Available online at www.sciencedirect.com

ScienceDirect

Procedia Chemistry 20 (2016) 40 – 44

Procedia Chemistry

11th Asian Conference on Chemical Sensors, ACCS 2015

Comparison of Clustering Methods in the Context of Chemical Structure Similarity Based Classification of VOCs

Azian Azamimi Abdullah[a, b], Md. Altaf-Ul-Amin[a], Naoaki Ono[a], Nurlisa Yusuf[b], Ammar Zakaria[b], Takaaki Nishioka[a] and Shigehiko Kanaya[a]*

[a]*Graduate School of Information Science, Nara Institute of Science and Technology, 8916-5, Ikoma, 630-0192, Nara, Japan*
[b]*Centre of Excellence for Advanced Sensor Technology, Universiti Malaysia Perlis, Jejawi, 02600 Perlis, Malaysia*

Abstract

This paper compares the performance of two clustering methods; DPClus graph clustering and hierarchical clustering to classify volatile organic compounds (VOCs) using fingerprint-based similarity measure between chemical structures. The clustering results from each method were compared to determine the degree of cluster overlap and how well it classified chemical structures of VOCs into clusters. Additionally, we also point out the advantages and limitations of both clustering methods. In conclusion, chemical similarity measure can be used to predict biological activities of a compound and this can be applied in the medical, pharmaceutical and agrotechnology fields.

Peer-review under responsibility of Universiti Malaysia Perlis

Keywords: Volatile organic compounds, Clustering; Chemical similarity; Chemometrics; Biological activity

1. Introduction

Volatile organic compounds (VOCs) are small molecules with diverse class of chemical compounds and have important role in chemical ecology and healthcare applications [1]-[3]. Mass-spectrometry and chemical-based sensors can be used to perform the detection of VOCs; however there are still limitations on the databases that accumulate information of VOCs until now. To meet this purpose, we have developed a KNApSAcK Metabolite Ecology Database, which contains the information of VOCs emitted

* Corresponding author. Tel.: +81-743-72-5952; fax: +81-743-72-5329.
E-mail address: skanaya@gtc.naist.jp

Peer-review under responsibility of Universiti Malaysia Perlis
doi:10.1016/j.proche.2016.07.006

by various organisms and their corresponding biological activities [4]. This database can be accessed online at http://kanaya.naist.jp/MetaboliteEcology/top.jsp. In order to investigate the relationships between VOCs and biological activities, it is important to determine the chemical structure similarity of VOCs [5]. Here, we calculate the chemical similarity measures between two chemical compounds by using Tanimoto coefficient and apply two clustering methods, namely DPClus algorithm [6] and hierarchical clustering [7] to cluster the chemical structures of VOCs. We compare the outcome of clustering results to find out a suitable algorithm for classification of VOCs.

2. Material and Methods

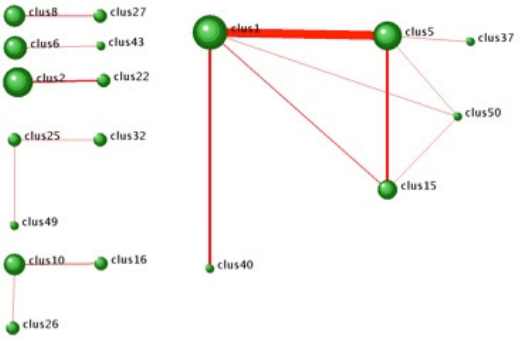
The fingerprint of a chemical compound is a binary vector indicating the substructures it contains. The similarity between two chemical compounds can be calculated as the number of bits that are common to the fingerprints representing those compounds [8], [9]. There are many coefficients that can be used for this purpose and in this paper, we choose Tanimoto coefficient to calculate such similarities [10]. Based on Tanimoto similarity measure, we applied two different clustering methods to classify the VOCs, which are DPClus clustering and hierarchical clustering.

A network of VOCs was constructed by selecting structurally highly similar VOC pairs for applying the DPClus algorithm. In DPClus, a network is considered as an undirected simple graph $G = (N, E)$ that consists of nodes N and edges E. Density d_k of any cluster k is the ratio of the number of edges present in the cluster $(|E|)$ and the maximum possible number of edges in the cluster $(|E|_{max})$. The cluster property of node n with respect to cluster k is represented by $cp_{nk} = E_{nk} / (d_k \times N_k)$, where N_k is the number of nodes in cluster k. E_{nk} is the total number of edges between the node n and each of the nodes of cluster k.

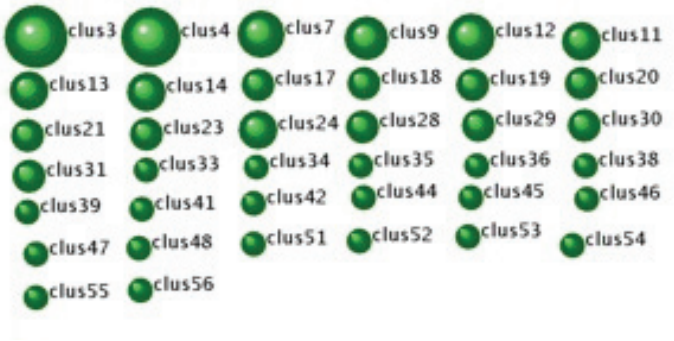
Meanwhile, hierarchical clustering involves creating clusters that have a predetermined ordering from top to bottom, or otherwise. Before any clustering is performed, it is required to determine the proximity matrix containing the distance between each point using a distance function. Then, the matrix is updated to display the distance between each cluster. In this study, we use 6 different methods depending on how the distance between each cluster is measured that is single, complete, average, centroid, median and Ward's method.

3. Results and Discussion

We applied both DPClus and hierarchical clustering to cluster all 341 VOCs that we accumulated in our database. In case of DPClus algorithm, we used 0.6 as input density d_k and 0.5 as input cluster property cp_{nk}. DPClus generated 56 clusters. Figure 1 (a) shows the interacted clusters while Fig. 1 (b) shows the independent clusters of DPClus.



(a)



(b)

Fig. 1. (a) Interacted clusters of DPClus. (b) Independent clusters of DPClus.

To be consistent, we extracted 50 clusters based on hierarchical clustering. The size of the biggest cluster generated by DPClus is 18 while in case of hierarchical clustering it is 98 (centroid's method) as shown in Fig. 2. It is also observed that in hierarchical clustering, there is some imbalance in the size of generated clusters. On the other hand, a few clusters generated by DPClus algorithm are in balanced size.

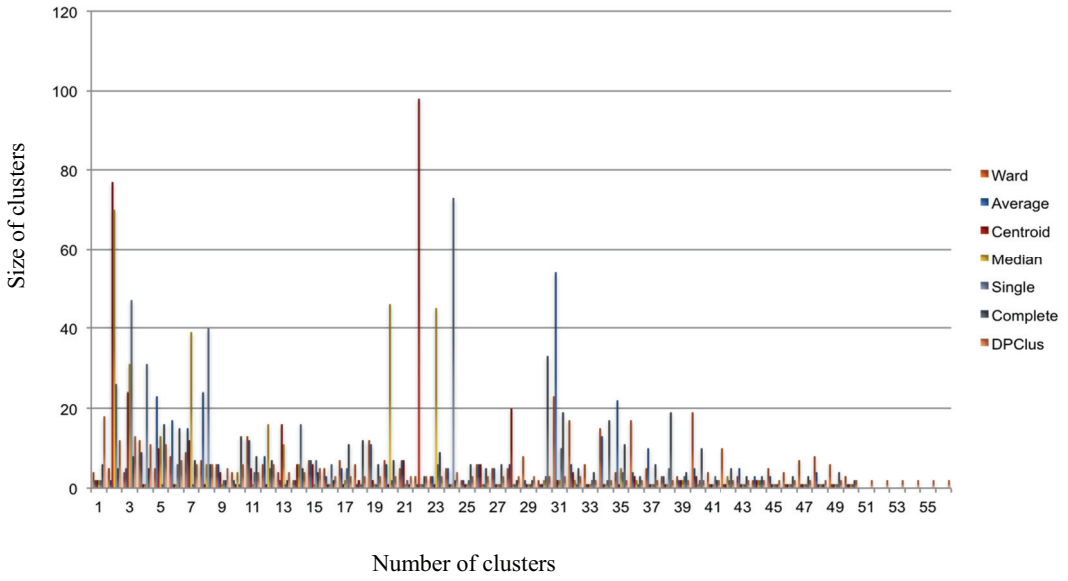


Fig. 2. Distribution of cluster size generated by DPClus and each linkage method of hierarchical clustering.

For comparison purpose, we also investigated how the generated clusters by these two clustering methods match to each other. To calculate how effectively DPClus generated clusters overlaps with hierarchical clusters, we use a matching score measure, *m* as follow:

$$m = \frac{i^2}{a \times b} \quad (1)$$

Here, a is the size of a cluster generated by DPClus, b is the size of a clusters generated by hierarchical method and i is the size of intersection set of a and b. The calculated matching score is ranged between 0 and 1, where value of 1 shows the maximum overlapping score between two generated clusters. The distribution of matching score between DPClus and each method of hierarchical clustering is shown as Fig. 3. From this figure, we can observe that the Ward's method of hierarchical clustering has the most matching clusters with DPClus algorithm while median has the least matching clusters with DPClus. Both DPClus and hierarchical methods generated clusters of VOCs with highly structural similarity and similar biological activity. For example, cluster 1 generated by DPClus algorithm contains 18 VOCs, which are terpenoids and their biological activities are anti-cholinesterase, antimicrobial and defense activities. These results somehow aligned with our previous results, where we have shown that structurally similar group of VOCs generated by hierarchical clustering correspond to similar biological functions by conducting statistical analysis involving hypergeometric distribution based p-values [2]. In the present work, we show that DPClus generated clusters are substantially similar to hierarchical clusters. Therefore, it can be concluded that DPClus generated clusters are also rich with VOCs having similar biological activities.

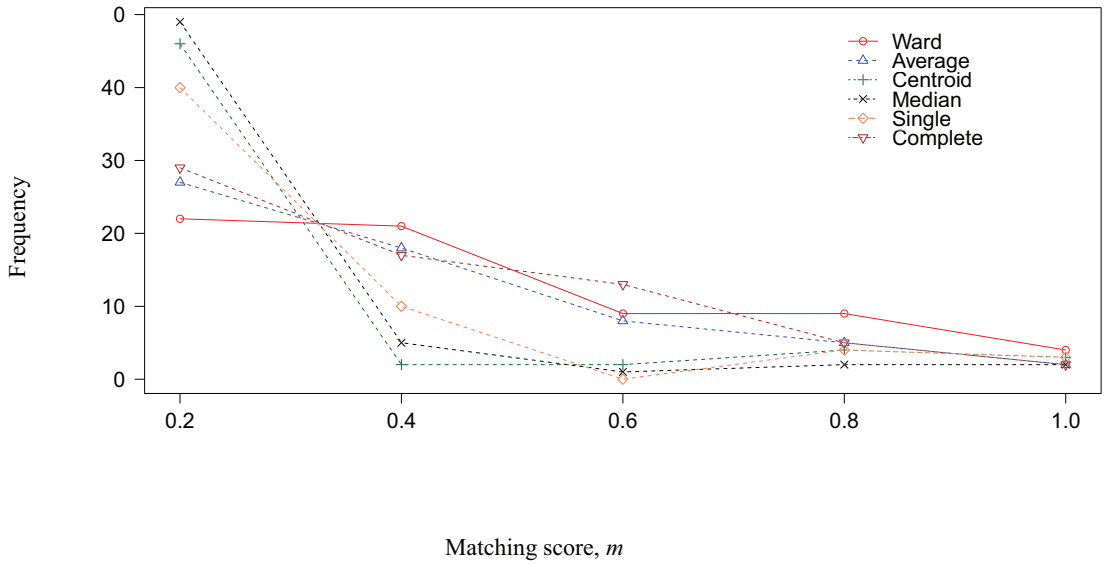


Fig. 3. Distribution of matching score between DPClus and each linkage method of hierarchical clustering.

4. Conclusion

In this paper, we discussed two different clustering methods, namely DPClus graph clustering and hierarchical clustering to cluster the chemical structures of volatile organic compounds (VOCs) using Tanimoto coefficient as chemical similarity measure. Additionally, we compared the performances of DPClus algorithm with 6 different methods of hierarchical clustering, which are single, complete,

average, centroid, median and Ward's method. We found that Ward's method has the most matching clusters with DPClus while median has the least matching clusters. Compared to hierarchical clustering, DPClus can give a better visualization of how generated clusters are interacted with each other and we found that VOCs belonging to the interacted clusters have similar chemical structure, which indicates possibilities of exhibiting similar biological activities.

Acknowledgements

This work is partly supported by the National Bioscience Database Center in Japan and NAIST Big Data Project. The authors would like to thank the Universiti Malaysia Perlis and Ministry of Education Malaysia for funding the postgraduate studies of the first author.

References

1. D. D. Rowan, "Volatile metabolites", Metabolites, 2011, 1, 41-63.
2. Y. Iijima, "Recent advances in the application of metabolomics to studies of biogenic volatile organic compounds (BVOC) produced by plants", Metabolites, 2014, 4, 699-721.
3. C. Lourenco and C. Turner, "Breath analysis in disease diagnosis: Methodological considerations and applications," Metabolites, 2014, 4, 465-498.
4. A. A. Abdullah, M. Altaf-Ul-Amin, N. Ono, et al., "Development and Mining of a Volatile Organic Compound Database," BioMed Research International, Article ID 139254, in press.
5. P. Willett, J. M. Barnard and G. M. Downs, "Chemical similarity searching," J. Chem. Inf. Comput. Sci., 1998, 38, 983-996.
6. M. Altaf-Ul-Amin, Y. Shinbo, K. Mihara et al., "Development and implementation of an algorithm for detection of protein complexes in large interaction networks," BMC Bioinformatics 2006, 7:207.
7. L. Kaufman, P. J. Rousseeuw, "Finding Groups in Data: An Introduction to Cluster Analysis (1 ed.)," New York: John Wiley, 1990, ISBN 0-471-87876-6.
8. R. Todeschini, V. Consonni, H. Xiang, J. Holliday, M. Buscema and P. Willett, "Similarity coefficients for binary chemoinformatics data: Overview and extended comparison using simulated and real data sets," Journal of Chemical Information and Modeling, 2012, 52(11), 2884-2901.
9. S. M. Arif, J. D. Holliday and P. Willett, "Comparison of chemical similarity measures using different numbers of query structures," Journal of Information Science, 2013, 39(1), 7-14.
10. Bajusz et al., "Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?" Journal of Cheminformatics, 2015, 7:20