Journal of Computational Science & Engineering 16 (2015) 554-558

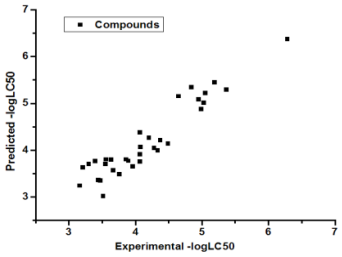
Available at http://www.asocse.org ISSN 1710-4068

QSAR model for the LC50 of nitrobenzenes

Fucheng Song[1]*, Cuizhu Ge[2], Huazhen Yang[1], Hongzong Si[1]

[1]Department of Public Health, Qingdao University, Qingdao, 266071, Shandong, China

[2]School of Pharmacy, Qingdao University, Qingdao, Shandong 266071, China



Abstract: **In the study of toxicology, animal experiments have given the toxicity of nitrobenzene compounds. Similarly, in the environment of nitrobenzene has potential harm to human body. In order to further study the characteristics of nitrobenzene, according to its physical and chemical properties, through the different method to establish QSAR model to predict the toxicity. In our study, a nonlinear quantitative structure–activity relationship model for the prediction of the LC50 of nitrobenzene was developed by the gene expression programming (GEP). This GEP model is established basis on four descriptors which were selected from the descriptor list by heuristic method (HM). A nonlinear, four-descriptor model based on GEP with mean-square error 0.065 and a correlation coefficient (R^2) 0.906 for the training set, mean-square error 0.043 and a correlation coefficient (R^2) 0.889 for the test set. The results indicate that this QSAR model has stable predictive capability of nitrobenzene toxicity and it provides a new method for toxicology research. The QSAR used in toxicology, will greatly reduce the harm of experimental animals.**

Keywords: Gene expression programming; LC50; Toxicology; Quantitative structure activity relationship; Nitrobenzene

Received 16 December 2014, Revised 20 January 2015, Accepted 24 January 2015

* Corresponding Author: Fucheng Song, 1032479764@qq.com

1. Introduction

With the rapid development of the modern chemical industry, new increasingly used for the synthesis of organic compounds, especially in the area of agriculture and biological medicine. New compounds are released into the environment, and then enter the body. So the impact on human health needs further study. To this title, we need to find more suitable for new method for evaluation of chemical toxicity [1].

The quantitative structure-activity relationship (QSAR) method is based on the chemical structure information and using quantitative parameters alone or combining with quantitative conventional parameters to construct the mathematical model, and using this model to forecast the toxicity of some unknown toxic compounds [2-6]. In the future, QSAR can aid the traditional animal in vivo tests in the studying of toxicology [7]. QSAR is a tool of investigation for the relationship between of physicochemical and biological behavior. However, hinder the development of this new field is the lacking of reliable experimental data of physicochemical and biological activity of compounds [8-10].

Nitrobenzene is a very important chemical raw materials and chemical intermediates that is widely used in pharmaceutical industry, chemical dyes, agriculture and other fields. Nitrobenzene is also internationally recognized as hazardous chemicals and one of the environment persistent organic pollutants. Nitrobenzene was deemed as the pro-electrophiles producing the corresponding potentially toxic C-nitroso-compounds [11]. Acute toxicity test showed that aquatic life is sensitive to nitrobenzene and toxic symptoms are earlier and more strongly [12]. The US EPA has its rules for priority monitoring because nitrobenzene are mutagenic and carcinogenic compounds [13]. To our knowledge, the QSAR study plays a key role for nitrobenzene compounds toxicity.

The gene expression programming (GEP) is a function of good analysis modeling software, in describing the molecular structure also depends on the CODESSA analysis software [14]. GEP has been successfully used in many areas, but seldom used in chemical analysis. The purpose of this study is to adopt the GEP method to predict the nitrobenzene compounds toxicity LC50 relationship with its molecular parameters. And then, a reliable QSAR model was set up. In the future, this model can be used to predict the toxicity of nitrobenzene compounds, and open up a new way for toxicology studies of nitrobenzene compounds.

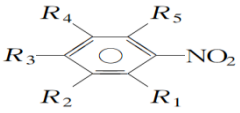
2. Data and Method

2.1. Data

The structures and LC50 values for 35 nitrobenzenes were taken from the literature [15] and are listed in Table 1. This research takes - logLC50 as indicators of toxicology. The greater of the value, the nitrobenzenes are more toxic. 35 compounds were stochastically fallen into 25 training set and 10 test set. QSAR model is built according to the training set, and used this QSAR model to predict the numerical test set.

2015 © The American Computational Science Society. All rights reserved.

Table 1 Experimental and predictive value –logLC50 of the nitrobenzene compounds (GEP)



Compounds	R_1	R_2	R_3	R_4	R_5	-logLC50 (exp)	-logLC50 (pre)	Resid.
1	Me	H	H	H	H	3.57	3.60	-0.03
2*	H	Me	H	H	H	3.63	3.54	0.09
3	H	H	Me	H	H	3.76	3.76	0.00
4	NO_2	H	H	H	H	5.45	5022	0.23
5	H	NO_2	H	H	H	4.38	4.27	0.11
6	H	H	NO_2	H	H	5.22	5.17	0.05
7	NO_2	Me	H	H	H	5.01	4.85	0.16
8	Me	H	H	NO_2	H	3.75	4.01	-0.26
9*	Me	H	NO_2	H	H	5.15	4.62	0.53
10*	Me	NO_2	H	H	H	3.99	4.25	-0.24
11	NO_2	H	Me	H	H	5.08	4.85	0.23
12	H	Me	H	NO_2	H	3.91	4.01	-0.10
13*	H	NO_2	H	NO_2	H	5.29	5.39	-1.00
14	H	H	H	H	H	3.02	3.73	-0.71
15	NH_2	H	H	H	H	3.70	3.46	0.24
16*	NH_2	H	H	NO_2	H	4.07	4.16	-0.09
17	H	H	OH	H	H	3.36	3.20	0.16
18	H	H	F	H	H	3.70	3.40	0.30
19	Me	NO_2	H	NO_2	H	4.88	4.68	0.20
20	NO_2	Me	NO_2	H	H	6.37	5.90	0.47
21	Me	NH_2	H	H	H	3.48	3.58	-0.10
22*	H	Me	NH_2	H	H	3.24	3.38	-0.14
23*	H	NH_2	Me	H	H	3.35	3.36	0.01
24	NH_2	Me	H	H	H	3.80	3.76	0.04
25	NH_2	H	Me	H	H	3.80	3.55	0.25
26*	NH_2	H	H	Me	H	3.79	3.66	0.13
27	OH	H	H	NH_2	H	3.65	3.81	-0.16
28	Me	H	H	NH_2	H	3.77	3.73	0.04
29*	OH	H	H	NO_2	H	4.04	4.04	0.00
30	Me	NH_2	H	NO_2	H	4.14	4.47	-0.34
31	Me	NH_2	NO_2	H	H	5.34	4.89	0.45
32*	NH_2	H	H	NO_2	Me	4.26	4.10	0.15
33	NH_2	NO_2	Me	H	H	4.21	4.18	0.03
34	NH_2	H	NO_2	Me	H	4.18	4.32	-0.14
35	Me	NO_2	H	NH_2	H	4.46	4.63	-0.17

The star '*' is test set.

2015 © The American Computational Science Society. All rights reserved.

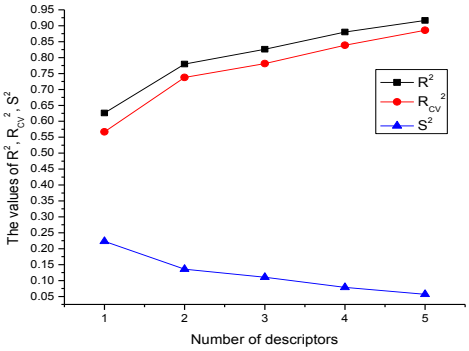


Figure 1. R^2, R^2cv and S^2 along with the change of the descriptors.

In this study, all nitrobenzene compounds are built molecular structural formula in Chemdraw software. Firstly, the geometry optimization is in the Hyperchem software, the calculation of molecular mechanics is with MM^+. The Semi-empirical AM1 method can get more accurate optimization. New data will be further precisely calculated in the MOPAC [16]. Then, the HIN files generated by geometrical optimization, the MNO files generated by MOPAC calculation and the -logLC50 values will be passed in CODESSA software to calculate the optimum molecular descriptors. There are five groups of descriptors: Constitutional descriptors, Geometrical descriptors, Electrostatic descriptors, Thermodynamic descriptors and Topological descriptors.

2.2. Molecular descriptors

Table 2 Corresponding physicochemical meaning and statistical parameters

Symbol	Physical-chemical meaning	Coefficient	Standard error	T-test
CIC_2	Complementary Information content (order 2)	0.038	0.006	5.72
MCIHO	Min coulombic interaction for a H-O bond	-9.39	1.36	-6.89
FNSA1	FNSA-1 Fractional PNSA [Zefirov's PC]	-5.79	1.33	-4.33
MNSEC	Max atomic state energy for a C atom	-2.94	0.714	-4.12

3. Results and Discussion

3.1. The HM model

487 descriptors were calculated and analyzed by the CODESSA program for all the nitrobenzene compounds. At the same time, the CODESSA software can establish a linear HM model. From the study, we can see that R^2, R^2cv and S^2 changes correspondingly with the increase of molecular descriptors number (Figure 1.). The values of R^2, R^2cv and S^2 tend to be stable while selecting 4 descriptors. To select the set of descriptors that are most relevant to the LC50 of nitrobenzene that the linear models with 4 variables were built. The corresponding physicochemical meaning and statistical parameters are in the table 2.

$$-\log LC50 = 0.038059CIC_2 - 9.3996MCIHO - 5.7996FNSA1 - 2.9463MNSEC \quad (1)$$

$R^2=0.880$ $R^2_{CV}=0.838$ $S^2=0.0787$ $F=55.1$

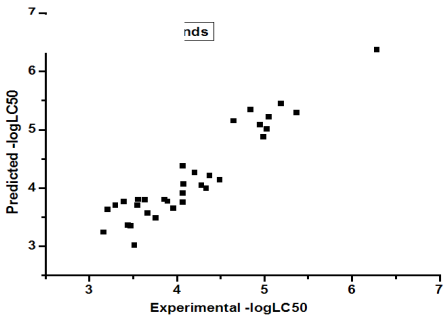


Figure 2. Experimental and predictive value –logLC50 by HM.

Figure 2. shows the linear plots of all compounds on the basis of the 4-parameter model.

There will be a co-linearity of molecular descriptors. In order to avoid the result, the correlation coefficient among the various molecular descriptors is not more than 0.8 [17]. Table 3 lists the correlation coefficient of the molecular descriptors which were calculated by SPSS software. Evidently, the correlation coefficients do not exceed 0.8. Four molecular descriptors are selected in the HM model.

IC is based on the theory of Shannom information of molecular descriptors. IC_2 is the second grade classification sphere near the radius of atomic, Eq. (2), (3). CIC_2 values are positively correlated to the toxicity of nitrobenzene. Toxicity of nitrobenzene will fortify while the values of CIC2 increase.

$$IC_2 = -\sum_{i=1}^{2} \frac{n_i}{n} \log_2 \frac{n_i}{n} \quad (2)$$

$$CIC_2 = \log_2 n - IC_2 \quad (3)$$

MCIHO describes the min coulombic interaction for H-O bond, which can reflect the coulombic interaction between the H and O atoms. Toxicity of nitrobenzene will descend while the values of MCIHO increase. The interaction between nitro O atoms with a wide range of H atoms in animals will reduce toxicity of nitrobenzene. But the increase of nitro can enhance the toxicity of nitrobenzene that related to the interaction between N and O atoms. Toxicities of the nitrobenzene 5, 8, 13 (table 1) are obviously greater than having no nitro compounds. Methemoglobin formed by the reaction of nitro with hemoglobin will cause hemolysis in body. At the same time，the nitro acting on the liver cells contributes to hepatic parenchymal lesions.

2015 © The American Computational Science Society. All rights reserved.

FNSA1 stands for Fractional partial negative surface area, which are one of Geometrical descriptors. FNSA1 is an important molecular descriptor in the study of the QSAR and QSPR, which can be elected by the semi-empirical method. In Eq. (1), FNSA1 values are negatively correlated to the toxicity of nitrobenzene. In CODESSA software, the formula for FNSA is as follows [18]:

$$FNSA1 = \frac{PNSA\ 1}{TMSA} \quad (4)$$

PNSA1-partial negatively charged molecular surface area.

TMSA-total molecular surface area.

Max atomic state energy for a C atom values are also negatively correlated to the toxicity of nitrobenzene. Nitrobenzene compounds with methyl will increase the max atomic state energy for a C atom. Toxicity of compound 1, 2, 3 with methyl lower than that of the other nitrobenzene compounds. Low toxicity of nitrobenzene compounds with methyl can also be obtained from the Eq. (1). Therefore, more evidences can demonstrate the reliability of the HM model.

Table 3 Correlation of molecule descriptors in the model

	CIC_2	MCIHO	FNSA1	MNSEC
CIC_2	1.00	-0.318	-0.040	-0.457
MCIHO		1.00	0.104	-0.172
FNSA1			1.00	-0.312
MNSEC				1.00

3.2. The GEP model

GEP model adopts the same molecular descriptors with HM model. The software automatic problem solver (APS) [19] allows the easy optimization of intermediate solutions and the easy testing of the evolved models against a test set. The elements {+,-, *, /, Mod, log, sin} are used as the function set, the elements {a, b, c, d} as the final set. Various setting parameters involved in the GEP are given in Table 4. The steps for GEP are shown in Figure 3. This article uses the sum of squared residuals as a fitness function:

$$Fitness = \sum_{j=1}^{N} (Y_j - \hat{Y}_j)^2 \quad (5)$$

where Yj and Yj are the real and predicted output values for the jth sample respectively; N: the number of nitrobenzene compounds.

In this run, made a very satisfactory solution in the 3th generation with a R^2 0.90 and S^2 0.043 for the training set and a R^2 0.89 and S^2 0.065 for the test set. Figure 4 and 5 showed the fitting situation of training set and test set. The ETs can be expressed as Eq. (6).

$$y = \mathrm{Mod}\{\mathrm{Mod}[\mathrm{Sin}(X_2 + X_3), (X_2 \times X_3 \times X_3)], X_3\} + \frac{X_2 \times X_3}{X_3 \times X_4} - \frac{\log X_3}{X_1} + \mathrm{Mod}\{[\mathrm{Sin}(2X_2 + X_4) - X_3], X_2\} + \mathrm{Mod}\left[\frac{\mathrm{Sin}\,(X_4+2X_2)}{X_3}, X_1\right] + \log[(\frac{X_3}{X_2} - X_3) \times X_2 + X_1] \quad (6)$$

Table 4 Parameters for the simple symbolic regression problem

Parameter names	Values
Generations number	100
Gene head size	8
Number of genes	5
Linking function	Addition
Generations without Change	200
Number of Tries	3
Max Complexity	5
Mutation	0.044
Inversion	0.1
IS Transposition	0.1
RIS Transposition	0.1
One-Point Recombination	0.3
Two- Point Recombination	0.3
Gene Recombination	0.1
Gene Transposition	0.1
Selection range	100
Precision	0.01

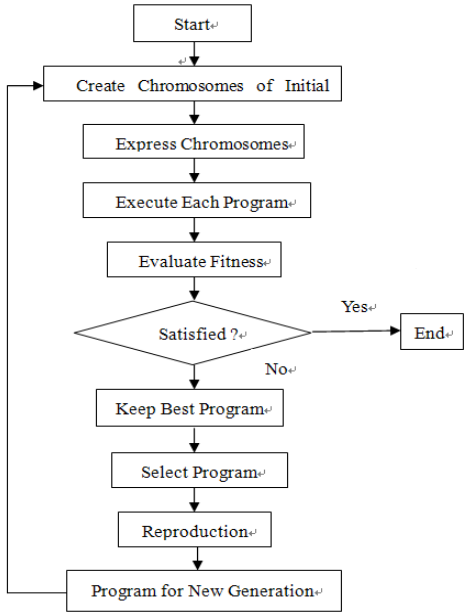


Figure 3. The brief flowchart of GEP

In the Eq. (6) X1, X2, X3, X4 represent variables of CIC2, MCIHO, FNSA1, MNSEC. GEP model can accurately predict the toxicity of nitrobenzene compounds with the error less than 0.1. However, the model for most of the toxicity of nitrobenzene predicted values are lower than the experimental value. This may be related to MCIHO molecular descriptors in the model that due to the excessive negative effect.

2015 © The American Computational Science Society. All rights reserved.

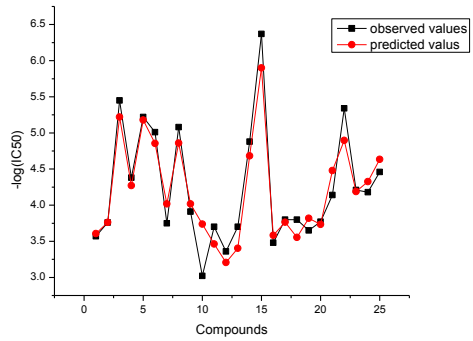


Figure 4. Fitting curve of experimental and predictive value -logLC50 in training set.

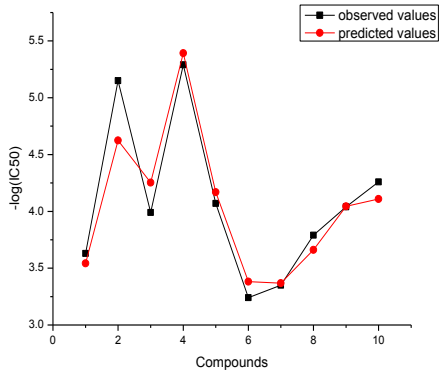


Figure 5. Fitting curve of experimental and predictive value -logLC50 in test set.

3.3. The GEP compares with HM

In GEP model, R^2 of the training set 0.906 and the test set 0.899 are both greater than R^2 of HM 0.880, but S^2 of the training set 0.043 and the test set 0.065 are both less than S^2 of HM 0.079. Therefore, the predicted ability of GEP is better than HM. This is because the best provided by this technique. This way is determined after controlling numerous preliminary models from the millions of linear and nonlinear models.

Above all, it is obvious that nitrobenzene compounds toxicity is influenced by physical and chemical properties. For example, the min coulombic interaction for H-O bond can descend the toxicity. Molecular descriptors selection is the priority among priorities of the GEP model building. In this study, the 4 molecular descriptors are the most important physical and chemical parameters for GEP model to predict the values of –logLC50.

4. Conclusion

Studies have shown that the GEP is very efficient tools to build a nonlinear model and a very effective to predict the toxicity of nitrobenzene compounds. This method can be used in the analysis of toxic compounds. At the same time, the models proposed could identify and provide some methods about what structural features are connected with the biological activity of these compounds. In toxicology, QSAR method can replace in vivo tests to predict toxicity of a kind of compound that could result in the reduction of numerous unnecessary acute in vivo toxicity tests. And this model also focuses in vivo testing on cases with potentially higher risks for humans.

Acknowledgements

This work was supported by Qingdao University Innovative Teaching Laboratory Research Project and the Shandong Province Department of Project: No. 2012YD18038 and No. 2012YD18042.

References

[1] M. Smiesko, E. Benfenati, Chem. Inf. Comput, 44 (2004) 976.
[2] I. A. Doichinoova, R. N. Natcheva, D.N. Mihailova, Eur J Med Chem, 29 (1994) 133.
[3] Y. H. Qi, L. Xu, Chem J Chinese Universities, 21 (2000) 219-221.
[4] H. Huang, X.D. Wang, Chemosphere, 53 (2003) 963-970.
[5] S.J. Ren,Chemosphere, 53 (2003) 1053-1069.
[6] S. Trohalaki, E. Gifford, R. Pachter, Computers and Chemistry, 24 (2000) 421-427.
[7] P. Freidig, S. Dekkers, M. Verwei, E. Zvinavashe, J. G. M. Bessems, J. J. M. van de Sandt, Toxicology Letters 170 (2007) 214-222.
[8] Andrey, Toropov, Alla P. Toropova, Chemosphere 124 (2015) 40-46.
[9] H. Gonz áez-D áz, S. Arrasate, A. J. G ómez-San, N. Sotomayor, E. Lete, Top. Med. Chem, 13 (2013) 1713-1741.
[10] D. A. Winkler, E. Mombelli, A. Pietroiusti, L. Tran, A. Worth, B. Fadeel, M. J. McCall, Toxicology 313 (2013) 15-23.
[11] V. K. Agrawala, P. V. Khadikar, Bioorganic & Medicinal Chemistry, 9 (2001) 3035-3040.
[12] P. Z. Lang, X. F. Ma, G. H. Lu, Chemosphere, 32 (1996) 1547.
[13] X. J. Cui, Z. Y. Zhang, X. Yuan, Z. M. Su, S.D. Liu, Computers and Applied Chemistry, 11 (2005) 1038-1040.
[14] H. Z. Si, T. Wang, K. J. Zhang, Z.D. Hua, B. T. Fan, Bioorganic & Medicinal Chemistry, 14 (2006) 4834-4841.
[15] L. H. Hall, E. L. Manynard, L. B. Kier, Environ Toxico Chem, 4 (1989) 431.
[16] J. P. P. Stewart, No. 455, Indiana University, Bloomington, IN, 1989.
[17] K. J. Zhang, S. Q. Sun, C. L. Chai, Chiese Journal of Analytical Chemitry, 35 (2007) 1263-1268.
[18] D. T. Stanton, P. C. Jurs, Anal. Chem, 62 (1990) 2323.
[19] http://www.gepsoft.com/gepsoft.

2015 © The American Computational Science Society. All rights reserved.