




# Graph Clustering Based on Chemical Similarity in Marine Compounds and Antibacterial Compounds

Edy Saputra Rusdi<sup>1</sup> , Nur Hilal A. Syahrir<sup>2</sup>, A. Muh. Amil Siddik<sup>1</sup>,  
Supri Bin Hj Amir<sup>1</sup>, and Wahyudi Rusdi<sup>3</sup>

<sup>1</sup> Department of Mathematics, Faculty of Mathematical and Natural Sciences, Universitas Hasanuddin, Makassar, Indonesia

edy\_saputra@sci.unhas.ac.id

<sup>2</sup> Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Sulawesi Barat, Sulawesi Barat, Indonesia

<sup>3</sup> Department of Sharia Economics, Faculty of Islamic Economics and Business, State Islamic Institute (IAIN) Sultan Amai Gorontalo, Gorontalo, Indonesia

**Abstract.** Many potential medicines have been discovered from marine natural products (MNPs). It indicates that marine compounds are essential sources in drug development and discovery. Although many marine compounds show particular biological activity, few are recorded as antibacterial compounds. Therefore, finding the potential compound as an antibacterial compound from a marine organism is still challenging. The aim of this study is to utilize a computational approach to discover potential antibacterial compounds from marine resources. The study focuses on employing the BiClusO algorithm for clustering based on the chemical similarity between marine compounds and antibacterial compounds. The results show that the number of clusters formed for marine biota compounds with antibiotic drug compounds is 4. Then the compounds of marine biota with antibiotic compounds formed 7 clusters. Finally, from these clusters, we obtain compounds that are predicted to have similar properties to antibacterial drugs or compounds. From 73 marine compounds, only Sarasinose J and (-)-Sarasinose K are predicted as potent antibacterial compounds.

**Keywords:** Antibacterial · BiClusO · Marine Compound

## 1 Introduction

Antibiotics are organic or inorganic substances that may either kill or limit the development of microorganisms, which can have an impact on their ability to survive [1]. Since Fleming made the initial antibiotic discovery in 1929, several classes of antibiotics have been discovered and developed globally and are used to treat illnesses in people, animals, and plants brought on by harmful bacteria [2]. The development of antibiotics is considered the greatest significant advancement in science and medicine throughout the 20<sup>th</sup> century. Significant declines in mortality and complications from serious infectious

illnesses including TB, syphilis, pneumonia, and gonorrhea have been brought about by the use of antibiotics in both human and animal medicine [3].

Although there are many medications of antibiotics have been discovered, resistance to antibiotics also has been increasing nowadays. Therefore, many drugs have become less effective since the bacteria can tolerate them. As a result, infectious diseases have become more complex to cure [4].

Finding unusual marine creatures, extracting chemicals, and conducting wet lab experiments to disclose the biological activity of marine compounds are just a few of the challenges that must be overcome. All of them are time-consuming and require many costs. Therefore, we utilize a computational approach in this study to shorten the research time and reduce the cost [5].

The researchers previously tested the predictive ability of various machine learning techniques, including random forest (RF), logistic regression (LR), tree gradient enhancer regression (GBRT), support vector machines (SVM), and multi-layer perception (MLP), create predictive models, for anti-bacterial compounds [6]. Many previous studies focus on developing supervised learning algorithms for predicting antibacterial compounds. This paper presents unsupervised learning methods to reveal the antibacterial activity in marine compounds. The unsupervised learning approach we used in this work is a graph clustering method called the BiClusO algorithm. The input of this model is the chemical similarity between compounds; the model's output is clusters of compounds.

## 2 Material and Methods

### 2.1 Source Data

In this study, we collected three sets of compounds: antibiotics compounds, antibiotic drug compounds, and marine compounds. A total of 1546 compounds related to antibiotics compounds were obtained from the PubChem database [7]. The compounds were searched by the keyword "antibiotic" in the search column of the PubChem website (<https://pubchem.ncbi.nlm.nih.gov/#query=antibiotic>). The 1546 compounds contain either drugs or non-drug compounds related to antibacterial activity. All compounds were read by SDF format of molecules used to generate fingerprints in further analysis.

The other database that we utilized in this study is the DrugBank database. We collected antibiotic drugs in the DrugBank database (<https://go.drugbank.com/>) [8]. We used only 32 antibiotics drugs [9] that had been identified and had the SDF files in PubChem.

The last set of compounds is the marine compounds set. We collected marine compounds specifically from the South Sulawesi Waters in Indonesia. The compounds were found in the literature [10]. Seventy-three compounds from 17 marine organisms were collected in the SSW area. We verified the compounds in PubChem, and similar to the other sets of compounds, we downloaded the SDF file for further analysis.

### 2.2 Method

In this paper, the research flow is divided into three, namely: compound structure database collection, similarity score, and graph clustering, which is presented in Fig. 1.

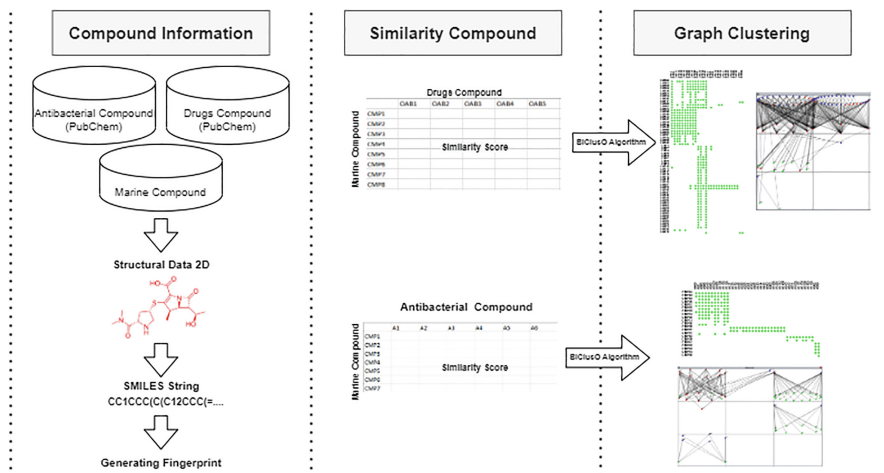


Fig. 1. Research Pipeline

### 2.3 Compound Structure Similarities

As the input in the BiClusO algorithm, we calculated the chemical similarity between compounds. In this work, we generate two types of chemical similarity matrices. The first is the matrix between 73 marine compounds and the 32 antibiotic drugs, and the second is between 73 marine compounds and 1546 antibacterial compounds. The chemical similarity is calculated based on the structural similarity measured by the sub-structure of each compound. The substructure is represented as a fingerprint in the R-Studio software, formed in binary variables 0 and 1. “0” indicates the absence of substructure in compounds, while “1” indicates the presence of substructures in compounds. The numerical measures to quantify this similarity are obtained by calculating the Jaccard coefficient ( $S_{A,B}$ ) [9] as follows:

$$S_{A,B} = z/[x + y - z] \quad (1)$$

where  $x$  is the number of similar bits in the first compound,  $y$  is the number of similar bits in the second compound, and  $z$  is the number of similar bits in both compounds.

### 2.4 BiClusO Algorithm

When a graph  $G = (V, E)$  can be partitioned into two subsets,  $V_1$  and  $V_2$ , each edge of  $G$  connects a vertex of  $V_1$  to a vertex of  $V_2$ , the graph is said to be bipartite. A bipartite graph can be represented by a binary matrix. BiClusO algorithm is able to generate a graph from a data matrix that represents a bipartite graph. Let a bipartite graph  $G = (V, E)$  be represented as binary matrix  $B$  of size  $|V_1| \times |V_2|$ . If  $(a, b) \in V_1$  presents any pair of row nodes, the following equations can be used to determine the association between these two nodes in terms of the Tanimoto coefficient and relation number:

$$TC = \frac{|M(a) \cap M(b)|}{|M(a) \cup M(b)|} \quad (2)$$

$$RN = |M(a) \cap M(b)| \quad (3)$$

where the neighbors of nodes  $a$  and  $b$  in set  $V$  are represented by the numbers  $M(a)$  and  $M(b)$ . We can construct a simple weighted graph using Tanimoto coefficients (TC) [11] and relation numbers (RN) [11]. Where the edge weights use the TC, the RN is used as a threshold to filter the edge, significantly reducing noise on the edges. From the remaining edges, a simple graph will be constructed. The DPclusO algorithm can easily separate densely connected regions as clusters from simple graphs. After clustering, a probability threshold will be used to list the nodes from set  $V$  in the cluster. Applying a similar procedure, the algorithm is repeated beginning with nodes from set  $V$ . Two-way biclustering is thus accomplished. Overlapping coefficient: Suppose that two biclusters are indicated by  $BC_1 = (a, c)$  and  $BC_2 = (b, d)$  where  $a \subseteq V_1$ ,  $b \subseteq V_1$ ,  $c \subseteq V_2$  and  $d \subseteq V_2$ . The following equation calculates the overlapping coefficient  $BCov$  [11] between  $BC_1$  and  $BC_2$ .

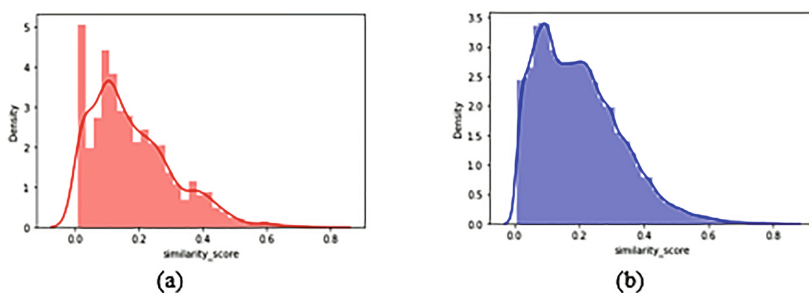
$$BCov = \frac{|a \cap b| |c \cap d|}{|a| |c| + |b| |d| - |a \cap b| |c \cap d|} \quad (4)$$

### 3 Results and Discussion

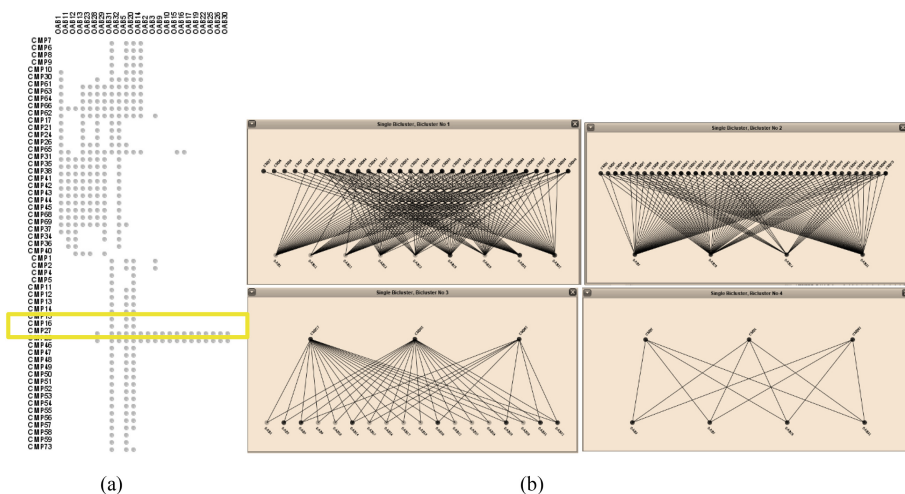
In this session, we compare data on marine compounds with drug compounds and marine compounds with antibacterial compounds. Assuming that the more similar the compound, the more similar its biological activity.

#### 3.1 Similarity of Marine Compounds and Drugs

The data used is data on marine compounds and antibiotic drug compounds. Then a chemical similarity matrix is made by calculating the structural similarity of the drug compounds with marine compounds. The size of the similarity matrix formed is  $73 \times 32$ . With an average similarity score of 0.17 in Fig. 2(a) and a standard deviation of 0.12.



**Fig. 2.** Density plot of similarity (a) marine compounds and drugs, (b) marine compounds and antibacterial compound



**Fig. 3.** (a) BiCluster Matrix of marine compounds and drugs, (b) 4 cluster of marine compounds and drugs

**Graph Clustering of Marine Compounds and Drugs.** Using a chemical compound similarity matrix of size  $73 \times 32$  as input to the BiClusO algorithm using  $RN = 4$  and  $TC = 0.33$  as the best threshold [12]. Then the parameters used, such as  $Density = 0.5$ ,  $CP = 0.5$ , and  $OV = 0.5$ , are the default parameters to form a cluster of simple graphs [13]. The smaller of the two highly overlapping clusters are discarded by filtering. In this case, the threshold value of the two clusters must have an  $OV$  value greater than or equal to the input value. The cluster merge function combines the two clusters and produces a large cluster. The overlap coefficient of Eq. (4) removes small biclusters from two biclusters. The join bicluster function combines the two biclusters and creates a larger bicluster [14].

The output of the BiClusO algorithm is a neighbor matrix which is then depicted in Fig. 3(a). Figure 3(b) formed 4 clusters that are independent of each other because of the parameters that have been set previously. For more details, see Table 1 below:

In Fig. 3(a), the size of the similarity matrix, which was initially  $73 \times 24$  after being processed with the BiClusO algorithm, becomes  $58 \times 40$ . Two marine compounds have the same structure as drug compounds, namely compounds with the initials  $CMP27((-)$ -Sarasinose J) and  $CMP28((-)$ -Sarasinose K), where each compound has similarities with 18 drug compounds, namely OAB (28, 31, 32, 5, 20, 14, 2, 3, 9, 10, 15, 16, 17, 19, 22, 25, 26 and 30).

**Table 1.** A cluster of marine compounds and drugs.

Cluster	Marine Compounds (in supplementary file1)	Drugs Compounds (in supplementary file2)
1	CMP7, CMP6, CMP8, CMP9, CMP10, CMP30, CMP61, CMP63, CMP64, CMP66, CMP62, CMP17, CMP21, CMP24, CMP26, CMP65, CMP31, CMP35, CMP38, CMP41, CMP42, CMP43, CMP44, CMP45, CMP68, CMP69, CMP37, CMP34, CMP36, CMP40	OAB1, OAB11, OAB12, OAB13, OAB23, OAB28, OAB29, OAB31, OAB32
2	CMP1, CMP2, CMP4, CMP5, CMP6, CMP7, CMP8, CMP9, CMP10, CMP11, CMP12, CMP13, CMP14, CMP15, CMP16, CMP26, CMP27, CMP28, CMP30, CMP46, CMP47, CMP48, CMP49, CMP50, CMP51, CMP52, CMP53, CMP54, CMP55, CMP56, CMP57, CMP58, CMP59, CMP61, CMP62, CMP63, CMP64, CMP65, CMP66, CMP73	OAB5, OAB20, OAB14, OAB31
3	CMP27, CMP28, CMP65	OAB2, OAB3, OAB5, OAB9, OAB10, OAB14, OAB15, OAB16, OAB17, OAB19, OAB20, OAB22, OAB25, OAB26, OAB28, OAB30, OAB31, OAB32
4	CMP2, CMP1, CMP62	OAB3, OAB5, OAB20, OAB31

### 3.2 Similarity of Marine Compounds and Antibacterial Compound

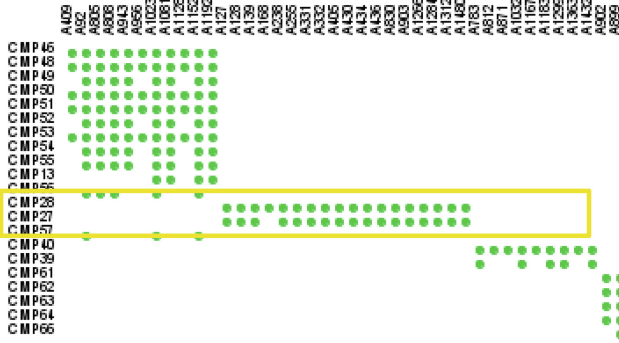
In this section, the data used for comparison is on marine compounds and antibiotic compounds, producing a matrix of chemical similarity by calculating the structural similarity of antibacterial compounds with marine compounds. The size of the similarity matrix formed is  $73 \times 1546$ . The average value of the similarity score is 0.19, as shown in Fig. 2(b) and the standard deviation is 0.13.

**Graph Clustering of Marine Compounds and Antibacterial Compound.** By entering the input similarity matrix of marine compounds and antibiotic compounds. Then applying the values of the RN, TC, Density, CP and OV parameters to the software [14], the results will be obtained in Table 2 and Fig. 4 below:

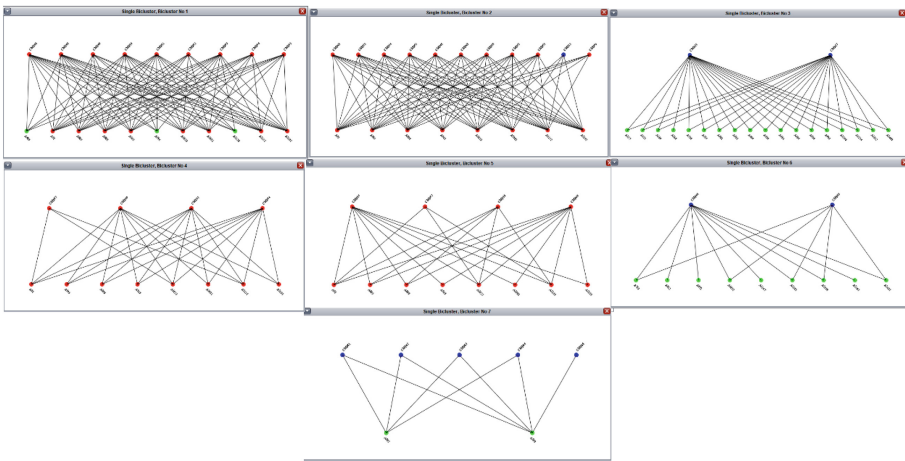
**Table 2.** Cluster of marine compounds and antibacterial compound

Cluster	Marine Compounds (in supplementary file1)	Drugs Compounds (in supplementary file2)
1	CMP46, CMP48, CMP49, CMP50, CMP51, CMP52, CMP53, CMP54, CMP55	A409, A92, A805, A808, A943, A956, A1023, A1081, A1128, A1152, A1192
2	CMP49, CMP52, CMP54, CMP55, CMP46, CMP48, CMP50, CMP51, CMP53, CMP13, CMP56	A92, A805, A808, A943, A1023, A1081, A1152, A1192
3	CMP28, CMP27	A127, A128, A139, A168, A238, A255, A331, A332, A405, A430, A434, A436, A830, A903, A1266, A1284, A1312, A1480
4	CMP55, CMP57, CMP56, CMP49	A92, A805, A808, A943, A1023, A1081, A1152, A1192
5	CMP55, CMP57, CMP56, CMP49	A92, A805, A808, A943, A1023, A1081, A1152, A1192
6	CMP40, CMP39	A783, A812, A871, A1032, A1167, A1183, A1299, A1363, A1432
7	CMP61, CMP62, CMP63, CMP64, CMP66	A902, A899

In Fig. 4(a), the matrix size, which was initially  $73 \times 1546$  after being processed with the BiClusO algorithm, becomes  $21 \times 40$ . There are two marine compounds with the most similar structure to medicinal compounds, namely compounds with the initials CMP27((-)-Sarasinoside J) with 17 antibiotic compounds and CMP28((-)-Sarasinoside K) with 18 antibiotic compounds. From the results of the similarity of chemical compound structures from marine compounds with drug compounds and antibiotic compounds, it is clear that there are compounds that have similarities, namely ((-)-Sarasinoside J with ((-)-Sarasinoside K.



(a)



(b)

**Fig. 4.** BiCluster Matrix of marine compounds and antibacterial compound, (b) 7 cluster of marine compounds and antibacterial compound

### 4 Conclusion

In this paper, the results show that the number of clusters formed for marine biota compounds with antibiotic drug compounds is 4. Then the compounds of marine biota with antibiotic compounds formed 7 clusters. Finally, from these clusters, we obtain compounds that are predicted to have similar properties with antibacterial drugs or compounds. From 73 marine compounds, only Sarasinocide J and (-)-Sarasinocide K are predicted as potent antibacterial compounds.

**Acknowledgments.** The author expresses his deepest gratitude to Hasanuddin University for assisting the author in the Internal Grant for Academic Advisor (PDPA) for 2022. With number 1476/UN.22/PT.01.03/2022.



## References

1. Fymat, Alain L.: Antibiotics and Antibiotic Resistance. *Biomedical Journal of Scientific & Technical Research*, 1(1) (2017).
2. Anh, H.Q., Le, T.P.Q., da Le, N., Lu, X.X., et al.: Antibiotics in surface water of East and Southeast Asian countries: A focused review on contamination status, pollution sources, potential risks, and future perspectives, *Science of The Total Environment*, 764 (2021).
3. Carvalho, I.T., Santos, L.: Antibiotics in the aquatic environments: A review of the European scenario, *Environment International*, 94 (2016).
4. Sarvananda, L., Premarathne, A. D.: The Growing Of Antibiotic Resistance: A Short Viewpoint, *Pharmaceutics and Pharmacology Research*, 5(3) (2022).
5. Durrant, J.D., Amaro, R.E.: Machine-learning techniques applied to antibacterial drug discovery, *Chemical Biology and Drug Design*, 85(1) (2015).
6. Li, W.-X., Tong, X., Yang, P.-P., Zheng, Y., Liang, J.-H., Li, G.-H., Liu, D., Guan, D.-G., Dai, S.-X.: Screening of antibacterial compounds with novel structure from the FDA approved drugs using machine learning methods, *Aging*, 14(3) (2022).
7. Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B.A., Thiessen, P.A., Yu, B., Zaslavsky, L., Zhang, J., Bolton, E.E.: PubChem in 2021: New data content and improved web interfaces, *Nucleic Acids Research*, 49(D1) (2021).
8. Wishart, D.S., Feunang, Y.D., Guo, A.C., Lo, E.J., Marcu, A., Grant, J.R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., Assempour, N., Iynkkaran, I., Liu, Y., Maclejewski, A., Gale, N., Wilson, A., Chin, L., Cummings, R., Le, Di., et al.: DrugBank 5.0: A major update to the DrugBank database for 2018, *Nucleic Acids Research*, 46(D1) (2018).
9. Siswanto, S., Syahrir, N.H.A.: Agglomerative Hierarchical Clustering Analysis In Predicting Antibacterial Activity Of Compound Based On Chemical Structure Similarity, *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, 16(4) (2022).
10. Hanif, N., Murni, A., Tanaka, C., Tanaka, J.: Marine natural products from Indonesian waters, *Marine Drugs*, 17(6) (2019).
11. Karim, M.B., Kanaya, S., Altaf-UI-Amin, M.: Implementation of BiClusO and its comparison with other biclustering algorithms, *Applied Network Science*, 4(1) (2019)
12. Karim, M.B., Kanaya, S., Amin, Md.A.-U: Comparison of BiClusO with Five Different Biclustering Algorithms Using Biological and Synthetic Data, *Complex Networks and Their Applications VII*, pp. 575-585. Springer International Publishing, Cham (2019).
13. Karim, M.B., Huang, M., Ono, N., Kanaya, S., Amin, M.A.U.: BiClusO: A Novel Biclustering Approach and Its Application to Species-VOC Relational Data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 17(6) (2020).
14. Karim, M.B., Kanaya, S., Altaf-UI-Amin, M.: DPCLusSBO: An integrated software for clustering of simple and bipartite graphs, *SoftwareX*, 16 (2021).

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

