Quantitative Structure–Activity Relationship for Prediction of the Toxicity of Phenols on *Photobacterium phosphoreum*

Xiaolin Li · Zunyao Wang · Hongling Liu · Hongxia Yu

Received: 19 September 2011 / Accepted: 20 April 2012 / Published online: 5 May 2012

Abstract Quantitative structure–activity relationships (QSAR) is an alternative to experimental toxicity testing and recommended by environmental protection agencies. In this background, an accurate and reliable QSAR model of 18 phenols for their toxicity to *Photobacterium phosphoreum* was developed using mechanistically interpretable molecular structural descriptors. The QSAR model was developed by stepwise multiple linear regression and the reliability of the model was evaluated by internal and external validation. The cross-validated correlation coefficient (q^2) was 0.7021, indicating good predictive ability for the toxicity of these phenols. The QSAR model suggests that the toxicity of the studied compounds mainly depends on the logarithm of octanol/water partition coefficient, dipole moment and the most negative atomic charge.

Keywords QSAR · Toxicity · Phenols · Multiple linear regression

Information regarding the toxicity of organic pollutants is very important in risk assessment of the chemicals. As one of many ways to fill toxicity data gaps, quantitative structure–activity relationships (QSAR) are powerful tools to rapidly estimate and predict biological activity (toxicity), and many QSAR models have been successfully developed (Jiang et al. 2010; Jing et al. 2010; Zeng et al. 2011).

Phenols have been widely used in the manufacture of plastics and industrial disinfectants, and used as chemical reagents in industrial processes (Habibi-Yangjeh et al. 2006). They can also be produced by degradation of more complex compounds (Verschueren 2001). Phenols are supposed to be one of the most toxic water pollutants as they are often carcinogenic (Della et al. 2001). Recently, some QSAR studies on toxicity of phenols have been carried out by employing different descriptors and modeling techniques (Su et al. 2008; Jin et al. 2010). The previous works have some shortcomings such as small dataset, without internal validation, not very precise descriptors and do not clarify the toxicity mechanism.

Photobacterium phosphoreum (P. phosphoreum) is a kind of luminescent bacterium in seawater, its character of luminous intensity change with toxic substance inhibition of growth (i.e., cell density) made it as an index for compound toxicity measuring and water quality monitoring. With the development of computer technology and quantum chemistry, more precise method for descriptors calculation, such as the ab initio calculation, is replacing the traditional semi-empirical method. The main aim of the present work is to develop a reliable QSAR model for modeling and predicting toxicity of various phenols towards *P. phosphoreum* using precise molecular descriptors and a multiple linear regression (MLR) algorithm. In addition, to better understand the toxicity mechanism of phenols at the molecular structure level, the structural characteristics affecting toxicity were gained by the descriptors that have clear physical and chemical meaning. This study may enrich the toxicity mechanism study of phenols.

Materials and Methods

A dataset of 18 phenols with 15 min *P. phosphoreum* toxicity reported by Yu et al. (2009) was used as the model

X. Li · Z. Wang · H. Liu · H. Yu (✉)
State Key Laboratory of Pollution Control and Resource Reuse,
School of the Environment, Nanjing University,
Nanjing 210046, People's Rebuplic of China
e-mail: hongxiayu01@yahoo.com.cn

dataset. Toxicity values were determined by MICROTOX test instrument (toxicity tester model DXY-2, made by the Institute of Soil Science, Chinese Academy of Sciences, Nanjing, P. R. China) when the bacteria were exposed to the tested phenols in 3% NaCl solution for 15 min. The EC_{50} was then calculated by linear interpolation method. The list of compounds along with their toxicity values (log EC_{50}, mol L^{-1}) are shown in Table 1.

The total dataset are randomly divided into training set and test set in a ratio of approximate 80:20 (15 and 3 compounds, respectively). The training set is used to construct QSAR model and the test set to validate the external prediction ability of the resulting QSAR model.

In order to calculate the molecular descriptors, structures of phenols were built using ChemDraw Ultra (version 11.0) and then the molecular structures were optimized using semiempirical quantum-chemical method AM1 Hamiltonian (Dewar et al. 1985) implemented in MOPAC software (version 6.0) to generate the energy-minimized conformations.

Fourteen frequently used physicochemical descriptors (including one hydrophobic, nine electronic, two thermodynamic, and two steric property descriptors) were considered for the QSAR development (Table 2). The descriptors are selected for their wide use in QSAR analyses, and potential to relate to reactive modes of toxic action. The log K_{ow} value was estimated using the EPI Suite (version 4.0, US-EPA). M_w and HOF values were calculated by Chem3D Ultra (version 11.0). The apparent acid dissociation constant (pK_a) value was estimated by SPARC (web version, accessed at http://sparc.chem.uga.edu). The other descriptors were computed by the Gaussian 03 program (Frisch et al. 2003) at the B3LYP/6-311G** level. Due to space limitations, only descriptors entered into QSAR model were shown in Table 1. The log K_{ow} for dataset was in the range of 0.24 and 2.94, the range of μ was between 0.926 and 5.792, and the range of q^- was between −0.470 and −0.311.

To select the relevant descriptors, multiple linear regression (MLR) analysis was performed for the toxicity dataset against all 13 descriptors using DPS software (Tang and Feng 2007) (version 9.50). In this analysis, stepwise regression was employed to identify the most important descriptors contributing to the toxicity of chemicals, for it can use F-tests to eliminate insignificant descriptor on every step of entering a new descriptor (Xu et al. 2006). According to the regression result, QSAR model was developed. The statistical quality of QSAR model was evaluated by correlation coefficient (R), Fischer ratio (F), probability factor related to F-ratio (p) and root mean square error (RMSE). RMSE is defined as follows:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(Y_{i\,\mathrm{exp.}} - Y_{i\,\mathrm{pred.}}\right)^2}{\mathrm{n}}} \qquad (1)$$

Table 1 The toxicity of phenols and values of the descriptors entered into QSAR model

No.	Chemical	log $EC_{50\ (\mathrm{obs.})}$	log $EC_{50\ (\mathrm{pred.})}$	Residual	log K_{ow}[b]	μ	q^-
1	Phenol[a]	2.72	2.97	−0.25	1.51	1.344	−0.363
2	Catechol	3.14	3.27	−0.13	1.03	2.486	−0.412
3	*p*-Nitrophenol	3.72	3.48	0.24	1.91	5.309	−0.341
4	*o*-Aminophenol	3.34	3.50	−0.16	0.60	1.376	−0.467
5	*p*-Chlorophenol	3.88	3.75	0.13	2.16	2.543	−0.358
6	*m*-Cresol[a]	3.31	3.62	−0.31	2.06	1.638	−0.363
7	Hydroquinone	3.14	2.66	0.48	1.03	2.655	−0.370
8	*o*-Cresol	3.35	3.58	−0.23	2.06	1.683	−0.360
9	*o*-Chlorophenol	3.43	3.31	0.12	2.16	0.926	−0.339
10	2,3-Dimethylphenol	3.60	4.24	−0.64	2.61	1.949	−0.361
11	4-*tert*-Butylcatechol	5.87	5.42	0.45	2.94	2.518	−0.411
12	2,4,6-Trinitrophenol	2.51	2.33	0.18	1.54	1.714	−0.315
13	*o*-Nitrophenol[a]	3.48	3.07	−0.41	1.91	5.730	−0.311
14	*m*-Nitrophenol	3.31	3.64	−0.33	1.91	5.792	−0.349
15	Resorcinol	2.22	2.48	−0.26	1.03	2.382	−0.360
16	*p*-Aminophenol	3.27	3.20	0.07	0.24	2.077	−0.470
17	*p*-Cresol	3.71	3.63	0.08	2.06	1.407	−0.365
18	2,4-Dichlorophenol	4.01	4.01	0.00	2.80	1.141	−0.336

[a] Test set

[b] Estimated in non-ionized form

Table 2 List of descriptors used in the development of QSAR model

Category of descriptors	Descriptor	Symbol	Unit
Hydrophobic	The logarithm of octanol/water partition coefficient	$\log K_{ow}$	
Electronic	Total energy	TE	eV
	Energy of the highest occupied molecular orbital	E_{HOMO}	eV
	Energy of the lowest unoccupied molecular orbital	E_{LUMO}	eV
	Dipole moment	μ	10^{-30} esu
	Electrophilicity index	ω	
	Average molecular polarizability	α	Debye
	The most positive hydrogen atomic charge	qH^{+}	e
	The most negative atomic charge	q^{-}	e
	The logarithm of acid dissociation constant	pK_a	
Thermodynamic	Standard heat of formation	HOF	kcal mol^{-1}
	Free energy	G^{0}	Hartree
Steric	Molecular weight	M_w	
	Molecular volume	V	cm^3 mol^{-1}

where n is the number of compounds in the model, $Y_{i\ \mathrm{exp.}}$ is the observed log EC_{50} value of the ith compound, $Y_{i\ \mathrm{pred.}}$ is the predicted value of the ith compound.

The reliability of the resulting QSAR model was explored using two different types of validation criteria: (1) internal validation by leave-one-out (LOO) cross-validation, (2) external validation by dividing the dataset into training and test set. As a statistical parameter, prediction error sum of squares (PRESS) can measure the accuracy of a modeling based on the cross-validation. Using the PRESS and SSY (sum of squares of deviations of the experimental values from their mean), the square of the cross-validated correlation coefficient (q^2) and the standardized predicted error sum of squares (S_{PRESS}) values, which can be used to evaluate the performance of validation, can be easily calculated. The equations that calculate the aforementioned parameters are presented below:

$$q^2 = 1 - \frac{\mathrm{PRESS}}{\mathrm{SSY}} = 1 - \frac{\sum_{i=1}^{n}\left(Y_{i\,\mathrm{exp.}} - Y'_{i\,\mathrm{pred.}}\right)^2}{\sum_{i=1}^{n}\left(Y_{i\,\mathrm{exp.}} - Y_{\mathrm{mean}}\right)^2} \tag{2}$$

$$S_{\mathrm{PRESS}} = \sqrt{\frac{\mathrm{PRESS}}{n}} \tag{3}$$

where $Y'_{i\,\mathrm{pred.}}$ is the predicted log EC_{50} value of the ith compound predicted by a model obtained without using the ith chemical, Y_{mean} is the mean observed value of all the compounds in the model.

Results and Discussion

Based on observed toxicity data and calculated values for the relevant molecular descriptors, QSAR models were developed by stepwise MLR analysis. After residual analysis and model tests, three descriptors were entered into the final regression equation. The QSAR model suitable for predicting toxicity values of phenols was obtained and had following specifications:

$$\log EC_{50} = -4.2642 + 1.1304 \log K_{ow} + 0.0950\,\mu - 14.8852\,q^{-} \tag{4}$$

$$\begin{aligned} n &= 15, R = 0.9293, F = 22.8926, \\ &p < 0.0000, \mathrm{RMSE}_{(\mathrm{training\ set})} = 0.2883, \\ &\mathrm{RMSE}_{(\mathrm{test\ set})} = 0.3300, q^2 = 0.7021, S_{\mathrm{PRESS}} = 0.4179. \end{aligned}$$

As can be seen from the Eq (4), a three-descriptor QSAR model was developed to describe toxicity of phenols to *P. phosphoreum*, with both good correlation ($R = 0.9293$ and $\mathrm{RMSE}_{(\mathrm{training\ set})} = 0.2883$) and high stability and prediction ability ($q^2 = 0.7021$ and $\mathrm{RMSE}_{(\mathrm{test\ set})} = 0.3300$). Based on QSAR model, the predicted values of the training and test sets are listed in Table 1 and the relationship between observed and predicted log EC_{50} values was depicted in Fig. 1. As can be seen, these data are uniformly distributed around the regression line, which suggests the obtained model has satisfied predictive ability. As shown in Eq (4), the toxicity of phenols have positive correlation with the logarithm of octanol/water partition coefficient (log K_{ow}) and dipole moment (μ), and negative correlation with the most negative atomic charge (q^{-}). The physical and chemical meaning of the three descriptors will be briefly described later.

In this work, variance inflation factors (VIF) are adopted to evaluate the collinearity of descriptors in the model. VIF is defined as:

$$\mathrm{VIF} = 1/r^2 \tag{5}$$

where r is the correlation coefficient of multiple regression between one independent variable and the others. If

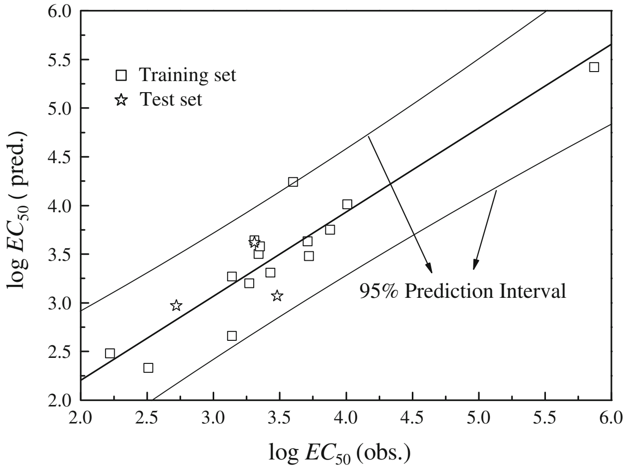

Fig. 1 Plot of observed log EC_{50} versus predicted log EC_{50}

Table 3 VIF, SR and *t* test value of descriptors in QSAR model

Descriptor	VIF	SR	*t* test value
log K_{ow}	1.5220	1.1239	8.1317
μ	1.0381	0.1637	1.4346
q^-	1.5589	−8.8448	6.0399

VIF = 1.0, no self-correlation exists among each variable; when VIF ranges from 1.0 to 5.0, the correlation equation is acceptable; if VIF > 10.0, the regression equation is unstable and recheck is necessary. As can be seen from Table 3, the VIF values of the three descriptors are all smaller than 5.0, indicating that there is no collinearity among the selected descriptors and the resulting model has good stability.

In order to distinguish the importance of each descriptor on toxicity of phenols, standard regression coefficients (SR) and *t* test values of the three descriptors are also listed in Table 3. As shown in Table 3, the absolute value of SR and *t* test value of log K_{ow} are 1.1239 and 8.1317, respectively, both larger than that of μ and q^-, which indicates that in this QSAR model, the influence of log K_{ow} on toxicity is stronger than that of μ and q^-.

By interpreting the molecular descriptors in the regression model, it is possible to gain some insight into structural features that are likely to govern toxicity of the studied compounds, which could be used to study toxicity of structurally-related compounds. There are three descriptors included in the regression model, which proved to be important features and make statistically-significant contribution to the model.

As indicated by the higher *SR* and *t*-test values, log K_{ow} appeared as the most significant descriptor for the derived QSAR model. Thus we can infer that the partition coefficient (log K_{ow}) is the most important descriptor for the toxicity. The developed model suggests that higher lipophilicity results in higher toxicity, which was in agreement with the previous reports of Aptula et al. (2005). Lipophilicity is very important for organic compounds to permeate, transport to and bioaccumulate in organisms. The diffusion of organic compounds across biological membranes is regulated by both the lipid membrane and the nonmoving aqueous solvent layer at both the inside and outside surfaces of the membrane (McKim et al. 1985). The log K_{ow} generally simulates an organic compound's lipophilicity, which is important in describing the passage of an organic compound through membranes. Organic compounds with higher log K_{ow} are more likely to exhibit transmembrane uptake into cells, enriching the intracellular concentration of the substance, where binding with receptors ultimately results in loss of membrane integrity and cell necrosis.

Dipole moment (μ) follows log K_{ow} in importance for its contribution to the output of this QSAR model. As an electronic descriptor, μ characterizes the average charge separation in a molecular system (Cronin and Livingstone 2004), and can represent the electronic information of compounds. Furthermore, μ can partially reflect molecular polarity. According to the model, μ has favorable contribution towards the toxicity value as evidenced by the positive regression coefficient. The higher μ value, the easier these phenols to participate in certain dipole–dipole or polar types of interaction with targets in cells, and leading to greater toxicity. The positive correlation between toxicity and μ was consistent with the previous reports of Zhang et al. (2008).

The most negative atomic charge (q^-) was also involved in toxicity model but had a smaller contribution compared to the other descriptors. The magnitude of q^- may characterize atomic charges, which are related to the reactive centers of chemicals (Niu and Yu 2004). Here q^- is negative, showing that decreasing the atomic charge produces stronger binding to the active site and therefore potentially enhancing toxicity. Moreover, chemicals with lower q^- have stronger electron-donating ability, and more easily form hydrogen bonds with target molecules.

The molecular descriptors used in QSAR model demonstrates that the mechanism underlying the toxicity of phenols is mainly related to their hydrophobic and electronic properties. The examination of the molecular descriptors can lead to a better understanding of the relation between structure and toxicity of the phenols. From the above discussion, it appears that hydrophobic interactions govern the toxicity of the phenols, and electronic interactions aid this course. In the first step, hydrophobic factor influences the process of membrane permeation and release of the phenols. In the second step, electrostatic interaction and other complicated electronic features influence interaction between phenol substances and target/receptors in cells.

In this study, a robust and stable MLR model was established for 18 phenols, the built model was assessed by internal and external validation, and the validations indicate that the QSAR model is adequate and satisfactory, and that the selected descriptors can provide an illustration of the contributing hydrophobic and electronic properties which are responsible for the toxicity of phenols. By interpreting the molecular descriptors in the regression model, we can conclude that increased log K_{ow} and μ, decreased magnitude of q^{-} are responsible for the greater toxicity of the studied compounds. The QSAR model developed in the present paper may be useful for providing insight into mechanisms of phenol toxicity to bacteria, as well as to other higher organisms.

Acknowledgments This work was supported by the National Natural Science Foundation of P. R. China (20737001, 20977046). All authors thank anonymous reviewers and editors for their valuable suggestions on revising and improving the work.

References

Aptula AO, Roberts DW, Cronin MT, Schultz TW (2005) Chemistry-toxicity relationships for the effects of di- and trihydroxybenzenes to *Tetrahymena pyriformis*. Chem Res Toxicol 18:844–854

Cronin M, Livingstone D (2004) Predicting chemical toxicity and fate. CRC Press, Boca Raton

Della GM, Monaco P, Pinto G, Pollio A, Previtera L, Temussi F (2001) Phytotoxicity of low-molecular-weight phenols from olive mill wastewaters. Bull Environ Contam Toxicol 67:352–359

Dewar MJS, Zoebisch EG, Healy EF, Stewart JJP (1985) AM1: A new general purpose quantum mechanical molecular model. J Am Chem Soc 107:3902–3909

Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR (2003) Gaussian 03, Revision A.1. Gaussian, Inc., Pittsburgh

Habibi-Yangjeh A, Danandeh-Jenagharad M, Nooshyar M (2006) Application of artificial neural networks for predicting the aqueous acidity of various phenols using QSAR. J Mol Model 12:338–347

Jiang L, Lin ZF, Hu XL, Yin DQ (2010) Toxicity prediction of antibiotics on luminescent bacteria, photobacterium phosphoreum, based on their quantitative structure-activity relationship models. Bull Environ Contam Toxicol 85:550–555

Jin B, Liu C, Jin Q (2010) Quantitative structure-activity relationship for heterogeneous phenol compounds using zero point energy. Chin. J Struct Chem 29:1353–1361

Jing GH, Li XL, Zhou ZM (2010) Quantitative structure-activity relationship (QSAR) study of toxicity of substituted aromatic compounds to *Photobacterium phosphoreum*. Chin. J Struct Chem 29:1189–1196

McKim J, Schmieder P, Veith G (1985) Absorption dynamics of organic chemical transport across trout gills as related to octanol-water partition coefficient. Toxicol Appl Pharm 77:1–10

Niu JF, Yu G (2004) Molecular structural characteristics governing biocatalytic oxidation of PAHs with hemoglobin. Environ Toxicol Phar 18:39–45

Su LM, Yuan X, Mu CF, Yang JC, Zhao YH (2008) Evaluation and QSAR study of joint toxicity of substituted phenols and cadmium to *Photobacterium phosphoreum*. Chem Res Chin. U 24:281–284

Tang QY, Feng MG (2007) DPS data processing system: Experimental design, statistical analysis, and data mining. Science Press, Beijing

Verschueren K (2001) Handbook of environmental data on organic chemicals, 4th edn. Wiley, New York

Xu HY, Yu QS, Zou JW, Wang YH, Wang HQ, Chen XS (2006) A QSRR study on the relative retention time of halogenated methyl-phenyl ethers. Chin. J Struct Chem 25:811–817

Yu RL, Lin XY, Hu GR (2009) The joint toxicity of phenols to *Photobacterium phosphoreum*. J Huaqiao U 30:549–552

Zeng M, Lin ZF, Yin DQ, Zhang YL, Kong DY (2011) A K_{ow}-based QSAR model for predicting toxicity of halogenated benzenes to all algae regardless of species. Bull Environ Contam Toxicol 86:565–570

Zhang HJ, Zhang JY, Zhu YM (2008) In vitro investigations for the QSAR mechanism of lymphocytes apoptosis induced by substituted aromatic toxicants. Fish Shellfish Immun 25:710–717