ICC

Iranian Chemical Communication

Original Research Article

Payame Noor University http://icc.journals.pnu.ac.ir

Prediction of boiling point and water solubility of crude oil hydrocarbons using sub-structural molecular fragments method

Faraidon Ghaderi, Saadi Saaidpour*

Department of Chemistry, Faculty of Science, Sanandaj Branch, Islamic Azad University, Sanandaj, Iran

Received: 31 October 2015, Accepted: 1 March 2016, Published: 1 March 2016

Abstract

The quantitative structure–property relationship (QSPR) method is used to develop the correlation between structures of crude oil hydrocarbons (80 compounds) and their boiling point and water solubility. Sub-structural molecular fragments (SMF) calculated from structure alone were used to represent molecular structures. A subset of the calculated fragments selected using stepwise regression (forward and backward steps) (SR) was used in the QSPR model development. Multiple linear regressions (MLR) are utilized to construct the linear prediction model. The prediction results agree well with the experimental values of these properties. The comparison results indicate the superiority of the presented models and reveal that it can be effectively used to predict the boiling point temperatures and water solubility values of crude oil hydrocarbons from the molecular structures alone. The stability and predictivity of the proposed models were validated using internal validation (leave one out and leave many out) and external validation. Application of the developed models to test set of 16 compounds demonstrates that the new model sare reliable with good predictive accuracy and simple formulation.

Keywords: Boiling point; water solubility; crude oil hydrocarbon; ISIDA-QSPR; prediction.

*Corresponding author: Saadi Saaidpour
Tel: +98 (87) 33184300, Fax: +98 (87) 33184301
E-mail: saadisaaidpour@gmail.com, sasaaidpour@iausdj.ac.ir

Introduction

Crude oil is a mixture of comparatively volatile liquid hydrocarbons. Crude oils are commonly characterized by the type of hydrocarbon compound, for example paraffins and naphthenes. The different hydrocarbon compounds will have different boiling point that in the refinery is separated by distillation. Boiling point and water solubility in crude oil hydrocarbons are important matters in the oil industry. The solubility of a substance is the amount of substance that will dissolve in a given amount of solvent. Solubility is a quantitative term that depends on the physical and chemical properties of the solute and solvent as well as on temperature and pressure. The water dissolved in crude oil can freeze and block the fuel line or pipe. For instance, dissolved water in the gas phase may form condensate, ice and gas hydrate which may lead to corrosion/erosion of pipelines, blockage of transfer lines, damage of compressor impeller, etc. [1–4]. The solubility of water in hydrocarbons, even at ambient temperatures, can have great practical importance [5]. The importance of the solubility of water in crude oil will increase in view of processing, safety, hazard, and environmental considerations focusing on product quality and equipment sustainability.

The boiling point is one of the main physicochemical properties used to characterize and identify compounds [6]. The first work applying Quantitative Structure–Property/Activity Relationships (QSPR/QSAR) to Boling Point was by Wiener [7]. Moreover, extensive efforts have been made to apply the structural information to fit experimental BP [8-12].In many physical-chemistry areas and organic compounds, it is increasingly necessary to translate those general relations into quantitative associations expressed in useful algebraic equations known as Quantitative Structure-Property (-Activity) Relationships (QSAR/QSPR) [13-14].Recently, the sub-structural molecular fragments (SMF) method has been widely performed to predict many properties [15-17]. In this paper, we applied the sub-structural molecular fragment (SMF) method and described the QSPR modeling of Boiling Point (BP) and water solubility (logSw) for crude oil hydrocarbons using multiple linear regression approach and fragmental descriptors in ISIDA (In SILico design and Data Analysis) software.

Materials and methodology

Data set

The initial dataset used in this study consists of 80 crude oil hydrocarbons and properties from Handbook of Physical Properties for Hydrocarbons and Chemicals [18]. In the present study, the total set of hydrocarbons is partitioned into a training set including 64 hydrocarbons and test set with 16 hydrocarbons (see **supplementary data**).

Computer and software

In this study, the implementations ware performed using computer programs on a Lenovo laptop computer with windows 7 operating system. At first, the molecular structures of all compounds were drawn by Chem Office program. Preoptimized using MM^+ molecular mechanics methods and final geometries of the minimum energy conformation were obtained by more precise optimization with the semi-empirical AM1 method (applying a root mean square gradient limit of 0.01 Kcal. mol^{-1}. $Å^{-1}$). Finally, SDF (Structure Data File) file of the resulted geometries compose by EdiSDF were put in to ISIDA/QSPR (version 5.76.003, 2010) to calculate substructural molecular fragments.

Molecular fragments

The ISIDA/QSPR program realizes the substructural molecular fragments (SMF) method [19-24]; it uses two types of topological descriptors (fragments): “atom/bond sequences”, and “augmented atoms”. Three sub-types of molecular fragments of AB, A and B are defined for each class. For the fragments I, they represent sequences of atoms and bonds (AB), of atoms only (A), or of bonds only (B). *Shortest* or *all paths* from one atom to the other are used. For each type of sequences, the minimal (n_{min}) and maximal (n_{max}) number of constituted atoms must be defined. Thus, for the partitioning I (AB, n_{min} - n_{max}), I(A, n_{min} - n_{max}) and I(B, n_{min} - n_{max}), the program generates “intermediate” sequences involving n atoms ($n_{min} \leq n \leq n_{max}$). In the current version of ISIDA/QSPR, $n_{min} \geq$ 2 and $n_{max} \leq$ 15. The number of sequences’ types of different length corresponding to n_{min} = 2 and n_{max} = 15 is equal to 105 for each of three sub-types AB, A and B, totally 315 types of sequences.

The key problem of any QSPR study is related to selection of pertinent descriptors to QSPR model. In ISIDA software, screening descriptors mainly follows three steps, namely filtering stage, forward stepwise pre-selection stage and backward stepwise selection stage. In the first stage, the program eliminates variables which have a small correlation coefficient with the property, and those highly correlated with other variables, which were already selected for the model. In the second stage, the suite of forward and backward stepwise algorithms has been used for variable pre-selection in ISIDA studies by the Variable Selection Suite (VSS) program. The final selection is performed using backward stepwise variable selection procedure based on the t statistic criterion.

QSPR model in ISIDA/QSPR

The modeled physical or chemical property *Y* can be quantitatively calculated accounting for contributions of fragments using linear, non-linear and fitting equations. In this study, descriptors calculated are based on linear equation (Equation 1).

$$Y = a_0 + \sum_i a_i N_i + \beta(1)$$

Where a_i is fragment contribution, N_i is the number of fragments of *i* type. The α_i term is fragment independent and β term is external descriptors (e.g., topological, electronic, etc.) by default β = 0. Contributions of α_i are calculated by minimizing a functional

$$\boldsymbol{U}[\boldsymbol{ai}] = \sum_{i=1}^{n} w_i \, (\boldsymbol{Y_{exp,i}} - \boldsymbol{Y_{pred,i}})^{\boldsymbol{2}} => \min(2)$$

Where *n* is the number of the compoundsin the training set, w_i the weight accounting for the accuracy of the experimental data, Y_{exp} and Y_{pred} are, respectively, experimental and calculated according to (equation 2) property values. The equation (1) represents the calculation of property *Y* by using additive contributions of fragments. The coefficients of the equation (1) being optimized at the training stage are then used to estimate *Y* values of the compounds from the test set or to screen external databases of real or virtual compounds.

A significant advantage of SMF method is the possibility to select during the training stage several best fit models (instead of a single QSPR model) related to different fragmentation schemes. A Consensus Model (CM) can be calculated by ISIDA/QSPR program which combines the information issued from several models. Using singular value decomposition method (SVD), ISIDA/QSPR fits the a_i terms in equations (3) and calculates corresponding statistical characteristics (correlation coefficient (R), standard deviation (s), Fischer's criterion (F), cross-validation correlation coefficient (Q), standard deviation of predictions (s_{PRESS}), Kubyni's criterion (FIT), RH-factor of Hamilton and matrix of pair correlations (co-variation matrix) for the terms(a_i) and performs statistical tests to select the best models. The predictive ability of the models is characterized by leave-one-out correlation coefficient $Q^2$and by leave-one-out standard deviation s_{PRESS}, as well as by dispersions of predicted values of averages over several models [25].

Model validation

All compounds in each initial data set can be randomly shuffled to avoid possible artificial ordering due to data preparation. Each initial data set was split into two sub-sets: training and test sets. The QSPR models were built on the training set followed by "prediction" calculations for the test set. So internal validation (Q^2_{loo}) and external validation (Q^2_{ext}) should be applied for evaluating the model. The internal validation of the model is necessary for robustness and possible high predictive power. In this research, we have applied the leave-one-out (LOO) for the internal validation, which is calculated according to the formula.

$$\boldsymbol{Q^2_{loo} = 1 - \frac{\sum_{i=1}^{training}(Y_i - \bar{Y}_i)^2}{\sum_{i=1}^{training}(Y_i - \bar{Y})^2}} (3)$$

Where Yi , $\bar{Y}$i and Y are the experimental, predicted, and averaged (over the entire training dataset) values of the samples in the training set.

$$Q^2_{ext} = 1 - \frac{\sum_{i=1}^{test}(Y_i - \bar{Y}_i)^2}{\sum_{i=1}^{test}(Y_i - \bar{Y})^2} (4)$$

Where Yi and $\bar{Y}$i are respectively experimental, predicted values of the test set. The other useful parameters named squared correlation coefficient (R^2) and root mean-squared error (RMSE) were also employed to evaluate the performance of developed models, which are important indicators for linear correlation between predicted and experimental data. They characterize an ability of the model to reproduce quantitatively the experimental data. R^2 is an indicator that measures the linear correlation degree between one variable and another. RMSE indicates the dispersion degree of the random error, which summarizes the overall error of the model.

$$R^2 = \frac{\sum_{i=1}^{n}(Y_{i,pred} - \bar{Y})^2}{\sum_{i=1}^{n}(Y_{i,exp} - \bar{Y})^2} (5)$$

$$RMSE = \left[\frac{1}{n}\sum_{i=1}^{n}(Y_{i,exp} - Y_{i,pred})^2\right]^{0.5} (6)$$

Where $Y_{i,exp}$ is the experimental property in the sample *i*, $Y_{i,pred}$ represented the predicted property in the sample *i* , $\bar{Y}$ is the mean of experimental property in the prediction set and n is the total number of samples in the prediction set[26, 27].

Results and discussion

A QSPR is a mathematical relationship between a property of chemical or is basically a statistical approach correlating the response property data with descriptors encoding chemical information, in this case Boling Point and water solubility in crude oil, and molecular fragments of the chemical.

We used the recently developed substructural molecular fragments (SMF) method which is based on the representation of the molecular graph by fragments and on the calculation of their contributions to a given property. In this paper, molecular fragments were as molecular descriptor and we chose the length of sequences respectively from two to eight. At the training stage, a linear model based on equation (1) involving types of fragments variables has been selected. Fragments contributions represented by the coefficients a_i in equation (1) provide one with helpful information concerning the hydrocarbons.

There are two types of descriptors in ISIDA/QSPR software: atom/bond sequences, and augmented atom, that three sub-types of descriptor of AB, A and B are defined for each class. For each type of Minimal (n_{min}) and Maximal (n_{max}) number of constituted atoms. In this study, fragments contribution illustrates 4 sequences containing 4 , 4 , 5 and 8 for boiling point, 4, 5, 6 and 8 for $logS_{w,}$ atom and linking bonds are selected that were described in the **Tables** (**1**) and (**2**). In the 80 hydrocarbons selected, 64 hydrocarbons were used as training set and 16 hydrocarbons for the test set. In this work, we study two properties of Boiling Point (BP) and water solubility (logSw) in crude oil hydrocarbons. The fragments contribution is calculated based on the following equation:

$$logSw\ or\ BP = A_0 + \sum(A_i \times N_i) \quad (7)$$

whereA_i is contributon of fragment i, N_i is the number of fragments and A_0 term is fragment independent. In the equation (7), we predict property (BP and logSw) for crude oil hydrocarbons with 4 fragments descriptors. The analysis of the fragments contribution in the **Tables**(**1**)and (**2**) illustrate that: 1) some of the fragments bring positive(C-C-C-H and H-C-C-H) or negative(H-C-C-C-H and C-C-C-C-C-C-C-C) contributions into boiling point 2) some of the fragments bring positive(H-C-C-H and H-C-C-C-H) or negative(H-C-C-C-C-H and C-C-C-C-C-C-C-C)contribution into logSw. It is seen that the greatest fragments coefficient have most effect on increasing property. The branched chain compounds have lower boiling points than the corresponding straight chain isomers. We have already observed that the boiling point of straight chain is related to the number of carbon atom in their molecules. Increased intermoleculars are related to the greater molecule-molecule contact possible for larger alkanes.

For example, the boiling point of n-octane (398.83 Kelvin) is higher than the boiling point of 3-ethylhexane (391.83

Kelvin).This is due to the fact that the branching of the chain makes the molecular most compact and thereby decreases the surface area. Therefore, force that is acting here are van Der Waals dispersion forces which are proportional to surface area. The water solubility of the branched hydrocarbon isomers is higher in all instances than for the normal hydrocarbons, as a result, logarithm solubility of the branched hydrocarbons are higher than normal hydrocarbons, because more branching will reduce the size of the molecule, making it easier to solvate.

Table1. Fragments contribution, coefficient (A_i), standard deviation and their t-Test ofthe Equation (7) for BP

No	**Variable (i)**	**Contribution (A_i)**	**Standard deviation**	**t-Test**
0	A_0	198.2671	8.590	23.07
1	C-C-C-H	4.5245	0.308	14.71
2	H-C-C-H	3.8259	0.195	19.67
3	H-C-C-C-H	-1.3124	0.250	5.24
4	C-C-C-C-C-C-C-C	-6.7028	1.300	5.14

Table 2. Fragments contribution, coefficient(A_i), standard deviation and their t-Test of the Equation (7) for logSw

no	**Variable (i)**	**Contribution (A_i)**	**Standard deviation**	**t-Test**
0	A_0	2.1170	0.0139	151.72
1	H-C-C-H	-0.0041	0.0003	13.61
2	H-C-C-C-H	-0.0025	0.0002	11.06
3	H-C-C-C-C-H	-0.0012	0.0001	10.27
4	C-C-C-C-C-C-C-C	0.0099	0.0021	4.77

The statistical parameters QSPR-MLR model is illustrated in the Table 3 for BP is R^2= 0.9966 and Q^2= 0.9919, for logSw is R^2= 0.9914 and Q^2= 0.9785. As regards, proximity values of Q^2 and R^2, prediction logSw and BP is reliable.

Table 3. Statistical parameters of QSPR/MLR models.

Statistical parameters for training and test sets	BP	Log S_w
Multiple correlation coefficient (train)	R^2= 0.9933	R^2= 0.9829
Squared Correlation coefficient LOO –CV (train)	Q^2_{loo}= 0.9919	Q^2_{loo}= 0.9785
Fischer's criterion (train)	F = 2186.8388	F = 851.8474
Standard deviation (train and test)	S= 6.5044,5.3594	SD= 0.010,0.0167
Root mean-squared error (train and test)	RMSE= 6.2452,5.013	RMSE= 0.0096,0.01559
Mean absolute error (train and test)	MAE=5.0799,4.428	MAE= 0.0070,0.0138
Squared Correlation coefficient external prediction (test)	Q^2_{ext}= 0.9951	Q^2_{ext}= 0.9486

The values of experimental, predicted and residuals data for training set and test set of BP and logS_W are shown in supplementary data. The Figures 1 and 2 show that the predicted values are in good agreement with experimental values. Compared with the predicted result of the training and test set, the squared determination coefficient (R^2) is very high and the prediction error is quite low.

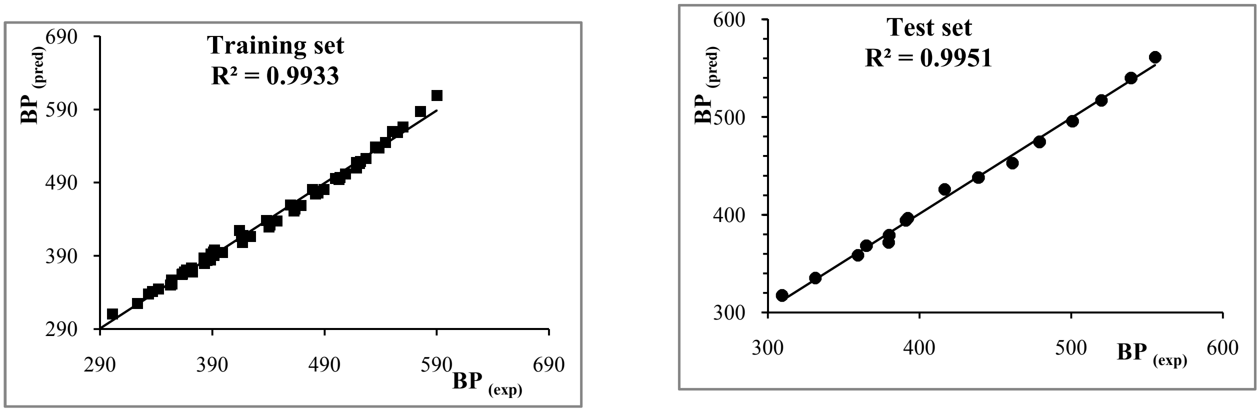

Figure 1. Experimental and predicted values of boiling point for training and test sets

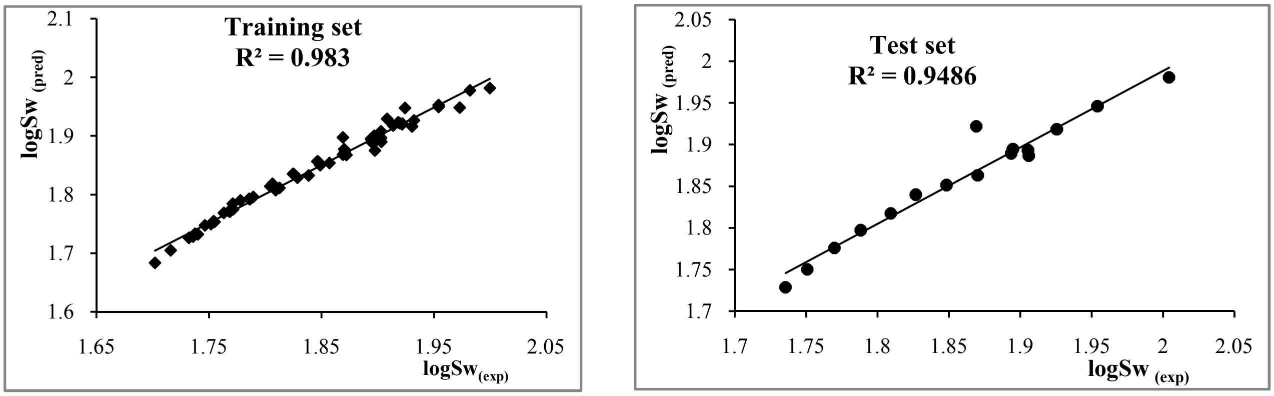

Figure 2. Experimental and predicted values of water solubilityfor training and test sets

Conclusion

The prediction of boiling point and water solubility are an important matter of oil and gas industry. In this present work, we developed modeling QSPR based on the fragment descriptors in ISIDA software. Models based on a fragment (SMF descriptors) had higher prediction ability. MLR modeling method was used to QSPR study of BP and $\log S_w$ data of 80 hydrocarbons in crude oil. The results illustrated that the satisfactory models were obtained, and the prediction errors were small. The results indicate that a strong correlation exists between BP and $\log S_w$ with fragments for hydrocarbons.

Acknowledgements

We wish to thank Prof. V. P. Solovev and Prof A. Varnek for their precious help in use of ISIDA-QSPR software.

References

[1] A. Chapoy, A.H. Mohammadi, D. Richon, B. Tohidi*, Fluid Phase Equillibr.*, **2004**, *220*, 113–121.

[2] J.H. Gary, G.E. Handwerck, *Petroleum Refining Technology and Economics*, 4th ed., **2001**.

[3] S. Mokhatab, W.A. Poe, J.G. Speight, *Handbook of Natural Gas Transmission and Processing*, **2006**.

[4] A. Chapoy, S. Mokraoui, A. Valts, D. Richon, A.H. Mohammadi, B. Tohidi, *Fluid Phase Equilibr.*, **2004**, *226*, 213–220.

[5] Y. Carl L, R. Preetam M, *Oil & Gas Journal*, **2010**, *108(46)*, 130 – 133.

[6] C.C. Rechsteiner, *McGraw-Hill*, **1982**.

[7] H. Wiener, *J. Am. Chem. Soc.*, **1947**, *69*, 17-20.

[8] D. Sola, A. Ferri, M. Banchero, L. Manna, S. Sicardi, *Fluid Phase Equillibr*., **2008** , *263(1)*, 33-42.

[9] J. Ghasemi, S. Saaidpour, *QSAR Comb. Sci.*, **2009**, *28*, 1245-1254.

[10] D. Abooali, M.A. Sobati, *Int. J. refrig.,* **2014** *, 40* , 282-293.

[11] Y.M. Dai, Z.P. Zhu, Z. Cao, Y.F. Zhang, J.I.Zeng, X.Li, *J. Mol.Graph., Model. ,***2013** *,44* , 113-119.

[12] I. Oprisiu, G. Marcou, D. Horvath, D.B. Brunel, F. Rivollet, A.Varnek, *Thermochim. Acta ,* **2013**, *553* , 60- 67.

[13] B. Chen, T. Zhang, T. Bond, Y. Gan, *J. Hazard. Mater.*, **2015**, 299, 260-279.

[14] J.A. Morrill, E.F.C. Byrd, *J. Mol.Graph. Model*, **2015**, *62*, 190-201.

[15] V.P. Solov'ev, A. Varnek *, J. Chem. Inf. Comp. Sci.*, **2003**, *43*, 1703-1719.

[16] V.P. Solov'ev, A.A. Varnek, *Russ. Chem. Bull., Internat. Edit.*, **2004**, *53*, 1434-1445.

[17] A. Varnek, G. Wipff, *J. Chem. Inf. Comp. Sci.*, **2002**,*42*, 812-829.

[18] Y. Carl L, *Houston: Gulf Publish-ing Co*, **2005**.

[19] A. Varnek, G. Wipff, V. P. Solovev, *Solvent Extr. Ion Exc.*, **2001**, *19*, 791-837.

[20] A. Varnek, G. Wipff, V. P. Solovev , A. F. Solotnov*, J. Chem. Inf. Comput. Sci.*, **2002**, *42*, 812-829.

[21] V.P. Solov'ev, N.V. Kireeva, A.Yu.Tsivadze, A.Varnek, *J. Struc. Chem.*, **2006**, *47*, 298-311.

[22]V.P. Solovev, A. Varnek, *Russ. Chem. Bull*., **2004**, *53*, 1434–1445.

[23] A. Varnek, V.P. Solovev, *Chem. High T.* Scr., **2005**, *8(5)*, 403–416.

[24] A. Varnek, D. Fourches, F. Hoonakker. V. P. Solovev, *J. Comput. Aided Mol. Des.,* **2005**, *19*, 693–703.

[25] S. Saaidpour, *Iran. J. Math. Chem.*, **2014**, *5(2)*, 127-142.

[26] S. Saaidpour, *Phys. Chem. Res.*, **2016**, *4 (1)*, 61-71.

[27] S. Saaidpour, *Orient. J. Chem.*, **2014**, *30(2)*, 793-802.