



Contents lists available at *Avicenna Publishing Corporation (APC)*

Asian Journal of Green Chemistry

journal homepage: <http://ajgreenchem.avicenna.pub/>



Original Research Article

Quantitative structure activity relationship (QSAR) study of octanol-water partition coefficients of some of environmental toxic of petroleum substances

Mehrdad Shahpar^{a,*}, Sharmin Esmaeilpoor^b

^a Director of Ilam Petrochemical Company

^b Department of Chemistry, Payame Noor University, P.O. BOX 19395-4697, Tehran, Iran

ARTICLE INFORMATION

Received: 3 August 2017

Received in revised: 1 September 2017

Accepted: 13 September 2017

Available online: 4 October 2017

DOI: 10.22631/ajgc.2017.94413.1010

KEYWORDS

Petroleum substances

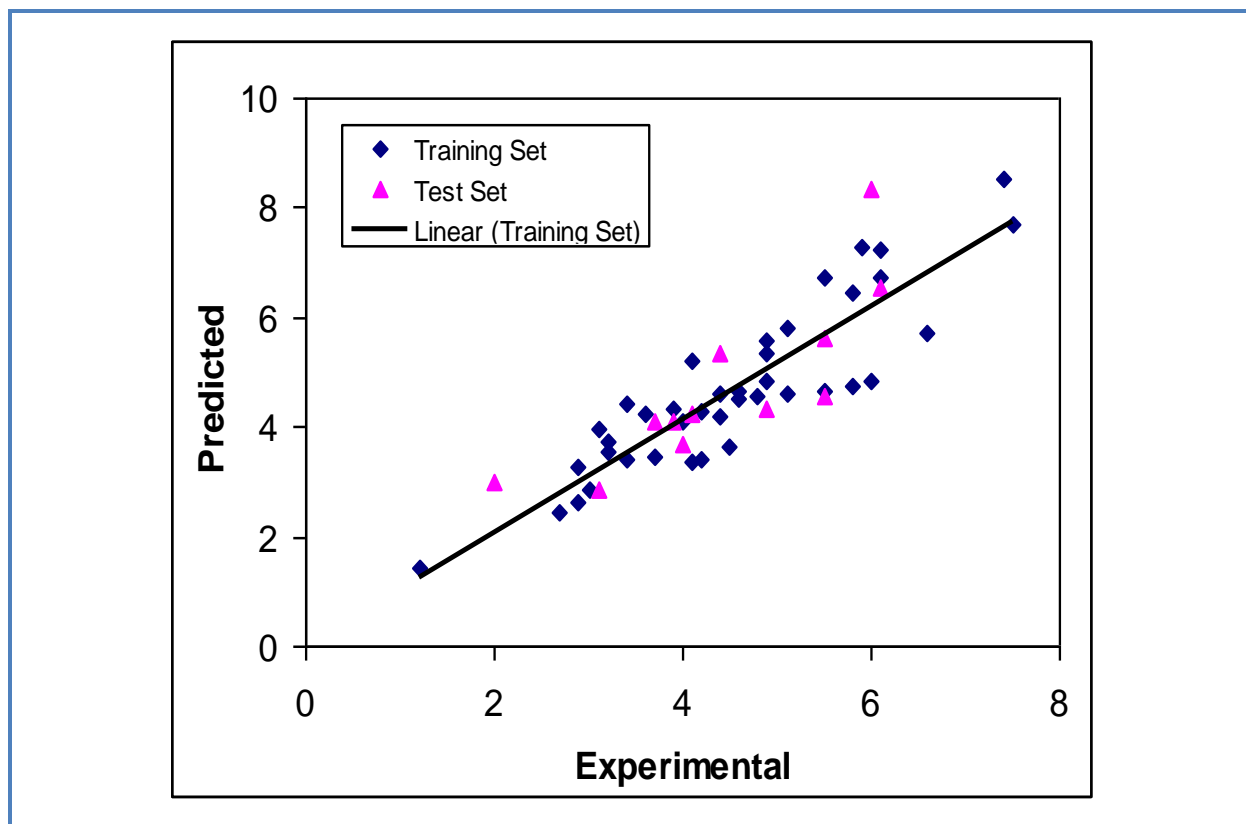
Genetic algorithm

QSAR

ABSTRACT

Life and its extraction fuels climate change. We performed studies upon an extended series of petroleum hydrocarbons, with octanol-water partition coefficients ($\log K_{ow}$), by using the quantitative structure-activity relationship (QSAR) methods that imply analysis of correlations and representation of models. A suitable set of molecular descriptors was calculated and the genetic algorithm (GA) was employed to select those descriptors, resulting in the best-fit models. The partial least squares PLS (PLS) was utilized to construct the linear QSAR model. The best GA-PLS model contains 27 selected descriptors in 10 latent variables space. The R^2 and RMSE for training and test sets were (0.827, 0.088) and (0.716, 0.185), respectively. Inspection of the results reveals a higher R^2 and lowers the RMSE value parameter for the data set GA-PLS. The GA-PLS linear model has good statistical quality with low prediction error. This is the first research on the QSAR which uses GA-PLS for the prediction octanol-water partition coefficients of some of the environmental toxic of the petroleum substances.

Graphical Abstract



Introduction

Petroleum hydrocarbon (PHC) is often used as a general term to describe a mixture of various organic compounds, mostly pure hydrocarbons, but also including low-polarity hetero-substituted saturates and aromatics, found in crude oil, bitumen and coal. Petroleum hydrocarbons are typically grouped by structure: saturates, olefins, aromatics, polar compounds (a wide variety of compounds containing sulfur, oxygen and nitrogen), and asphaltenes [1, 2].

Due to their derivation from the natural crude oils and the refining processes used in their production, petroleum substances are complex mixtures of hydrocarbons and various compositions. Reflecting the properties of the constituent hydrocarbons, they are hydrophobic and exhibit low solubility in water. However, since the individual hydrocarbon components vary considerably in molecular weight or 'carbon number', and also vary in their degree of unsaturation, the individual hydrocarbons exhibit a range of water solubility. When adding incremental amounts of a complex petroleum substance in water, a point will be reached at which the solubility limit of the least soluble component is exceeded and the remaining components will then partition between the water and the undissolved hydrocarbon phase. As a result, the composition of the total dissolved hydrocarbons will

be different from the composition of the parent substance. Thus, the strict definition of 'water solubility' does not apply to petroleum substances [3].

As well as being hydrophobic, hydrocarbons are lipophilic, resulting in high octanol-water partition coefficients (or $\log K_{ow}$ values). Hydrocarbons cause toxicity in aquatic organisms by a mechanism referred to as 'non-polar narcosis', or 'general' (non-specific) toxicity [3]. The values of $\log K_{ow}$ for individual hydrocarbons increase with increasing carbon number within homologous series of generic types. Substances with the same carbon number show decreasing $\log K_{ow}$ values with increasing levels of unsaturation. Quantitative structure activity relationships (QSAR), relating $\log K_{ow}$ values of single hydrocarbons to toxicity, show that water solubility decreases more rapidly with increasing K_{ow} than does the concentration causing effects [4]. This relationship varies somewhat with species, but it follows that there is a $\log K_{ow}$ limit for hydrocarbons, above which they will not exhibit acute toxicity; this limit is at a $\log K_{ow}$ value of about 4 to 5 [3, 4]. It has been confirmed experimentally that for fish and invertebrates, paraffinic hydrocarbons with a carbon number of 10 or higher ($\log K_{ow} > 5$) show no acute toxicity and that alkylbenzenes with a carbon number of 14 or greater ($\log K_{ow} > 5$) similarly show no acute toxicity [5]. From these well-demonstrated solubility 'cut-offs' for acute toxicity of hydrocarbon substances, which directly relate to their physico-chemical properties, it is clear that the same should hold for complex petroleum substances. As discussed later, individual components may contribute to the 'total toxicity' of the mixture even if individually, they do not reach a toxic concentration. A number of reports, dealing with QSAR octanol-water partition coefficients calculation of several compounds, have been published in the literature [6-8].

In particular, GA is frequently used as search algorithms for variable selection in chemometrics and QSAR. GA is a stochastic method to solve the optimization problems defined by fitness criteria, applying the evolution hypothesis of Darwin and different genetic functions, i.e. crossover and mutation [9, 10]. Partial least square (PLS) is the most commonly used multivariate calibration method [11, 12]. PLS regression is a recent technique that generalizes and combines features from principal component analysis and multiple regressions. It is particularly useful when we need to predict a set of dependent variables from a (very) large set of independent variables (i.e., predictors). It originated in the social sciences but became popular first in chemometrics (i.e., computational chemistry) due in part to Herman's son Svante and in sensory evaluation. In the present study, GA-PLS was employed to generate QSAR model that correlate the structure of petroleum substances; with observed K_{ow} .

Experimental

Data set

The octanol-water partition coefficients ($\log K_{ow}$) of the 52 environmental toxic of petroleum substances was taken from literature [13] is shown in Table 1. Schem 1 shows Chemical Structure of a few petroleum substances.

Selection of descriptors

All structures were drawn with the HyperChem software (version 6). Optimization of molecular structures was carried out by semi-empirical AM1 method using the Fletcher-Reeves algorithm until the root mean square gradient of 0.01 was obtained. Since the calculated values of the electronic features of molecules will be influenced by related conformation. In the current research an attempt was made to use the most stable conformations. Some electronic descriptors such as polarizability, dipole moment and orbital energies of LUMO and HOMO were calculated by the HyperChem software. Also optimized structures were used to calculate 1497 descriptors by DRAGON software version 3.

One of the challenging parts in developing models is choosing suitable parameters encoding different aspects of the molecular structure. A large number of structural descriptors can be calculated using existing software's such as dragon. However, nowadays the main problem is choosing the most adequate and interpretable parameters needed for developing the models among a large number of them. To reduce the original pool of descriptors to an appropriate size, objective descriptor reduction was performed using various criteria. Reducing the pool of descriptors eliminates those descriptors which contribute either no information or whose information content is redundant with that of other descriptors present in the pool. Any descriptor that had identical or zero values for greater than 90% of the compounds was eliminated.

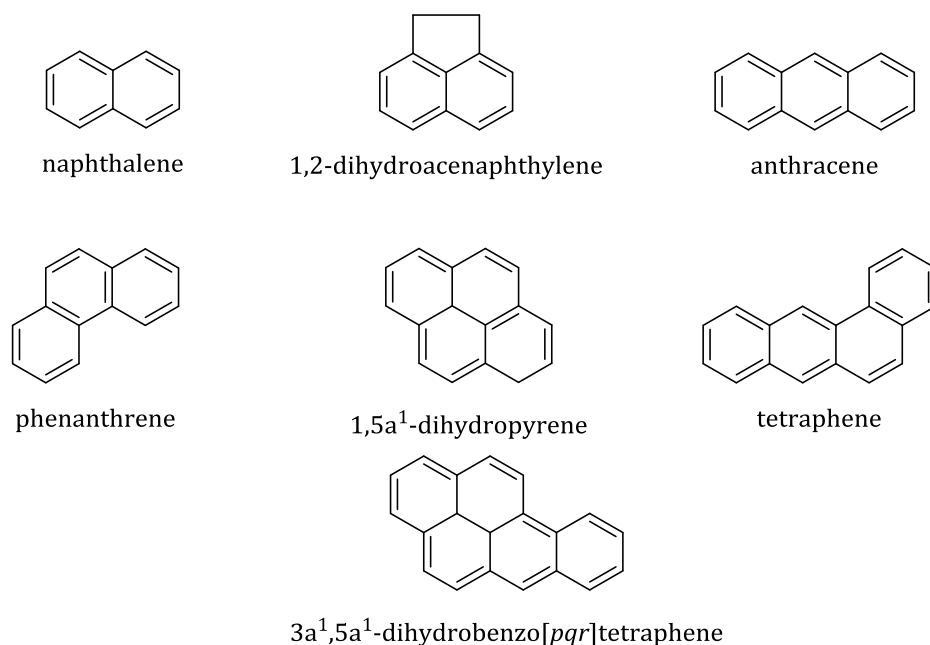
Table 1. The data set and the corresponding observed K_{ow} value

Entry	Name	K_{ow}
	Training Set	
1	2,3 dimethylheptane	4.6
2	n-nonane	4.8
3	Pentamethylheptane	5.9
4	Dimethyldecane	6.0
5	n-dodecane	6.1
6	dodecane	7.4
7	Cyclohexane	3.4
8	Decalin (cis)	1.2

9	Decalin (trans)	4.2
10	135 trimethylbenzene	4.4
11	toluene	2.7
12	styrene	2.9
13	Ethyl benzene	3.0
14	p-xylene	3.1
15	m-xylene	3.2
16	Methyl styrene	3.4
17	1,3,5 trimethylbenzene	3.6
18	isopropyl benzene	3.7
19	diphenyl methane	4.0
20	m-diethylbenzene	4.1
21	diisopropyl benzene	4.9
22	phenyl-xylylmethane	5.1
23	m-terphenyl	5.5
24	diethylbiphenyl	5.8
25	dibenzyltoluene	6.6
26	acenaphthylene	2.9
27	naphthalene	3.2
28	methylnaphthalene	3.9
29	isopropyl naphthalene	4.6
30	propyl naphthalene	4.9
31	triisopropyl naphthalene	7.5
32	fluorene	4.1
33	acenaphthene	4.2
34	dibenzothiophene	4.4
35	phenanthrene	4.5
36	fluoranthene	4.9
37	me-phenanthrene	5.1
38	benzo(a)anthracene	5.5
39	chrysene	5.8
40	benzo(a)pyrene	6.1

Training Set

1	2,2,4,6,6 pentamethylheptane	6.0
2	Tetrahydronaphthalene	4.0
3	o-xylene	3.1
4	benzene	2.0
5	1,3,5-trimethylbenzene	3.7
6	biphenyl	4.1
7	o-terphenyl	5.5
8	ethylnaphthalene	3.9
9	diisopropyl naphthalene	6.1
10	anthracene	4.4
11	pyrene	4.9
12	me-phenanthrene	5.5



Scheme 1. Chemical Structure of a few petroleum substances

Computer hardware and software

All calculations were run on a HP Laptop computer with AMD Turion 64X2 processor with windows XP operating system. The optimizations of molecular structures were done by the HyperChem 7.0 (AM1 method) and descriptors were calculated by Dragon Version 3.0 software's. Cross validation, GA-PLS and other calculation were performed in the MATLAB (Version 7, Mathworks, Inc.) environment.

Genetic algorithm

Genetic algorithm is a stochastic method to solve the optimization problems, defined by fitness criteria applying to the evolution hypothesis of darwin and different genetic functions, i.e., crossover and mutation. Compared to the traditional search and optimization procedures, GA is robust, global and generally more straightforward to apply to situations where there is little or no a priori knowledge about the process to be controlled. Since GA does not require derivative information or a formal initial estimate of the solution region and because of the stochastic nature of the search mechanism, it is capable to search the entire solution space with a greater probability of finding the global optimum.

Partial least squares

PLS is a linear multivariate method for relating the process variables X with responses Y . PLS can analyze data with strongly collinear, noisy, and numerous variables in both X and Y [19]. PLS reduces the dimension of the predictor variables by extracting factors or latent variables that are correlated with Y while capturing a large amount of the variations in X . This means that PLS maximizes the covariance between matrices X and Y . In PLS, the scaled matrices X and Y are decomposed into score vectors (t and u), loading vectors (p and q), and residual error matrices (E and F):

$$X = \sum_{i=1}^a t_i p_i^T + E \quad (1)$$

$$Y = \sum_{i=1}^a u_i q_i^T + F$$

Where a is the number of latent variables. In an inner relation, the score vector t is linearly regressed against the score vector u .

$$U_i = b_i t_i + h_i \quad (2)$$

Where b is regression coefficient that is determined by minimizing the residual h . It is crucial to determine the optimal number of latent variables and cross validation is a practical and reliable way to test the predictive significance of each PLS component. There are several algorithms to calculate the PLS model parameters. In this work, the NIPALS algorithm was used with the exchange of scores [12].

N-octanol/water partition coefficient

N-octanol/water partition coefficient (K_{ow}) is defined as the ratio of the concentration of a chemical in *n*-octanol and water at equilibrium at a specified temperature.

$$K_{ow} = \text{Concentration in Octanol} / \text{Concentration in water} \quad (3)$$

Values of K_{ow} are unitless and usually expressed as $\log K_{ow}$, a relative indicator of the tendency of an organic compound to adsorb to soil and living organism. Also effect of ionic strength on the octanol-water partition coefficients is as follows:

$$KOS (\text{salt}) = Kow \times 10^S \quad (4)$$

Where (S) is Séchenov constant. LogK_{ow} are generally inversely related to water solubility and directly proportional to molecular weight of a substance. Under REACH, K_{ow} or logK_{ow} is not required for inorganic substances. K_{ow} or logK_{ow} is a key input parameter in environmental modelling tools to estimate environmental exposure levels.

Results and discussion

Results of the GA-PLS model

To reduce the original pool of descriptors to an appropriate size, the objective descriptor reduction was performed using various criteria. Reducing the pool of descriptors eliminates those descriptors which contribute either no information or whose information content is redundant with other descriptors present in the pool. The remained descriptors were employed to generate the model with the GA-PLS program. The best model is selected on the basis of the highest square correlation coefficient (R^2) and root mean square error (RMSE) of prediction and simplicity of the model. These parameters are probably the most popular measure of how well a model fits the data. The best GA-PLS model contains 27 selected descriptors in 10 latent variables space. The R^2 and RMSE for training and test sets were (0.827, 0.088) and (0.716, 0.185), respectively. For this in general, the number of components (Latent variables) is less than number of independent variables in PLS analysis. The PLS model uses higher number of descriptors that allow the model to extract better structural information from descriptors to result in a lower prediction error. The statistical parameters R^2 and RMSE were obtained for proposed model. Each of the statistical parameters mentioned above were used for assessing the statistical significance of the QSAR model.

Inspection of the results reveals a higher R^2 and lowers RMSE value parameter for the data set GA-PLS. The GA-PLS linear model has good statistical quality with low prediction error. A plot of predicted K_{ow} versus experimental K_{ow} values by GA-PLS for is shown Figure 1. Obviously, there is a close agreement between the experimental and predicted K_{ow} and the data represent a very low scattering around a straight line with respective slope and intercept close to one and zero. The key strength of PLS, unlike regression analysis, is their ability to flexible mapping of the selected features by manipulating their functional dependence implicitly. The statistical parameters reveal the high predictive ability of PLS model. The whole of these data clearly displays a significant improvement of the QSAR model consequent to nonlinear statistical treatment. Obviously, there is a close agreement between the experimental and predicted RT and the data represent a very low scattering around a straight line with respective slope and intercept close to one and zero. As can be seen in this section, the PLS is more reproducible octanol-water partition coefficients.

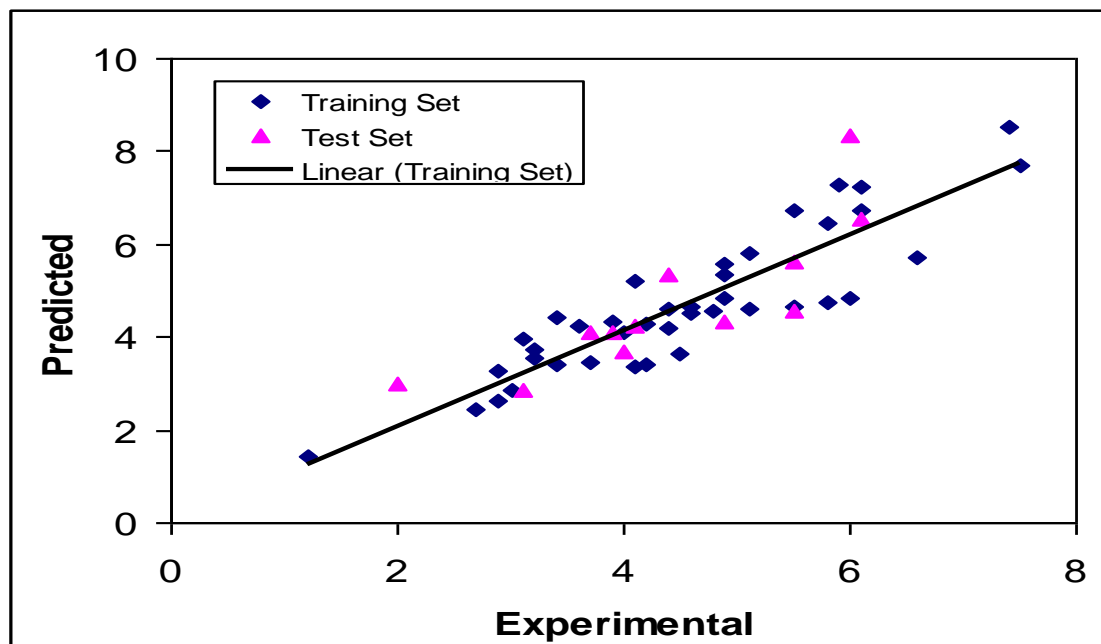


Figure 1. Plot of predicted K_{ow} obtained by GA-PLS against the experimental values

The $\log K_{ow}$ is a very important parameter for predicting the distribution of a substance in various environmental compartments (water, soil, air, biota, etc). Substances with high $\log K_{ow}$ values tend to adsorb more readily to organic matter in soils or sediments because of their low affinity for water. Chemicals with very high $\log K_{ow}$ values (i.e. >4.5) are of greater concern because they may have the potential to bio-concentrate in living organisms. For above reason, n-octanol/water partition coefficient (K_{ow}) is used as a screening test for bio-accumulation test. The assumption behind this is that the uptake of an organic substance is driven by its hydrophobicity. For organic substances with a $\log K_{ow}$ value below 4.5 it is assumed that the affinity for the lipids of an organism is insufficient to exceed the bio-accumulation criterion i.e. a BCF value of 2000. For some groups of chemicals, such as metals and surface active compounds, $\log K_{ow}$ is not a valid descriptor for assessing the bioaccumulation potential. Information on bioaccumulation of such substances should therefore take account of other descriptors or mechanisms than hydrophobicity.

Interpretation of descriptors

In the chromatographic retention of compounds in the stationary phase two important types of interactions contribute to the chromatographic retention of the compounds: the induction and dispersion forces. The dispersion forces are related to steric factors, molecular size, shape and branching, while the induced forces are related to the dipolar moment, which should stimulate dipole-induced dipole interactions.

Constitutional descriptors are most simple and commonly used descriptors, reflecting the molecular composition of a compound without any information about its molecular geometry. Number of C atoms, the average bond order of a C atom and the minimum atomic state energy for a C atom quantify the bond strength between the C atoms. A molecule locked in a rigid conformation due to strong intramolecular interactions is in fact less free to move and is expected to have a higher boiling point.

The hydrogen bonding is a measure of the tendency of a molecule to form hydrogen bonds. This is related to number of Hydrogen atoms (nH). Hydrogen-bonding may be divided into an electrostatic term and a polarization/charge transfer term.

The geometrical descriptors are suitable for complex-behaved properties, because they take into account the 3D-arrangement of atoms without ambiguities (as those appearing when using chemical graphs), as well as they do not depend on the molecular size and thus they are applicable to a large number of molecules with great structural variance, which have a characteristic common to all of them.

The GETAWAY (GEometry, Topology, and Atom-Weights Assembl Y) descriptors try to match 3Dmolecular geometry provided by the molecular influence matrix and atom relatedness by molecular topology, with chemical information by using different atomic weights. These descriptors are quickly computed from the atomic positions of the molecule atoms (hydrogens included).

The geometrical descriptors are suitable for complex-behaved properties, because they take into account the 3D-arrangement the atoms without ambiguities (as those appearing when using chemical graphs), as well as they do not depend on the molecular size and thus they are applicable to a large number of molecules with great structural variance, which have a characteristic common to all of them.

Gravitational index (G_2) (bond-restricted) is a geometrical descriptor that reflecting the mass distribution in a molecule and defined as Eq. (5)

$$G_2 = \sum_{a=1}^A \left(\frac{m_i \cdot m_j}{r_{ij}^2} \right)_a \quad (5)$$

Where m_i and m_j are the atomic masses of the considered atoms; r_{ij} the corresponding interatomic distances; and A the number of all pairs of bonded atoms of the molecule. This index is related to the bulk cohesiveness of the molecules, accounting, simultaneously, for both atomic masses (volumes) and their distribution within the molecular space. This index can be extended to any other atomic property different from atomic mass, such as atomic polarizability, atomic, van der Waals volume etc.

Topological descriptors are based on a graph representation of the molecule. They are numerical quantifiers of molecular topology obtained by the application of algebraic operators to matrices representing molecular graphs and whose values are independent of vertex numbering or labeling. They can be sensitive to one or more structural features of the molecule such as size, shape, symmetry, branching and cyclicity and can also encode chemical information concerning atom type and bond multiplicity.

Although these descriptors are often successful in rationalizing $\log K_{ow}$ of petroleum substances, they cannot account for conformational changes and they do not provide information about electronic influence through bonds or across space. For that reason, quantum chemical descriptors are used in developing QSAR.

Quantum chemical descriptors were defined in terms of atomic charges and used to describe electronic aspects both of the whole molecule and of particular regions, such atoms, bonds, and molecular fragments. They include thermodynamic properties (system energies) and electronic property (HOMO energy). The HOMO as an electron donor represents the ability to donate an electron. The HOMO energy plays a very important role in the nucleophilic behavior and it represents molecular reactivity as a nucleophile [13].

From the above discussion, it can be seen that the particle size, hydrogen bonding and electrostatic interactions are the likely three factors controlling the $\log K_{ow}$ of these compounds. All the descriptors involved in the model, which have explicit physical meaning, may account for the structure responsible for the $\log K_{ow}$ of these compounds.

Model validation and statistical parameters

The applied internal (leave-group-out cross validation (LGO-CV)) and external (validation set) validation methods were used for the predictive power of models. In the leave-group-out procedure one compound was removed from the data set, the model was trained with the remaining compounds and used to predict the discarded compound. The process was repeated for each compound in the data set. The predictive power of the models developed on the selected training set is estimated on the predicted values of validation set chemicals. The data set should be divided into three new sub-data sets, one for calibration and prediction (training), and the other one for validation sets. The calibration set was used for model generation. The prediction set was applied deal with overfitting of the network, whereas validation set which its molecules have no role in model building was used for the evaluation of the predictive ability of the models for external set [14, 15].

In the other hand by means of training set, the best model is found and then, the prediction power of it is checked by validation set, as an external data set. In this work, from all 52 components, 30 components are in calibration set, 10 components are in prediction set and 12 components are in validation set). The result clearly displays a significant improvement of the QSAR model consequent to non-linear statistical treatment and a substantial independence of model prediction from the structure of the validation molecule. In the above analysis, the descriptive power of a given model has been measured by its ability to predict partition of unknown petroleum substances.

For the constructed models, some general statistical parameters were selected to evaluate the predictive ability of the models for log K_{ow} values. In this case, the predicted log K_{ow} of each sample in prediction step was compared with the experimental acidity constant.

Root mean square error (RMSE) is a measurement of the average difference between predicted and experimental values, at the prediction step. RMSE can be interpreted as the average prediction error, expressed in the same units as the original response values. Its small value indicates that the model predicts better than chance and can be considered statistically significant. The RMSE was obtained by the following formula:

$$RMSE = \left[\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \right]^{\frac{1}{2}} \quad (6)$$

The other statistical parameter was relative error (RE) that shows the predictive ability of each component, and is calculated as:

$$RE(\%) = 100 \times \left[\frac{1}{n} \sum_{i=1}^n \frac{(y_i^{\wedge} - y_i)}{y_i} \right] \quad (7)$$

The predictive ability was evaluated by the square of the correlation coefficient (R^2) which is based on the prediction error sum of squares and was calculated by following equation:

$$R^2 = \frac{\sum_{i=1}^n (y_i^{\wedge} - \bar{y})}{\sum_{i=1}^n (y_i - \bar{y})} \quad (8)$$

Where y_i is the experimental log K_{ow} in the sample i , \hat{y}_i represented the predicted log K_{ow} in the sample i , \bar{y} is the mean of experimental log K_{ow} in the prediction set and n is the total number of samples used in the validation set.

The main aim of the present work was to assess the performances of GA-PLS, for modeling the log K_{ow} of compounds. The procedures of modeling including descriptor generation, splitting of the data, variable selection and validation were the same as those performed for modeling of the log K_{ow} of petroleum substances.

Conclusion

In this research, an accurate QSAR model for estimating the octanol-water partition coefficients of environmental toxic of petroleum substances was developed by employing the GA-PLS technique. This model has good predictive capacity and excellent statistical parameters. It is easy to notice that there was a good prospect for the GA-PLS application in the QSAR modeling. It can also be used successfully to estimate the K_{ow} for new compounds or for other compounds whose experimental values are unknown. This is the first research on the QSAR which uses GA-PLS for prediction octanol-water partition coefficients of some of environmental toxic of petroleum substances.

References

- [1]. Mehrotra K.N., Upadhyaya S.K. *Recl. Trav. Chim. Pays-Bas.*, 1987, **106**:625
- [2]. Nazaré M., Couto P.F.S., Basto M.P., Vasconcelo, M.T. *Chemosphere*, 2011, **84**:1052
- [3]. Abdel-Shafy H.I., Mansour M.S.M. *J. Pet.*, 2016, **25**:107
- [4]. Guarino C., Spada V., Sciarrillo R. *Chemosphere*, 2017, **170**:10
- [5]. Wu M., Li W., Ick W., Ye X., Kost D., Chen L. *Chemosphere*, 2017, **169**:124
- [6]. Papa E., Dearden J.C., Gramatica P., *Chemosphere*, 2007, **67**:351
- [7]. Sharma D., Narasimhan B., Kumar P., Jalbout A. *Eur. J. Med. Chem.*, 2009, **44**:1119
- [8]. Kar S., Roy K., *J. Hazard. Mater.*, 2010, **177**:344
- [9]. Noorizadeh H., Noorizadeh M., *Med Chem Res.*, 2014, **13**:264
- [10]. Van Dijck G., Van Hulle M.M. *Chemom. Intell. Lab. Syst.*, 2011, **107**:318
- [11]. Noorizadeh H., Farmany A., *Environ Sci Pollut Res.*, 2011, **8**:148
- [12]. Ribeiro J.S., Augusto F., Salva T.J.S., Thomaziello R.A., Ferreira M.M.C. *Anal. Chim. Acta.*, 2009, **634**:172
- [13]. Todeschini R., Consonni V. *Handbook of Molecular Descriptors*, Wiley/VCH, Weinheim, 2000
- [14]. Deeb O., *Chemom. Intell. Lab. Syst.*, 2010, **104**:181

[15]. Pran Kishore D., Balakumar C., Raghuram Rao A., Partha Pratim R., Roy K. *Bioorg. Med. Chem. Lett.*, 2011, **21**:818

How to cite this manuscript: Mehrdad Shahpar*, Sharmin Esmaeilpoor. Quantitative structure activity relationship (QSAR) study of octanol-water partition coefficients of some of environmental toxic of petroleum substances. *Asian Journal of Green Chemistry*, 2017, 1, 116-129. DOI: [10.22631/ajgc.2017.94413.1010](https://doi.org/10.22631/ajgc.2017.94413.1010)