Journal of Computational Methods in Molecular Design, 2016, 6 (3):64-69

ISSN : 2231- 3176
CODEN (USA): JCMMDA

Semi-empirical (PM3) Based Insillico Prediction of Acute Toxicity of Phenols

Ibraheem Wasiu Aderemi[1]*, Jangber Zaphania Nicodemus[2], Ojilere Chileziemanya Juliet[1] and Ezekiel Malgwi Anjili[1]

[1]Department of Chemistry, Ahmadu Bello University, Zaria Nigeria
[2]Department of Veterinary Anatomy, Ahmadu Bello University, Zaria Nigeria

---

ABSTRACT

A toxicity data set of 58 phenols to Tetrahymena pyriformis expressed as pEC_{50} (Log to base 10 of EC_{50}) was taken from literature. 70% (41 phenols) of the data was used as training set while 30% (17 phenols) was used as test set. Muti-linear Regression equations were built using the experimental pEC_{50}as dependent variable and the various molecular descriptors as independent variables. The best Quantitative structure-toxicity relationship (QSTR) model hinted that the toxicity of phenol was dominantly influenced by octanol-water partition coefficient (XlogP) and moment of inertia (MOMI) descriptors. The results of the statistical analysis of the two parameter model include; n = 41, LOF score = 0.079, R^2 = 0.6691, $R^2_{adj.}$ = 0.6517, Q^2_{LOO} = 0.6260, F-value = 38.42. The generated QSTR model has been proven to possess statistical significance, high predictive power and wide applicability domain. Thus, it is recommended for environmental risk assessment of phenols.

Keywords: QSTR, Phenols, Toxicity, Tetrahymena pyriformis, XlogP, MOMI.

INTRODUCTION

Phenols compounds are basic materials for industry production, which are commonly used in chemical synthesis. They can spread through air and water, with strong carcinogenicity, teratogenicity and mutagenicity [1-2], which will cause great damage to environment, plants, animals and human health. In view of the health implication associated with the pollution of the environment with these compounds, their quantitative risk assessment becomes increasingly important in the modern society and is slowly incorporated into legislation of different countries [3]. For instance, the European Union (EU) has introduced the Registration, Evaluation and Authorization of Chemicals (REACH) program for assessment of human and environmental risk of all chemicals that are produced or imported in the amount greater than 1 ton per year [3]. It is clear that if such a risk assessment is performed purely experimentally, it would require a huge amount of resources as well as time. Therefore, the introduced program encouraged the use of QSTR modeling and other alternatives especially for the risk assessment of chemicals that are produced or imported in smaller quantities [3]

QSTR modelling is based on the idea that all the information related to a molecule can be derived from its chemical nature by means of parameters that encode or describe different molecular features, and these parameters, or descriptors, can be correlated to a particular chemical or biochemical activity, i.e

Activity = F(structure) = $F(\cdot)$ [4].

The aim of this research is to construct a rational Quantitative structure-toxicity relationship model for predicting the acute toxicity of phenols.

MATERIALS AND METHODS

A toxicity data set of 41 phenols to *Tetrahymena pyriformis* expressed as pEC_{50} (Log to base 10 of EC_{50}) was taken from literature. The entire data set and their respective pEC_{50} are presented in Table 1. 70% (41 phenols) of the data was used as training set while 30% (17 phenols) was used as test set

2.1 Molecular descriptors

All computations were performed by using *Spartan' 14 software* (Spartan wave function, 2014). The geometries of all 58 phenols were optimized with Semi-empirical (PM3). *Padel descriptortool kit* was used to compute the molecular descriptors of each optimized molecules.

2.2 Statistical analysis

Genetic Function Approximation (GFA) is a widely-used statistical analysis method. In this paper, *Material Studio* statistical software was employed to realize the GFA analysis. The linear relationship between the toxicity data of the compounds and their structure parameters was fitted by GFA-multiple stepwise regression method at 95% confidence intervals.

It is a distinctive characteristic of GFA that it could create a population of models rather than a single model. GFA algorithm, selecting the basic functions genetically, developed better models than those made using stepwise regression methods [5] and then, the models were estimated using the "lack of fit" (LOF), which was measured using a slight variation of the original Friedman formula in equation 1, so that best model received the best fitness score [6].

$$LOF = SSE / (1 - \frac{C+dp}{M})^2 \qquad \text{Eqn. 1}$$

Table 1: Experimental pEC_{50} of the phenols studied and their CAS registry number

Compound	Name	**CAS#**	Toxicity
1	4-Fluorophenol	000371-41-5	0.02
2	2-Cyanophenol	000611-20-1	0.03
3	5-Fluoro-2-hydroxyacetophenone	000394-32-1	0.04
4	2,4-Dimethylphenol	000105-67-9	0.07
5	2-Hydroxyacetophenone	000118-93-4	0.08
6	2,5-Dimethylphenol	000095-87-4	0.08
7	3,5-Dimethylphenol	000108-68-9	0.11
8	4'-Hydroxypropiophenone	000070-70-2	0.12
9	2,3-Dimethylphenol	000526-75-0	0.12
10	3,4-Dimethylphenol	000095-65-8	0.12
11	2-Ethylphenol	000090-00-6	0.16
12	2-Chlorophenol	000095-57-8	0.18
13	4-Hydroxy-2-methylacetophenone	000875-59-2	0.19
14	4-Ethylphenol	000123-07-9	0.2
15	3-Ethylphenol	000620-17-7	0.23
16	2,3,6-Trimethylphenol	002416-94-6	0.28
17	2,4,6-Trimethylphenol	000527-60-6	0.28
18	2-Hydroxy-5-methylacetophenone	001450-72-2	0.31
19	2-Bromophenol	000095-56-7	0.33
20	5-Bromo-2-hydroxybenzyl alcohol	002316-64-5	0.34
21	2,3,5-Trimethylphenol	000697-82-5	0.36
22	2-Chloro-5-methylphenol	000615-74-7	0.39
23	4-Allyl-2-methoxyphenol	000097-53-0	0.42
24	2-Hydroxybenzaldehyde	000090-02-8	0.42
25	2,6-Difluorophenol	028177-48-2	0.47
26	4-Cyanophenol	000767-00-0	0.52
27	4-Propyloxyphenol	018979-50-5	0.52
28	4-Chlorophenol	000106-48-9	0.55
29	5-Methyl-2-nitrophenol	000700-38-9	0.59
30	2-Bromo-4-methylphenol	006627-55-0	0.6
31	2,4-Difluorophenol	000367-27-1	0.6

32	3-Isopropylphenol	000618-45-1	0.61
33	2-Chloro-4,5-dimethylphenol	001124-04-5	0.69
34	4-Butoxyphenol	000122-94-1	0.7
35	4-Chloro-2-methylphenol	001570-64-5	0.7
36	3-tert-Butylphenol	000585-34-2	0.73
37	4-Chloro-3-methylphenol	000059-50-7	0.8
38	4-Iodophenol	000540-38-5	0.85
39	2,2'-Biphenol	001806-29-7	0.88
40	4-tert-Butylphenol	000098-54-4	0.91
41	3,4,5-Trimethylphenol	000527-54-8	0.93
42	4-sec-Butylphenol	000099-71-8	0.98
43	2,4-Dichlorophenol	000120-83-2	1.04
44	4-Chloro-3-ethylphenol	014143-32-9	1.08
45	2-Phenylphenol	000090-43-7	1.09
46	3-Chloro-4-fluorophenol	002613-23-2	1.13
47	6-tert-Butyl-2,4-dimethyIphenol	001879-09-0	.16
48	4-Chloro-3,5-dirnethyIphenol	00088-04-0	1.2
49	4-Cyclohexylphenol	001131-60-8	1.56
50	3,4-Dinitrophenol	000577-71-9	0.27
51	2,6-Dinitrophenol	000573-56-8	0.54
52	2,6-Dichloro-4-nitrophenol	000618-80-4	0.63
53	2,5-Dinitrophenol	000329-71-5	0.95
54	4-Bromo-2-fluoro-6-nitrophenol	000320-76-3	1.62
55	2-Amino-4-nitrophenol	061702-43-0	0.47
56	2,6-Diiodo-4-nitrophenol	000305-85-1	1.71
57	3-Fluoro-4-nitrophenol	000394-41-2	0.94
58	4-Hexyloxyphenol	018979-55-0	1.64

2.3 Validation and evaluation

Testing the stability, predictive power and generalization ability of the models is a very important step in QSAR study. As for the validation of predictive power of a QSAR model, two basic principles (internal validation and external validation) are available. The cross-validation is one of the most popular methods for internal validation. In this paper, the internal predictive capability of the model was evaluated by leave-one-out cross-validation (Q^2_{LOO}). Q^2_{LOO} of 0.5 and above is an indication that the QSAR model is robust and highly predictive [7].

RESULTS AND DISCUSSION

3.1 QSAR models and analysis

The best three GFA-MLR models are depicted my Models 1, 2, and 3. Based on the model with the least lack of fit (LOF) score, Model 1 was selected as the optimum QSAR model for predicting the toxicity of phenols.

Model 1:

$$pEC_{50} = 0.282582519\ XlogP + 0.000315219\ MOMI - 0.212957547$$

n = 41, LOF score = 0.079, R^2 = 0.6691, $R^2_{adj.}$ = 0.6517, Q^2_{LOO} = 0.6260, F-value = 38.42

Model 2:

$$pEC_{50} = 0.338953390\ XlogP + 0.000747736MOMI - 0.046790507\ WV.mass - 0.056557175$$

n = 41, LOF score = 0.081, R^2 = 0.7076, $R^2_{adj.}$ = 0.6839, Q^2_{LOO} = 0.6544, F-value = 29.85

Model 3:

$$pEC_{50} = 0.003370712\ Mw + 0.335758244\ XlogP - 0.554134121$$

n = 41, LOF score = 0.085, R^2 = 0.6454, $R^2_{adj.}$ = 0.5900, Q^2_{LOO} = 0.5900, F-value = 34.58

The definition of the descriptors in the models include; **MOMI** = moment of inertia along, **Mw** = molecular weight, **WV. Mass** = non-directional WHIM weighted by atomic mass, **XlogP** = measure of octanol-water partition coefficient. In the equation, *n* is the number of compounds, R^2 is the multiple correlation coefficient, R^2_{adj} is adjusted R^2, F stands for significance of regression

The high coefficient of determination (R^2) is an indication that the model explained a very high percentage of the response variable (descriptor) variation, high enough for a robust QSAR model. The high adjusted R^2 (R^2_{adj}) value and its closeness in value to the value of R^2 implies that the model has excellent explanatory power to the descriptors in it. Also, the high and closeness of Q^2 value to R^2 revealed that the model was not over-fitted. F value judges the

overall significance of the regression coefficients. The high F value of the model is an indication that the regression coefficients are significant. The high predictability of model 1 is also evidenced by the low residual values observed in Table 2 which gives the comparison of observed and predictedpEC_{50} of the molecules.

A good predictive ability of the model 1 for the training and test set compounds is depicted by Fig. 1 which gives the plot of predicted values of the test and training sets and their experimental values. Also, Fig. 2 gives the residual plot of the optimum model. Most of the calculated residuals are distributed on two sides of the zero line, a conclusion may be drawn that there is no systematic error in the development of the present model.

Based on model 1, the main factors that could impact the biological toxicity of phenol include **XlogP** (octanol-water partition coefficient) and **MOMI** (moment of inertia of a molecule). According to statistic learning theory, comparing the importance of each parameter entails the knowledge of the standardize coefficient of them in the regression equation. The bigger the absolute value of the standardized coefficient, the greater the influence of the parameter. In the equation, the standardized coefficient of **XlogP** and **MOMI** are 0.2825 and 0.0003, respectively. The two parameters describes the lipophilicity and molecular resistance to changes in the rotation direction. Their positive coefficient implies that the **EC_{50}** of the studied molecules increases with the increase in values of these descriptors in the molecule.

Table 2: Comparison of actual pEC_{50} and pred. pEC_{50} of the Training set using model 1

Compound	Actual pEC_{50}	Predicted pEC_{50}	Residual
2	0.0300	0.1090	-0.0790
3	0.0400	0.3260	-0.2860
4	0.0700	0.3800	-0.3100
6	0.0800	0.4408	-0.3607
7	0.1100	0.3887	-0.2787
8	0.1200	0.3072	-0.1872
12	0.1800	0.3796	-0.1996
13	0.1900	0.4446	-0.2546
15	0.2300	0.3673	-0.1373
17	0.2800	0.4705	-0.1905
18	0.3100	0.4590	-0.1490
19	0.3300	0.4149	-0.0849
20	0.3400	0.4760	-0.1360
21	0.3600	0.5306	-0.17064
23	0.4200	0.7625	-0.3425
24	0.4200	0.2127	0.2073
26	0.5200	0.0799	0.4401
27	0.5200	0.6399	-0.1199
28	0.5500	0.3408	0.2092
29	0.5900	0.6621	-0.0721
30	0.6000	0.5963	0.0038
31	0.6000	0.2270	0.3730
32	0.6100	0.6474	-0.0374
34	0.7000	0.9484	-0.2484
36	0.7300	0.7986	-0.0686
37	0.8000	0.4224	0.3776
38	0.8500	0.6249	0.2251
40	0.9100	0.8133	0.0967
41	0.9300	0.5909	0.3391
42	0.9800	0.9062	0.0738
43	1.0400	0.5976	0.4424
44	1.0800	0.5907	0.4893
45	1.0900	0.8465	0.2435
47	1.1600	0.9849	0.1751
49	1.5600	1.2435	0.3165
50	0.2700	0.4468	-0.1768
51	0.5400	0.9504	-0.4104
53	0.9500	0.7658	0.1842
55	0.4700	0.3085	0.1615
56	1.7100	1.7330	-0.0230
58	1.6400	1.6750	-0.03450

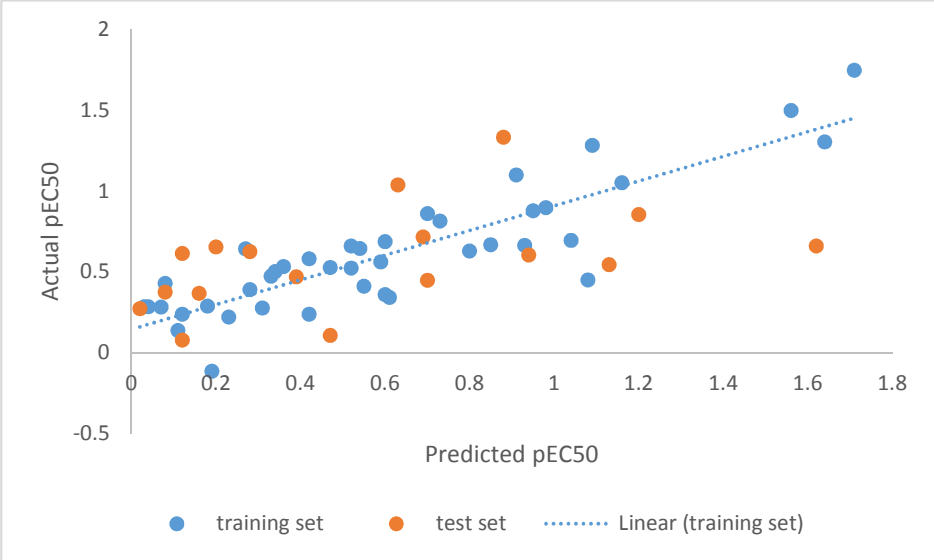

Fig. 1: Comparison between the predicted and experimental values of pEC_{50}

Fig. 2: Plot of the residuals *versus* the experiment pEC_{50}values of Model 1

CONCLUSION

According to the QSAR study, EC_{50} of halogenated phenols to *tetrahymena pyriformis* increases with the values of the descriptors; **XlogP** (octanol-water partition coefficient) and **MOMI** (moment of inertia of a molecule). Validation of the optimum model shows that it has good stability and great predictive power and as such can be of immense help in predicting the acute toxicity of phenols.

REFERENCES

[1] Y B Zang. *Chinese Agricl. Sci. Bull.*, **2012**, 28, 282-285.
[2] P R Zhan;H T Wang;Z X Chen Z.X.J. of Agro-Environ. Sci., **2008**, 27, 801-804.
[3] S Martin;T Alexandru; B Balaban;V Marjan; A M Pompe.*Commun. Math. Comput. Chem.*,**2016**, 75**,** 559-582

[4] C Luís;G. Branco;V S M Gonçalo; J A Carrera;M Ignacio;F Raquel;A Carlos A.M. Afonso. Retrieved Jan. 09, **2016** from *http://cdn.intechopen.com/pdfs/13913.pdf*.
[5] W Wu; C Zhang; W Lin;Q Chen; X Guo;Y Qian.*PLoSONE*, **2015**, 10, 3-16
[6] J F Friedman,*Multivariate Adaptive Regression Splines, Technical Report No. 102*, Stanford University, **1990.**
[7] V Ravichandran; R Harish; J Abhishek; S Shalini;P V Christapher; K A Ram. *Inter. J. of Drug Design and Discovery*, 2, **2011**, 511-519.