BAREKENG: Journal of Mathematics and Its Application
December 2022 Volume 16 Issue 4 Page 1441–1452
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol16iss4pp1441-1452

AGGLOMERATIVE HIERARCHICAL CLUSTERING ANALYSIS IN PREDICTING ANTIBACTERIAL ACTIVITY OF COMPOUND BASED ON CHEMICAL STRUCTURE SIMILARITY

Siswanto [1*], Nur Hilal A. Syahrir[2]

[1]Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Hasanuddin
Jl.Perintis Kemerdekaan KM.10, Makassar, 90245, Indonesia.
[2]Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Sulawesi Barat
Jl. Prof. Dr. Baharuddin Lopa, Majene, 91412, Indonesia.

Corresponding author's e-mail: [1] siswanto@unhas.ac.id*

Abstract. *Resistance to antibiotics is increasing to alarmingly high levels. As antibiotics are less effective, more infections are becoming more complex and often impossible to treat. Numerous antibiotics discovered in marine organisms show that the marine environment, which accounts for over half of the world's biodiversity, is a huge source of novel antibiotics. This resource must be explored to identify next-generation antibiotics. This research aimed to predict antibacterial activity in marine compounds using a computational approach to reduce the cost and time of finding marine organisms, extracting, and testing numerous unknown marine compounds' bioactivities. We used a simple unsupervised learning approach to predict the biological activity of marine compounds, called agglomerative hierarchical clustering. We mixed antibiotic drug data in DrugBank Database and chemical compound data from marine organisms in the literature to compile our dataset. We applied five linkage methods in our dataset and compared the best method by assessing internal validation measurement. We found that the Ward with squared dissimilarity matrix is the best method in the dataset, and ten compounds from 73 compounds of marine compounds are determined as potential marine compounds which have antibacterial activity.*

Keywords: *AGNES, antibiotic, computational approach, in-silico, natural compound, ward.*

Article info:

Submitted: 6th August 2022 *Accepted: 8th November 2022*

How to cite this article:
Siswanto and N. H. A. Syahrir, " AGGLOMERATIVE HIERARCHICAL CLUSTERING ANALYSIS IN PREDICTING ANTIBACTERIAL ACTIVITY OF COMPOUND BASED ON CHEMICAL STRUCTURE SIMILARITY ", *BAREKENG: J. Math. & App.*, vol. 16, iss. 4, pp. 1441-1452, Dec., 2022.

1. INTRODUCTION

Antibiotics are the first solution to fighting bacterial infections that threaten human health. However, when antibiotics are consumed over time, microorganisms acquire the capacity to tolerate them and evolve to prevail over antibiotics [1]. This condition is popularly known as antibiotic resistance. As antibiotics become less effective, many infections, including pneumonia, TB, blood poisoning, gonorrhea, and foodborne diseases, are becoming more challenging and occasionally impossible to treat [2]. Thus, antibiotic resistance is increasing to alarmingly high levels.

Natural products are an essential source of leads for drug development and have played a vital role in identifying and developing antibacterial agents [3]. The ocean is one of the potent natural product sources since natural products from the sea have shown unique molecular structures and promising biological activities. Numerous antibiotics have been discovered in marine organisms, such as bacteria, fungi, algae, sponges, cnidarians, arthropods, echinoderms, and mollusks [4]. It indicates that the marine environment, which accounts for almost half of the world's biodiversity, is a vast source of novel antibiotics and has to be investigated to find new generation antibiotics.

Although marine organisms are vast sources of antibiotic drug development, marine sources are mostly untapped resources where new antibiotic molecules can be found. Finding bioactive compounds from marine environments is never easy; there are many obstacles to overcome, such as finding rare marine resources, reviving inhospitable organisms outside of the marine environment, separating novel compounds from known ones, revealing the function of MNPs, and maximizing their pharmacological use [5]. In the drug discovery process, millions of extracts from marine organisms are synthesized and tested later for activity against various target infectious diseases. That process is time-consuming and requires a high cost since producing one "lead" compound, 50.000–100.000 active molecules might be required.

This research aimed to predict antibacterial activity in marine compounds using a computational approach. This method can reduce the cost and time of finding marine organisms, extracting, and testing numerous marine compounds whose bioactivities are still unknown, particularly their's antibacterial activity. There is much previous research using a computational approach in predicting biological activity, such as [6] an implemented classification algorithm, a particularly the Naïve-Bayesian classifier, in predicting bioactive activity. They combined the Similarity Ensemble Approach (SEA) and the maximum Tanimoto similarity. In the other study, [7] modified the Naïve Bayes algorithm to reveal structure−activity relationships. Those previous studies used supervised modeling to classify various kinds of biological activity. In this research, we used a simple unsupervised learning approach to predict the antibacterial activity of marine compounds. We used agglomerative hierarchical clustering analysis based on chemical similarity as an input in the algorithm. Chemical similarity is used in this research based on the hypothesis that two compounds with similar chemical structures probably have similar bioactivities and bind related target proteins [8]. Compared to previous studies which use a classification model, one of the advantages of this study is the ability to know related targets based on the synthetic drugs in a similar cluster. It can be useful for discovering the natural compounds' activities and mechanisms of action.

2. RESEARCH METHODS

2.1 Data Sources

In this research, we collected chemical structure data from marine compounds and antibiotic drug compounds. The marine compounds are the compounds from the waters of South Sulawesi Province (SSW), Indonesia. All chemicals of marine compounds were collected from the literature [8] and were identified in the PubChem database (https://pubchem.ncbi.nlm.nih.gov/) [9]. In addition, antibiotic drug compounds were collected from the DrugBank database (https://go.drugbank.com/) [10]. Both marine and antibiotics compounds were converted from their PubChem CID to an SDF file using the ChemmineR package in R [11]. We then used the fingerprint of all molecules as raw data in quantifying chemical similarity.

2.2 Methods

2.2.1 Compound Similarity Measures

Many computational techniques use chemical similarity in pharmaceutical research to find new molecules [12]. Chemical or compound similarity is the structural or functional closeness involving chemical components, molecules, or compounds. It usually is represented in numerical scores as an effort to quantify chemical structural similarity in compounds. Several formulas are provided in computing compound similarity. One of the most popular measures of molecular similarities is the Tanimoto coefficient or Jaccard coefficient. The Tanimoto is defined as the fraction of features in common between two molecules relative to the total number of features present in either molecule. The similarity between a pair of compounds calculated by the Tanimoto formula:

$$CS = c/[a + b - c] \tag{1}$$

where a is the number of on-bits in the first compound, b is the number of on-bits in the second compound, and c is the number of bits in both compounds. The 2D structure of compounds was converted to a binary number using 1024-bit fingerprints, unique substructures generated by fragmenting each molecule.

2.2.2 Agglomerative Hierarchical Clustering Method

Hierarchical clustering is one type of clustering method which assigns a set of objects into groups called clusters. Generally, hierarchical clustering falls into two types: 1) Divisive approach. Initially, all data points lie in a single cluster and are split recursively as one descends the hierarchy; 2) Agglomerative approach. Unlike the divisive approach, each data point begins in its cluster, and pairings of clusters are combined as one ascends the hierarchy [13]. This study used an agglomerative hierarchical approach where the data point was defined as a cluster and combined existing clusters at each step. We used different methods for this approach:

a. Complete Linkage
The basic principle in the single linkage is the distance between two clusters to be the maximum distance between any single data point in the first cluster and any single data point in the second cluster [14]. Based on this definition of distance between clusters, the two clusters are combined at each stage of the process with the smallest complete linkage distance. The distance D_{ij} between two clusters C_i and C_j is the maximum distance between two points x and y, with $x \in C_i$ and $y \in C_j$:

$$D_{ij} = \max_{\mathrm{x} \in C_i, y \in C_j} d(x, y) \tag{2}$$

where n_i is the number of elements in clusters C_i and n_j is the number of elements in clusters C_j. This method tends to form clusters with the same variance, tiny ones.

b. Average Linkage
In this method, the distance between two clusters was defined as the average distance between data points in the first cluster and data points in the second cluster [14]. The distance D_{ij} between two clusters C_i and C_j is the mean of the distances between the pair of points x and y, where $x \in C_i$ and $y \in C_j$:

$$D_{ij} = \sum_{\mathrm{x} \in C_i, y \in C_j} \frac{d(x, y)}{n_i \times n_j} \tag{3}$$

where n_i is the number of elements in clusters C_i and n_j is the number of elements in clusters C_j.

c. Centroid Method
In this method, the distance between two clusters is the distance between the two mean vectors of the clusters in the centroid method. The two clusters with the smallest centroid distance are combined at each process stage [14]. The distance D_{ij} between two clusters C_i and C_j is the squared euclidean distance between the gravity centres of the two clusters. It is defined as:

$$D_{ij} = \left\| \bar{x}_i - \bar{x}_j \right\|^2 \tag{4}$$

where $\bar{x}_i$ is the mean vectors in clusters C_i and $\bar{x}_j$ is the mean vectors in clusters C_j.

d. Ward's Method

This method is an ANOVA-based approach that does not define a direct distance measurement between two data points or clusters. In Ward's method, one-way univariate ANOVA is implemented for each variable, with clusters representing the groups. In each step, the method calculates the incremental sum of squares and pairs of clusters with minimum cluster distance [15]. Assume that there are three clusters called C_i, C_j and C_k including n_i, n_j and n_k as the number of rows (or columns). Clusters C_j and C_k are aggregated to form a new single cluster called C_l.

The distance between cluster C_i and the new cluster C_l is calculated as:

$$D_{il} = a \times D_{ij} + b \times D_{ik} - c \times \ D_{jk} \tag{5}$$

where

$$a = \frac{n_i + n_j}{(n_i + n_j + n_k)} \tag{6}$$

$$b = \frac{n_i + n_k}{(n_i + n_j + n_k)} \tag{7}$$

$$c = \frac{n_i}{(n_i + n_j + n_k)} \tag{8}$$

Ward.D1 and Ward.D2 algorithms are found in the literature and available in software packages that produce different results when applied to the same distance matrix D. Ward.D2 method also uses the minimum variance method; however, dissimilarities are squared before clustering.

2.2.3 Cluster Validation Measures

Validity measurements endeavor to determine how correctly the clusters represent the data. There is a proliferation of validity metrics, and various assessments frequently generate divergent results. This research uses three internal validity measures: Dunn Index, Average Silhouette Width, and Calinski-Harabasz Index.

a. Dunn Index

J. Dunn proposed Dunn's Index (DI) in 1974 as an index based on cluster element distance. The Index is obtained by calculating the ratio between the minimal intercluster distance and to maximal intracluster distance [16]. It follows:

$$D(\mathcal{C}) = \frac{\min\limits_{1 \le i < j \le q} \left(\min\limits_{i \in C_k, j \in C_l} \ D(C_i, C_j) \right)}{\max\limits_{1 \le k \le q} diam(C_k)} \tag{9}$$

where $D(C_i, C_j)$ is the dissimilarity function between two clusters C_i and C_j and $diam(C)$ is the diameter of a cluster. The diameter is the maximum distance between observations in a cluster C_k, which may be considered a measure of cluster dispersion. If the data set contains compact and well-separated clusters, the distance between the clusters is expected to be large, and the diameter of the clusters is expected to be small. Therefore, the Dunn index should be maximized.

b. Average Silhouette Width

The silhouette width is the average of each observation's silhouette value. The silhouette value measures the degree of confidence in the clustering assignment of a particular observation, with well-clustered observations if values are near one and poorly clustered observations if values are near -1 [16].

For observation i, it is defined as:

$$S(i) \ = \frac{b_i - a_i}{max(b_i, a_i)} \tag{10}$$

where a_i is the average distance between i and all other observations in the same cluster, and b_i is the average distance between i and the observations in the "nearest neighbouring cluster," i.e.

$$a_i = \frac{1}{n\big(C(i)\big)} \sum_{j \in C(i)} dist(i,j) \tag{11}$$

and,

$$b_i = \min_{C_k \in \mathcal{C} \setminus C(i)} \sum_{j \in C_k} \frac{dist(i,j)}{n(C_k)} \tag{12}$$

where $C(i)$ is the cluster containing observation i, $dist(i,j)$ is the Euclidean between observations i and j, and $n(C)$ is the cardinality of cluster C. The silhouette width thus lies in the interval [−1, 1] and should be maximized.

c. Calinski-Harabasz Index

This index is an evaluation index based on the degree of cluster dispersion [17]. It is defined by:

$$\mathbf{CH(C)} = \frac{\boldsymbol{B(C)(N - C)}}{\boldsymbol{W(C)(C - 1)}} \tag{13}$$

where C is the corresponding number of clusters, $B(C)$ is the inter-cluster divergence, also called the inter-cluster covariance, $W(C)$ is the intra-cluster divergence, also called the intra-cluster covariance, and N is the number of samples. The $B(C)$ defined as:

$$\boldsymbol{B(C)} = \left(\sum_{c=1}^{C} \boldsymbol{a_c \|\bar{x}_c - \bar{x}\|^2} \right) \tag{14}$$

and, the $W(C)$ follows the equation:

$$\boldsymbol{W(C)} = \left(\sum_{c=1}^{C} \sum_{C(j)=c} \left\| \boldsymbol{\bar{x}_j - \bar{x}_c} \right\|^2 \right) \tag{15}$$

The high degree of cluster dispersion, the larger the $B(C)$ is. The closer the relationship is in the cluster, the smaller the $W(C)$ is. The better clustering effect occurs when the ratio is higher, and the value of the CH index is higher.

3. RESULTS AND DISCUSSION

3.1. Chemical Similarity

Determining the structural similarity of two chemical compounds is necessary for discovering prospective drug development molecules. We generated a compound chemical similarity matrix of 105 compounds in this study. Seventy-three compounds originated from marine compounds, specifically from 17 marine biotas in the waters of South Sulawesi. A list of the marine compounds can be found in [12] with the SSW code of province. Many marine compounds' bioactivities are still undetermined, while the rest (32 compounds) are antibiotics drugs whose bioactivities are known and pass clinical testing. The list of 32 antibiotic drugs can be seen in Table 1.

Table 1. List of Antibiotics Drugs Obtained from DrugBank Database

ID	Drugbank Id	Name	ID	Drugbank Id	Name
D1	DB00027	Gramicidin D	D17	DB00684	Tobramycin
D2	DB00199	Erythromycin	D18	DB00759	Tetracycline
D3	DB00207	Azithromycin	D19	DB00798	Gentamicin
D4	DB00260	Cycloserine	D20	DB00803	Colistin
D5	DB00314	Capreomycin	D21	DB00826	Natamycin
D6	DB00400	Griseofulvin	D22	DB00955	Netilmicin
D7	DB00415	Ampicillin	D23	DB01045	Rifampicin
D8	DB00446	Chloramphenicol	D24	DB01053	Benzylpenicillin
D9	DB00452	Framycetin	D25	DB01082	Streptomycin
D10	DB00479	Amikacin	D26	DB01172	Kanamycin
D11	DB00512	Vancomycin	D27	DB01190	Clindamycin

ID	Drugbank Id	Name	ID	Drugbank Id	Name
D12	DB00595	Oxytetracycline	D28	DB01201	Rifapentine
D13	DB00615	Rifabutin	D29	DB01220	Rifaximin
D14	DB00626	Bacitracin	D30	DB01421	Paromomycin
D15	DB00646	Nystatin	D31	DB02703	Fusidic Acid
D16	DB00681	Amphotericin B	D32	DB08874	Fidaxomicin

The chemical similarity of antibiotic medicines and marine chemicals is assessed based on the hypothesis that two compounds with similar chemical structures have similar bioactivities and bind functionally related target proteins [8]. Using the Tanimoto formula, we obtained a chemical similarity matrix represented in a heat map (see Figure 1).

The range of chemical similarity scores in this measurement is from 0 to 1. The compound with high similarity scores has a high value and vice versa. Therefore, we can see that the red area represents the compound pair with high structure similarity, in contrast to the dark blue area, where the pair of the compound has a low similarity. At the same time, the yellow one indicates the compound pairs with moderate similarity. As shown in the below heatmap, the proportion of red is smaller than the yellow and dark blue. In other words, many pairs of compounds have lower than high similarity. In further analysis, the similarity of the compounds to all compounds as a row becomes input in clustering analysis. The clustering is conducted to group the compounds with high similarity and separate the compound with low similarity based on the chemical similarity score.

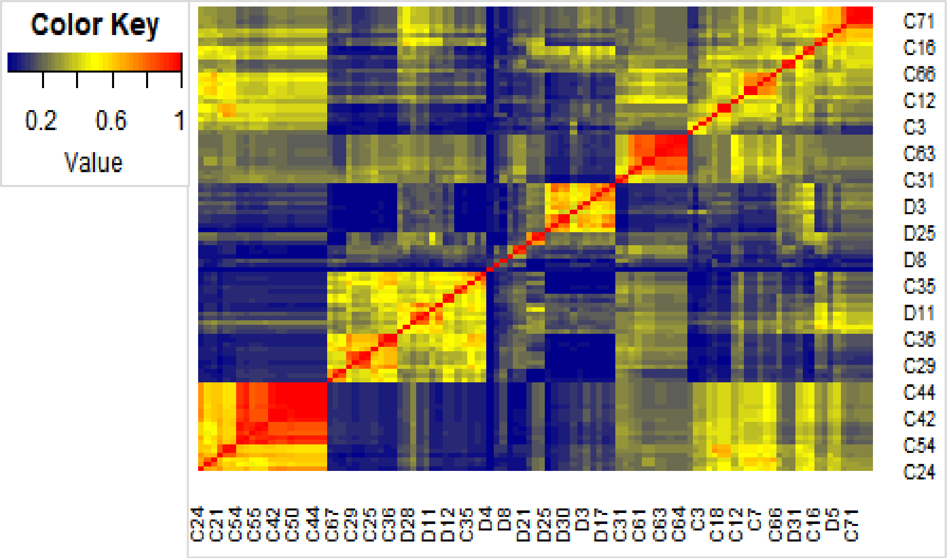


Figure 1. Chemical Similarity Heatmap

3.2. Cluster Evaluation and Comparison

Several linkage methods in agglomerative hierarchical clustering analyses are applied to group the compounds based on their similarity to all other compounds. We also implemented internal clustering validation to compare the various methods and identify the optimal number of clusters. Three internal cluster validation metrics were used and produced different results. We calculated the Dunn Index, Average Silhouette Width, and Calinski Harabasz Index from two to 20 clusters using agglomerative hierarchical clustering analysis in five linkage methods.

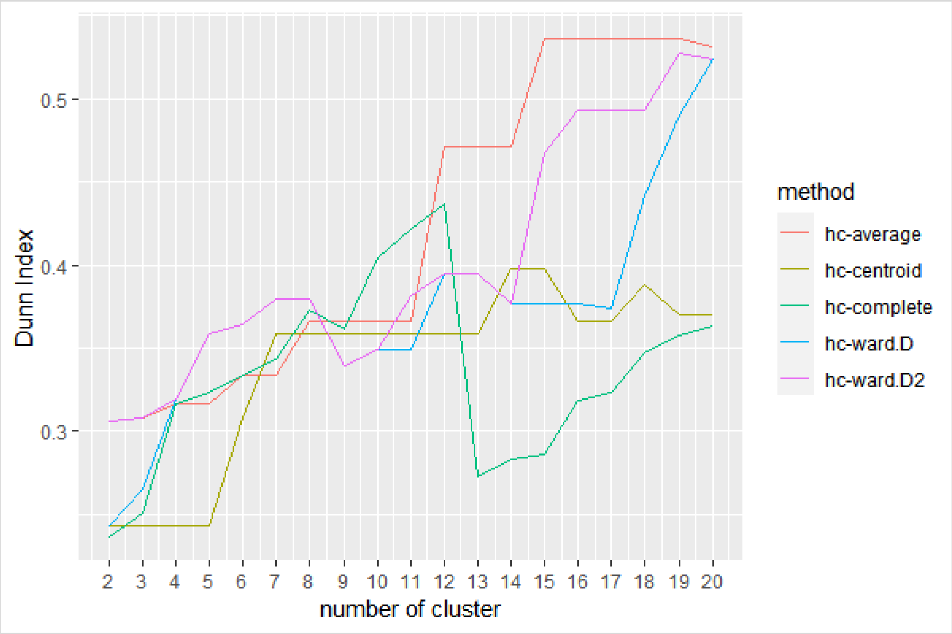


Figure 2. Dunn Index of Agglomerative Clustering

The second measurement of validation is the silhouette coefficient. The average silhouette width demonstrates a linear relationship with the cluster number. As shown in Figure 3, the average silhouette width in $n = 20$ is the highest of the three methods. If we extend the cluster's upper limit, the average silhouette still tends to increase and decrease after around $n = 30$. The method with the highest Dunn-Index depends on the cluster number since no single linkage method has the highest index Dunn in every cluster. Thus, according to this picture, for $n < 10$, we indicate Ward.D2 with $n = 9$ as the best method based on Average Silhouette Width.

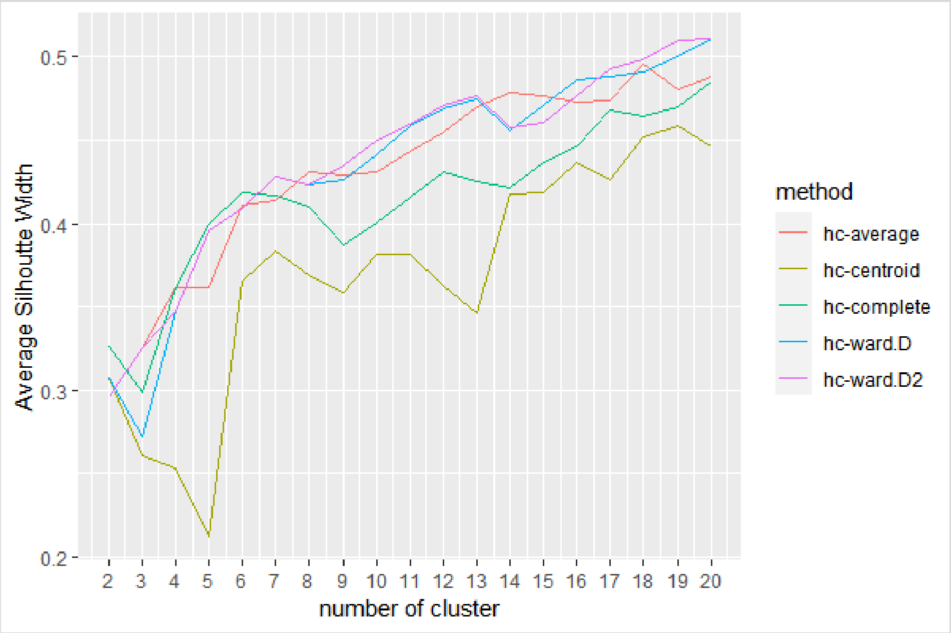


Figure 3. Average Silhouette Width of Agglomerative Clustering

The last internal cluster validation applied in this study is the Calinski – Harabasz (CH) Index. The higher the CH index, the better the clustering is (see Figure 4). Based on the graph, we obtain that Ward.D1 method is suitable for $n < 6$. However, if $n > 6$, Ward.D2 is the best method. Thus, according to the CH index graph, for $n < 10$, we indicate Ward.D2 with $n = 9$ as the best method based on Average Silhouette Width. We further applied Ward.D2 with $n = 9$ in clustering chemical compounds based on chemical similarity.

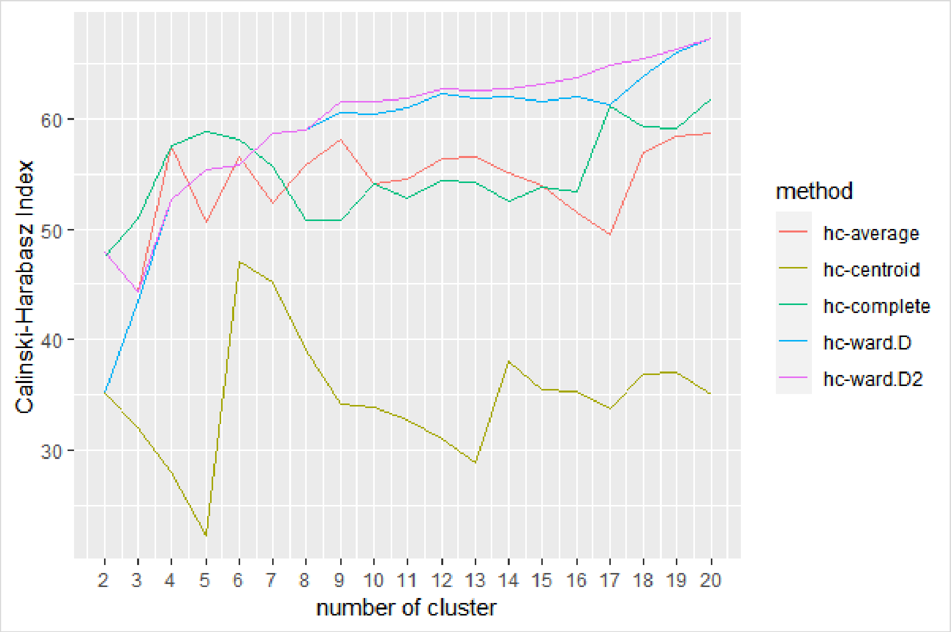
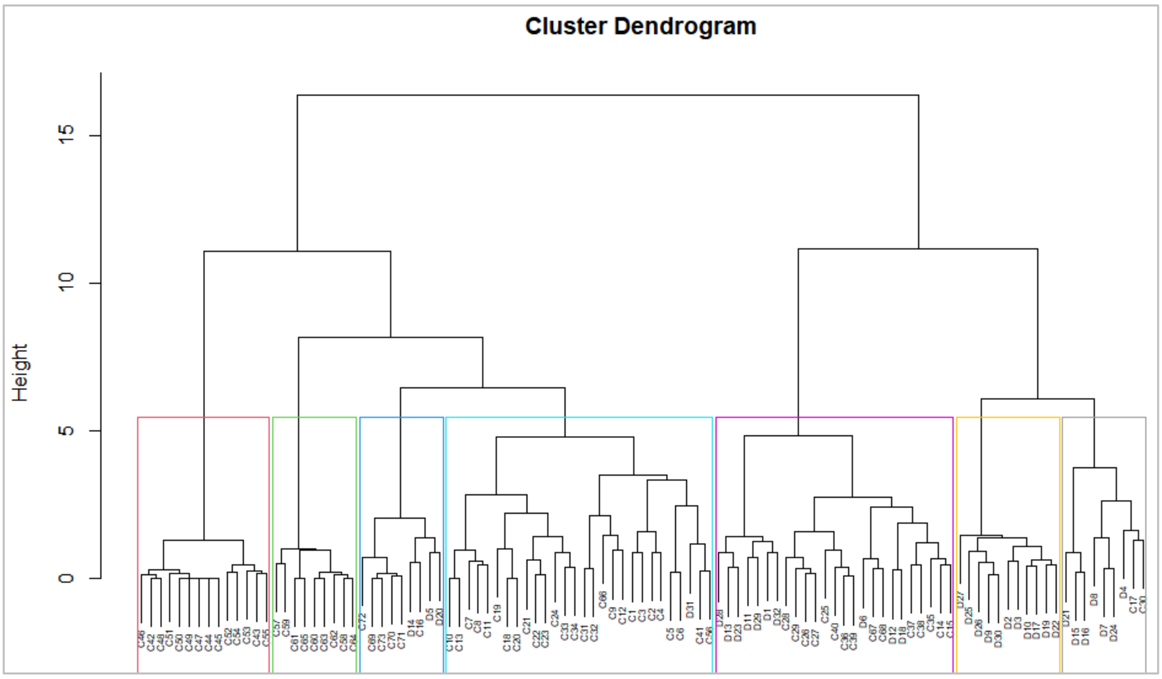


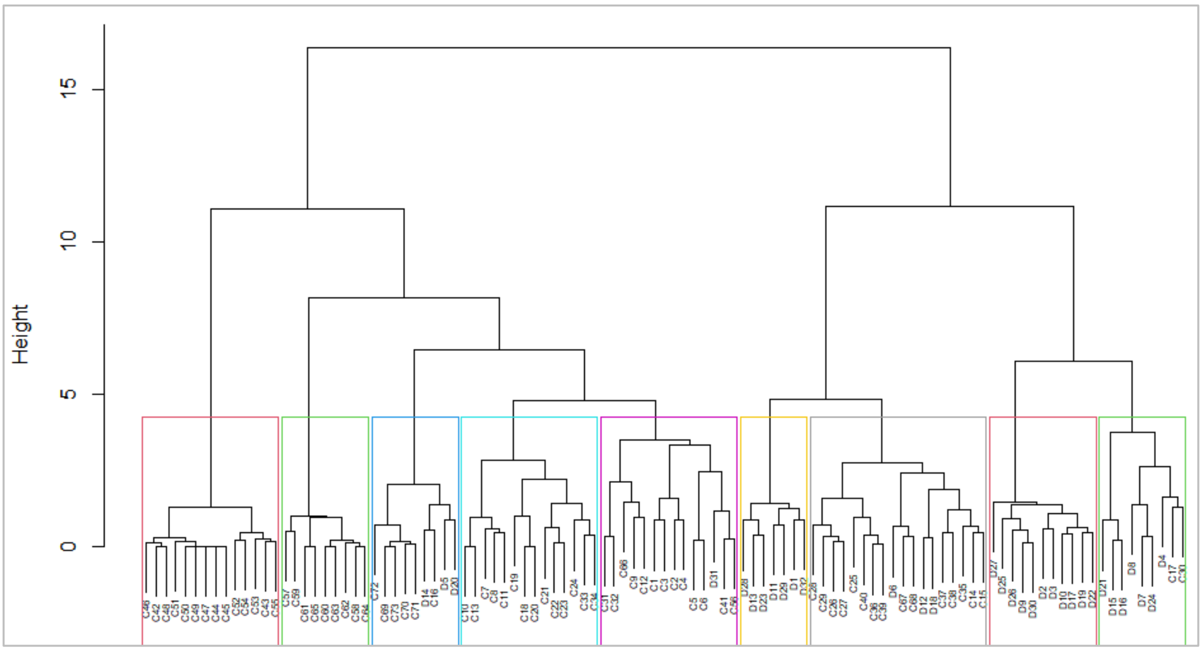
Figure 4. Calinski-Harabasz Index of Agglomerative Clustering

3.3. Clustering of the Chemical Compound Results

Based on the internal validation results, we applied agglomerative hierarchical clustering using Ward.D2 for $n = 7$ and $n = 9$. The dendrograms of clustering are shown in Figure 5a and Figure 5b, respectively. Both the ward methods are summarized in Table 2. The number of antibiotics drugs clustered well in both ward methods; however, Ward.D2 for $n = 9$ conducted more groups for antibiotics drugs. Cluster 1 in $n = 7$ is divided into two groups in $n = 9$ (clusters 1 and 5). In further analysis, we use the nine clusters to predict the potential marine compounds for antibacterial activity since the antibiotics drugs are more clustered well.



a)



b)

Figure 5. a) Dendrogram of Ward.D2 Method for $= 7$; b) Dendrogram of Ward.D2 Method for $n = 9$

As demonstrated in Table 2, it can be seen clearly that cluster 1 and cluster 2 contain only antibiotics drugs. All the antibiotics drugs in those clusters have no similarities with 73 marine compounds. Thus, those clusters are not necessary to be explored further and are classified as impotent clusters

Table 2. Number of Compounds in each Cluster

Ward D.2, $n = 7$				Ward D.2, $n = 9$			
cluster	**number of compounds**	**number of antibiotics drugs**	**number of marine compounds**	**cluster**	**number of compounds**	**number of antibiotics drugs**	**number of marine compounds**
1	25	10	15	1	7	7	
2	11	11		2	11	11	
3	9	7	2	3	9	7	2
4	9	3	6	4	9	3	6
5	28	1	27	5	18	3	15
6	14		14	6	14	1	13
7	9		9	7	14		14
				8	14		14
				9	9		9
Total	**105**	**32**	**73**		**105**	**32**	**73**

Similar to clusters 1 and 2, clusters 7, 8, and 9 contain only one type of compound. Those three clusters contain 37 compounds clustered with only marine compounds. There are no antibiotics drugs assigned in those clusters. So, we can conclude that none of the marine compounds in these clusters can have antibacterial activities based on this clustering of chemical similarity (See Table 3).

Table 3. List of most Potent, Moderately Potent, and Impotent Clusters and their Compounds

Cluster	Category	Marine Compounds	Similar antibiotics drugs
1	Impotent cluster	-	Seven drugs (D1, D11, D13, D23, D28, D29, D32)
2	Impotent cluster	-	11 drugs (D10, D17, D19, D2, D22, D25, D26, D27, D3, D30, D9)
3	Most potent cluster	**2** compounds (C17, and C30)	Seven drugs (D15, D16, D21, D24, D4, D7, and D8)
4	Moderately potent cluster	**6** compounds (C16,C69,C70,C71,C72, C73)	Three drugs (D14, D20, and D5)
5	Moderately potent cluster	**15** compounds (C14, C15, C25,C26, C27, C28, C29, C35, C36, C37, C38, C39, C40, C67, C68)	Three drugs (D12, D18, and D6)
6	Least potent cluster	**13** compounds (C1, C12, C2, C3,C31, C32, C4, C41, C5, C56, C6, C66, C9)	One drug (D31)
7	Impotent cluster	**14** compounds (C10, C11, C13,C18, C19, C20, C21, C22, C23, C24, C33, C34, C7, C8)	-
8	Impotent cluster	**14** compounds (C42, C43, C44, C45, C46, C47, C48, C49, C50, C51, C52, C53, C54, C55)	-
9	Impotent cluster	**9** compounds (C57, C58, C59, C60, C61, C62, C63,C64,C65)	-

Clusters 3, 4, 5, and 6 are the main clusters that can be analyzed to find the potent antibacterial activity from marine compounds since some antibiotic drugs are clustered with marine compounds. Those four clusters are divided into three categories: most potent cluster (cluster 3), moderately potent cluster (cluster 4 and cluster 5), and least potent cluster (cluster 6). The most potent cluster produces the most potent compound, the moderately potent cluster produces a moderate potent compound, and the least potent cluster produces the least potent compound of marine compounds (Figure 3).

Two marine compounds in cluster 3 are (+)-Helicascolide (C17) and 3,5-Dibromo-4methoxyphenethyl amine (C30). Those two compounds are the most potent marine compounds since the marine compound lies in a cluster where antibiotics drugs dominate the group. In cluster 4 and cluster 5, the proportion of marine compounds is higher than antibiotic drugs. Thus, screening marine compounds can be applied to filter the most potent compound in those two groups. One of the ways to filter the potent marine compound in this study is to look at the nearest neighbor compound that can easily be seen in the dendrograms (See Figure 6). In cluster 4, after identifying the nearest compound from antibiotics drugs in Figure 6, we found that Callyaerin G (C16) and Cadiolide B (C14) are the marine compounds with high similarity with Capreomycin (D5) and Colistin. Thus, the two marine compounds can be classified as moderately antibacterial potent compounds. In cluster 5, using the previous method, we found four compounds (Latonduine A (C67), Latonduine B (C68), Naamine G (C37), and Naamine F (C38)) that are similar to antibiotics drugs (D6 and D18).

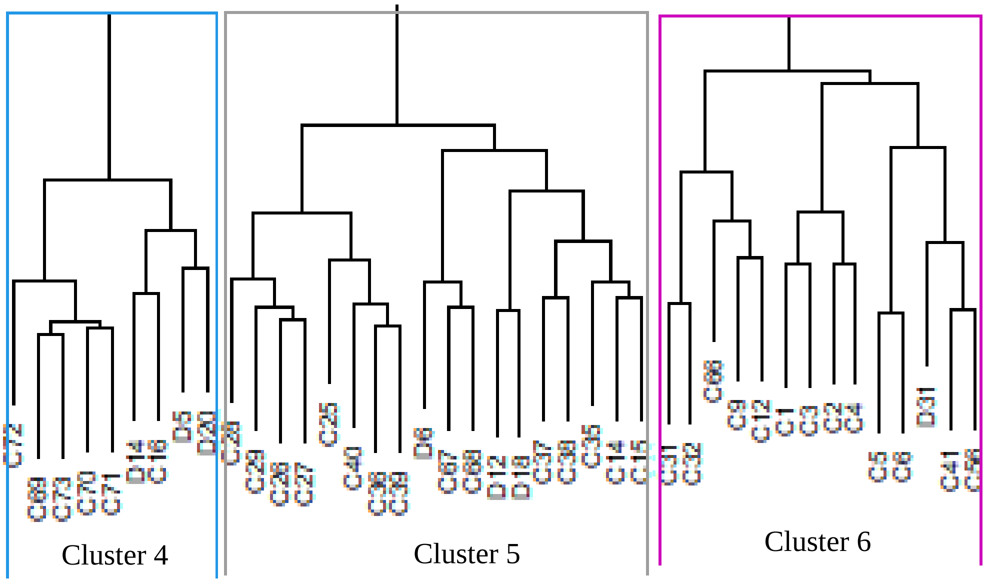


Figure 6. Detailed Dendrograms of Clusters 4, 5, and 6

As the proportion of marine compounds is more significant than antibiotics drugs in cluster 6, those marine compounds in this cluster are the least potent antibacterial compound. However, the most similar compound with the "one" antibiotics drug in cluster 6 still has the chance to be the least potent compound. The compound was identified using the dendrogram (See Figure 6). In dendrograms, the compound with close distance with antibiotics drug are (–)-Sarasinoside J (C41) and (–)-Sarasinoside K (C56). Thus, both compounds are classified as the least potent antibacterial compound.

4. CONCLUSIONS

Agglomerative hierarchical analysis can successfully group the compounds based on chemical similarity. The formed clusters can predict the antibacterial activity by combining antibiotic drug compounds with unknown biological activity. We found that the Ward.D2 with squared dissimilarity matrix is the best method in the dataset, and ten compounds from 73 compounds of the marine compound are determined as potential marine compounds to have antibacterial activity. The result of this study can be used as the first screening of researchers before applying different methodologies such as molecular docking analysis, molecular dynamics simulation, QSAR analysis, or in-vivo analysis to know the ability of compounds to antibacterial activity. This method also can be improved by adding various features as input in clustering modeling and improving the hierarchical clustering method to gain a better cluster.

REFERENCES

[1] L. Verstraete, B. Van den Bergh, N. Verstraeten, and J. Michiels, “Ecology and evolution of antibiotic persistence,” *Trends in Microbiology*, vol. 30, no. 5. pp. 466–479, 2022, doi: 10.1016/j.tim.2021.10.001.

[2] L. Sarvananda and A. D Premarathne, "The Growing Of Antibiotic Resistance: A Short Viewpoint," *Pharm. Pharmacol. Res.*, vol. 5, no. 3, pp. 01–02, 2022, doi: 10.31579/2693-7247/068.

[3] H. Li *et al.*, "Discovery of Marine Natural Products as Promising Antibiotics against Pseudomonas aeruginosa," *Mar. Drugs*, vol. 20, no. 3, 2022, doi: 10.3390/md20030192.

[4] T. B. Ng, R. Chi, F. Cheung, J. H. Wong, and A. A. Bekhit, "Antibacterial products of marine organisms," 2015, doi: 10.1007/s00253-015-6553-x.

[5] G. Zhang, J. Li, T. Zhu, Q. Gu, and D. Li, "Advanced tools in marine natural drug discovery," *Curr. Opin. Biotechnol.*, vol. 42, pp. 13–23, 2016, doi: 10.1016/j.copbio.2016.02.021.

[6] J. J. Irwin, G. Gaskins, T. Sterling, M. M. Mysinger, and M. J. Keiser, "Predicted Biological Activity of Purchasable

Chemical Space," *J. Chem. Inf. Model.*, vol. 58, no. 1, pp. 148–164, 2018, doi: 10.1021/acs.jcim.7b00316.

[7] P. V. Pogodin, A. A. Lagunin, A. V. Rudik, D. S. Druzhilovskiy, D. A. Filimonov, and V. V. Poroikov, "AntiBac-Pred: A Web Application for Predicting Antibacterial Activity of Chemical Compounds," *J. Chem. Inf. Model.*, vol. 59, no. 11, pp. 4513–4518, 2019, doi: 10.1021/acs.jcim.9b00436.

[8] V. Periwal *et al.*, "Bioactivity assessment of natural compounds using machine learning models trained on target similarity between drugs," *PLoS Comput. Biol.*, vol. 18, no. 4, p. e1010029, 2022.

[9] S. Kim *et al.*, "PubChem in 2021 : new data content and improved web interfaces," vol. 49, no. November 2020, pp. 1388–1395, 2021, doi: 10.1093/nar/gkaa971.

[10] D. S. Wishart *et al.*, "DrugBank 5.0: a major update to the DrugBank database for 2018," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D1074–D1082, 2018.

[11] Y. Cao, T. Backman, K. Horan, and T. Girke, "ChemmineR: Cheminformatics Toolkit for R." Citeseer, 2014.

[12] N. Hanif, A. Murni, C. Tanaka, and J. Tanaka, “Marine natural products from Indonesian waters,” *Mar. Drugs*, vol. 17, no. 6, 2019, doi: 10.3390/md17060364.

[13] L. Billard and E. Diday, "Agglomerative Hierarchical Clustering," *Clustering Methodology for Symbolic Data*. pp. 261–316, 2019, doi: 10.1002/9781119010401.ch8.

[14] A. M. Jarman, "Hierarchical cluster analysis: Comparison of single linkage, complete linkage, average linkage and centroid linkage method," *Georg. South. Univ.*, 2020.

[15] S. Miyamoto, R. Abe, Y. Endo, and J.-I. Takeshita, "Ward method of hierarchical clustering for non-Euclidean similarity measures," in *2015 7th International Conference of Soft Computing and Pattern Recognition (SoCPaR)*, 2015, pp. 60–63.

[16] T. Gupta and S. P. Panda, "Clustering validation of CLARA and k-means using silhouette & DUNN measures on Iris dataset," in *2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon)*, 2019, pp. 10–13.

[17] J. Baarsch and M. E. Celebi, "Investigation of internal validity measures for K-means clustering," in *Proceedings of the international multiconference of engineers and computer scientists*, 2012, vol. 1, pp. 14–16.