Bulgarian Chemical Communications, Volume 47, Number 4 (pp. 1197 – 1203) 2015

QSAR study of halogenated benzene bioaccumulation factors in fish

F. Chen[1,2], N. Li[1,2], D. Yang [1,2,3*], Y. Zhou[1,2]

[1]*Key Laboratory of Marine Bio-resources Restoration and Habitat Reparation in Liaoning Province, Dalian Ocean University, 116023, Dalian, China*
[2]*Key Laboratory of North Mariculture, Ministry of Agriculture, Dalian Ocean University, 16023, Dalian, China*
[3]*College of Life science and Technology, Dalian University of Technology, 116021, Dalian, China*

Received April 7, 2015

Certain elements or compounds, that are difficult to decompose, accumulate in organisms. The phenomenon of higher compound concentrations in organisms than in the environment is called bio-concentration. This research studies 21 bio-concentration factors (B_{CF}) of halogenated benzene in fish. Here, 14 halogenated benzene molecule parameters are randomly selected as a training set and the remaining parameters are considered as a testing set to calculate halogenated benzene molecule descriptors. A prediction model of the quantitative structure–activity relationship between bio-concentration and molecular descriptors is built using the multiple linear regression method. The independent variable with serious collinearity is eliminated to obtain an optimal prediction model (R^2=0.919). The prediction data of linear regression are also obtained using the molecular parameters of the test set. A support vector machine (SVM) is used to predict the result of the test set using the training set as study samples. The best method for predicting the bio-concentration factor of halogenobenzene in fish is determined by comparing the prediction accuracy of the two methods. The final results indicate that the model officially built by stepwise regression can effectively predict $B_{CF.}$ SVM is more accurate in predicting B_{CF}. SVM possesses better predictive capability in issues with a small amount of samples.
Key words: Bio-concentration factor, halogenobenzene, linear regression, quantitative structure-activity relationship, prediction model, support vector machine.

INTRODUCTION

Halogenated benzenes are compounds that feature replacement of one or more hydrogen atoms in benzene by halogen atoms; these compounds are widely distributed in the air, soil, groundwater, surface water, and the sea [1]. Halogenated benzenes are widely used in chemical, electronic, and pharmaceutical industries as chemical raw materials. They have become the focus of many researchers because of their strong carcinogenic, teratogenic, and mutagenic effects, toxicity, persistence, and bio-concentration [2]. The concentration of halogenated benzenes in the environment stepwise increases through the food chain. The more high-end creatures in the food chain, the more harmful are the side effects. This phenomenon is called bio-concentration.

The bio-concentration factor is the ratio between the balance concentration of a compound in an organism and its balance concentration in water. B_{CF} is generally used to express bio-concentration. Bio-concentration is associated with the food chain, where different creatures are closely linked by eating and being eaten. While the concentrations of some chemicals may be harmless to some creatures in the environment, they can become harmful to human health through biological enrichment and biomagnifications in the food chain [3,4]. Therefore, studying the bio-concentration factor of toxic halogenated benzenes is significant. Studying the physical and chemical properties of halogenated benzenes may also provide a means of predicting bio-concentration.

The quantitative structure–activity relationship (QSAR) between the molecular structures of halogenated benzenes and their bio-concentration is the primary subject of this study. The essence of the QSAR model is to obtain information from the structure-bio concentration factor model through a sufficient amount of data and establish the molecular structure of halogenated benzenes and their enrichment factor using multiple linear regression (MLR) [5,6]. The law of large numbers of classical statistical mathematics states that the statistical law will be accurately known and the results of the fitting model can fully reflect the real law only when the known sample number approaches infinity. In actual problems such as chemicals, drug design, and environment protection, obtaining the known sample number is often difficult. Vapnik proposed statistical learning theory by studying mathematical theory for over 30 years and developed the support vector machine on the basis of statistical learning theory [7, 8]. SVM,

* To whom all correspondence should be sent:
E-mail: dzyang1979@hotmail.com

which includes SV classification and SV regression, has been successfully applied in language recognition, facial recognition, text recognition, and drug design [9, 10, 11]. SVM is not only effective at simplifying structures but also presents outstanding capability for all technology types, especially in the field of generalization, as confirmed by many previous experiments [12, 13]. Therefore, we used SVM to verify small sample descriptors that have been screened out in the present work. A liable mathematical prediction model is further established using SVM to provide a theoretical basis for analyzing the biological toxicity of halogenated benzene compounds.

MATERIALS AND METHODS

Data set

This study employed experimental data of 21 halogenated benzene bio-concentration factors described in ref. [14]. The SMILES format of halogenated benzene was searched to obtain descriptor information and construct the MLR model. PaDEL-Descriptor software was used to obtain the molecular descriptors of halogenated benzene [15]. The 3D descriptors of halogenated benzene were calculated to select the most suitable descriptor parameter by MLR. The QSAR model between the descriptor and B_{CF} was finally built. The logarithm lgB_{CF} was used to express the bio-concentration in a more scientific manner because the different B_{CF} values of halogenated benzene largely vary. The higher the value of lgB_{CF}, the higher is the bio-concentration capability of the compound. Here, 14 molecular parameters were randomly selected as a training set for modeling, and the remaining 7 parameters were classified as a testing set. The optimal prediction model was determined by MLR of the parameters.

Method

MLR

MLR provision is the classic QSAR modeling method that establishes the forecast method by analyzing two or more independent variables and a dependent variable correlation analysis [16]. A linear relationship for the calculation of the correlation between the independent and dependent variables cannot determine which descriptors exactly predict and establish the mathematical model using the selection of the 3D descriptor parameters. Therefore, this study uses stepwise regression method to establish the halogenated benzene bio-concentration factor B_{CF} covariant relationship with molecular descriptors. The mathematical function expression of the MLR equation is as follows:

$$Y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n, \quad (1)$$

In eq. (1) Y is the concentration fact; α_0 is a constant; $x_1 \ldots x_n$ are molecular descriptors; and $\alpha_1 \ldots \alpha_n$ are regression coefficients.

SVM

SVM was used to validate the results and compare them with MLR forecasting to verify the accuracy of MLR in reducing errors and make more accurate predictions. SVM may be completely described by the training set and kernel function [13]. Finding the optimal hyper plane, which is the plane that separates all samples with the maximum margin, is an essential principle of SVM [17, 18]. This plane helps to improve the predictive ability of the model and lowers the error that may occasionally occur during classification.

Fig. 1 illustrates the optimal hyper plane; here, "red" indicates the samples of type 1 and "blue" represents the samples of type 1.

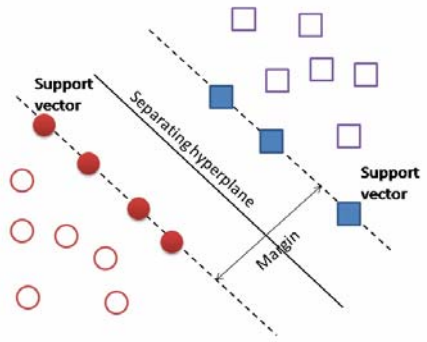

Fig. 1.Support vectors determining the position of the optimal hyper plane.

The SVM method can be applied to the function fitting problem. SVM was applied in this study to obtain the regression formula as follows:

$$Y = f(x) = \sum_{i=1}^{n} (T_i^* - T_i)(I_i \times I) + b, \quad (2)$$

The nonlinear problem must be transformed into a linear problem using the kernel function method to project the original data to a high-dimensional feature space. The following formula is then obtained:

$$Y = f(x) = \sum_{i=1}^{n} (T_i^* - T_i)K(I_i \times I) + b, \quad (3)$$

RESULTS

QSAR model of the stepwise regression method

The 3D molecular descriptors of halogenated benzene compounds are studied through 3D autocorrelation. We describe the bio-concentration factor by using the descriptors as charged partial surface area, length over breadth, and weighed holistic invariant molecule [19,20,21]. Vv and LOBMAX are filtered from the existing molecular descriptors by the stepwise regression method. Vv is used to describe the weighed holistic invariant molecula [19], and LOBMAX is used to describe the length over breadth. The QSAR model between B_{CF} and the molecular structure descriptors is established according to the minimal relative standard deviation and the maximum principle of correlation coefficient as:

$$Y = 2.132 + 0.245 \times Vv - 1.029 \times LOBMAX, \quad (4)$$

First, the sample size of the regression model is four times larger than the independent variable based on data with statistical regularity. Second, the correlation coefficient of the regression equation established in Table 1 is R^2=0.919, which indicates that the model provides good correlation. The model test value of F is 74.294, and the significance level is less than 0.05.

Table 1. Statistical universe of the regression equation.

R	R^2	R^2_{adj}	F	Sig.
0.965	0.931	0.919	74.294	0.000

The equations of each variable test value are listed in Table 2. The results of the t test show that the coefficients of the statistic equation are 2.675, 7.535, and −2.557. The probability of the t test is 0, which means that the regression equation matches the standard. The VIF values are 1.69 and 1.69 when the VIF value is larger than 10. This result indicates that the independent variables have strong collinearity. The independent variables at this point are consistent with the collinearity inspection, so it can be considered as linear fitting.

Table 2. Determination of the regression equation.

	Beta	t	Sig.	VIF
Constant	2.132	2.675	0.022	
Vv	0.245	7.535	0.000	1.690
LOBMAX	−1.029	−2.557	0.027	1.690

Tables 3 and 4 show the SMILES of the halogenated benzene compounds, descriptor parameters, experimental values, calculated values, and residual errors. Fourteen of the halogenated benzene lgB_{CF} were selected as the training set, and 7 were selected as the test set. The absolute value of the minimum residual error is 0.01, while the absolute value of the maximum residual error is 2.06. The ratios of the experimental values are 0.54% and 67.8%, respectively.

Fig. 2 shows the solid dots indicating the training set and the hollow dots indicating the test set. The experimental and predicted values of the training and testing set molecules are similar to some extent.

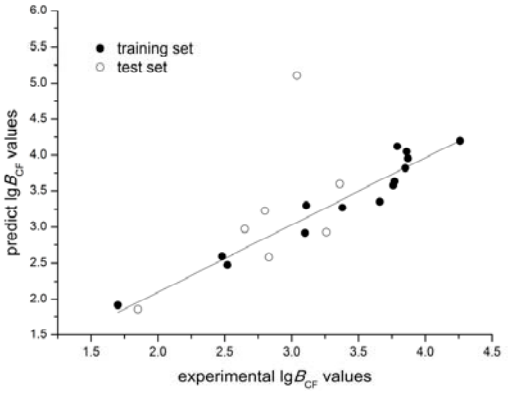

Fig.2. Comparison of the experimental and predicted values

SVM prediction of the test set results

The selected compound structure description parameters should have significant structural characteristics in the SVM application to build a matching model and help predict the compound properties. We obtained Vv and LOBMAX in this study using the stepwise regression method of molecular descriptors with 14 descriptors taken as training samples for SVM learning. The lgB_{CF} predicted by the SVM is obtained by inputting the descriptor parameters of the test set into SVM.

Seven of the halogenated benzene lgB_{CF} were selected as the test set. The absolute value of the minimum residual error is 0.15, whereas the absolute value of the maximum residual error is 1.22. The ratios of the experimental values are 8% and 40%, respectively.

MLR and SVM results

The appropriate horizontal ordinate range is selected and the prediction scatter plot of the MLR test set is described as follows:

Fig. 3 shows precise matching results, and the predicted and experimental values are close. Figs. 3b and 3c reveal the residual, experimental, and predicted relationships. The residual values are relatively low.

The SVM test results are highly similar to those of MLR. SVM can generate a fairly analogical and precise result compared with the MLR test results. In summary, both MLR and SVM can be suitably applied for determining halogenated benzene bioaccumulation factors in fish.

Table 3. MLR prediction of the training set parameters of the bio-concentration factor of halogenated benzene compounds in fish.

Name	Vv	LOBMAX	SMILES	lgB_{CF}		
				experimental	predicted	residual
1,2,3,4-tetrachlorobenzene	11.04	1.17	c1(c(c(ccc1Cl)Cl)Cl)Cl	3.77	3.63	0.14
1,2,3-trichlorobenzene	9.63	1.15	c1(c(cccc1Cl)Cl)Cl	3.11	3.31	□0.20
1,2,4,5-tetrachlorobenzene	10.81	1.17	c1(c(cc(c(c1)Cl)Cl)Cl)Cl	3.76	3.58	0.18
1,2-dichlorobenzene	7.90	1.43	c1(c(cccc1)Cl)Cl	2.48	2.59	□0.11
1,3,5-trichlorobenzene	9.94	1.26	c1(cc(cc(c1)Cl)Cl)Cl	3.38	3.27	0.11
1,4-dichlorobenzene	7.38	1.42	c1(ccc(Cl)cc1)Cl	2.52	2.47	0.05
hexachlorobenzene	14.12	1.35	c1(c(c(c(Cl)c(c1Cl)Cl)Cl)Cl)Cl	4.26	4.20	0.06
2,4,5-Trichlorotoluene	12.51	1.21	c1(c(cc(Cl)c(c1)Cl)Cl)C	3.87	3.96	□0.09
1,2,3,4,5-Pentachlorobenzene	12.63	1.15	c1(c(c(cc(c1Cl)Cl)Cl)Cl)Cl	3.86	4.05	□0.19
1,2,4,5-Tetrabromobenzene	13.17	1.20	c1(c(cc(Br)c(c1)Br)Br)Br	3.79	4.12	□0.33
1,2,4-Tribromobenzene	11.06	1.44	c1(c(ccc(c1)Br)Br)Br	3.66	3.36	0.30
1,3,5-Tribromobenzene	12.15	1.25	c1(cc(cc(c1)Br)Br)Br	3.85	3.82	0.03
Bromobenzene	6.73	1.81	c1(ccccc1)Br	1.7	1.92	□0.22
1,2-dibromobenzene	9.17	1.42	c1(c(cccc1)Br)Br	3.1	2.92	0.18

Table 4. MLR prediction of the test set parameters of the bio-concentration factor of halogenated benzene compounds in fish.

Name	Vv	LOBMAX	SMILES	lgB_{CF}		
				experimental	predicted	residual
1,2,3,5-tetrachlorobenzene	11.26	1.26	c1(c(cc(Cl)cc1Cl)Cl)Cl	3.36	3.60	□0.24
1,2,4-trichlorobenzene	9.29	1.44	c1(c(ccc(c1)Cl)Cl)Cl	3.26	2.93	0.33
1,3-dichlorobenzene	7.96	1.08	c1c(cccc1Cl)Cl	2.65	2.97	□0.32
Chlorobenzene	6.16	1.73	c1(ccccc1)Cl	1.85	1.86	□0.01
1,3-dibromobenzene	9.28	1.14	c1c(cccc1Br)Br	2.8	3.23	□0.43
Hexabromobenzene	17.76	1.34	c1(c(c(c(Br)c(c1Br)Br)Br)Br)Br	3.04	5.10	□2.06
1,4-dibromobenzene	8.30	1.54	c1(ccc(Br)cc1)Br	2.83	2.58	0.25

Table 5. SVM prediction of the test set parameters of the bio-concentration factor of halogenated benzene compounds in fish.

Compound	Experimental	Predicted	Residual
1,2,3,5-tetrachlorobenzene	3.36	3.77	□0.41
1,2,4-trichlorobenzene	3.26	3.1	0.16
1,3-dichlorobenzene	2.65	2.48	0.17
Chlorobenzene	1.85	1.7	0.15
1,3-dibromobenzene	2.8	3.1	□0.3
Hexabromobenzene	3.04	4.26	□1.22
1,4-dibromobenzene	2.83	2.48	0.35

a) a)

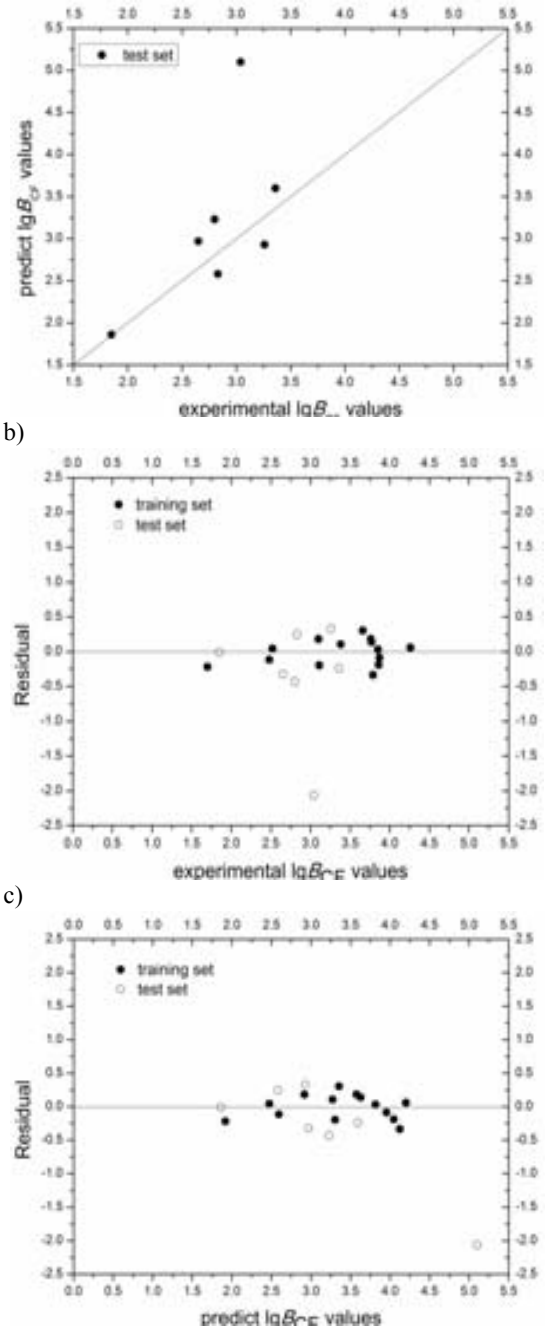

Fig. 3. Test results of the MLR model. a) experimental *vs* predicted values; b) experimental *vs* residual values; and c) predicted *vs* residual values.

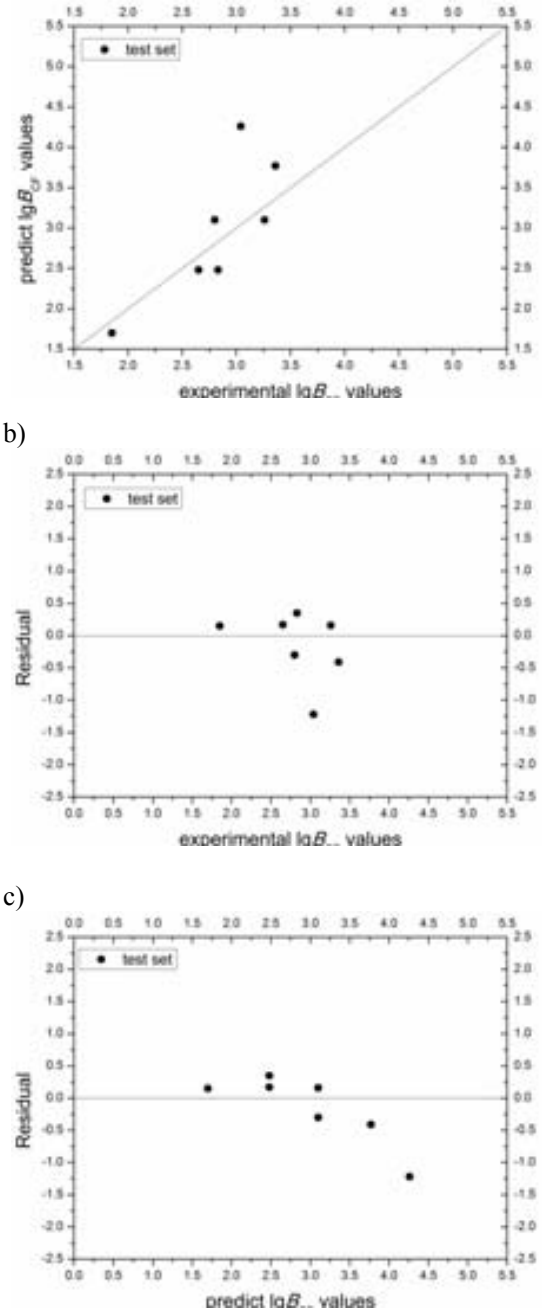

Fig. 4. Test results of the SVM model. a) experimental *vs* predicted values; b) experimental *vs* residual values; and c) predicted *vs* residual values.

CONCLUSIONS

In this work, we used the SVM and MLR stepwise analysis methods. The principle of the MLR stepwise analysis method considers the size of the contribution of each variable to the dependent variable by introducing the variables one at a time and simultaneously inspecting the previously introduced variable. The capability of discriminating a previously introduced variable becomes non-significant as a new variable is introduced. This new variable is removed when a discriminant variable is remarkable, and stepwise regression ends when no important variable for the rest of the variables can be introduced to the discriminant [22]. The equation must conform to the test, the parameters of the obtained descriptors must be remarkable, and the training set correlation coefficient must be R^2=0.919 to eliminate collinearity independent variables in the model and ensure that the model has strong predictive ability. The test set molecules are substituted into the equation to obtain the predicted values. The residual distributions of the training and test sets are even. This model has two independent variables, thereby highlighting the influence of Vv and LOBMAX on the B_{CF} of the halogenated benzenes.

Given that SVM has the advantage of solving the problems of small samples, this study also uses SVM to validate the results of MLR. The Vv and LOBMAX descriptors of the 14 groups of training set data are used as learning samples, and the remaining 7 groups of test data are substituted into SVM to obtain the predicted values. The residual value of the prediction obtained from SVM is more stable than that obtained from MLR. The root mean square error (RMSE) of the SVM prediction is 0.42, whereas that of the MLR prediction is 0.94. These values illustrate that the SVM model prediction error is smaller and the model obtained through this method is more stable. The residual value of the MLR prediction of hexabromobenzene is 67.8%, whereas that predicted by SVM is 40.1%. These values show that hexabromobenzene cannot fit the QSAR model; however, in this study, we still retained the full information of hexabromobenzene to authenticate results.

In conclusion, the statistical results of the stepwise method are suitable (training set R^2=0.919), and the prediction ability of the model is outstanding. We can obtain more precise prediction values by using SV Minstead of MLR. The model can better specify the descriptors of LOB and WHIM. Thus, we can accurately predict the ultimate bio-concentration factor. The model obtained provides guidance for future research on halogenated benzene concentration factors. The method proposed in this work may be an important means of halogenated benzene compounds concentration research in the fields of chemistry, drug design, and environment protection.

Acknowledgements: *This work was funded by the National Marine Public Welfare Research Project (Nos. 201305002 and 201305043), the Natural Science Foundation of Dalian (No. 2012J21DW014), and the Project of Marine Ecological Restoration Technology Research on the Penglai 19-3Oil Spill Accident (No.201422).*

REFERENCES

1. L. H. Hall, E. L. Maynard, L. B. Kier, *Environ. Toxicol. Chem*., **5**, 333 (1986).
2. H. Bahadar, S. Mostafalou, M. Abdollahi, *Toxicol. Appl. Pharmacol.*, **276**, 83 (2014).
3. S. M. Choi, S. Y. Kwan, C. M. Wong, *Microbial*, **53**, 54 (1996).
4. P. G. Hatcher, P. A Mcgillivary, *Environ. Sci. Technol.,* **13**, 1225 (1979).
5. G. Piir, S. Sild, U. Maran, SAR. *QSAR. Environ. Res.*, **24,** 175 (2013).
6. E. Papa, J. C. Dearden, P. Gramatica, *Chemosphere,* **67**, 351 (2007).
7. C. Cortes, V. Vapnik, *Mach. Learn*., **20**, 273 (1995).
8. V. Vapnik, Statistical Learning Theory, New York: Wiley, 1998.
9. R. Burbidge, M. Trotter, B. Buxton, S. Holden, *Comput. Chem.,* **26,** 5 (2001).
10. E. Byvatov, U. Fechner, J. Sadowski , G. Schneider, *J. Chem. Inf. Comput. Sci.,* **43**, 1882 (2003).
11. V. Wan, W. M. Campbell, *Proceedings of the 2000(SiPS) IEEE.*, **2**, (2014).
12. P. Bartlett, J. Shawe-Taylor, Generalization performance of support vector machines and other pattern classifiers, Advances in Kernel Methods - Support Vector Learning, MIT Press, 1999.
13. B. Sholkopf, K. Sung, C. J. C Burges, F. Girosi, P. Niyogi, T. Poggio, V. Vapnik, *IEEE. Trans. Signal. Process.*, **45,** 2758 (1997).
14. C. J. Feng, X. H. Du, *Journal of Shenzhen University Science and Engineering*, **31,** 96 (2014).
15. C. W. Yap, *J. Comput. Chem.*, **32**, 1466 (2011).
16. R. M. Rose, M. St. J. Warne, R. P. Lim, *Environ. Contam. Toxicol.,* **34**, 248 (1998).
17. X. Zhong, J. Li, H. Dou, S. J. Deng, G. F. Wang, Y. Jiang, Y. Wang, Z. Zhou, L. Wang, F. Yan, *PLoS ONE*, **8**, p. e69434 (2013).
18. Y, Shen, Z, He, Q. Wang, Y. Wang, in IEEE International Instrumentation and Measurement Technology Conference (I2MTC), Graz, AUSTRIA 2012, p. 1977.
19. R. Todeschini, P. Gramatica, 3D QSAR in Drug

Design, Springer, Netherlands, 1998, p. 355.
20. R. Todeschini, V. Consonni, Molecular Descriptors for Chemoinformatics John Wiley & Sons., New York, 2009.
21. D. T. Stanton , P. C. Jurs, *Anal. Chem*., **62**, 2323 (1990).
22. A. Ciampi, J. Thiffault, *Comput. Stat. Data Anal*, **4**, 185 (1986).

QSAR-ИЗСЛЕДВАНЕ НА БИО-КОНЦЕНТРИРАНЕТО НА ХАЛОГЕНИРАН БЕНЗЕН В РИБИ

Ф. Чен[1,2], Н. Ли[1,2], Д. Яан[1,2,3*], Й. Жоу[1,2]

[1]*Ключова лаборатория по възстановяване на морски биоресурси и подобряване на жилищната среда в провинция Ляонинг, Океаноложки университет в Далиан, Китай*
[2]*Ключова лаборатория по северни морски култури, Министерство на земеделието, Океаноложки университет в Далиан, Китай*
[3]*Колеж за науки за живота и технологии, Университет в Далиан, Китай*

Постъпила на 5 април, 2015 г.

(Резюме)

Някои натрупани съединения се разграждат трудно в организмите. Явлението на повишена концентрация в организмите спрямо в околната среда се нарича био-концентриране. В тази работа са изследвани 21 фактори на био-концентриране (B_{CF}) на халогениран бензен в риби. Четиринадесет молекулни параметри на халогенираните бензени са произволно избрани като базова мрежа, а останалите се смятат като зависими променливи за изчисляван на молекулни дескриптори. Създаден е модел за предсказване на отношението количествена структура-активност между био-концентрирането и молекулните дескриптори чрез множествена линейна регресия. Елиминирани са независимите променливи със значима взаимна линейна връзка за постигането на оптимални предвиждания с коефициент на корелация R^2=0.919. Предсказаните данни от линейната регресия са получени и чрез молекулните параметри от зависимите променливи. Използван е поддържащ вектор (SVM) за предсказване на зависимите променливи от базовите данни като проби. Най-добрият метод за предсказване на фактора на био-концентриране на халогено-бензени в риби се определя чрез сравняване на точността на предвиждане на двата метода. Крайният резултат показва, че моделът, построен чрез степенна регресия може ефективно да предскаже $B_{CF.}$ Прилагането на SVM може да е по-точно в предвиждането на B_{CF}, особено при малък брой проби.