Journal of Hazardous Materials 161 (2009) 74–80

Contents lists available at ScienceDirect

Journal of Hazardous Materials

journal homepage: www.elsevier.com/locate/jhazmat

A quantitative structure property relationship for prediction of solubilization of hazardous compounds using GA-based MLR in CTAB micellar media

Jahan B. Ghasemi*, Azizeh Abdolmaleki, Noushin Mandoumi

Chemistry Department, Faculty of Sciences, Razi University, Kermanshah, Iran

ARTICLE INFO

Article history:
Received 19 February 2008
Received in revised form 13 March 2008
Accepted 13 March 2008
Available online 26 March 2008

Keywords:
Micellization
Quantitative structure property relationship
Cationic micelle
Multivariate linear regression
Hazardous compound

ABSTRACT

QSPR studies for estimating the incorporation organic hazardous compounds in cationic surfactant (CTAB) were developed by application of the structural descriptors and multiple linear regression (MLR) method. Various structure-related descriptors were studied in order to derive information on hydrophobic, electronic and steric properties of solute molecules. Theoretical molecular descriptors selected by genetic algorithms-procedure were followed to predict log K_s values by a stepwise-MLR method. A simple model with low standard errors and high correlation coefficients was selected. It was also found that MLR method could model the relationship between solubility and structural descriptors perfectly. The proposed methodology was validated using full cross validation and external validation using division of the available data set into training and test sets. The squared regression coefficient of prediction for the MLR model was 0.9624. The results illustrated that the linear techniques such as MLR combined with a successful variable selection procedure are capable to generate an efficient QSPR model for predicting the solubility of different compounds. The proposed model can be used adequately for the prediction and description of the solubility of organic compounds in micellar solutions.

1. Introduction

It is very interesting for practical purposes that aqueous micellar solutions can be used instead of more dangerous and toxic organic solvents in a wide range of industrial applications. They provide a simple path for fractionation and concentration of environmental and biological samples under mild conditions. In fact, they allow performing the efficient and selective removal of organic and inorganic hazardous compounds from aqueous streams by simply adding a more, safe, cheap and versatile component so-called amphiphile. The micellar-enhanced ultrafiltration (MEUF) and cloud-point procedure have been applied for preconcentration and removal of several organic pollutants including pesticides, herbicides, aromatic hydrocarbons, PCBs, aliphatic alcohols, aromatic amines, phenols and chlorophenols from aqueous samples [1–4].

Microheterogeneous micellar solubilization environment could play an important role in determining the nature and relative magnitudes of the various factors that contribute to the solubilization [1]. Surfactants can also solubilize materials into solvents other than water. Even when surfactant aggregation does not occur or the aggregation number is small in a particular solvent in the absence of other material, the addition of solvent-insoluble material, such as water, may give rise to aggregation with consequent solubilization of the additive [5]. In this way, surfactant micelles can enhance the sensitivity and can bring about changes in solubility, pK_a, chemical equilibria, reaction rates and mechanisms, spectral distributions and intensities and the stereoselectivity of some chemical processes [6–9]. Therefore, surfactant organized assemblies have great potential application to many processes of technological interest and in analytical chemistry such as separation science. Micellar solutions has proposed for the development of extraction, purification and preconcentration processes according to the ability of micelles to solubilize different compounds [10–12].

One of the most successful approaches for the prediction of chemical properties, starting solely with molecular structural information, is modeling of quantitative structure–activity/property relationships (QSAR/QSPR). The QSPR model provides significant additional insight into the relationship between the molecular structure and fundamental processes and phenomena in chemistry. Such a data processing strategy is useful in describing the relationship between chemicals molecular structures and analytical parameters. The concept that there exists a close relationship between bulk properties of compounds and their molecular structure allows one to provide a clear connection between the macroscopic and the microscopic properties of matter [13]. Many published QSPR models are based on correlations with experimen-

* Corresponding author. Tel.: +98 831 835 8077; fax: +98 831 836 9572.
E-mail address: Jahan.ghasemi@gmail.com (J.B. Ghasemi).

0304-3894/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jhazmat.2008.03.089

tal data, mainly with octanol/water partition coefficients (K_{oc}) and water solubility (S_w), others on molecular structure descriptors and much has been written about solubilization in micellar media. Very little is actually known in any quantitative sense about the relationship between the molecular structure of a neutral solute and its solubility in a given detergent micelle [7,14]. Furthermore, several investigations have developed empirical relationships between the analytical parameters (i.e. cmc or Kraft point) and the structural features of surfactants. However, all of these have been limited to the homologous series of surfactants [15–17].

Therefore, the development of a theoretical model for calculation of the K_s for a diverse set of compounds seems to be necessary. In order to study the influence of the molecular structure on micellar solubility, it is desirable to develop the large possible set of solutes with reliable K_s measurements. Due to the limited amount of micellar solubility data in the literature, this effort focused on solubility a heterogeneous set of neutral compounds in cationic surfactant (CTAB). If structure-property relationships can be generally developed for this subset, it provides a good base to expand the study to examine other micellar media. For the cationic surfactants, solute hydrogen bond acidity favors incorporation into the micelle. CTAB as an ionic surfactant provide a polarizable solubilization environment. It has a higher electrostatic surface potential and a greater degree of counter ion binding. A positively charged micellar surface should provide additional solubilization of orientations of the interfacial water in which the hydrogens point away from the micellar surface [1,5].

2. Experimental

2.1. Data set

The experimental data utilized in this work consists of the micellar solubility (K_s), for 40 solutes in CTAB were reported by Frank Quina et al. [1] (Table 1). The data set is heterogeneous, and includes practically all the principal functional groups present in pesticides, herbicides and various organic pollutants. The standard experimental conditions adopted for the K_s values were ambient temperature (20–30 °C) at low extents of solute incorporator in the absence of significant amounts of added electrolyte or other additives. Some of the chemicals in the literature database have more than one K_s value and the result of being derived from different sources; in these cases were randomly selected. The modeled data were expressed in logarithmic units ($\log K_s$), for chemicals with a $\log K_s$ range of −0.39 to 4.06.

Table 1
Experimental K_s values for incorporation of nonionic solutes in CTAB

No.	Solute	K_s (M^{-1})	No.	Compound	K_s (M^{-1})
Gases and aliphatic hydrocarbons					
1	Oxygen	0.72	4	Propane	33
2	Methane	1.9	5	Cyclohexane	500
3	Ethane	8.7			
Halocarbons					
6	Dichloromethane	5.8	8	Tetrachloromethane	100
7	Chloroform	26	9	1-Iodobuthane	45
Aromatic hydrocarbons and derivatives					
10	Benzene	36	15	Ethylbenzene	154
11	Nitrobenzene	46	16	Anisole	37
12	Chlorobenzene	104	17	Naphthalene	1500
13	Bromobenzene	157	18	Biphenyl	7200
14	Toluene	83	19	Anthracene	5340
Alcohols					
20	1-Propanol	0.5	24	*tert*-Butyl alcohol	1
21	2-Propanol	0.4	25	1-Hexanol	26
22	1-Butanol	2.9	26	Cyclohexanol	6.2
23	Butan-2-ol	1.4	27	Benzyl alcohol	26
Phenols					
28	Phenol	68	29	*p*-Cresol	170
Aldehydes and ketones					
30	Benzaldehyde	15	32	Propiophenone	49
31	Acetophenone	18			
Carboxylic acids and derivatives					
33	Benzoic acid	140	36	Ethyl benzoate	33
34	Benzamide	10	37	*p*-Methyl benzoic acid	63
35	Benzonitrile	15			
Amines					
38	*p*-Toluidine	42	40	Ethyl *p*-aminobenzoate	250
39	Aniline	22			

2.1.1. Descriptor generation and variable selection

The strategy used in this study consists of four fundamental stages: (a) selection of data set, (b) molecular descriptor generation, (c) GA-variable selection and (d) regression analysis. The derivation of theoretical molecular descriptors proceeds from the chemical structure of the compounds. In order to calculate the theoretical descriptors, all molecular structures were constructed with the aid of ChemDraw Ultra version 9.0 computational chemistry software and were optimized using Allinger's MM2 force filed, the semi-empirical AM1 algorithm by MOPAC and further optimization was done using PM3 methods by default on the 3D-structure of molecules in Chem3D Ultra version 9.0 (ChemOffice 2005, Cambridge Soft Corporation) software media.

A total of 54 molecular descriptors of differing types based on 3D structure were calculated to describe compound structural diversity. The descriptors calculated accounts three important properties of the molecules: (a) thermodynamic, (b) electronic and (c) steric, as they represent the possible molecular interactions which determined the solubility of the studied molecules. These descriptors used as input variables for variable selection by genetic algorithm. In our study, a genetic algorithm procedure was used for selection of descriptors using the PLS Toolbox (version 2.0, Eigenvector Company, USA). The GA is implemented in MATLAB (version 7.0, MathWorks, Inc.). For deriving the QSPR model, the GA analysis was begun with multiple linear regression (MLR)-regression method for population size of 64 and mutation rate 0.003. Other parameters summarized in Table 2.

MLR analysis was performed by the SPSS software, (SPSS Ver. 11.5, SPSS Inc.) by using stepwise method for model building.

2.1.2. Data processing

In pre-reduction step, the calculated descriptors were searched for constant values for all molecules and those detected were removed. To decrease the redundancy existed in the descriptors data matrix, the correlation among descriptors and with the $\log K_s$ of the molecules was examined and collinear descriptors (i.e. $R^2 > 0.94$) were detected. Among the collinear descriptors, one with the highest correlation with $\log K_s$ retained and the others were removed from the data matrix. After these steps, 27 descriptors

Table 2
Parameters of genetic algorithm GA

Cross-validation	Random subset
Number of subsets	4
Window width	3
Initial term%	20%
Maximum generation	100
Convergence (%)	80
Cross-over	Double

Table 3
Correlation matrix for selected variables by GA

	log *K*	DPLL[a]	HomoE[b]	log *P*[c]	MP[d]	RepE[e]
log *K*	1					
DPLL	−0.19805	1				
HomoE	−0.10216	0.222757	1			
log *P*	0.860989	−0.23142	−0.02406	1		
MP	0.577207	0.477531	0.007778	0.355147	1	
RepE	0.76541	0.145376	−0.29033	0.650662	0.823401	1

[a] Dipole length.
[b] Homo energy.
[c] Octanol/water partition coefficient.
[d] Melting point.
[e] Repulsion energy.

were retained for next analysis step. Thus, 27 molecular descriptors underwent subsequent variable selection. The genetic algorithm was applied to the input set of these molecular descriptors for each chemical of the studied and the related response. The GA model searching is performed on the different populations of models which can run separately. As a result, a total of five theoretical descriptors were obtained for each of the 40 compounds in the data set which is presented as the correlation matrix in Table 3. Correlation coefficients are a measure of how closely two values (descriptor and property) is related to each other by a linear relationship. If a descriptor has a correlation coefficient of 1, it describes the property exactly. A correlation coefficient of zero means the descriptor has no relevance. It is clear that the appeared descriptors in the MLR model are not highly correlated. To further evaluate the probable collinearity between selected descriptors ridge traces using different lambda values were sketched using ridge function of the Statistics Toolbox 3.0 of MATLAB.

In order to perform examination of the final model, the available data set was split into a training set and external prediction sets through activity sampling. Thus, by ordering the chemicals according to their ascending experimental values, selecting the most and the least active, and taking each chemical from the set in the prediction set. More than 25% of the total data set to be used after model development for the external validation. The training set of 29 compounds with $\log K_s$ values in the range of −0.39 to 4.06, was used to adjust the parameters of the model, and the test set of 11 compounds with $\log K_s$ in the range of 0.76–3.85 was used to evaluate its predictive ability. The underlying goal at this step is to ensure that both the training and the prediction sets separately span the whole descriptor space occupied by the entire data set, and that the chemical domain in the two data sets is not too dissimilar**.**

2.2. Molecular modeling

A major step in constructing the QSPR models is finding a set of molecular descriptors that represent variation in the structural properties of the molecules. The modeling and prediction of the physicochemical properties of organic compounds is an important objective in many scientific fields [18–20]. MLR regression is a linear technique that can determine the relative importance of descriptors, are usually used to generate QSPR models. Also, this model has been successfully employed in various aqueous solubility predictions studies. MLR method provides equation linking the structural features to the $\log K_s$ of the compounds:

$$\log K_s = a_0 + a_1\mathbf{d}_1 + a_2\mathbf{d}_2 + \cdots + a_n\mathbf{d}_n \quad (1)$$

where the intercept (a_0) and the regression coefficients of the descriptors (a_i) are determined by using the least-squares method. $\mathbf{d}_i$ has the common definition, variable or descriptor in this case, the elements of this vector are equivalent numerical values of a 3D structures of the molecules or structural descriptors [21,22]. Here, we used MLR analysis on the molecular descriptors that have been resulted in GA variable selection procedure. The GA-algorithm applied in this paper uses a binary representation as the coding technique for the given problem; the presence or absence of a descriptor in a chromosome is coded by 1 or 0. The GA performs its optimization by variation and selection via the evaluation of the fitness function (RMSECV). The algorithm used in this paper is an evolution of the algorithm described in Ref. [23], whose parameters are reported in Table 2. We obtained a five-descriptor subset, which keeps most interpretive information for $\log K_s$. In the next step, the SPSS software was applied for each chemical compound and the related response, in order to extract the best set of molecular descriptors for construction a simpler model.

3. Results and discussion

In order to find a relationship between $\log K_s$ and the structural features of the chemicals, the molecular descriptors that take into account different structural features was used. We used many different types of molecular descriptors, as the modeling input variables, in order to have the possibility of catching all the relevant structural features related to the studied response. As is general the cases, the choice of descriptors are crucial in order to set up a reasonable model. Since collinear variables degrade the performances of the models obtained by MLR analysis, attempts were made to detect and remove collinear descriptors. Currently, variable selection methods such as genetic algorithm and ant colony optimization are available which represent much better results in comparison with stepwise regression [24–27].

It has already shown that genetic algorithm (GA) can be successfully used [28]. The application of the genetic algorithm-variable subset selection procedure provides a large set of possible models with nearly equivalent predictive performance [29]. The models are based on a variety of descriptors, reflecting the different aspects of molecular structure [30].

3.1. Evaluation and correlation analysis

The MLR technique was performed on the molecules of the training set. We began the analysis by employing the GA technique to search for the optimal linear model containing the best molecular descriptors. The selected descriptors by GA were used to develop a MLR model. The stepwise multiple regression analysis was employed on the training data set to establish the quantitative regression model. Stepwise-MLR is a popular technique that have used on the training data set to select the most appropriate descriptors [31]. After regression analysis, a few suitable models were obtained among which the best model was selected and was employed for prediction. The best equation is selected on the basis of the highest multiple correlation coefficient (R^2) and simplicity of the model. Squared regression coefficient (R^2) is probably the most popular measure of how well a regression model fits the data. Another useful method to evaluate the appropriateness of the selected descriptor is the investigation of the collinearity between descriptors using ridge lambda traces. To this end the ridge traces using different values of the lambda was constructed. The optimum value of the lambda was 0.01967 and the ridge regression coefficients for DPLL, HomoE, log *P*, MP and RepE are −0.1009, 0.0020, 0.5695, 0.0640 and 0.0006, respectively. These results showed that the descriptors HomoE and RepE are collinear and have no considerable statistical impact on the final equation. Multiple linear regression analysis provided a useful equation that can be used to predict of compounds based on these parameters. This QSPR

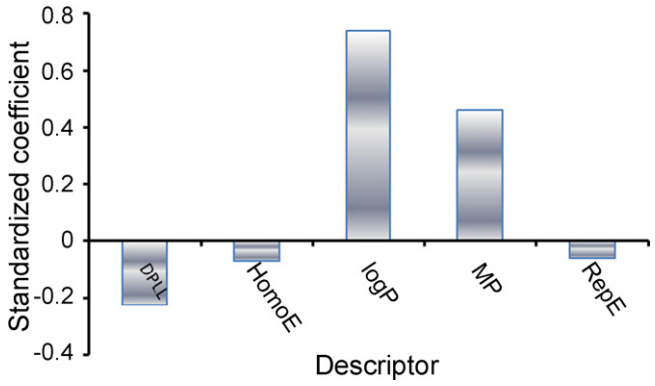

Fig. 1. Standardized coefficients versus descriptors in MLR model.

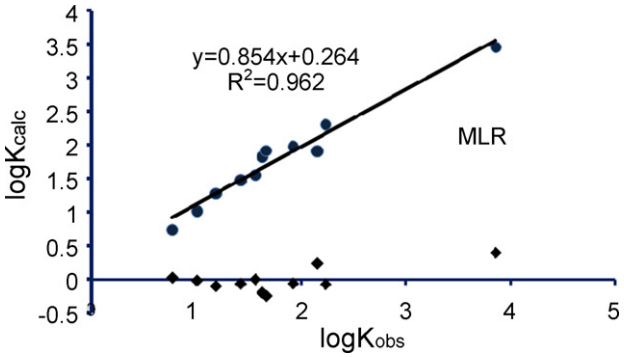

Fig. 2. Log K_s estimated by MLR modeling versus experimental values and residual versus experimental log K_s in MLR model.

Table 4
Standardized coefficients and variable inflation factors (VIF) values for selected descriptors of the MLR model

Source	Standardized coefficients	VIF
log P	0.71420	1.522
MP	0.4246	1.183
DPLL	−0.2373	1.376

model for the solubility of the compounds includes three molecular descriptors. The result obtained from the multivariate combinations is shown in Eq. (2).

$$\log K_s = -1.1522\,(\pm 0.2901) + 0.0070\,(\pm 0.0015)\,\mathrm{MP} + 0.8089\,(\pm 0.0897)\log P - 0.1262\,(\pm 0.0454)\,\mathrm{DPLL} \quad (2)$$

In the proposed model, two of the three variables (log P and MP) are related to the thermodynamic properties of the molecules and DPLL is related to electronic structure of molecules (e.g. the partial charge distribution or the electronegativities of atoms).

Positive values in the regression coefficients reveal that the indicated descriptor contributes positively to the value of log K_s, whereas negative values indicate that the greater values of the descriptor have the lower value of the log K_s. In other words, increasing the DPLL will decrease log K_s and increasing the log P and MP increases extent of log K_s of the organic compounds. Fig. 1 shows the effect of log P, MP and DPLL for the QSPR study of organic solutes in CTAB. Table 4 shows the three independent variables corresponding to MLR model with standardized coefficients and VIF values. The standardized regression coefficient reveals the significance of an individual descriptor presented in the regression model. The greater the absolute value of a coefficient, the greater

Table 5
Comparison of experimental and predicted values of log K_s for test set by MLR model

No.	Exp. (log K)	MLR model	
		Pred. (log K)	RE (%)[a]
6	0.763428	0.735717	−3.62977
33	1	1.016339	1.633874
14	1.176091	1.277663	8.636375
7	1.414973	1.480145	4.605885
10	1.556303	1.553416	−0.18549
37	1.623249	1.824273	12.38403
34	1.919078	1.980928	3.222899
32	2.146128	1.908906	−11.0535
28	2.230449	2.303061	3.255471
9	1.662758	1.913134	15.05786
18	3.857332	3.458646	−10.3358

[a] Relative error.

Table 6
Statistical parameters obtained by applying the MLR and PLS methods to the test set

Parameter	MLR
RMSEP	0.169
REP (%)	9.561
SEP	0.176
R^2	0.9624
NDS[a]	3

[a] Number of descriptors.

the weight of the variable in the model. Thus, the magnitudes and signs of the coefficients should be compatible with the process of transferring a solute and its cavity from bulk water to the micellar pseudophase [1]. Experimental versus predicted values (Table 5) for log K_s values and the residuals (predicted log K_s − experimental log K_s) values, obtained by the MLR modeling, were shown graphically in Fig. 2. This plot shows a good correlation of observed versus calculated solubility data for modeled dependent variables of chemicals. Multiple regression analysis of these data yielded excellent fit (R^2 = 0.9624) for chemicals in CTAB micelle. The model obtained is quite successful, bearing in mind the great variety of functionality: amino, alkyl, hydroxyl, carbonyl, carboxyl groups and aromatic rings. Furthermore, this data have been measured by different groups in the diverse methodology and the lack of rigorous identity of experimental conditions. Methods include solubilization at saturation (SOL), cmc depression (CMC), micellar liquid chromatography (MLC), ultrafiltration (UF) and solute vapor pressure techniques (SVP) [1].

The agreement observed between the predicted experimental values and the random distribution of residuals about zero mean in Fig. 2 confirms the good predictive ability of MLR modeling. For the constructed model, four general statistical parameters were selected to evaluate the prediction ability of the model for solubility. Table 6 shows the statistical parameters for the compounds obtained by applying the MLR method to the test set. The statistical parameters root mean squares error of prediction (RMSEP), relative error of prediction (REP)% and standard error of prediction (SEP)

Table 7
Total variance explained obtained by factor analysis

Factor	Initial eigenvalues	
	Variance percent	Cumulative percent
1	46.20	46.20
2	28.54	74.74
3	18.76	93.50
4	5.45	98.96
5	1.04	100

Extraction method: unweighted least squares.

was obtained for proposed MLR model. Also, the results obtained from factor analysis for training data show that three numbers of factors can explain more than 90% of the variance observed in data (Table 7).

Each of the statistical parameters mentioned above were used for assessing the statistical significance of the QSPR model. Additionally the developed MLR-QSPR model was also checked for multicollinearity problem by the calculation of correlation matrix and variance inflation factor values (VIF) using SPSS program (Table 5). As shown by correlation matrix, $\log P$ and repulsion energy (RepE) show a good correlation (0.86 and 0.76, respectively) with $\log K_s$. It can be seen from this table that RepE and melting point (MP) has a high 0.82 intercorrelation. Furthermore, RepE show a relatively high correlation (0.65) with $\log P$. In both cases, each descriptor encodes different aspects of molecular structure. Thus, RepE is a redundant descriptor that its presence does not improve the results. Then the model suffers from the defect due to collinearity. VIF values greater than five indicates that information of descriptors can be hidden by correlation of descriptors [32,33].

3.2. Model validation

Validation is a crucial aspect of any QSAR/QSPR modeling [34]. The accuracy of proposed MLR model was illustrated using the evaluation techniques such as leave one out (LOO) cross validation procedure and validation through an external test set.

3.2.1. Cross validation technique

Cross validation is a popular technique used to explore the reliability of statistical models. Based on this technique, a number of modified data sets are created by deleting in each case one or a small group (leave-some-out) of objects. For each data set, an input–output model is developed, based on the utilized modeling technique. Each model is evaluated, by measuring its accuracy in predicting the responses of the remaining data (the ones that have not been utilized in the development of the model) [35]. In particular, the LOO procedure was utilized in this study. A QSPR model was then constructed on the basis of this reduced data set and subsequently used to predict the removed data point. This procedure was repeated until a complete set of predicted was obtained. The statistical significance of the screened model was judged by the correlation coefficient (Q^2), error of the estimation (0.38) and the F-statistic (89.97). The predictive ability was evaluated by the cross validation coefficient (Q^2 or R^2_{cv}) which is based on the prediction error sum of squares (PRESS) and was calculated by following equation:

$$R^2_{cv} \equiv Q^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_{i/i})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (3)$$

where $\hat{y}_{i/i}$ denotes the response of the ith object estimated by using a model obtained without using the ith object. Using this approach, the model had a LOO Q^2 of 0.87. This value of Q^2 ($Q^2 > 0.5$) can be considered as proof of the high predictive ability of the model. However, this assumption is in many cases incorrect and can be that exist the lack of the correlation between the high LOO Q^2 and the high predictive ability of QSAR/QSPR models has been established and corroborated recently [34]. Thus, the high value of LOO Q^2 appears to be necessary but not sufficient condition for the models to have a high predictive power. These authors stated that an external set is necessary. As a next step, further analysis was also followed for chemical property of the new set of compounds using the developed QSPR model.

3.2.2. Validation through the external validation set

Validating QSPR with external data (i.e. data not used in the model development) is the best method of validation. However the availability of an independent external validation set of several compounds is rare in QSPR. Thus, the predictive ability of a QSPR model with the selected descriptors was further explored by dividing the full data set. The predictive power of the regression model developed on the selected training set is estimated on the predicted values of prediction set chemicals. A training set (29 solutes) of compounds was used to refine the model and a prediction set (11 solutes) of randomly selected chemicals was chosen to test the model. Experimental and predicted values for $\log K_s$ of prediction set and the relative error values, RE (%), obtained by the MLR modeling were shown in Table 5. A value of R^2 near one indicates a perfect linear fit.

3.3. Interpretation of descriptors

The micellization potential of cationic surfactants covering a diverse range of structures is found to be well modeled by a combination of three parameters consisting of electronic and thermodynamic properties. The best three-parameter equation obtained for the prediction of solubility for an unknown compound. The most important descriptors in this correlation are the *thermodynamic* and *electronic* features in the compound, the factors that influence the solubility of each species. The QSPR developed indicated that lipophilicity ($\log P$) of the solutes, melting point (MP) of the molecules and Dipole length (DPLL) of the solutes; the factors that influence the solubility of each species and satisfactorily describes the solubility of structurally different solutes. Comparison of the standardized regression coefficient of the descriptors appearing in MLR model shows that the $\log P$ of the molecules has the largest effect on the K_s of the cationic surfactant (CTAB).

The models for predicting micellar solubility of neutral solutes based on calculated descriptors and electrostatic interaction have previously been developed using PLS and MLR regression in our research group [30]. The equation derived in the mentioned work, has presented a QSPR model for solubility in an anionic surfactant (SDS). We reported that the solubility of the organic chemicals was mainly controlled by the hydrophobicity of the compounds with the electronic property of minor role. In agreement with our previous study, the hydrophobicity descriptor ($\log P$) plays an important role in micellar solubility. The octanol/water partition coefficient, usually as its logarithm, $\log P$ characterizes the effectiveness of hydrophobicity of the compounds. It is a very important indicator of transport and permeation through membranes, interaction with biological receptors and enzymes, toxicity, and biological potency. In environmental sciences the hydrophobicity is often used to predict solubility, the bioconcentration factor, and the organic adsorption coefficient (K_{oc}). Although hydrophobicity is a key concept in surfactant science, to date $\log P$ has not found much application in this field [17,30,36]. The fact that similar descriptors have been reported to correlate with partition coefficients of different compounds suggests that this correlation model has wider applications [37]. The positive standardized coefficient for $\log P$ parameter is in accordance with physical considerations; compounds with higher hydrophobicities have stronger interactions with the medium and thus enhances the solubility of chemicals.

The melting point (MP) is a fundamental physical property of compounds, which has been wide used for the calculation of other physicochemical properties such as vapor pressure, aqueous solubility, transporting within organism and phase equilibrium properties. It is generally accepted that solubility of a compound is strongly correlated with its melting point. Prediction models

for solubility that include the melting point as a descriptor often result in acceptable accuracy of the solubility predictions [30,38]. If the forces holding the molecule in the crystal are strong, then the solubility and vapor pressure will be low. Conversely, the melting point will be high, as the melting point is a measure of the energy required to disrupt the crystal lattice. For organic compounds, the dominant intermolecular force affecting the melting point is intermolecular hydrogen bonding. Compounds with intramolecular hydrogen bonding normally exert less intermolecular attraction and, therefore, have a lower melting point than their intermolecular hydrogen bonded analogues. Melting point also affects the toxicity of a compound. As noted above, melting point affects solubility, and solubility controls toxicity. If a compound is only poorly soluble, its concentration in the aqueous environment may be too low for it to exert a toxic effect. Therefore, the melting point of a compound depends mainly on the molecular size and symmetry, as well as on intermolecular interactions. This physical property for a crystal is governed by the hydrogen bonding ability of the molecules, the molecular packing in crystals (effects from molecular shape, size, and symmetry), and other intermolecular interactions such as charge transfer and dipole–dipole interactions in the solid phase [30,39].

The remaining descriptor, dipole length (DPLL) is an electronic descriptor. In particular, electronic parameters are considered important in establishment of QSAR models and are helpful to quantify different types of intermolecular and intramolecular interactions, as these interactions are usually responsible for properties of chemical and biological systems [30]. Dipole length is the electric dipole moment divided by the elementary charge. Electric dipole is a vector quantity, which encodes displacement with respect to the centre of gravity of positive and negative charges in a molecule. Dipole length encodes information about the charge distribution in molecules and is important for modeling polar interactions. Large substituents decrease DPLL value which is not desirable [30,40]. The relative insensitivity of K_s to solute dipolarity is attributed to the fact that molecules with significant dipolarity preferentially solubilize at or near the micellar–water interface in what is often described as an "alcohol-like" medium [1,30].

With considering the heterogeneity of the solutes with respect to molecular structure, size, hydrogen-bonding affinities and polarity, it is remarkable that a single model is adequate to correlate the data. Finally, since the coefficients of the regression equation appear to be chemically reasonable, it should be provide a novel means of exploring the relationship between detergent molecular structure and nature of the corresponding micellar solubilization microenvironment. Hence, we conclude that using physicochemical descriptor correlated the solubility data a good correlation is obtainable for solubility in cationic micellar media.

4. Conclusion

The QSPR model provides significant additional insight into the relationship between the molecular structure and fundamental processes and phenomena in chemistry. Such a data processing strategy is useful in describing the relationship between chemical molecular structures and analytical parameters. Therefore, the successful description of the micellar solubility presented with a few physicochemical significant molecular descriptors for diverse chemical compounds in a cationic surfactant (CTAB). These are simple to calculate, providing a rapid and accurate method for estimation and description of solubility behavior in micellar solutions. After GA-variable selection, stepwise-MLR analysis was followed to develop a model for predicting the solubility in aqueous micellar solution. The descriptors involved in the correlations reflect both the intermolecular and intramolecular interactions. In agreement with previous studies, it was found, the incorporation solutes in aqueous micellar solutions ($\log K_s$) are primarily determined by the hydrophobic part of molecule. Our study indicates the quantitative relationship between structure and property is a tool for the quantitation of physicochemical properties of solutes and for the prediction such as micellar solubility. Based on obtained results, it seems that quantitative structure activity/property relationships (QSAR/QSPR) could be quite useful for understanding processes that involve the transfer of solutes between two phases.

References

[1] F.H. Quina, E.Q. Alonso, J.P.S. Farah, Incorporation of nonionic solutes into aqueous micelles: a linear solvation free energy relationship analysis, J. Phys. Chem. 99 (1995) 11708–11714.
[2] D.K. Taylor, R. Carboneln, J.M. DeSimone, Opportunities for pollution prevention and energy efficiency enabled by the carbon dioxide technology platform, Annu. Rev. Energy Environ. 25 (2000) 115–146.
[3] J.F. Brennecke, M.A. Stadtherr, A course in environmentally conscious chemical process engineering, Comput. Chem. Eng. 26 (2002) 307–318.
[4] S.P. Beaudoin, C.S. Grant, R.G. Carbonell, Removal of organic films from solid surfaces using aqueous solutions of nonionic surfactants. 1. Experiments, Ind. Eng. Chem. Res. 34 (1995) 3307–3317.
[5] M.J. Rosen, Surfactant and Interfacial Phenomena, 3rd ed., John Wily & Sons, New York, 2004, p. 190.
[6] W.L. Hinze, Solution Chemistry of Surfactants, Plenum Press, New York, 1979.
[7] E. Pramauro, A. Bianco Prevot, Solubilization in micellar systems. Analytical and environmental applications, Pure Appl. Chem. 67 (4) (1995) 551–559.
[8] R. Codony, M.D. Prat, J.L. Beltraňn, Study on partition equilibria of metal complexes innon-ionic micellar solutions from spectrophotometric data, Talanta 52 (2000) 225–232.
[9] A. Stoyanova, A. Alexiev, Surfactants and kinetic determinations of microelements, Trakia J. Sci. 3 (2005) 1–9.
[10] S. Shahmirani, E. Vasheghani Farahani, J. Ghasemi, Development of a model to predict partition coefficient of organic pollutants in cloud point extraction process, Anal. Chim. 96 (2006) 327–338.
[11] M.H. Abraham, H.S. Chdha, J.P. Dixon, C. Rafols, C. Treiner, Hydrogen bonding. Part 40. Factors that influence the distribution of solutes between water and sodium dodecylsulfate micelles, J. Chem. Soc. Perkin Trans. 2 (1995) 887–894.
[12] C.O. Rangel-Yagui, A. Pessoa-Jr, L.C. Tavares, Micellar solubilization of drugs, J. Pharm. Pharm. Sci. 8 (2) (2005) 147–163.
[13] J. Ghasemi, S. Saaidpour, S.D. Brown, QSPR study for estimation of acidity constants of some aromatic acids derivatives using multiple linear regression (MLR) analysis, Theochem 805 (2007) 27–32.
[14] P.V. Khadikar, D. Mandloi, A.V. Bajaj, Sh. Joshi, SAR study on solubility of alkanes in water and their partition coefficients in different solvent systems using PI index, Bioorg. Med. Chem. Lett. 13 (2003) 419–422.
[15] P.D.T. Huibers, V.S. Lobanov, A.R. Katritzky, D.O. Shah, M. Karelson, Prediction of critical micelle concentration using a quantitative structure-property relationship approach. 1. Nonionic surfactants, Langmuir 12 (1996) 1462–1470.
[16] M. Jalali-Heravi, E. Konuze, Use of quantitative structure-property relationships in predicting the Kraft point of anionic surfactants, Int. Electr. J. Mol. Des. 1 (2002) 410–417.
[17] D.W. Roberts, Application of octanol/water partition coefficients in surfactant science: a quantitative structure-property relationship for micellization of anionic surfactants, Langmuir 18 (2002) 345–352.
[18] O. Ivanciuc, T. Ivanciuc, A. Balaban, Quantitative structure-property relationship study of normal boiling points for halogen-oxygen, Tetrahedron 54 (1998) 9129–9142.
[19] O. Ivanciuc, T. Ivanciuc, P.A. Filip, D. Cabrol-Bass, Estimation of the liquid viscosity of organic compounds with quantitative structure-property model, J. Chem. Inf. Comput. Sci. 39 (1999) 515–524.
[20] M. Citra, Estimating the pK_a of phenols, carboxylic acids and alcohols from semi-empirical quantum chemical methods, Chemosphere 38 (1999) 191–206.
[21] J. Ghasemi, S. Saaidpour, QSPR prediction of aqueous solubility of drug-like organic compounds, Chem. Pharm. Bull. 55 (2007) 669–674.
[22] J. Ghasemi, S. Asadpour, A. Abdolmaleki, Prediction of gas chromatography/electron capture detector retention times of chlorinated pesticides, herbicides, and organohalides by multivariate chemometrics methods, Anal. Chim. Acta 588 (2007) 200–206.
[23] J.H. Holland, Genetic algorithms, Sci. Am. 267 (1992) 66–72.
[24] A.S. Barros, D.N. Rutledge, Genetic algorithm applied to the selection of principal components, Chemom. Intell. Lab. Syst. 40 (1998) 65–81.
[25] C. Yin, X. Liu, W. Guo, T. Lin, X. Wang, L. Wang, Prediction and application in QSPR of aqueous solubility of sulfur-containing aromatic esters using GA-based MLR with quantum descriptors, Water Res. 36 (2002) 2975–2982.
[26] S.S. Liu, H.L. Yin, L.S. Wang, VSMP: a novel variable selection and modeling method based on the prediction, J. Chem. Inf. Comput. Sci. 43 (2003) 964–969.
[27] S.S. Liu, C.S. Yin, L.S. Wang, Combined MEDV–GA–MLR method for QSAR of three panels of steroids, J. Chem. Inf. Comput. Sci. 42 (2002) 749–756.

[28] R. Leardi, A.L. Gonzales, Genetic algorithms applied to feature selection in PLS regression: how and when to use them, Chemom. Intell. Lab. Syst. 41 (1998) 195–207.
[29] P. Gramatica, E. Giani, E. Papa, Statistical external validation and consensus modeling: a QSPR case study for K_{oc} prediction, J. Mol. Graph. Mod. 25 (2007) 755–766.
[30] J. Ghasemi, A. Abdolmaleki, S. Asadpour, F. Shiri, Prediction of solubility of non-ionic solutes in anionic micelle (SDS) using a QSPR model, QSAR Comb. Sci. 27 (2008) 338–346.
[31] G. Melagraki, A. Afantitis, H. Sarimveis, P.A. Koutentis, J. Markopoulos, O. Igglessi-Markopoulou, Identification of a series of novel derivatives as potent HCV inhibitors by a ligand-based virtual screening optimized procedure, Bioorg. 15 (2007) 7237–7247.
[32] G. Choudhary, C. Karthikeyan, N.S. Moorthy Hari Narayana, S.K. Sharma, P. Trivedi, QSAR analysis of some cytotoxic thiadiazinoacridines, Int. Elec. J. Mol. Des. 4 (2005) 793–802.
[33] H.D. Cho, S.K. Lee, B.T. Kim, K.T. No, Quantitative structure–activity relationship (QSAR) study of new fluorovinyloxyacetamides, Bull. Korean Chem. Soc. 22 (2001) 388–394.
[34] J. Acevedo-Martínez, J.C. Escalona-Arranz, A. Villar-Rojas, F. Téllez-Palmero, R. Pérez-Rosés, L. González, R. Carrasco-Velar, Quantitative study of the structure–retention index relationship in the imine family, J. Chromatogr. A. 1102 (2006) 238–244.
[35] A. Afantitis, G. Melagraki, H. Sarimveis, P.A. Koutentis, J. Markopoulos, O. Igglessi-Markopoulou, A novel QSAR model for predicting induction of apoptosis by 4-aryl-4H-chromens, Bioorg. Med. Chem. 14 (2006) 6686–6694.
[36] A.J. Leo, Calculating log P_{oct} from structures, Chem. Rev. 93 (1993) 1281–1306.
[37] A.R. Katritzky, V.S. Lobanov, M. Karelson, QSPR: the correlation and quantitative prediction of chemical and physical properties from structure, Chem. Soc. Rev. 24 (1995) 279–287.
[38] N. Sun, H.X. Uezhong, K. Dong, X. Zhang, H.H. Lu, S. Zhang, Prediction of the melting points for two kinds of room temperature ionic liquids, Fluid Phase Equilib. 246 (2006) 137–142.
[39] A.R. Katritzky, R. Jain, A. Lomaka, R. Petrukhin, U. Maran, M. Karelson, Perspective on the relationship between melting points and chemical structure, Cryst. Growth. Des. 1 (2001) 261–265.
[40] R. Todeschini, V. Consonni, R. Mannhold, H. Kubinyi, H. Timmerman (Series Editor), Handbook of Molecular Descriptors, Weinheim, Wiley–VCH, 2000.