MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **71** (2014) 305-321

ISSN 0340 - 6253

Predicting Critical Micelle Concentration by Using Stepwise – MLR and PLS as a Variable Selection Mix Method

Reza Behjatmanesh–Ardakani[1,*], Seyed Mohsen Mirhosseini[2], Fatemah Ghaderiyeh–Mahmood Abadi[3]

[1]*Dep. of Chemistry, Payame Noor University, PO Box 19395-3697, Tehran, Iran*

[2]*Dep. of Statistics, Yazd University, Yazd, Iran*

[3]*Dep. of Chemistry, Payame Noor University, PO Box 19395-3697, Tehran, Iran*

(Received June 23, 2012)

Abstract

In this paper, multiple linear regression (MLR) and partial least squares regression (PLS) models are used to estimate critical micelle concentration of anionic surfactants. RHF/6-31G* level of theory is used for collecting quantum chemical descriptors. In addition, a small number of topological descriptors are also utilized. The best descriptors are selected by PLS, the stepwise and Enter methods of MLR. The determination coefficient (R^2) and mean square of error (*MSE*), for the best model of training set are 0.989 and 0.007, respectively. Validation is guaranteed by calculation of determination coefficient for prediction set (R^2_{pred}), that is higher than 0.98. In addition to the above gas phase model, conductor–like polarizable continuum model (CPCM) is also used to calculate all descriptors in the solution phase. This new model has R^2 0.981.

Introduction

Surfactants are versatile products used in variety industries such as detergents, emulsifiers, drilling muds and the flotation agents. Nowadays, these are applied in high-technology areas such as micro-electronics [1, 2] and nanotechnology [3]. At a critical concentration named as

* Corresponding author
Email address: reza_b_m_a@yahoo.com, behjatmanesh@pnu.ac.ir

CMC, surfactants aggregate to form micelles. Micelles are in equilibrium whit smaller cluster sites and monomers [1,2]. One particular important point in industry is CMC determination/prediction.

One of the methods for the prediction of chemical properties such as CMC based only on molecular structural information is quantitative structure–activity/property relationships (QSAR/QSPR) modeling. QSAR/QSPR models are mathematical equations relate chemical structure to a property. QSAR/QSPR modeling is found to be extremely helpful in prediction of activity/property of new chemical compounds. For a QSAR or QSPR modeling following items should be done: (1) descriptor generation, (2) splitting data to training and prediction (or validation) sets, (3) variable selection, (4) finding appropriate model between selected variables and activity/property and (5) model validation. A large number of descriptors can be generated by existing codes. These parameters categorized as geometrical, topological, physicochemical and electronic descriptors [4,5]. Choosing adequate descriptors for QSAR/QSPR studies is difficult and challenging. To overcome this problem a powerful variable selection technique is needed [6,7].

In the present work, the best variables among electronic and topological descriptors are selected using PLS and MLR and then a linear MLR model is developed to predict the *LogCMC* of some surfactants. Our method shows a considerable improvement for regression models.

Methods

Six steps are used for estimation *LogCMC* of anionic surfactants by QSPR modeling. These contain 1) generation of the files containing the chemical structures in a computer- readable format, 2) optimization of molecule geometries with an *ab initio* method, 3) computing electronic and topological descriptors, 4) selection of descriptors by PLS and MLR to study the relationship between structural features and the *LogCMC*, and 6) building of the model.

Software

HyperChem version 7. 0 program [8] is used to draw the molecular structures. The structures of the compounds are pre-optimized by the semi-empirical method PM3. The descriptors are then calculated by the Gaussian 98 software [9]. MLR regression is performed by the SPSS 15.0 package [10], and PLS is performed by Minitab 15.0 program [11].

Data set

CMC values of anionic surfactants, are taken from References 1, 12 and 13. The compounds contain different types of structures rather than represent particular class of molecules. Table 1 shows the list of compounds in which long and short, straight chain and branched, aliphatic and aromatic and unsaturated carboxylate, sulfonate and sulfate surfactants are present. The data set is split into a training set and a prediction set. The prediction set of 10 surfactants is selected randomly. The training set of 31 compounds, with *LogCMC* values in the range of -0.333 to -3.635, is used to adjust the parameters of the model [12,13].

Molecular modeling and structural descriptors

Standard *ab initio* molecular orbital calculation is carried out using the Gaussian 98 software. All geometries of the species are fully optimized at the RHF/6-31G(d) level of theory. The electronic descriptors for each species are calculated in both gas and solution phases. However, gas and solution phases descriptors are considered separately by defining two different models. Conductor-like polarizable continuum model (CPCM) is used to calculate properties in solution [14]. These calculations are done using SCRF=CPCM keyword.

PLS modeling

In partial least squares regression (PLS) a multivariate approach is used by which molecular properties (variables) are shown by a *X*-matrix, which is then related to a response matrix, *Y*. PLS uses an approach to reduce the number of independent variables by linear combinations of original *x*-values. By this trick, collinearity problem is solved and better regression equations are obtained. The matrix consisting of the *A* scores (number of principal components) is denoted by *T* and the corresponding matrix of loadings is denoted by *P*. Set of input matrices can be written as [5, 6]:

$$X = TP^{t} + E \tag{1}$$

$$y = Tq + f \ . \tag{2}$$

Table 1. Comparison between experimental and calculated *LogCMC* values the for training set with equation (15). Experimental data are for 40 °C.

No	Formula	*LogCMC*(Exp.)	*LogCMC*(Pred.)	Residuals
1	n-$C_8H_{17}COO^-K^+$	-0.333	-0.294	-0.039
2	n-$C_{10}H_{21}COO^-K^+$	-0.936	-0.96	0.024
3	n-$C_{12}H_{25}COO^-K^+$	-1.538	-1.56	0.022
4	n-$C_{14}H_{29}COO^-K^+$	-2.137	-2.160	0.023
5	n-$C_8H_{17}SO_3^-Na^+$	-0.79	-0.915	0.125
6	n-$C_{10}H_{21}SO_3^-Na^+$	-1.4	-1.317	-0.083
7	n-$C_{12}H_{25}SO_3^-Na^+$	-1.943	-1.935	-0.008
8	n-$C_{14}H_{29}SO_3^-Na^+$	-2.602	-2.685	0.083
9	n-$C_{15}H_{31}SO_3^-Na^+$	-3.139	-2.984	-0.155
10	n-$C_8H_{17}SO_4^-Na^+$	-0.866	-0.898	0.044
11	n-$C_{10}H_{21}SO_4^-Na^+$	-1.481	-1.485	0.004
12	n-$C_{11}H_{23}SO_4^-Na^+$	-1.783	-1.780	-0.003
13	n-$C_{12}H_{25}SO_4^-Na^+$	-2.063	-2.077	0.014
14	n-$C_{13}H_{27}SO_4^-Na^+$	-2.367	-2.374	0.007
15	n-$C_{14}H_{29}SO_4^-Na^+$	-2.620	-2.672	0.052
16	n-$C_8H_{17}OOC(CH_2)_2SO_3^-Na^+$	-1.312	-1.418	0.106
17	n-$C_{10}H_{21}OOC(CH_2)_2SO_3^-Na^+$	-1.883	-1.913	0.03
18	n-$C_{12}H_{25}OOC(CH_2)_2SO_3^-Na^+$	-2.523	-2.452	-0.071
19	$C_{10}H_{21}CH(CH_3)SO_3^-Na^+$	-1.827	-1.949	0.122
20	$C_8H_{17}CH(C_3H_7)SO_3^-Na^+$	-1.635	-1.593	-0.042
21	$C_7H_{15}CH(C_4H_9)SO_3^-Na^+$	-1.548	-1.521	-0.027
22	$C_6H_{13}CH(C_5H_{11})SO_3^-Na^+$	-1.442	-1.469	0.027
23	$C_7H_{15}CH(C_7H_{15})SO_3^-Na^+$	-2.144	-2.263	0.119
24	$C_6H_{13}CH(CH_3)SO_4^-Na^+$	-0.745	- 0.593	-0.152
25	$C_8H_{17}CH(CH_3)SO_4^-Na^+$	-1.305	-1.134	-0.171
26	$C_5H_{11}CH(C_5H_{11})SO_4^-Na^+$	-1.081	-1.129	0.048
27	$C_{11}H_{23}CH(CH_3)SO_4^-Na^+$	-2.187	-2.115	-0.072
28	$C_9H_{20}CH{=}CHCH_2SO_3^-Na^+$	-1.886	-1.914	0.028
29	$C_{11}H_{24}CH{=}CHCH_2SO_3^-Na^+$	-2.569	-2.51	-0.059
30	Para $C_8H_{17}C_6H_4SO_3^-Na^+$	-1.907	-1.969	0.062
31	$C_{17}H_{35}SO_3^-Na^+$	-3.635	-3.581	-0.054

The matrix *E* and vector *f* contain residuals for *X* and *y*, respectively, and vector *q* is loading for *y*. PLS regression is obtained by maximizing the covariance between *y* and all possible linear functions of *X*. The regression coefficient vector used in the linear PLS predictor can be computed using the equation:

$$\hat{b} = \hat{w}(\hat{P}^T\hat{w})^{-1}\hat{q}\,. \qquad (3)$$

where the $\hat{w}$ is the matrix of loading weights(5, 6).

MLR modeling

Multiple linear regression (MLR) is another model that considers the relation between descriptors (x) and response variable (y) to be shown by linear equation [5, 6]:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + e_i \quad (4)$$

β's are unknown constants named as regression coefficients that should be estimated during modeling and e_i is residual. If errors are normal with constant variance and independent, the best regression coefficients will obtain by the method of Least squares [15]. The regression coefficients may be made comparable with each other by standardizing appropriately.

Evaluation of MLR model is done by cross-validation techniques. Two main cross-validation methods consist of leave-one-out (LOO) and leave-more-out (LMO). In these methods training set is modified by removing random small group molecules in each step and then the model is evaluated by measuring its accuracy in predicting the responses of the deleted group. A successful model must have an ability to predict not only the property of internal molecules (internal validation) but also of the external sources (external validation). So, applying only LOO-CV or LMO-CV is not sufficient to evaluate the predictive ability of a model [16].

Besides above cases, an important point in a statistical model is outliers. There is a criterion for identity outliers. The $|e_i|$ is computed and examined for each compound. If the absolute value of any $|e_i|$ is greater than $3\sqrt{MSE}$,this indicates that the sample "i" observation should be carefully scrutinized as a possible outlier [15].

Statistical parameters

For the constructed models, eight general statistical parameters are selected to evaluate the ability of the models. The *PRESS* (predicted residual sum of squares) statistic appears to be the most important parameter for a good estimate of the real predictive error of the models. Its small value indicates that the model predicts better than chance and can be considered statistically significant. It is calculated by following equation [15]:

$$PRESS = \sum_{i=1}^{n} (e_i)^2 = \sum_{i=1}^{n} \left[y_i - \hat{y}_{(i)} \right]^2 . \quad (5)$$

where y_i is the experimental *LogCMC* of the anionic surfactant in the sample *i*, $\hat{y}_i$ represented the predicted *LogCMC* of the anionic surfactant in the sample *i*, and *n* is the total number of samples used in the prediction set.

Cross-validated R^2_{CV} (or R^2_{pred}) explains variance in prediction:

$$R^2_{CV} = R^2_{pred} = \frac{PRESS}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \quad (6)$$

$\bar{y}$, is the mean of experimental *LogCMC* in the prediction set. The coefficient of determination (R^2) indicates the quality of fit and is calculated as:

$$R^2 = 1 - \frac{SSE}{S_{total}} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \quad (7)$$

The fourth statistical parameter is the adjusted R^2:

$$R^2_{Adj} = \bar{R}^2 = 1 - \frac{SSE/(n-p)}{S_{total}} = 1 - \frac{n-1}{n-p}(1-R^2). \quad (8)$$

where, *n* is the number of members of the training set and *p* is the number of parameters involved in the correlation. R^2 increases with additional predictors and is somewhat sensitive to changes in *n* and *p*. In particular, in small samples, if *p* is large relative to *n*, there is a tendency for R^2 to be artificially high. The adjusted R^2 corrects for the artificiality introduced when *p* approaches *n* through the use of a penalty function which scales the result [15].

The next statistical parameter is mean of squares error (*MSE*):

$$MSE = \frac{SSE}{n-p} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-p}. \quad (9)$$

MSE represents the standard distance data values far from the regression line. For a given study, the better the equation predicts the response, the lower *MSE*. *F*-ratio test is the most well-known statistical tests, this is defined as the ratio between the *MSR* and the *MSE*:

$$F = \frac{MSR}{MSE}. \quad (10)$$

F is compared to the critical value F_{crit} for the corresponding degrees of freedom. High values of the *F*-ratio test indicate reliable models. Associated with each predictor variable x_j is a number denoted by VIF_j, called the variance inflation factor for x_j, which is defined as [17]:

$$VIF_j = \frac{1}{1-r^2_{x_j}}. \quad (11)$$

where r is the correlation coefficient of one independent variable against others. Large VIF_j values imply strong correlation. It has been suggested, as a rule of thumb, that values of VIF_j greater than 10.0 may be considered large enough to suspect serious multicollinearity [9].

The next statistical parameter is Durbin-Watson that indicate residuals are independent or not. 'Independent' means that the individual e_{ij} are randomly distributed and not influenced by an external factor.

Increasing the number of measurements increases the precision of the prediction. According to equation (12), increasing the number of measurements (n) decreases SSE and increase R^2 and F according to equations (7) and (10).

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad (12)$$

This is not always the best approach. Since for large samples the time and cost increase. When using smaller sample sizes and the broader distribution, we have more precision: the standard deviation x is larger and residuals variance is smaller in equations (13) and (14).

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{S_x^2}, \qquad \sigma^2 = const \quad (13)$$

$$Var(\hat{\beta}) = (X'X)^{-1}\sigma^2 \quad (14)$$

where S^2_x, σ^2 and $Var(\hat{\beta}_1)$, are standard deviation x, residuals variance and coefficients variance respectively.

Results and discussion

PLS analysis

Figure 1 shows R^2 and R^2_{LOO} as a function principle components (*PCs*). It shows that minimum six *PCs* needed to obtain accurate PLS model. The statistical parameters of models are given in Table 2. The score plot, Figure 2, is a scatterplot of the x-scores from the sixth and fifth components in the model. The sixth components explain most of the variance in the predictors, then the configuration of the points on this plot reflects the original multidimensional configuration of our data. In this study, the score plot reveals that samples 5 and 26 may have high leverage values. These samples may appear as outliers or leverage points on other plots. The coefficient plot, Figure 3, displays the relative importance of the

variables for modeling the response. As seen in Figure 3, the most important variables used for modeling the LogCMC in PC_6 are: the total dipole moment (μ(g)) [18], the polarizability (α) [19], softness (S) [20], the minimum natural atomic orbital taken from the results of natural population analysis (NAO_{min}) [21] all in gas phase, the total dipole moment (μ(aq)) in aqueous phase, the bond number (B) and the number of atoms (A) which are all negatively correlated to the response. It is interesting to note that E_{LUMO}, [22] the Gibbs energy (ΔG), the maximum atomic charge on the atom C taken from the results of natural population analysis (q_C^+nbo) [23], the atomic charge on the atom C 1 taken from the results of population analysis based on molecular electrostatic potential (mep_C1) [24] and the HOMO/LUMO energy fraction (fH/L) in the gas phase, the electronegativity (χ) [25], the maximum atomic charge on the atom C based on Mulliken population analysis (q_C^+(aq)) in aqueous phase and the total change in the Gibbs free energy (ΔG_{total}) display a positive correlation to the response. A high value of these variables denotes a large response.

Table 2. Results from the PLS analysis

Components	*x* variance	R^2	*PRESS*	R^2_{LOO}
1	0.3377	0.5822	9.4907	0.4016
2	0.5085	0.8950	23.9825	0
3	0.6191	0.9702	64.3433	0
4	0.7504	0.9837	31.5275	0
5	0.8690	0.9864	11.1089	0.2996
6	0.9226	0.9883	0.4178	0.9736

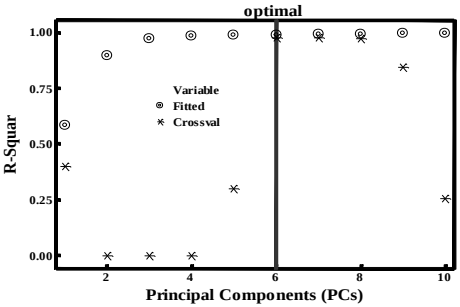

Figure 1. The model selection plot that is a scatterplot of the R^2 and R^2_{LOO} values as a function of the number of components.

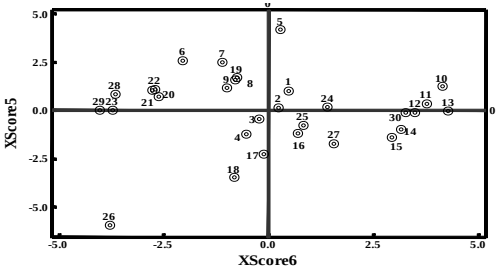

Figure 2. The score plot that is a scatterplot of the *x*-scores from the sixth and fifth components in the model

PC_6 predicts cross-validated value (R^2_{LOO}) of 97.36%, of variances in the *LogCMC* of the studied molecules, see Table 2. This indicates that variables are sufficient for modeling the response, as is visualized in Figure 4. Also low *PRESS* value indicates that PC_6 is excellent, see Table 2.

MLR analysis

As a first check, we investigate the normal distribution of residuals. In Figure 5 frequency is drown against standardized residuals and Figure 6 shows standardized predicted values *LogCMC* plotted against standardized residuals. Based on Figure 5, it is clear that the residuals have normal distribution. In Figure 6, the data in this plot are randomly scattered about the horizontal line showing that the residuals variance is constant.

An aberration in stepwise regression is obtaining chance correlation models. This is occurred when the ratio of the number of samples (or molecules) to the number of original variables (or descriptors) is very low. To develop QSPR models, a new approach is used. In this approach the common descriptors between PLS and stepwise MLR methods are considered as inputs for ENTER MLR method. Five descriptors are common between two methods containing total dipole moment (μ(g)), the minimum natural atomic orbital between carbon atoms taken from the results of natural population analysis (NAO_{min}) both in the gas phase, total dipole moment (μ(aq)), the maximum atomic charge on the atom C based on Mulliken population analysis (q_C^+(aq)) in aqueous phase and the bond number (B). NAO_{min} is the sum of core, valence and Rydberg natural orbitals calculated with the keyword POP=NPA in Gaussian 98. These five descriptors are used as inputs of a new ENTER MLR analysis.

Following equation is a result of ENTER method:

$$LogCMC=3.978-0.031\mu(g)-0.344NAO_{min}\,(g)-0.069B. \qquad (15)$$

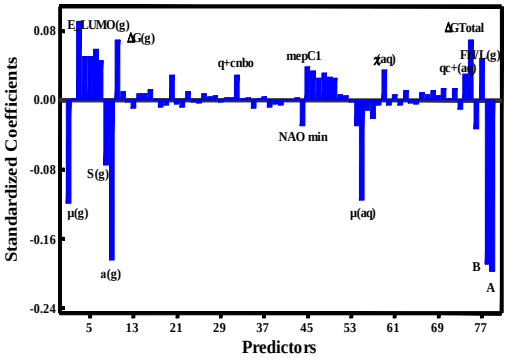

Figure3. The coefficient plot that is a projected scatterplot showing the standardized coefficients for each predictor.

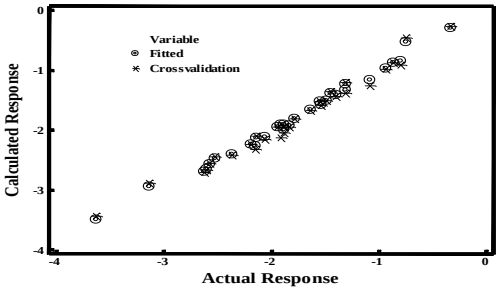

Figure 4. Calculated versus observed *LogCMC* values in PLS model**.**

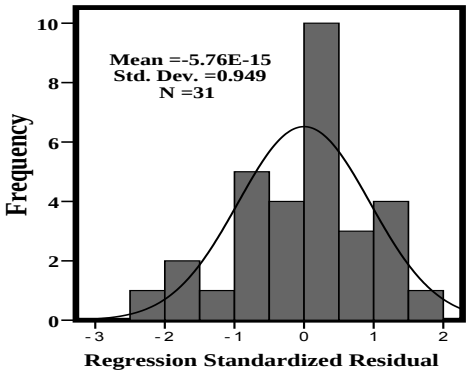

Figure 5. Frequency against standardized residuals.

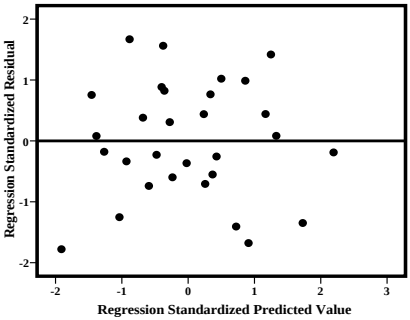

Figure 6. Standardized predicted values of *LogCMC* against standardized residuals.

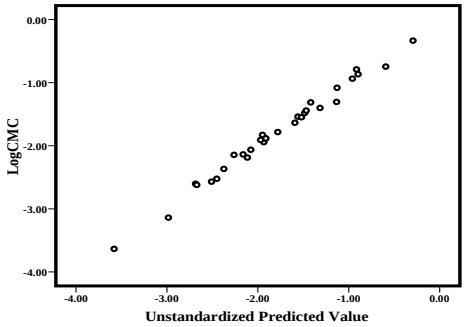

Figure 7. the plot of experimental *LogCMC* values versus predicted *LogCMC* values with equation (15).

The standardized coefficients, *VIF* and the *p*-values for above model are all presented in Table 3. The usual limit used in the interpretation of a *p*-value is 0.05 (or 5%). If *p*-value is less than 0.05, it is reasonable to believe that the observed results are not due to random variations. Furthermore, Table 3 shows that the *VIF*s are less than 2. This means that the variables are weakly correlated to each other.

The standardized coefficients of the model corresponding to the three dependent variables allow comparing the relative weight of the variables in the model. As seen, the resulting model has three significant descriptors. As is obvious, the *B* and the total dipole moment and the NAO_{min} have a negative coefficients. These show that the *LogCMC* increases with decreasing *B*, dipole moment and NAO_{min}.

According to the standardized coefficients (see Table 3), the most significant descriptor appearing in regression equation (15) is the descriptor *B* , which is the simplest variable**,**

defined as the number of bonds in the molecule. The next important descriptor is the total dipole moment, $\mu(g)$. Dipole moment is the measure of polarity of a polar covalent bond. It is defined as the product of charge on the atoms and the distance between the two bonded atoms. The total dipole moment, however, reflects only the global polarity of a molecule. The next important descriptor is the natural atomic orbitals (NAO_{min}) which is obtained from the diagonalization of density matrix of a given molecule.

It is clear that the micellization process occurs in a solution phase. However, our above model relates two variable calculated in the gas phase to the CMC. To obtain a model with aqueous phase calculated descriptors, we repeat previous approach except that all quantum variables are calculated from keyword SCRF=CPCM. This keyword models solvent molecules as a conductor-like polarizable continuum model [14]. All parameters for CPCM have the default values in the Gaussian 98 program. After searching the common descriptor between PLS and stepwise-MLR, following model is obtained based on ENTER-MLR and the common descriptors:

Table 3. The standardized coefficients, *VIF* and the *p*-values for the MLR model based on descriptors of equation (15)

Source	Value standardized	*P*-value	*VIF*	Lower bound (95%)	Upper bound (95%)
B	-0.659	0.00	1.812	-0.408	-0.063
μ *(g)*	-0.404	0.00	1.845	-0.036	-0.027
NAO_{min}	-0.253	0.00	1.243	-0.075	-0.28

$$LogCMC=1.956-0.0290\mu(aq)+0.426q^{+}_{C}(aq)-0.0700B. \quad (16)$$

Analyzing equation (16) it's observed that the maximum atomic charge on the atom C based on Mulliken population analysis (q_C^{+}(aq)) in aqueous phase has a positive coefficient for *Log CMC*. The q_C^{+}(aq) is indicating an influence of the interactions of molecules and polar compounds increase CMC. The standardized coefficients, *VIF* and the *p*-values are all presented in Table 5.

We employed a two-step validation protocol. The data set was divided into training and test sets. The model is first validated internally using the training set. The training set was applied for fitting of the line, whereas test set for which its molecules have no role in model building was used for the evaluation of the prediction ability of the model. MLR model was developed for anionic surfactants using the SPSS version 15.0 software. The training set consisted of 31 molecules and the test set consisted of 10 molecules. The statistical results (R^2, R^2_{Adj}, *F*, *MSE*, *PRESS* and *D. W.*) are summarized in Table 7. Our statistical analysis gives excellent results

based on three descriptors for these *LogCMC* and the model with 3 descriptors has no outliers with absolute deviation exceeding $3\sqrt{MSE}$, see Table 1.

Table 4. The values of total bond number, *B*, the dipole moment in gas, $\mu(g)$ the dipole moment in aqueous phase, $\mu(aq)$, the minimum natural atomic orbitals NAO_{min} and maximum net atomic charges on C atom, $q^{+}_{C}(aq)$

No	*B*	$\mu(g)$	$\mu(aq)$	NAO_{min}	$q^{+}_{C}(aq)$
1	27	21.488	24.105	5.0389	0.752
2	34	27.325	30.07	5.0387	0.752
3	40	33.243	36.4312	5.0386	0.759
4	46	39.219	42.3577	5.0392	0.756
5	30	19.532	22.4652	6.4217	-0.3
6	36	19.187	19.6619	6.4173	-0.298
7	42	25.640	26.5388	6.4208	-0.3
8	48	36.340	39.2674	6.4212	-0.3
9	51	39.243	42.1605	6.4213	-0.3
10	31	21.140	22.892	6.0236	0.016
11	37	26.660	28.482	6.0233	0.016
12	40	29.476	31.3009	6.0234	0.017
13	43	32.323	34.1623	6.0232	0.017
14	46	35.188	37.0421	6.0229	0.017
15	49	38.084	39.8952	6.0232	0.017
16	40	29.209	31.7342	4.9946	0.84
17	46	31.703	34.0295	4.9994	0.804
18	52	35.519	36.6377	5.0132	0.818
19	42	26.092	29.6932	6.4208	-0.3
20	42	14.776	15.5748	6.4180	-0.296
21	42	12.506	13.105	6.4157	-0.294
22	42	10.844	13.6726	6.4165	-0.303
23	51	16.360	19.1761	6.4155	-0.299
24	31	13.444	16.209	5.8413	0.169
25	37	17.401	19.5305	5.8484	0.142
26	40	10.738	13.3957	5.8402	0.146
27	46	28.868	31.404	5.8473	0.148
28	41	29.987	33.3476	6.1642	-0.189
29	47	35.766	39.106	6.1639	-0.188
30	44	26.394	30.2347	6.0496	0.031
31	57	45.087	----	6.42124	----

Table 5. The standardized coefficients, *VIF* and the *p*-values for the MLR model based on descriptors of equation (16)

Source	Value standardized	*P*-value	*VIF*	Lower bound (95%)	Upper bound (95%)
B	-0.694	0.00	1.520	-0.077	-0.063
μ (aq)	-0.402	0.00	1.653	-0.034	-0.024
q^{+}_{C}(aq)	0.273	0.00	1.241	0.329	0.523

Table 6. Coefficient of correlation of *Log CMC* and selected descriptors.

	LogCMC	*μ(g)*	NAO_{min}	*B*	*μ(aq)*	$q^{+}_{C}(aq)$
LogCMC	1					
μ(g)	-0.736	1				
NAO_{min}	-0.281	-0.222	1			
B	-0.942	0.590	0.178	1		
μ(aq)	-0.673	0.995	-0.317	0.509	1	
$q^{+}_{C}(aq)$	0.223	0.300	-0.998	-0.119	0.305	1

Calculation of prediction ability

Validation is a crucial aspect of any QSPR modeling. The most popular validation criteria are leave-one-out (R^2_{LOO}) and leave-five-out (R^2_{LFO}). The results are presented in Table 7. The proposed models are evaluated for prediction by cross-validation as well as using an external test set. The test set consisted of 10 anionic surfactants. The results are presented in Table 8 for equations (15) and (16). A chance model has low ability to reproduce *y* variable of the external test set molecules. The models show high external prediction ability. Figure 7 shows the plot of experimental *LogCMC* values versus predicted *LogCMC* values with equation (15).

Previous works on anionic surfactants

There are reports of QSPR models to predict the *LogCMC* of anionic surfactants. Katritzky *et al.* correlated the *LogCMC* for a data set of 119 sulfates and sulfonates with QSPR approach. The regression equation (equation (17)) included three descriptors (i*) t-sum-KH* is the sum of Kier and Hall index of zero order over all hydrophobic tails, (ii) *TDM* is the AM1 calculated total dipole moment of the molecule, (iii) *h-sum RNC* is the sum of the relative number of carbon atoms over all hydrophilic heads. The determination coefficient R^2 = 0.94, Fisher criterion *F* = 597, and *MSE*= 0.0472 [13].

$$LogCMC = -0.314(\pm 0.010)t\text{-}sum\ KH - 0.034(\pm 0.003)TDM - 1.45(\pm 0.18)h\text{-}sum\text{-}RNC + 1.89(\pm 0.11). \quad (17)$$

Zhang et al. modeled 98 *LogCMC* values for carboxylate, sulfates and sulfonates surfactants and found the regression equation (equation (18)) with three descriptors (R^2 = 0.980, *F* =

1505.23, *MSE* = 0.0107, R^2_{CV} = 0.978). In equation (18) N_T represents the total atom number, μ is the *ab initio* calculated total dipole moment of the molecule and $Q_{C\ max}$ represents the maximum net atomic charges on C atom [12].

$$LogCMC=1.89(\pm0.0671)-0.0697(\pm0.00151)N_T-0.0323(\pm 0.0015)\mu+0.381(\pm0.0305)Q_{Cmax}. \quad (18)$$

Table 7. Statistical parameters obtained by applying the MLR method to the training set.

Eq	R^2	R^2_{Adj}	F	*MSE*	*PRESS*	*D.W.*	R^2_{LOO}	R^2_{LFO}
15	0.989	0.987	778.274	0.007	0.247	2.445	98.44	97
16	0.981	0.978	439.568	0.009	0.338	2.52	97.26	97

Table 8. Comparison between experimental and predicted *LogCMC* values of external test set for the MLR model based on descriptors of equations (15) and (16).

No	Formula	*LogCMC (exp.)*	Predicted values *LogCMC(pred.)*		Residuals	
			Eq(15)	Eq(16)	Eq(15)	Eq(16)
1	$C_6H_{13}SO_3^-Na^+$	-0.496	-0.330	-0.332	-0.166	-0.164
2	$C_{13}H_{27}SO_3^-Na^+$	-2.421	-2.372	-2.368	-0.049	-0.053
3	$C_9H_{19}CH(C_2H_5)SO_3^-Na^+$	-1.730	-1.703	-1.723	-0.027	-0.007
4	$C_6H_{13}CH(C_6H_{13})SO_4^-Na^+$	-1.714	-1.672	-1.718	-0.042	0.004
5	$C_7H_{15}CH(C_6H_{13})SO_4^-Na^+$	-2.013	-1.915	-1.963	-0.098	-0.05
6	$C_9H_{19}CH(C_4H_9)SO^-_4Na^+$	-2.171	-1.897	-1.935	-0.274	-0.236
7	$C_{10}H_{21}CH(C_3H_7)SO_4^-Na^+$	-2.288	-1.993	-2.021	-0.295	-0.267
8	$C_{11}H_{23}CH(C_2H_5)SO_4^-Na^+$	-2.367	-2.249	-2.264	-0.118	-0.103
9	$C_{12}H_{25}CH(CH_3)SO_4^-Na^+$	2.481	-2.243	-2.254	-0.238	-0.227
10	$ParaC_7H_{15}C_6H_4SO_3^-Na^+$	-1.582	-1.675	-1.707	0.093-	-0.125

Comparison of equation (15) with equations (17) and (18) suggests that our QSPR equation has a better statistics in comparison with previous results, because the R^2_{Pred} and R^2 values of equation (15) are larger than that of previous results, and the *MSE* value is obviously lower than that of previous results.

Conclusions

We have mixed PLS and stepwise-MLR techniques to select descriptors for ENTER-MLR modeling. This approach has been repeated for the calculation of descriptors in two gas and solution phases. The results showed that the models in both above phases can accurately predict the CMCs.

References

[1] M. J. Rosen, *Surfactants and Interfacial Phenomena*, Wiley, New Jersey, 2004.

[2] B. Jonsson, B. Lindman, K. Holmberg, B. Kronberg, *Surfactants and Polymers in Aqueous Solution*, Wiley, New York, 1999.

[3] V. Uskoković, M. Drofenik, Reverse micelles: Inert nano-reactors or physico-chemically active guides of the capped reactions, *Adv. Colloid. Interfac.* **133 (**2007) 23–34.

[4] R. Todeschini, V. Consonni, *Handbook of Molecular Descriptors*, Willy–VCH, Weinheim, 2000.

[5] T. Isaksson, T. Fearn, T. Davies, T. Næs, *A User-Friendly Guide to Multivariate Calibration and Classification*, NIR Publications, Charlton Mill, 2004.

[6] R. G. Brereton, *Chemometrics — Data Analysis for the Laboratory and Chemical Plant*, Wiley, Atrium, 2003.

[7] B. Hemmateenejad, M. Yazdani, QSPR models for half-wave reduction potential of steroids: A comparative study between feature selection and feature extraction from subsets of or entire set of descriptors, *Anal. Chim. Acta.* **634** (2008) 27–35.

[8] HyperChem 7, Molecular Mechanics and Quantum Chemical Calculations Package, HyperCube, 2002.

[9] M. J. Frisch, G. W.Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, V. G. Zakrzewski, J. A. Montgomery, R. E. Stratmann, J. C. Burant, S. Dapprich, J. M. Millam, A. D. Daniels, K. N. Kudin, M. C. Strain, O. Farkas, J. Tomasi, V. Barone, M. Cossi, R. Cammi, B. Mennucci, C. Pomelli, C. Adamo, S. Clifford, J. Ochterski, G. A. Petersson, P. Y. Ayala, Q. Cui, K. Morokuma, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. Cioslowski, J. V. Ortiz, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. Gomperts, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, C. Gonzalez, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, J. L. Andres, C. Gonzalez, M. Head-Gordon, E. S. Replogle, and J. A. Pople, Gaussian *98*, Revision A.6, Gaussian, Inc., Pittsburgh, 1998.

[10] SPSS for windows, Statistical Package for IBM PC, SPSS, 2006.

[11] Minitab Inc, *MINITAB* Statistical Software, Release 14 for Windows, State College, Pennsylvania, 2003.

[12] X. Li, G. Zhang, J. Dong, X. Zhou, X. Yan, M. Luo, Estimation of critical micelle concentration of anionic surfactants with QSPR approach, *J. Mol. Struct. (Theochem)* **710** (2004) 119–126.

[13] 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. 2. Anionic surfactants, *J. Colloid. Interf. Sci.* **187** (1997) 113–120.

[14] S. Miertus, J. Tomasi, Approximate evaluations of the electrostatic free energy and internal energy changes in solution processes, *Chem. Phys.* **65** (1982) 239–2248.

[15] D. C. Montgomerey, E. A. Peck, G. G. Vinig, *Introduction to Linear Regression Analysis*, Wiley, New York, 2001.

[16] T. Yang, C. M. Breneman, S. M. Cramer, Investigation of multi–modal high–salt binding ion–exchange chromatography using quantitative structure–property relationship modeling, *J. Chromatogr. A* **67** (2007) 1561–1569.

[17] Y. Pan, J. Jiang, Z. Wang, Quantitative structure–property relationship studies for predicting flash points of alkanes using group bond contribution method with back-propagation neural network, *J. Hazard. Mater.* **147** (2007) 424–430.

[18] K. Hemelsoet, V. Van Speybroeck, M. Waroquier, How useful are reactivity indicators for the description of hydrogen abstraction reactions on polycyclic aromatic hydrocarbons? *Chem. Phys. Lett.* **444** (2007) 17–22.

[19] A. Hinchliffe, H. J. Soscún, Ab initio studies of the dipole moment and polarizability of azulene in its ground and excited singlet states, *Chem Phys Lett* **412** (2005) 365–368.

[20] P. W. Ayers, Using reactivity indicators instead of the electron density to describe Coulomb systems, *Chem. Phys. Lett.* **438** (2007) 148–152.

[21] L. Xiao–Hong, T. Zheng–Xin, Z. Xian–Zhou, Natural bond orbital (NBO) population analysis of para–substituted S–nitrosothiophenols, *J. Mol. Struct. (Theochem)* **900** (2009) 50–58.

[22] R. M. Issa, M. K. Awad, F. M. Atlam, Quantum chemical studies on the inhibition of corrosion of copper surface by substituted uracils, *Appl. Surf. Sci.* **255** (2008) 2433–2441.

[23] A. E. Reed, F. Weinhold, Natural atomic orbitals and natural population analysis, *J. Chem. Phys.* **78** (1983) 4066–4072.

[24] C. J. Crastol, E. D. Stevens, Use of electrostatic potentials to study non-bonded intramolecular interactions in 1,8-disubstituted naphthalenes with carbonyl groups as electrophilic substituents, *J. Mol. Struct. (Theochem)* **454** (1998) 51–59.

[25] W. J. Fan, R. Q. Zhang, S. Liu, Computation of large systems with an economic basis set: Structures and reactivity indices of nucleic acid base pairs from density functional theory, *J. Comput. Chem.* **28** (2006) 967–974.