new data matrix D are explained by the first PCs. In Table 1, the predicted pEC50 values of EVPCR method are include. The plots of the predicted pEC50 obtained by EVPCR model against the observed pEC50 are showed in Fig.5. Table 6 showed that, the K-PCR method was include in EVPCR procedure. The statistical quality of EVPCR was similar to K-PCR.

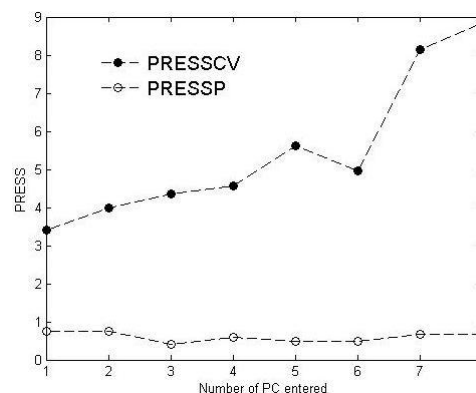


Fig.4—The variation of the $PRESS_{CV}$, $PRESS_P$ as a function of number of entered PC in the EVPCR model

Table 6—Results of EVPCR in the presence of different entered PCs

	R^2_{cv}	$PRESS_{CV}$	R^2_t	$PRESS_T$	R^2_p	F	MAE_p
PC1	0.874	3.408	0.898	0.740	0.519	159.225	0.341
PC1+PC2	0.853	3.974	0.898	0.746	0.515	75.198	0.342
PC1+PC2+PC3	0.839	4.366	0.917	0.411	0.732	58.610	0.258
PC1+PC2+PC3+PC4	0.831	4.570	0.919	0.586	0.619	42.467	0.312
PC1+PC2+PC3+PC4+PC5	0.792	5.620	0.941	0.481	0.687	44.294	0.268
PC1+PC2+PC3+PC4+PC5+PC6	0.817	4.958	0.951	0.495	0.678	41.985	0.277
PC1+PC2+PC3+PC4+PC5+PC6+PC7	0.699	8.141	0.952	0.659	0.571	34.030	0.303
PC1+PC2+PC3+PC4+PC5+PC6+PC7+PC8	0.673	8.855	0.952	0.663	0.569	27.361	0.311

vulgaris toxicity by EVPCR

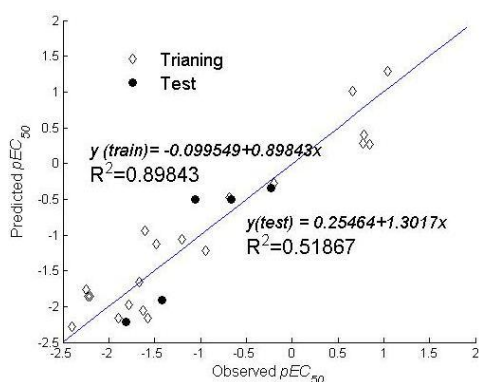


Fig.5—Scatter plot of the predicted VS. observed *Chlorella vulgaris* toxicity by EVPCR

The correlation coefficient between each one of the PCs and pEC50 is represented in the last column of Table 5. As it is showed in Table 5, the changes in the correlation on coefficients are not in the same trend not as the eigenvalues. PCs with higher correlation coefficients have greater information about the variation in the pEC50. The order of decreasing of the correlation coefficient of PCs with pEC50 is:

PC1>PC5>PC3>PC6>PC4>PC7>PC8>PC2. The PC1 has high correlation with pEC50 (0.948). The difference between the eigenvalues and correlation

coefficient is that the eigenvalues explain the variances in the descriptors data matrix which can be represent by the corresponding eigenvector, the correlation coefficient is a part of the variance in the dependent variables (pEC₅₀) which is explained by a eigenvalue. The PC2 has high information content about the descriptors, but do not has useful information about pEC₅₀. This can explain why no improvement was achieved after adding PC2 to PC1 in the EVPCR model. Therefore, it is reasonable to select the PCs based on their information contents about dependent variables instead of their information contents about independent variable. In other words, the CPPCR is superior to EVPCR method.

The results of CRPCR are showed in table 7. The significance of variables in table 7 is the same as the table 6. The plot of PRESSCV and PRESSP against the number of PCs are shown in Figure 8. The values of PRESSCV and PRESSP are decreased by increasing number of PCs up to 4, next, the PRESSCV and PRESSP are gradually increased. It's means than 4 PCs (PC1+PC5+PC3+PC6) could well be used to develop the PCR model. From the fourth row of Table 7, the R² of training, prediction data set and LOO cross validation are 0.949, 0.79 and 0.907 respectively, which suggest at least 87.4% of

variances of pEC₅₀ are explained by that four PCs selected by correlation ranking procedure. In Table 1, the predicted pEC₅₀ values of EVPCR method are included. The plots of the predicted pEC₅₀ obtained by CRPCR model against the observed pEC₅₀ are showed in Figure 9. All the above results tell the statistical quality of the results obtained by CRPCR are better than those obtained by EVPCR model.

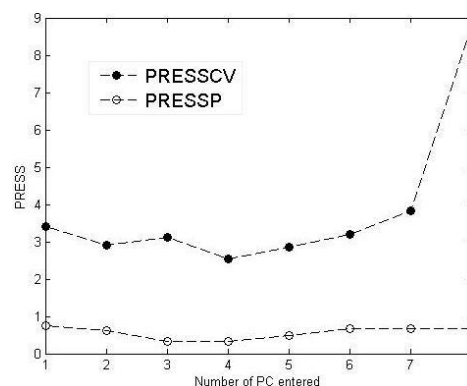


Fig.6—The variation of the PRESSCV, PRESSP as a function of number of entered PC in the CRPCR model

Table 7—Results of CRPCR in the presence of different entered PCs

	R ² _{cv}	PRESS _{CV}	R ² _t	PRESS _p	R ² _p	F	MAE _p
PC1	0.874	3.408	0.898	0.740	0.519	159.225	0.341
PC1+PC5	0.893	2.901	0.920	0.625	0.593	97.906	0.281
PC1+PC5+PC3	0.884	3.131	0.938	0.342	0.778	81.058	0.209
PC1+PC5+PC3+PC6	0.907	2.527	0.949	0.323	0.790	69.276	0.218
PC1+PC5+PC3+PC6+PC4	0.895	2.844	0.951	0.495	0.678	54.246	0.277
PC1+PC5+PC3+PC6+PC4+PC7	0.882	3.200	0.952	0.663	0.569	43.001	0.303
PC1+PC5+PC3+PC6+PC4+PC7+PC8	0.859	3.829	0.952	0.665	0.567	34.105	0.311
PC1+PC5+PC3+PC6+PC4+PC7+PC8+PC2	0.673	8.855	0.952	0.663	0.569	27.361	0.311

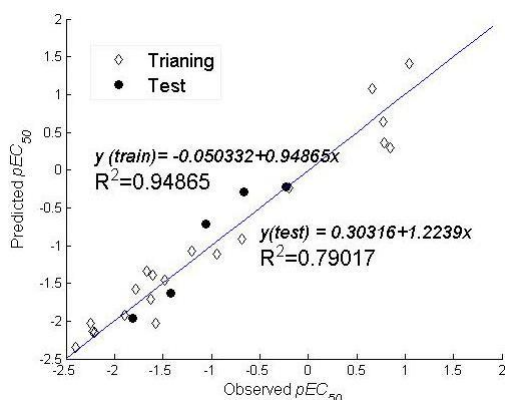


Fig.7—Scatter plot of the predicted VS. observed *Chlorella vulgaris* toxicity by CRPCR

Conclusions

The optimized geometries, based on the 96 hr-EC₅₀ of 25 aromatic hydrocarbons with *C. Vulgaris* were carried out at the B3LYP/6-311G** level by density functional theory (DFT) calculation. The MLR and three types PCR were applied to build the prediction of pEC₅₀. The results demonstrated that all model could describe the toxicity of those organic compounds which similar to this paper's. The comparison of the results obtained from four different methods showed that the CRPCR gives the best accuracy.

Acknowledgements

The research was supported by the National Natural Science Foundation of China (NSFC, No.41606138), the Special Funds of Basic Research of Central Public Welfare Institute (2016Z01-02).

References

- Lin, X.M., Liu, W.X., Chen, J.L., Xu, S.S., and Tao, P., Distribution and ecological risk assessment of polycyclic aromatic hydrocarbons in surface sediments from Bohai Sea, China. *Acta Secintiae Circumstantiae*, 25(2005):70-75.
- Ou, D.N., Multi-media Distribution and Sources Identification of Polycyclic Aromatic Hydrocarbons (PAHs) in the Yangtze Estuarine and Coastal Ecosystem. Shanghai: East China Normal University. 2007, pp:255-270.
- Zeiger, E., Carcinogenicity of mutagens: Predictive capability of the Salmonella mutagenesis assay for rodent carcinogenicity. *Cancer Research*, 47(1987):1287-1296.
- Bocskay, K.A., Tang, D., Orjuela M.A., and Liu X., Chromosomal aberrations in cord blood are associated with prenatal exposure to carcinogenic polycyclic aromatic hydrocarbons. *Cancer Epidemiology Biomarkers Pre*, 14(2005):506-511.
- Jacobs, I.S., and Bean, C.P., "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp.271-350.
- Sijm, H.M.D.T., Middelkoop J., and Vrisekoop K., Algal density-dependent bioconcentration factors of hydrophobic chemical. *Chemosphere*, 31(9)1995:4001-4012.
- Ma, J.Y., Xu, L.G., Wang, S.F., Zheng, R.Q., Jin, S.H., Huang, S.Q., and Huang, Y.J., Toxicity of 40 herbicides to the green alga. *Ecotox. Environ. Saf*, 51(2002):128-132.
- Smiesko, M., Benfenati, E., Predictive models for aquatic toxicity of aldehydes designed for various model chemistries. *J. Chem. Inf. Comput. Sci*, 44(2004):976-984.
- Schmitt, H., Altenburger, R., Jastorff, B., and Chüürmann, G., Quantitative structure-activity analysis of the algae toxicity of nitroaromatic compounds. *Chem. Res. Toxicol*, 13(2000):441-450.
- Knauer, K., Lampert, C., and Gonzalez-Valero, J., Comparison of in vitro and in vivo acute fish toxicity in relation to toxicant mode of action. *Chemosphere*, 68(2007):1434-1441.
- Roy K., and Ghosh G., QSTR with extended to pochemical atom indices. 2. Fish toxicity of substituted benzenes. *J. Chem. Inf. Comput. Sci*, 44(2004):559-567.
- Papa, E., Villa, F., and Gramatica, P., Statistically validated QSARs, based on theoretical descriptors, for modeling aquatic toxicity of organic chemicals in Pimephales promelas (fathead minnow). *J. Chem. Inf. Model*, 45(2005):1256-1266.
- Yuan, H., Wang, H.Y., and Cheng, Y.Y., Local and global quantitative structure-activity relationship modeling and prediction for the baseline toxicity. *J. Chem. Inf. Model*, 47(2007):159-169.
- Montgomery, D.C., and Peck, E.A., *Introduction to Linear Regression Analysis*, Wiley, New York, 1982.
- Jolliffe, I. T., *Principal Component Analysis*, Springer, New York, 1986.
- Kalivas, J. H., and Lang, P.M., *Mathematical Analysis of Spectral Orthogonality*, Marcel Dekker, New York, 1994.
- Puchwein, G., Selection of Calibration Samples for Near-Infrared Spectrometry by Factor Analysis of Spectra. *Anal. Chem*, 1988, 60, 569-573.
- Wang, L.M., Chen g, Y., and Yang, N., *Application regression analysis*, Fudan university press, Shanghai, 2008.
- Hemmateenejad, B., and Shamsipur, M., Quantitative structure-electrochemistry relationship study of some organic compounds using PC-ANN and PCR. *Internet Electron J Mol Des*, 2004;4:316-34.
- J Sutter, M., Kalivas, J. H., and Lang, P.M., Which Principal Components to Utilize for Principal Component Regression, *J. Chemometr*, 1992, 6, 217-225.
- Sun, J., A Correlation Principal Component Regression Analysis of NIR Data, *J. Chemometr*, 1995, 9, 21-29.
- Guillard, R.R.L., and Ryther, J.H., Studies of marine planktonic diatoms I. *Cyclotella nana* Hustedt, and *Detonula confervacea* (Cleve) Gran. *Can J Microbiol*, 8(1962):229-239.
- Sheng, L.F., Jiang, J. G., and Ling, Q. S., Test of toxicity of nitryl-aromatic hydrocarbon to *Dunaliella salina*. *Marine Sciences*, 3(2006):15-17.
- Frisch, M.J., Trucks, G.W., Schlegel, H.B., Scuseria, G.E., Robb, M.A., Cheeseman, J.R., Zakrzewski, V.G., Montgomery, Jr.J.A., Stratmann, R.E., Burant, S., Dapprich, J.M., Millam, A.D., Daniels, K.N., Kudin, M.C., Strain, O., Farkas, J., Tomasi, V., Barone, M., Cossi, R., Cammi, B., Mennucci, C., Pomelli, C., Adamo, S., Clifford, J., Ochterski, G.A., Petersson, P.Y., Ayala, Q., Cui, K., Morokuma, D.K., Malick, A.D., Rabuck, K., Raghavachari, J.B., Foresman, J., Cioslowski, J.V., Ortiz, A.G., Baboul, B.B., Stefanov, G., Liu, A., Liashenko, P., Piskorz, I., Komaromi, R., Gomperts, R.L., Martin, D.J., Fox, T., Keith, M.A., Al-Laham, C.Y., Peng, A., Nanayakkara, M., Challacombe, P.M.W., Gill, B., Johnson, W., Chen, M.W., Wong, J.L., Andres, C., Gonzalez, M., Head-Gordon, E.S., Replogle, J.A., Pople. *Raghava-03, Revision A1, Gaussian, Inc., Pittsburgh*, 2003.
- Maldonado, A.G., Doucet, J.P., Petitjean, M., and Fan, B. T., Molecular similarity and diversity in chemoinformatics: from theory to application. *Mol. Diversity*, 10(2006):39-79.
- Xia B.B, Liu, K.P., Gong, Z.G., Zheng B., Zhang, X.Y., and Fan, B.T., Rapid toxicity prediction of organic chemicalsto *Chlorellavulgaris* using quantitative structure-activity relationships methods. *Ecotoxicology and Environmental Safety*, 72(2009):787-794.
- Zhao, W.Y., *Econometrics tutorial*. Shanghai university of financial and economical press, Shanghai, 2010, pp.107-121.
- Xie, Z.H., *Statistic analysis and application of Matlab: 40 cases analysis*. Press of Beijing University of Aeronautics and Astronautics, Beijing, 2010, pp.251-287.
- Li, C.X., *Biometrics. the fourth edition*. Science Pre, Beijing, 2008.