Application of Quantitative Structure-Property Relationship (QSPR) Models for the Predictions of Critical Micelle Concentration of Gemini Imidazolium Surfactants

Ely Setiawan¹, Mudasir² and Karna Wijaya³

¹,²,³Department of Chemistry, Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada, Indonesia
¹Department of Chemistry, Faculty of Mathematics and Natural Sciences, Universitas Jenderal Soedirman, Indonesia
¹Austrian-Indonesian Centre for Computational Chemistry (AIC), Department of Chemistry, Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada, Indonesia

E-mail: ely.setiawan@mail.ugm.ac.id

Abstract. The critical micelle concentration (cmc) is important indexes to determine the performance of surfactants quantitatively. In this study, based on the molecular descriptors, which calculated, by Dragon software, were proposed a quantitative structure-property relationship analysis for prediction of cmc of gemini imidazolium surfactants. In the present study, the associative neural networks (ASNN) and multiple linear regression analysis (MLRA) technique were used to interpret the chemical structural functionality (molecular descriptors) against the cmc of gemini imidazolium surfactants. The models were validated rigorously through 5-fold cross-validation. The ASNN method showed to be better than the MLRA method in terms of the internal and the external prediction accuracy with high statistical quality \( r^2 = 0.96, q^2 = 0.95, \) RMSE = 0.08 and MAE = 0.04 respectively. The developed QSPR models are publicly available on the web site at https://ochem.eu/model/11583111 for ASNN model and https://ochem.eu/model/43858671 for MLRA model. These models can be applied to predict the cmc of new gemini imidazolium surfactants.

1. Introduction

Gemini surfactants, as a new class of surface-active agent molecules, have attracted the increasing interest of academic research groups over the past decade. These compounds are composed of two surfactants molecules which have two hydrophilic head groups and two hydrophobic groups, linked by a spacer. In case of cationic surfactants, hydrophilic head group of these compound is positively charged nitrogen atom with two hydrophobic groups (long alkyl chain), linked by a flexible or rigid spacer. In comparison with conventional surfactants, gemini surfactants appear to be better physicochemical properties such as lower critical micelle concentration (cmc), higher efficient to reducing surface tension of water, higher adsorption efficiency, and better solubilizing, wetting, and foaming properties [1]. Currently, imidazolium ring bearing cationic surfactants have been widely investigated because of their unique properties and possibility of use in several applications, including gene delivery agent [2], corrosion inhibitors [3], drug delivery agent [4], antimicrobial and antifungal agent [5]. Due to the existence of the imidazolium headgroup, gemini imidazolium surfactants has
several advantages over other Gemini cationic surfactants, for example; they exhibit a stronger tendency toward self-aggregation [6].

One of the most common reasons for the development of Gemini surfactants is that their cmc values are much lower than the corresponding monomeric surfactants. As is well known, the cmc of the surfactant is related to its molecular structure [7]. Usually, the cmc of surfactants are determined by the experimental techniques, such as conductometry, tensiometry, NMR spectroscopy, calorimetry, and fluorescence emission spectroscopy [8]. Recently, QSAR were also used to predict the cmc of surfactants [9]. QSAR is effective in relating the structural (descriptors) to physicochemical properties. This method shows it can be a predictor tool for various surfactants properties [10]. In this work, quantitative structure-property relationships (QSPR) modeling was used to predict the cmc of Gemini imidazolium surfactants. Various statistical methods are used to build QSPR models like multilinear regressions or artificial neural networks [11].

Although many QSPR models have been proposed for cmc, even for Gemini surfactants [12], to the best of author’s knowledge, there are no earlier reports of QSPR models for cmc prediction for Gemini imidazolium surfactants. In present study, we developed new QSPR models for prediction of cmc of Gemini imidazolium surfactants. The QSPR models were developed using associative neural networks (ASNN) [13] and multilinear regression analysis (MLRA) methods. The quality of QSPR models was evaluated using the 5-fold cross-validation method [14].

2. Materials and methods

2.1. Data set and descriptors calculation

In this study, the initial set of 70 Gemini imidazolium surfactants with known cmc were obtained from several published papers [15, 16, 17, 18, 19]. These compounds were randomly divided into 57 compounds for a training set and 13 compounds for a test set (validation set). Structural formulas of Gemini imidazolium surfactants were drawn using MarvinSketch, converted in SMILES (Simplified Molecular Input Line Entry System) format and stored with their corresponding cmc (-log cmc or pCMC) in the OCHEM platform [20]. All Gemini imidazolium surfactants structures were cleaned by the Chemaxon standardizer and structural coordinates were optimized using CORINA [21].

The descriptors were calculated using Dragon v6 software [22], which covers 4885 molecular descriptors including topological indices, information indices, drug-like indices, ring descriptors, functional group counts, walk and path counts, atom-centered fragments, molecular properties. In order to reduce the interference of multicollinearity, all of the descriptors were filtered before the development of QSPR models. In this study, we used the pairwise decorrelation method for descriptor selection and the duplicated descriptors with the pairwise correlation of more than 0.95 were eliminated.

2.2. Development of QSPR models

In this study, we used the Online CHEmical Modeling environment (OCHEM, http://www.ochem.eu) to build QSPR models for predicting cmc of Gemini imidazolium surfactants. OCHEM is a user-friendly web-based platform that aims to reduce the amount of work a QSAR/QSPR researcher should perform to obtain a predictive model. OCHEM provides tools for analysis and modification of the QSAR/QSPR models. Furthermore, models can be published on the Web and publicly used by others.

We developed the QSPR models with associative neural networks (ASNN) and multiple linear regression analysis (MLRA) technique. ASNN is the combination of an ensemble of feed-forward neural networks and the kNN (k-nearest neighbors) method. The ASNN corrects the bias in the ensemble by utilizing the correlation between ensemble responses as a measure of distance. In the present study, the ASNN method employed three-layered back propagation neural network with the input layer was equal to the number of molecular descriptors calculated by Dragon software, the number of neuron in the hidden layer was five and one neurons in the output layer (pCMC). The
network is trained using the SuperSAB algorithm with 1000 iterations and a set of 64 networks in the ensemble.

**Table 1.** SMILES notation and experimental and calculated pCMC value of gemini imidazolium surfactants by using the result of QSPR model obtained in this study

| Compound | SMILES notation | pCMC | Experimental | ASNN | MLRA |
|----------|-----------------|------|--------------|------|------|
| GIS_1    | CN1C=C[N+](COCCOC[N+])2=CN(C=C)C=C2=Cl1.[Cl-],[Cl-] | 2.15 | 2.18 | 2.24 |
| GIS_2    | CCN1C=C[N+](COCCOC[N+])2=CN(C(C)C=C2=Cl1.[Cl-],[Cl-] | 2.23 | 2.30 | 2.19 |
| GIS_3    | CCCCN1C=C[N+](COCCOC[N+])2=CN(CCCC)C=C2=Cl1.[Cl-],[Cl-] | 2.41 | 2.38 | 2.42 |
| GIS_4    | CCCCCCN1C=C[N+](COCCOC[N+])2=CN(CCCCCC)C=C2=Cl1.[Cl-],[Cl-] | 2.58 | 2.57 | 2.53 |
| GIS_5    | CCCCCCCCN1C=C[N+](COCCOC[N+])2=CN(CCCCCCC)C=C2=Cl1.[Cl-],[Cl-] | 2.65 | 2.67 | 2.65 |
| GIS_6    | CCCCCCCCCN1C=C[N+](COCCOC[N+])2=CN(CCCCCCCC)C=C2=Cl1.[Cl-],[Cl-] | 2.72 | 2.72 | 2.74 |
| GIS_7    | CCCCCCCCCCN1C=C[N+](COCCOC[N+])2=CN(CCCCCCC)C=C2=Cl1.[Cl-],[Cl-] | 2.81 | 2.83 | 2.86 |
| GIS_8    | CCCCCCCCCCN1C=C[N+](COCCOC[N+])2=CN(CCCCCC)C=C2=Cl1.[Cl-],[Cl-] | 2.92 | 2.93 | 2.87 |
| GIS_9    | CCCCCCCCCNNCN1C=C[N+](COCCOC[N+])2=CN(CCCCCC)C=C2=Cl1.[Cl-],[Cl-] | 3.04 | 3.03 | 3.02 |
| GIS_10   | CCCCCCCCCCNN1C=C[N+](COCCOC[N+])2=CN(CCCCC)C=C2=Cl1.[Cl-],[Cl-] | 3.15 | 3.16 | 3.15 |
| GIS_11   | CCCCCCCCCCCCCN1C=C[N+](COCCOC[N+])2=CN(CCCCCC)C=C2=Cl1.[Cl-],[Cl-] | 3.34 | 3.34 | 3.31 |
| GIS_12   | CCCCCCCCCCCCCCN1C=C[N+](COCCOC[N+])2=CN(CCCCCC)C=C2=Cl1.[Cl-],[Cl-] | 3.52 | 3.5 | 3.43 |
| GIS_13   | CN1C=C[N+](COCCOC[N+])2=CN(CC)C=C2=Cl1.[Cl-],[Cl-] | 2.18 | 2.29 | 1.68 |
| GIS_14   | CCN1C=C[N+](COCCOC[N+])2=CN(C(C)C=C2=Cl1.[Cl-],[Cl-] | 2.26 | 2.26 | 2.30 |
| GIS_15   | CCCCCCN1C=C[N+](COCCOC[N+])2=CN(CCCCCC)C=C2=Cl1.[Cl-],[Cl-] | 2.61 | 2.58 | 2.57 |
| GIS_16   | CCCCCCCCN1C=C[N+](COCCOC[N+])2=CN(CCCCCC)C=C2=Cl1.[Cl-],[Cl-] | 2.68 | 2.66 | 2.67 |
| GIS_17   | CCCCCCCCCNN1C=C[N+](COCCOC[N+])2=CN(CCCCCC)C=C2=Cl1.[Cl-],[Cl-] | 2.75 | 2.73 | 2.81 |
| GIS_18   | CCCCCCCCCNN1C=C[N+](COCCOC[N+])2=CN(CCCCCC)C=C2=Cl1.[Cl-],[Cl-] | 2.84 | 2.83 | 2.86 |
| GIS_19   | CCCCCCCCCNN1C=C[N+](COCCOC[N+])2=CN(CCCCCC)C=C2=Cl1.[Cl-],[Cl-] | 2.95 | 2.95 | 2.87 |
CCNN(C)=C2)=C1.[Cl]-[Cl-]

GIS_43
CCCN(C)=C(N)+[CCOCOCOC][N]+2=CN(CCC
CCOCOC)=C=C2)=C1.[Cl]-[Cl-]

GIS_44
CCCN(C)=C(N)+[CCOCOCOC][N]+2=CN(CCC
CCOCOC)=C=C2)=C1.[Cl]-[Cl-]

GIS_45
CCCN(C)=C(N)+[CCOCOCOC][N]+2=CN(CCC
CCOCOC)=C=C2)=C1.[Cl]-[Cl-]

GIS_46
CCCN(C)=C(N)+[CCOCOCOC][N]+2=CN(CCC
CCOCOC)=C=C2)=C1.[Cl]-[Cl-]

GIS_47
CCCN(C)=C(N)+[CCOCOCOC][N]+2=CN(CCC
CCOCOC)=C=C2)=C1.[Cl]-[Cl-]

GIS_48
CCCN(C)=C(N)+[CCOCOCOC][N]+2=CN(CCC
CCOCOC)=C=C2)=C1.[Cl]-[Cl-]

GIS_49
CCCN(C)=C(N)+[CCOCOCOC][N]+2=CN(CCC
CCOCOC)=C=C2)=C1.[Cl]-[Cl-]

GIS_50
CCCN(C)=C(N)+[CCOCOCOC][N]+2=CN(CCC
CCOCOC)=C=C2)=C1.[Cl]-[Cl-]

GIS_51
CCCN(C)=C(N)+[CCOCOCOC][N]+2=CN(CCC
CCOCOC)=C=C2)=C1.[Cl]-[Cl-]

GIS_52
CCCN(C)=C(N)+[CCOCOCOC][N]+2=CN(CCC
CCOCOC)=C=C2)=C1.[Cl]-[Cl-]

GIS_53
CCCN(C)=C(N)+[CCOCOCOC][N]+2=CN(CCC
CCOCOC)=C=C2)=C1.[Cl]-[Cl-]

GIS_54
CCCN(C)=C(N)+[CCOCOCOC][N]+2=CN(CCC
CCOCOC)=C=C2)=C1.[Cl]-[Cl-]

GIS_55
CCCN(C)=C(N)+[CCOCOCOC][N]+2=CN(CCC
CCOCOC)=C=C2)=C1.[Cl]-[Cl-]

GIS_56
CCCN(C)=C(N)+[CCOCOCOC][N]+2=CN(CCC
CCOCOC)=C=C2)=C1.[Cl]-[Cl-]

GIS_57
CCCN(C)=C(N)+[CCOCOCOC][N]+2=CN(CCC
CCOCOC)=C=C2)=C1.[Cl]-[Cl-]

GIS_58
CCCN(C)=C(N)+[CCOCOCOC][N]+2=CN(CCC
CCOCOC)=C=C2)=C1.[Cl]-[Cl-]

GIS_59
CCCN(C)=C(N)+[CCOCOCOC][N]+2=CN(CCC
CCOCOC)=C=C2)=C1.[Cl]-[Cl-]

GIS_60
CCCN(C)=C(N)+[CCOCOCOC][N]+2=CN(CCC
CCOCOC)=C=C2)=C1.[Cl]-[Cl-]

GIS_61
CCCN(C)=C(N)+[CCOCOCOC][N]+2=CN(CCC
CCOCOC)=C=C2)=C1.[Cl]-[Cl-]

GIS_62
CCCN(C)=C(N)+[CCOCOCOC][N]+2=CN(CCC
CCOCOC)=C=C2)=C1.[Cl]-[Cl-]

GIS_63
CCCN(C)=C(N)+[CCOCOCOC][N]+2=CN(CCC
CCOCOC)=C=C2)=C1.[Cl]-[Cl-]

GIS_64
CCCN(C)=C(N)+[CCOCOCOC][N]+2=CN(CCC
CCOCOC)=C=C2)=C1.[Cl]-[Cl-]
Multiple linear regression (MLR) is one of the earliest methods used for constructing QSPR/QSAR models, but it is still one of the most commonly used one to date, because of simplicity and easily interpretable model. The selected descriptors were employed with MLR to develop the linear model of the property of interest, which takes the form:

$$Y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_nx_n$$

(1)

in this equation Y is the predicted property value of the dependent variable, $x_1$ to $x_n$ represent the specific descriptors, while $b_1$ to $b_n$ represent the coefficient of those descriptors, $b_0$ is the intercept of this equation.

2.3. Model validation and evaluation

The models developed in this paper were evaluated by 5-fold cross-validation and externally validated by an external validation set independent from the model training. All QSPR models were evaluated with the squared correlation coefficient, $r^2$, cross-validation coefficient, $q^2$, root mean squared error (RMSE) and mean absolute error (MAE). QSPR models are considered effective at the value of $r^2>0.6$, $q^2>0.5$, with low RMSE and MAE value, can be used to assess the cmc of new gemini imidazolium surfactants [23].

3. Results and Discussion

In this study, we developed QSPR models for cmc of gemini imidazolium surfactants prediction were generated by ASNN and MLRA along with molecular descriptors which calculated by Dragon v6 packages implemented at the OCHEM platform. As a result, there were two models generated by combination; ASNN model and MLRA model.

| Table 2. Statistical quality of the QSPR model |
|---------------------------------------------|
| Model | Set        | n  | $r^2$   | $q^2$   | RMSE  | MAE   |
|-------|------------|----|---------|---------|-------|-------|
| ASNN  | Training   | 57 | 0.96 ± 0.02 | 0.95 ± 0.02 | 0.08 ± 0.02 | 0.04 ± 0.01 |
|       | Test       | 13 | 0.91 ± 0.04 | 0.85 ± 0.10 | 0.20 ± 0.10 | 0.07 ± 0.04 |
| MLRA  | Training   | 57 | 0.91 ± 0.02 | 0.89 ± 0.04 | 0.12 ± 0.02 | 0.07 ± 0.01 |
|       | Test       | 13 | 0.91 ± 0.05 | 0.85 ± 0.09 | 0.17 ± 0.10 | 0.08 ± 0.04 |

Table 2 clearly shows the better statistical quality performance of the ASNN model compared to MLRA model, with the value of $r^2$, $q^2$, RMSE and MAE were 0.96, 0.95, 0.08 and 0.04, respectively. The ASNN method represents an innovative method to calculate nonlinear models between structure and property (cmc) of gemini imidazolium surfactants. The correlation of the experimental pCMC of gemini imidazolium surfactants with the calculated of ASNN model is shown in figure 1.
The MLRA model was built with the same number of descriptors as ASNN model. Equation 2 is obtained by MLRA method. The five molecular descriptors with the highest significant contribution to the pCMC of gemini imidazolium surfactants are MW (molecular weight), SpMax_AEA(dm) (leading eigenvalue from augmented edge adjacency mat. weighted by dipole moment), GATS1e (Geary autocorrelation of lag 1 weighted by Sanderson electronegativity), MATS6v (Moran autocorrelation of lag 6 weighted by van der Waals volume), and Eig04_AEA(dm) (eigenvalue n. 4 from augmented edge adjacency mat. weighted by dipole moment). The model equation indicates that the pCMC of gemini imidazolium surfactants has a direct relationship with MW, and has an inverse relationship with SpMax_AEA(dm), GATS1e, MATS6v, and Eig04_AEA(dm).

\[
p\text{CMC} = 16.3 + 0.00356 \times \text{MW} - 3.83 \times \text{SpMax}_\text{AEA(dm)} - 1.6 \times \text{GATS1e} - 0.483 \times \text{MATS6v} - 1.05 \times \text{Eig04}_\text{AEA(dm)}
\]  

(2)
According to the statistical quality of MLRA model (table 2), the squared correlation coefficient ($r^2$) and cross-validation coefficient, $q^2$ are 0.91 and 0.89 for the training set, respectively. The RMSE and MAE value (0.12 and 0.07, respectively) are lower enough to indicate successful predictions of the QSPR model developed by MLRA. The graph of the experimental versus predicted cmc of gemini imidazolium surfactants of MLRA model is shown in figure 2.

4. Conclusion
ASNN and MLRA methods were used to build QSPR model of the cmc of gemini imidazolium surfactants using molecular descriptors which calculated by Dragon v6 packages implemented at the OCHEM platform for the first time. The results showed that the predicted values from both techniques are close to the true values. High correlation coefficients for ASNN and MLRA model along with low prediction errors demonstrated the ability of the models to predict cmc of gemini imidazolium surfactants. The predictive ability of the ASNN and MLR models was tested by 5-fold cross-validation method. The results show that the ASNN model predicts cmc better with the value of $r^2$, $q^2$, RMSE and MAE were 0.96, 0.95, 0.08 and 0.04, respectively. The ASNN model produces high statistical quality and low prediction error model compared with MLRA model. The developed QSPR models are publicly available on the web site at https://ochem.eu/model/11583111 for ASNN model and https://ochem.eu/model/43858671 for MLRA model. These models can be applied to predict the cmc of new gemini imidazolium surfactants.

References
[1] Menger F M Keiper J S 2000 Angew. Chem., Int. Ed. 39 1906–20
[2] Sunitha S Reddy P S Prasad R B N Kanjilal S 2011 Eur. J. Lipid Sci. Technol. 113 756–762
[3] Zhou T Yuan J Zhang Z Xin X Xu G 2016 Colloid Surf., A 575 57–65
[4] Rodrigues M Calpena A C Amabilino D B Garduno-Ramírez M L Perez-Garcia L 2014 J. Mater. Chem. B 2 5419–29
[5] Wang L Qin H Ding L Huo S Deng Q Zhao B Meng L Yan T 2014 J. Surfactants Deterg. 17 1099–1106
[6] Kamboj R Singh S Bhadani A Kataria H Kaur G. 2012 28 11969–78
[7] Goronja J Pejic, N Lezâic, A J 2016 J. Chem. Educ. 93 1277–81
[8] Lin C E Wang T Z Chiu T C Hsueh C C 1999 J. High Resolut. Chromatogr. 22 265–270
[9] Roy K Kabir H 2013 Chem. Eng. Sci. 87 141–151
[10] Hu J Zhang X Wang Z 2010 Int. J. Mol. Sci. 11 1020–47
[11] Abrougui K Gabsi K Mercatoris B Khemis CAmami R Chehaibi S 2019 Soil Tillage Res. 190 202–208
[12] Mozrymas A 2017 Colloid Polym. Sci. 295 75–87
[13] Tetko I V 2002 J. Chem. Inf. Comput. Sci. 42, 717–728
[14] Tropsha A Gramatic P Gombar V K 2003 QSAR Comb. Sci. 22 69–77
[15] Bhadani A Singh S Sakai H Abe M 2016 Synthesis and properties of heterocyclic cationic gemini surfactants. In Encyclopedia of Biocolloid and Biointerface Science; Ed Ohshima H (John Wiley Sons Inc.) chapter 43 pp 539–553
[16] Ren Y Liu H Yao X Liu M Hu Z Fan B 2006 J. Colloid Interface Sci. 302 (2) 669–672
[17] Pałkowski L Błaszczynski J Skrzypczak A Błaszczak J Kozakowska K Wroblewska J Kozusko S, Gospodarek E Krysinski J Slowinski R 2014 Chem Biol Drug Des 83 278–288
[18] Patial P Shaheen A Ahmad I 2014 J. Surfact. Deterg. 17:253–260
[19] Wang L Liu J Huo S Deng Q Yan T Ding L Zhang C Meng L Lu Q 2014 J. Surfact. Deterg. 17 1107–16
[20] Sushko I Pandey A Novotarskyi S Körner R Rupp M Teetz W Tanchuk V 2011 J. Comput.-Aided Mol. Des. 25 (6) 533–554
[21] Kovalishyn V Abramenko N Kopernyk I Charochkina L Metelytsia L Tetko I V Peijnenburg W Kustov L 2018 Food Chem. Toxicol. 112 507–517
[22] Todeschini R Consonni V 2000 *Handbook of Molecular Descriptors* (Weinheim, WILEY-VCH) p 667
[23] Tropsha, A 2010 *Mol. Inf.* 29 476–488