QSAR modeling for predicting the antifungal activities of gemini imidazolium surfactants against *Candida albicans* using GA-MLR methods

Ely Setiawan1*, Karna Wijaya2, Mudasir Mudasir2

1Department of Chemistry, Universitas Jenderal Soedirman, Purwokerto, Indonesia.
2Department of Chemistry, Universitas Gadjah Mada, Yogyakarta, Indonesia.

**ARTICLE INFO**

Received on: 25/11/2020  
Accepted on: 02/02/2021  
Available online: 05/04/2021

**Key words:**  
*Candida albicans*, gemini imidazolium surfactant, genetic algorithm, Mordred, multiple linear regression, QSAR.

**ABSTRACT**

This report presents a Quantitative Structure--Activity Relationships (QSAR) analysis of gemini imidazolium surfactants against *Candida albicans*. Mordred software is used to calculate various types of molecular descriptors. The data set contains 70 structures of gemini imidazolium surfactants and is divided into training set (75%) and test set (25%) to perform cross-validation step. Genetic algorithm technique combined with multiple linear regression method (GA-MLR) was used to investigate the correlation between molecular descriptors and antifungal activity of gemini imidazolium surfactants. As a result, the best GA-MLR model consisting of two topological descriptors (GATS4se and BalabanJ) exhibits good fitting and internal validation with $R^2 = 0.9073$, $Q^2_{LOO} = 0.8941$, and $Q^2_{LMO} = 0.8908$. Also, it was confirmed by the external validation procedure with $R^2_{test} = 0.8988$ and RMSE$_{test} = 0.3557$, indicating that the obtained model was robust, reliable, and strong to predict the antifungal activity of gemini imidazolium surfactants. The GA-MLR-QSAR could be a useful tool for the initial development and design of novel gemini imidazolium surfactant as antifungal agents.

**INTRODUCTION**

The genus *Candida* is responsible for about 80% of infectious fungi in a hospital environment and is a relevant cause of bloodstream infections (Kabir et al., 2012). The most invasive species, which are responsible for severe cases of candidiasis, include *Candida albicans*, *Candida parapsilosis*, *Candida glabrata*, and *Candida krusei*. The most reported clinical pictures related to candidiasis are cutaneous-mucous, visceral, and allergic (Fisher-Hoch and Hutwahner, 1995; Pfäler and Diekema, 2007).

Antifungal agents are an option for treating oral candidiasis, but the availability of antifungal drugs is less than antibacterial agents because eukaryotic fungal organisms are the same as mammalian cells, which makes selecting suitable antifungal targets a problem (Mayer et al., 2013). In addition, several antifungal drugs become resistant, including fluconazole, ketoconazole, and itraconazole (Lewis et al., 2012). Therefore, the development of new antifungal drugs to treat cases of candidiasis is important.

In recent years, cationic gemini surfactants that contain two head groups with two aliphatic chains connected by spacers have been exploited (Menger and Keiper, 2000; Rosen and Tracy, 1998). They exhibit lower cytotoxicity, better surface properties, and better ability to bind negatively charged substance correspond to the monomer surfactants in the same conditions (Brycki et al., 2017; Sharma et al., 2017; Shukla and Tyagi, 2006). Cationic gemini surfactants also have good antimicrobial activity (Bao et al., 2017; Tatsumi et al., 2014). Several studies have shown that the antimicrobial activity of the cationic gemini surfactant imidazolium chloride depends on its structure, for example, the length of the alkyl groups (Rath and Bai, 2016). The hydrophobic portion (alkyl chains) of the imidazolium cationic gemini surfactant interacts with the cell membrane, inducing cell membrane damage leading to cell lysis and death (Doležal et al., 2016). The results of the research by Palkowski et al. © 2021 Ely Setiawan et al. This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/).
Yu synthesized 70 gemini surfactants in test set; MAE

R

2

train

In fitting; Friedman lack of fit (LOF), internal validation (\( Q_{\text{LOO}}^2 \) Leaving-One-Out (LOO) cross-validation coefficient; \( Q_{\text{LMO}}^2 \) Leaving-Many-Out (LMO) cross-validation coefficient; RMSE\(_{\text{cv}}\), Root Mean Square Error of cross-validation), and external validation (\( R_{\text{test}}^2 \) the squared correlation coefficient of test set; RMSE\(_{\text{test}}\), Root Mean Square Error in test set; MAE\(_{\text{test}}\), Mean Absolute Error in test set). Also, the applicability domain (AD) was assessed using the Williams plot (Gramatica, 2007).

RESULTS AND DISCUSSION

The aim of this study was to find the correlation between structural parameters of gemini imidazolium surfactants and the antifungal activity against \( \text{C. albicans} \). Based on the basic principles of QSAR analysis, the structural parameters of a compound are expressed by the molecular descriptors. To obtain those molecular descriptors, first, we have drawn the 3D model of each gemini imidazolium surfactant and then optimized the geometry by employing Hartree–Fock (ab initio) method with 3-21G basis set. We used the equilibrium geometry as input on Mordred software to calculate the molecular descriptors. As result, 835 molecular descriptors were generated.

The variable selection process using the GA technique followed by the MLR method led to a GA-MLR model which consists of two molecular descriptors. The obtained model is shown in Eq. 1. Table 2 shows the statistical parameters of the obtained model. The antifungal activity (pMIC) prediction results for gemini imidazolium surfactants of this model are summarized in Table 1. The correlation graph between the predicted and experimental antifungal activity (pMIC) shown in Figure 1 also shows a slope close to 1. This means that the resulting model can provide a good level of prediction.

\[
pMIC = 23.1793 + 0.628 \times \text{GATS4se} - 0.7791 \times \text{BalabanJ}. \quad \text{Eq. 1}
\]

The coefficients of the molecular descriptors in Eq. 1 suggest that the 2D autocorrelation descriptor, namely, Geary coefficient of lag 4 weighted by Sanderson EN (GATS4se), and the topological of Balaban index (BalabanJ) are the most influence descriptors to the antifungal activity of imidazolium gemini surfactants. The positive coefficient of GATS4se descriptor indicates that an increase in GATS4se leads to an increase in the antifungal activity of gemini imidazolium surfactants. The topological of BalabanJ has a negative coefficient, which indicates that an increase in BalabanJ leads to a decrease in the antifungal activity of gemini imidazolium surfactants.
The correlation matrix between selected descriptors shows that the correlation between GATS4se and BalabanJ is very low (Table 3). This indicates that there is no significant intercorrelation among the descriptors used in the development of the model. Additionally, the residual predicted pMIC using Eq. 1 versus the experimental value of pMIC is shown in Figure 2. All residual predicted pMIC values are located between 1 and −1, which indicates that Eq. 1 has good accuracy and reliability for predicting the antifungal activity of gemini imidazolium surfactants against C. albicans.

Based on the validation parameters of the model (Table 2), the model satisfies the requirements made by Golbraikh et al. (2002) and Roy et al. (2012). The values of $R^2$ (0.9073) and $Q^2_{LOO}$ (0.8941) were reasonable, showing that the model was
significant and robust to predict the antifungal activity of gemini imidazole surfactants. Consider that the values of the difference between $R^2$ and $Q^2_{LOO}$ ($R^2 - Q^2_{LOO} = 0.0132$) are within the limit suggested by Kiralj and Ferreria (2009) which is an indication that the model does not have data overfitting. The low value of the LOF parameter (LOF = 0.1397) implies a good fit model with no current overfitting in the model.

Validation of the final model consists mainly of internal and external validation. The LOO and LMO cross-validation procedures were used for internal validation. According to Table 2, the model was stable and internally good for predictive results, because the values of the difference between $R^2_{LOO}$ and $Q^2_{LOO}$ are comparable to $R^2$. To validate the accuracy of the prediction model, it has been confirmed that the statistical parameters contribute to the external validity of the model. The coefficient of determination value ($R^2_{test} = 0.8988$) and error parameters value (RMSE$_{test} = 0.3557$ and MAE$_{test} = 0.2963$) of the test set showed that the obtained model is good and reliable.

After generating and evaluating the model, AD was employed to confirm that the obtained model can be considered reliable. Williams plot or leverage approach was used to measure the influence of descriptors on the model (Gramatica, 2007). The leverage value ($h_i$) shows the distance of a compound from the centroid of $X$, which is defined as

$$hi = Xi (X^T X)^{-1} X_i,$$

where $X$ is the characteristic matrix of the training set. The critical leverage value ($h^*$) is defined as

$$h^* = 3 (\frac{p+1}{n}),$$

where $p$ is the number of descriptors in the model and $n$ is the total number of compounds in the training set. As shown in Figure 3, all compounds lied within the domain of applicability which lower than the threshold leverage ($h^* = 0.170$). This indicated that no compounds in the dataset fell outside of the AD as an outlier.

The Y-scrambling method was employed as randomization tests to confirm that there was no random correlation between antifungal activity and selected descriptors (Rücker et al., 2007). This criterion is shown by the average value of $R^2_{Yscr}$ and $Q^2_{Yscr}$, which are both lower than $R^2$ and $Q^2$ of the model. In this work as shown in Figure 4, the values of $R^2$ and $Q^2$ of the model are higher than the $R^2_{Yscr}$ and $Q^2_{Yscr}$ values, which indicates that the model is not derived from random correlation.

<table>
<thead>
<tr>
<th>Table 2. Validation parameters of model.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fitting criteria</td>
</tr>
<tr>
<td>$R^2$</td>
</tr>
<tr>
<td>RMSE$_{train}$</td>
</tr>
<tr>
<td>LOF</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 3. Variable correlation matrix of model.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Descriptor</td>
</tr>
<tr>
<td>GATS4se</td>
</tr>
<tr>
<td>BalabanJ</td>
</tr>
</tbody>
</table>

Figure 1. Plot of experimental and predicted value of pMIC for training and test set.

Figure 2. Plot of residuals of predicted pMIC determined using Eq. 1 versus experimental pMIC.

Figure 3. Plot of experimental and predicted value of pMIC for training and test set.
CONCLUSION

It is concluded, through this work, that the GA-MLR analysis showed that the two highlighted descriptors play the role of antifungal activity of gemini imidazolium surfactants, namely, $\text{GATS4se}$ (Geary coefficient of lag 4 weighted by Sanderson EN) and BalabanJ (topological of Balaban index). The obtained QSAR model is significant and robust, does not show random correlation, and has a strong predictive ability ($R^2 = 0.9073$, $Q^2_{\text{LOO}} = 0.8941$, and $R^2_{\text{test}} = 0.8988$). The AD indicates that most structures are adequately represented by the chemical space of the model. Thus, the values of the predicted activities can be considered reliable.

ACKNOWLEDGMENTS

This project was financially supported by Universitas Gadjah Mada (UGM) through a Rekognisi Tugas Akhir (RTA) program in 2020.

CONFLICT OF INTEREST

The authors declared that they have no conflicts of interest.

AUTHOR CONTRIBUTIONS

All authors made substantial contributions to conception and design, acquisition of data, or analysis and interpretation of data; took part in drafting the article or revising it critically for important intellectual content; agreed to submit to the current journal; gave final approval of the version to be published; and agree to be accountable for all aspects of the work. All the authors are eligible to be an author as per the international committee of medical journal editors (ICMJE) requirements/guidelines.

ETHICAL APPROVALS

This study does not involve experiments on animals or human subjects.


How to cite this article: