*International Research Journal of Environmental Sciences*_____ ISSN 2319–1414
Vol. 7(1), 22-27, January **(2018)** *Int. Res. J. Environmental Sci.*

Chemometric modeling of depruration rate constants of polycyclic aromatic hydrocarbons in mussels (Elliptiocomplanata)

Vandana Pandey
Department of Chemistry, Kurukshetra University, Kurukshetra-136119, Haryana, India
vandana_p71@rediffmail.com

Available online at: www.isca.in, www.isca.me
Received 30th October 2017, revised 13th January 2018, accepted 19th January 2018

Abstract

This paper describes the QSAR study to predict the depuration rate constants (k_d) of a series of polycyclic aromatic hydrocarbons (PAHs) for mussels, Elliptiocomplanata. The reported Logk_d values of 26 compounds have been mapped linearly by means of stepwise multiple linear regression and non-linearly by artificial neural network trained with Levenberg-Marquardt (LM) algorithm, using molecular descriptors derived online from mole-db software. Descriptors selected by SW-MLR were used to develop non-linear model. The models were validated for predictability by both internal and external validation. Both linear and non-linear models satisfy the criteria of external validation as recommended by Golbraikh and Trospha. Comparison of quality of best ANN (R^2 = 0.96) model with SW-MLR (R^2 = 0.94) model showed that ANN trained with robust LM algorithm has better predictive power. Applicability domain analysis has also revealed that the suggested models have acceptable predictability.

Keywords: Polycyclic aromatic hydrocarbons, biomagnification factor, QSAR, artificial neural network.

Introduction

PAHs (Polycyclic Aromatic Hydrocarbons) are class of more than 100 different organic compounds consists of three or more fused benzene rings containing only carbon and hydrogen[1]. They are produced by incomplete combustion or high-pressure processes occurring naturally by forest fires and volcanoes, but most PAHs in surrounding air are formed during incomplete combustion of coal, wood, petroleum, petroleum products, or oil, burning of polypropylene, or polystyrene and motor vehicle exhaust[2,3]. These are one of the most common persistent organic pollutants (POP) in water bodies, thereby pose a threat, not only to water ecosystem but also a human health risk as some PAH are known carcinogens[4,5]. The toxic effects of PAHs have prompted monitoring their pollution in water bodies. Direct monitoring of these trace pollutants is time consuming, complex and expensive in aquatic environment due to their extremely low concentration in water bodies. Therefore, for monitoring of such types of chemicals using biological indicators in water bodies have been well established[6].

Mussels like *Elliptiocomplanata*. (Bivalvia: Unionidae) are commonly used in environmental monitoring program in order to access water contamination. For chemicals risk assessment, Bioaccumulation – the accumulation of chemicals and other pollutants in living organisms - is an important parameter. In order to estimate biomagnification factor (BMF), depuration rate constant (k_d) is an important kinetic parameter for giving information about time required for the polluted mussels to reach a steady state in the environment and characterizing the depuration process[7-9]. Calculation of depuration constant from experimental results is expensive and time consuming. Therefore, to decrease the experimental cost and to fill the data gap of organic pollutant, QSAR can be used as an alternative approach.

QSAR models are mathematical equations, constructing linear and non-linear relationship between experimental activity and chemical structure presented in the form of descriptors[10]. Conventional Multiple linear regression (MLR), principal component regression (PCR) and partial least squares (PLS) regression are the most commonly used linear method in QSAR modeling. Recently, there has been a great deal of interest in artificial neural network (ANN) in many areas of science and technology such as engineering, biology, and in the field of cognitive sciences. The applications of the ANN also appear in several areas of sciences including the investigation of QSAR, as ANNs enable the investigation of complex non-linear relationships[11-15].

Numerous studies regarding QSAR modeling for predicting depuration rate constants k_d values in mussels based on the octanol/water partition coefficient (Kow) of PAHs have been reported[16,17]. In addition, Wu *et al*. have constructed QSAR models using quantum chemical descriptors and partial least squares (PLS) regression, to predict the depuration rate constants (kd) of polycyclic aromatic hydrocarbons (PAHs) for mussels, Elliptio complanata[18,19].

The purpose of the present study is to further enhance the quality of QSAR by finding suitable representative descriptors from a large pool of descriptors and developing linear and non-

linear QSAR models to map better relationship between descriptors and the depuration rate constants of polycyclic aromatic hydrocarbons in mussels, Elliptiocomplanata.

Methodology

All calculations presented in this work were carried out on a personal computer with a Window XP operating system. SPSS software[20] was used for SW-MLR analysis. ANN calculations were performed with Matlab software[21]. In the first step, the values of depuration rate constants (Log k_d) of 26 polycyclic aromatic hydrocarbons compounds in mussels, Elliptiocomplanata, were taken from literature[18] and used as dependent variable. Online resource MOLE db –Molecular Descriptor Data Base, developed by Milano Chemometrics and QSAR research group[22] was used to calculate molecular descriptors for each PAH. Total 1124 descriptors were generated for each molecule including constitutional, topological descriptors, connectivity indices, information indices, 2 Dautocorrelations, Burden Eigen values descriptors, Eigen value based descriptors, geometrical descriptors, WHIM, Getaway, functional group counts, atom-centered fragments and molecular properties. Because of large no of descriptors pool, the calculated descriptors were first analyzed to check the existence of constant and near constant variables, which were removed. Further variable-selection for the QSAR modeling was carried out by stepwise linear regression method. The best multiple linear regressions identified contained three descriptors gives information about linear relationship between selected PAHs and their Log k_d values. For internal validation Y-randomization test was performed by randomly shuffling the dependent variable while keeping the independent variables as it is. All three discriptors selected by SW-MLR method were used as input for generating ANN models to obtain non-linear models. Proposed QSAR models were also validated by an external prediction (validation and test) set, as recommended by Golbraikh and Tropsha[23]. According to Golbraikh et al, an ideal splitting leads to a prediction set in which each of its members is close to at least one point of the training set. For this purpose, dataset was divided into three subsets: training set, validation set and test setto improve generalization. The applicability domain was assessed by the normalized mean Euclidean distance value for each compound. Euclidean AD is based on distance scores calculated by the Euclidean distance norms. At first, normalized mean distance score for training set compounds ranges from 0 to 1(0=least diverse, 1=most diverse training set compound). were calculated and then normalized mean distance score for test set were calculated, and those test compounds with score outside 0 to 1 range are said to be outside the applicability domain[24-26].

Results and discussion

As the aim was to give advantage of the descriptors used in the present study over the already reported quantum chemical descriptors[18], large pool of descriptors was generated for selected compounds using MOLEdb facilities available online. For success of any QSAR study, an important key factor is the selection of appropriate molecular descriptors. To fulfill this purpose, descriptor selection was carried out by SW-MLR method. Many models were generated by using SW-MLR methods. The goodness of the correlation is tested by the regression coefficient (R^2), the standard error of the estimate (SEE) and the F-test[27]. The best selected model contained three descriptors namely AMR, BEHm6 and H2e and resulted in a strong correlation to experimental pIC_{50} values (R^2=0.947, SEE=0.063, R^2_{adj}=0.935, Q^2_{Loo}=0.867). The best linear equation for this QSAR is presented in Equation-1.

$$\text{Log } k_d = -0.11356 - 0.01409\text{AMR} + 0.2385\text{BEHm6} - 0.24231\text{H2e} \quad (1)$$

In the above equation, AMR = Ghose-Crippen molar refractivity (molecular properties), The Ghose-Crippen molar refractivity (AMR) is calculated using a similar group contribution approach. BEHm6 = Highest eigenvalue number 6 of Burden matrix / weighted by atomic masses (Burden Eigenvalue Descriptors) and H2e = H autocorrelation of lag 2 / weighted by Sanderson electro negativity (Getaway descriptors). It can be seen from the equation (1) that higher score on BEHm6 predicts higher score on Log k_d value but higher score on other two variable predicts lower score on Log k_d value. Annova table showed that for overall regression F=83.816 with a probability well below 0.05. So the regression is significant. The robustness of this model was checked by Y–randomization test by generating fifty random models (average r^2 value is 0.202). The low randomized r^2 values indicate that the results obtained from the linear mapping by SW-MLR method were not due to a chance correlation or structural dependency of the training set. The conventional calculated Logk_d values for the compounds from SW-MLR method along with experimental values are listed in the Table-1.

For implementing fully connected, three-layer, computational neural networks the SW-MLR selected three descriptors were used as the input neuron of the network, whereas, Log k_d values of PAH's were used as output value. A feed forward network trained with Levenberg-Marquardt algorithm was used, in which mean squared error (MSE) was applied as the performance function. The Levenberg–Marquardt (LM) algorithm is basically a fastest modern second-order Hessian-based algorithm for nonlinear least squares optimization. Before training, the input and output vectors were scaled to [-1, 1]. Data set was divided in such a way that ratio of training, validation and test set was 0.7, .15 and .15 respectively. The transfer function in the first layer was tan-sigmoid and the output layer transfer function was linear. To select the number of neuron in the hidden layer, MSE value for the validation set was calculated with changing number of nodes in the hidden layer. The optimum number was determined by trial and error procedure ranging from 2-6, keeping in mind that number of compounds were 26 and number of input descriptors were three[28].

Table-1: Dataset, experimental and calculated logkd values by ANN and SW-MLR methods.

S No	CAS No	$Logk_d$(obs)	SW-MLR	ANN[a]	Normalised mean distance
1	91-20-3	-0.654	-0.6896064	-0.6524	0.563281395
2	90-12-0	-0.604	-0.6753805	-0.6509	0.32665884
3	581-42-0	-0.577	-0.6649341	-0.6507	0.132902401
4	208-96-8	-0.734	-0.6733343	-0.655	0.152877598
5	132-64-9	-0.635	-0.664853	-0.6609	0.176655396
6	2245-38-7	-0.746	-0.7852656	-0.7671	0.016607247
7	1730-37-6	-0.903	-0.8380767	-0.8598	0
8	132-65-0	-0.793	-0.7294359	-0.7658	0.01635404
9	85-01-8	-0.768	-0.7474079	-0.7563	4.75E-04
10	120-12-7	-0.747	-0.7312833	-0.7067	5.59E-04
11	206-44-0	-0.901	-0.8617818	-0.8565	0.039404305
12	56-55-3	-1.034	-1.021169	-1.0219	0.140725898
13	218-01-9	-1.078	-1.0123179	-1.0083	0.140745363
14	205-99-2	-1.082	-1.1191149	-1.1341	0.323748751
15	192-97-2	-1.138	-1.1754742	-1.1511	0.32369865
16	198-55-0	-1.376	-1.2750366	-1.2336	0.324077077
17	191-24-2	-1.223	-1.2441194	-1.2133	0.638028746
18	191-07-1	-1.3	-1.3844086	-1.2673	1
prediction set					
19	91-57-6*	-0.686	-0.6607787	-0.6372	0.326761969
20	83-32-9*	-0.625	-0.6523099	-0.6215	0.19462709
21	832-69-9*	-0.858	-0.8541074	-0.8484	0.023551112
22	50-32-8*	-1.122	-1.0849908	-1.1109	0.324812421
23	92-52-4**	-0.672	-0.7255193	-0.6711	0.17938446
24	86-73-7**	-0.721	-0.7360152	-0.7334	0.07973057
25	129-00-0**	-0.786	-0.8503359	-0.8508	0.039599993
26	53-70-3**	-1.163	-1.2601169	-1.2203	0.681019289

a=Calculated activity data by 3-4-1 ANN architecture; *= compounds in validation set; **= Compounds in test set.

The fitting quality of the ANN models having different node in the hidden layer along with best SW-MLR model is estimated by the coefficient of determination(R^2), root mean square error of calculation (RMSE) and mean absolute per cent error (MAPE) presented in the Table-2. As can be seen from Table-2, statistical parameters values of ANN having different nodes are better than those of SW-MLR for the prediction (validation and test set) set. This is believed to be due to the non-linear capabilities of the ANN model.

On the basis of highest R^2 value, lowest RMSE and MAPE value results, 3-4-1 architecture was selected and the calculated Log k_d values are presented in the Table-1. To estimate the predictive power of selected linear and non-linear QSAR models various statistical parameters for the prediction set were also calculated[29] as recommended by Golbraikh and Tropsha and the results are presented in the Table-3. Both models have satisfied the requirement of the value of various parameters as recommended by Golbraikh and Tropsha.

Figure-1 shows the plot of experimental Log k_d values of selected PAHs against values calculated using SW-MLR and ANN (3-4-1) method. The graphical representation also confirms superiority of best selected ANN model over SW-MLR model.

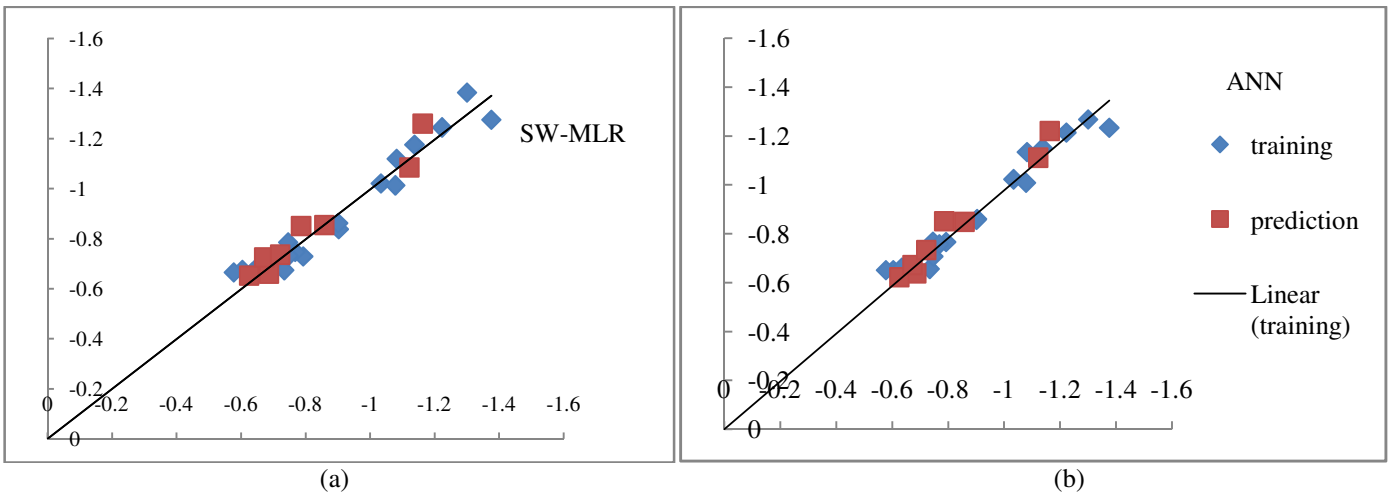
Table-2: Statistical results for the prediction set using different ANN models and SW-MLR method.

	SW-MLR	3-2-1	3-3-1	3-4-1	3-5-1	3-6-1
R2	0.958	0.968	0.972	0.974	0.95	0.954
RMSE	0.044	0.038	0.035	0.035	0.0431	0.046
MAPE	4.28	4.07	3.21	3.1	3.57	4.32

Table-3: Golbraikh and Tropsha acceptable model parameters for linear and non-linear model.

parameter	Threshold value	SW-MLR	NN(3-4-1)(LM)	
r^2	$r^2>0.6$	0.9549	0.9656	
$r_0^2-r'^2_0$	$\|r_0^2-r'^2_0\|<0.3$	0.00146	0.006151	
k	0.85<k<1.15	0.969	0.986	
k'	0.85<k'<1.15	1.02	1.01	

r^2 -squared correlation coefficient between the predicted and observed activities; r^{20}-coefficient of determination for linear regressions with intercepts set to zero, i.e.(predicted versus observed activities); r'^{20} -coefficient of determination for linear regressions with intercepts set to zero (observed versus predicted activities); k and k'-slopes of the above mentioned two regression lines.



(a) (b)

Figure-1: Plot between experimental and (a) SW-MLR, (b) ANN calculated Log k_d values for all dataset.

The outcomes from applicability domain analysis by Euclidean distance method are quite satisfactory within the normal distribution range and normalized mean distance values are reported in the Table-1. None of the compound was found to have normalized mean distance values outside the limiting values (0=least diverse, 1=most diverse training set compound). For both the training and test set, suggested linear and non-linear models have acceptable predictability.

Conclusion

Chemometric methods are successfully used in modeling and predicting the depuration rate constant of compounds. In the present study, a QSAR model was built linearly by SW-MLR method and non-linearly by ANN trained with LM algorithm to calculate depuration rate constants (k_d) of a series of polycyclic aromatic hydrocarbons (PAHs) for mussels, Elliptiocomplanata, starting with large set of descriptor. The Y-randomization technique as well as external prediction indicated that the models were significant, robust and have good predictability. The results of this work indicate the ANN is a promising tool for establishing non-linear relationship between selected PAHs and their Log k_d values.

References

1. Zedeck M.S. (1980). Polycyclic aromatic hydrocarbons: a review. *J Environ PatholToxicol*, 3(5-6), 537-567.

2. Cherng S.H., Lin S.T. and Lee H. (1996). Modulatory effects of polycyclic aromatic hydrocarbons on the mutagenicity of 1-nitropyrene: a structure-activity relationship study. *Mut Res*, 367(4), 177-185.

3. Lewtas J., Walsh D., Williams R. and Dobias L. (1997). Air pollution exposure DNA adduct dosimetry in humans and rodents: evidence for non-linearity at high doses. *Mut Res*, 378(1-2), 51-63.

4. Levin W., Wood A., Chang R., Ryan D., Thomas P., Yagi H. and Conney A. (1982). Oxidative metabolism of polycyclic aromatic hydrocarbons to ultimate carcinogens. *Drug metabolism reviews*, 13(4), 555-580.

5. Morrissey C.A., Bendell-Young L.I. and Elliott J.E. (2005**).** Identifying sources and biomagnification of persistent organic contaminants in biota from mountain streams of southwestern British Columbia, Canada. *Environ. Sci. Technol*., 39(20), 8090-8098.

6. Galassi S., Bettinetti R., Neri M.C.,Jeannot R., Dagnac T., Bristeau S., Sakkas V., Albanis T., Boti V., Valsamaki T., Falandysz J. and Schulte-Oehlmann U. (2008). A multispecies approach for monitoringpersistent toxic substances in the Gulf of Gdao´ sk (Baltic sea). *Ecotoxicol.Environ. Saf.*, 69(1), 39-48.

7. Morrison H., Lazar R., Haffner G.D. and Yankovich T. (1995). Elimination rate constants of 36 PCBs in zebra mussels (Dreissena polymorpha) and exposure dynamics in the Lake St. Clair-Lake Erie corridor. *Canadian journal of fisheries and aquatic sciences*, 52(12), 2574-2582.

8. Uno S., Shiraishi H., Hatakeyama S. and Otsuki A. (1997). Uptake and depuration kinetics and BCFs of several pesticides in three species of shellfish (Corbicula leana, Corbicula japonica, and Cipangopludina chinensis): comparison between field and laboratory experiment. *Aquatic toxicology*, 39(1), 23-43.

9. Weisbrod A.V., Burkhard L.P., Arnot J., Mekenyan O., Howard P.H., Russom C. and Lutz C. (2007). Workgroup report: review of fish bioaccumulation databases used to identify persistent, bioaccumulative, toxic substances. *Environmental Health Perspectives*, 115(2), 255.

10. Leo A. and Hoekman D.H. (1995). Exploring QSAR.:. Fundamentals and applications in chemistry and biology *An American Chemical Society Publication*.

11. Manallack D.T. and Livingston D.J. (1999). Neural network indrug discovery: have they lived up their promise?. *Eur. J. Med.Chem*., 34(3), 195-208.

12. Maier H.R. and Dandy G.C. (2000). Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. *Environmental Modelling & Software*, 15, 101-124.

13. Kaiser K.L. (2003). Neural networks for effect prediction in environmental and health issues using large datasets. *Molecular Informatics*, 22(2), 185-190.

14. Turner J.V., Maddalena D.J. and Cutler D.J. (2004). Pharmacokinetic parameter prediction from drug structure using artificial neural networks. *International journal of pharmaceutics*, 270(1), 209-219.

15. Hecht D., Cheung M. and Fogel G.B. (2008). QSAR using evolved neural networks for the inhibition of mutant PfDHFR by pyrimethamine derivatives. *Biosystems*, 92(1), 10-15.

16. Russell R.W. and Gobas F.A. (1989). Calibration of the freshwater mussel, Elliptio complanata, for quantitative biomonitoring of hexachlorobenzene and octachlorostyrene in aquatic systems. *Bulletin of environmental contamination and toxicology*, 43(4), 576-582.

17. Gewurtz S.B., Drouillard K.G., Lazar R. and Haffner G.D. (2002). Quantitative biomonitoring of PAHs using the Barnes mussel (Elliptio complanata). *Archives of environmental contamination and toxicology*, 43(4), 0497-0504.

18. Li F., Liu X., Zhang L., You L., Wu H., Li X. and Yu J. (2011). QSAR studies on the depuration rates of polycyclic aromatic hydrocarbons, polybrominated diphenyl ethers and polychlorinated biphenyls in mussels (Elliptio complanata). *SAR and QSAR in Environmental Research*, 22(5-6), 561-573.

19. Wu D., Liu X., Wang L., Wang L., Xu M., Sun T. and Zhou J. (2009). QSARs on the depuration rate constants of polycyclic aromatic hydrocarbons in Elliptio complanata. *Molecular Informatics*, 28(5), 537-541.

20. SPSS I. (2012). Statistics for windows, version 20.0. *IBM Corp*., Armonk NY.

21. Matlab R. (2013). Version 8.1. 0.604 (R2013a). *Natrick, Massachusetts: The MathWorks Inc*.

22. Ballabio D., Manganaro A., Consonni V., Mauri A. and Todeschini R. (2009). Introduction to MOLE DB-on-line molecular descriptors database. *MATCH Commun Math Comput Chem*, 62, 199-207.

23. Alexander G. and Alexander T. (2002). Beware of Q2. *J Mol Graph Model*, 20(4), 269-276.

24. Sahigara F., Mansouri K., Ballabio D., Mauri A., Consonni V. and Todeschini R. (2012). Comparison of different approaches to define the applicability domain of QSAR models. *Molecules*, 17(5), 4791-4810.

25. Gramatica P. (2007). Principles of QSAR models validation: internal and external. *Molecular Informatics*, 26(5), 694-701.

26. Roy K., Kar S. and Ambure P. (2015). On a simple approach for determining applicability domain of QSAR models. *ChemometrIntell Lab Syst.,* 145, 22-29.

27. Snedecor G.W. and Cochran W.G. (1967). Statistical methods, Oxford and IBH publishing Co. pvt. Ltd., New Delhi, 381-418.

28. Andrea T.A. and Kalayeh H. (1991) Applications of neural networks in quantitative structure-activity relationships of dihydrofolate reductase inhibitors. *J. Med. Chem*., 34(9), 2824-2836.

29. Roy K., Das R.N., Ambure P. and Aher R.B. (2016). Be aware of error measures. Further studies on validation of predictive QSAR models. *Chemometrics and Intelligent Laboratory Systems*, 152, 18-33.