Classification of Microorganism Species Based on Volatile Metabolite Contents Similarity

Azian Azamimi Abdullah1*, Amira Najihah Abdul Aziz1, Shigehiko Kanaya2 and Satya Ranjan Dash3

1School of Mechatronic Engineering, Universiti Malaysia Perlis, Pauh Putra Campus, 02600 Arau, Perlis, Malaysia
2Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma, 630-0192, Nara, Japan
3School of Computer Applications, KIIT University, India

*azamimi@unimap.edu.my

Abstract. Microorganism species can become pathogenic and cause bacterial infection, result from the imbalance in microbial ecosystem between host and microbe. Microorganisms emit secondary metabolites, known as volatile metabolites or organic compounds (VOCs) for various functions such as intra- or inter-species interactions, defense and attraction. Currently, VOCs are widely used as a biomarker for human diseases. This research is aimed to identify the relationship between microorganism species and volatile metabolite compound from the collected species and VOCs emitted organism data by using unsupervised machine learning approaches such as hierarchical clustering and graph-clustering method. Supervised machine learning methods also been used to classify the microorganism pathogenicity such as support vector machine (SVM) and random forest (RF). These data are collected from KNApSAcK and mVOC database where it provides most of the microorganism species and metabolites contents. From the collected data, there are in total of 1088 VOCs emitted by 517 microorganism species. As a result, the application of machine learning methods enable us to identify the relationship of species with their emitted VOCs and classify the microorganism species into their own pathogenicity.

1. Introduction

Microorganism species can become pathogenic and cause bacterial infection, result from the imbalance in microbial ecosystem between host and microbe [1]. Presence of specific genes that contain pathogenic characteristic may cause even the non-pathogenic species became virulence through gene transfer [2]. A variety of microorganism species can cause disease or infection where there are five main types of pathogenic organisms, which are viruses, bacteria, fungi, protozoa and worm. For a particular organism, metabolites are essential to the metabolism process where it promotes in growth, reproduction and defense against pathogen [3]. Basically, metabolites can be classified into primary metabolite, which involved in growth and development and secondary metabolite, which is important for ecological function such as antimicrobial activities and defense against pathogens.

Volatile organic compounds (VOCs) comprise a chemically diverse class of low molecular weight organic compounds having an appreciable vapor pressure under ambient conditions [4]. Because of this characteristic, it allows them to serve as signaling molecules act as identification in plant, food, predators or habitat [5]. The VOCs produced by plant will attract pollinators to disperse
the seed to pollinate and provide defense mechanism against disease [6]. For humans and animals, volatiles act as scents and contribute to flavorants of food and as disease biomarkers [7], [8].

The recent classification study are based on metabolite contents is regarding the plant classification. The study focused on the global metabolite content of plants which resulting in classification with natural phylogenetic and chemosystematics pattern [9]. Therefore, a new study should be done focusing on the classification of microorganism species based on the volatile metabolite contents to distinguish the pathogenic and non-pathogenic microorganism instead being done by the previous research. Machine learning has become one of the popular approach in solving complex problem and applications such as self-driving cars, fraud detection and medical prediction. Machine learning can be divided into supervised and unsupervised machine learning. Unsupervised machine learning is when the input of similar types is grouped together without the use of training data to specify which group it belongs to. Previously, this technique has been applied to analyze the relationships between 3D structure and biological activities by using network-based approach [10]. Supervised machine learning is task of training using function that requires input data and target representing the desired output.

In this study, the microorganism species will be classified based on the volatile metabolite contents similarity to determine the pathogenicity of the species. Clustering analysis method such as hierarchical clustering and graph clustering is used to determine the relationship between the species and their VOCs. For the classification process, supervised machine learning is used to distinguish the pathogenicity between species based on their VOCs by using support vector machine (SVM) and random forest (RF) as the classifier.

2. Proposed Methods

2.1. Data extraction

The microbial volatile metabolite data were acquired from KNAPSAcK Core DB, which is a part of the KNAPSAcK Family DB Pre-processing and mVOC Database, which is based on extensive literature search for microbial volatile organic compounds (mVOCs) [11]. The KNAPSAcK Core DB provides a comprehensive species-metabolite relationship, which is connected via the scientific name of each species and this database can be accessed online [12]. The mVOC Database is an online search platform database, which contains information about mVOC and their emitting organism [13], [14].

2.2. Clustering of species based on VOCs similarity

2.2.1 Hierarchical Clustering

Hierarchical agglomerative clustering method is implemented, which it start by putting each observation into its own separate cluster where all the distances between all the observation are being examined and pairs together with the closest ones to form a new cluster and this process continues until only single cluster remain [15]. The result of the clustering is represented in the form of dendogram. The following distance matrix is used to determine the Euclidean distances between species. Let this matrix be called M and \( M_{ik} = 1 \) if the species \( i \) is related to the \( k \)th VOC or otherwise \( M_{ik} = 0 \) [7]. Based on the Euclidean distance, the R, an open source programming language is used to analyse the clustering. In this method, four different ways is performed to determine the distance between two clusters. Complete linkage clustering is used to find the possible maximum distance between two points, which belong to two different clusters. The average and all possible pairwise distances for points belonging to two different clusters, mean linkage clustering is applied. The centroid linkage clustering is used to find the centroid of each cluster and the distances between centroids of two clusters. Ward linkage clustering is used to minimize the total variance within the cluster where the cluster with minimum distance between pair of cluster is merged.

\[
d_{ij} = \sqrt{\sum_{k=1}^{n} (M_{ik} - M_{jk})^2}
\]
2.2.2 Graph Clustering Based on DPClus

The DPClus is a graph clustering software, which has been developed based on a clustering algorithm that mainly focused on detection of protein complex in interaction networks [16]. This graph clustering software can extract high-density connected nodes as a cluster [7]. The input to the algorithm is the associated matrix of an interaction network and the outputs are protein complexes. The complexes are determined by way of finding clusters [16]. The DPClus presents an algorithm for detection of large interaction network of protein complex. In this network, a node represents a protein while the edge represents an interaction [17]. There are three important inputs in the DPClus algorithm, which are the network, a value of minimum density (d_k) and a minimum value for cluster property (cpnk). The value for density, d_k and cluster property, cpnk should be within the range of: 0<d_k≤1 and 0<cpnk≤1.

2.3. Classification of species based on VOCs Similarity

Supervised machine learning approach is the task of training using function that requires input data and target representing the desired output. For this classification process, microorganism species and their VOCs will represent the input data while the pathogenicity status will represent desired output. During classification process, supervised machine learning is used to separate the pathogen based on binary VOC input data. To ensure the accuracy of the classifier, two machine-learning methods were applied, which are random forest and support vector machine. The RF and SVM algorithm was used to generate classifiers and estimate the classification performance. For a given pathogen class, the classifiers’ sensitivity, precision and mean accuracy towards that class are given by (2), (3) and (4), respectively:

\[
\text{Sensitivity} = \frac{\text{number of correct predictions in that class}}{\text{number of experiments in that class}} \times 100\% \quad (2)
\]
\[
\text{Precision} = \frac{\text{number of correct predictions in that class}}{\text{number of predictions as that class}} \times 100\% \quad (3)
\]
\[
\text{Mean accuracy} = 1 - \frac{\text{number of incorrectly predicted pathogens}}{\text{total number of experiments}} \quad (4)
\]

2.3.1 Support vector machine (SVM)

SVM is one of the supervised machine learning methods where the purpose is for classification and regression, which are based on the concept of decision planes or hyperplane that define decision boundaries [18]. Separating line defines boundary in decision plane where it is the one that separates between a set of object having a different class of membership.

2.3.2 Random forest (RF)

Random forest is supervised machine learning method where every observation will be fed into a large number of decision tree. This method operate by constructing a multiple decision tree at training time and put into class or mean prediction of the individual trees. Random subspace method is the algorithm used for this machine learning. In RF classifier, the higher the number of trees in the forest gives the high accuracy results [19]. Variable importance can be used to rank the random forest ranking the importance variables in a regression or classification problem.

3. Results and Discussion

3.1. Clustering of species based on VOCs similarity

From the data extracted, 1088 VOCs emitted by 517 microorganisms are being analyzed to confirm the relationship between them. Two unsupervised learning methods were implemented to determine the relation, which are hierarchical clustering and graph clustering.

Hierarchical clustering is used to analyze the relationship between the microorganism species with their VOCs. There are four different linkage methods in hierarchical clustering were used, which are Ward linkage, Complete linkage, Centroid linkage and Mean linkage clustering to analyze and observe
which one out of the four methods will group the pathogenicity of microorganism based on their VOCs correctly. From the result, Ward Linkage Clustering are more preferable to be used in this study to obtain the relation between species and their VOCs. This is because Ward clustering grouping the microorganism by minimizing the variance between cluster and merged the cluster together if they have minimum distance, where in this case the similarity between VOCs causing the distance to be minimized and the microorganism being grouped according to their VOCs similarity. The Euclidean distance is used to calculate the distance between species based on VOCs where the information of microorganism species and its emitted VOCs has been converted into $517 \times 1088$ binary matrix. Here, the dendrogram is cut into 50 clusters to represent each microorganism into their own cluster.

To obtain more information regarding the relation of the microorganism with the mVOCs, high-density cluster is applied where the graph-clustering algorithm DPClus is used. Figure 1 shows the graph clustering result where the green node represents the cluster of microorganism while the red edges represent the interaction between clusters. DPClus clustering is set to the following setting: density value, $d_{in}$ was set to 0.6, cluster property value, $c_{pin}$ was set to 0.5, the minimum cluster size was set to 2, and the mode is set to non-overlapping. From Figure 1, the size of a green node is proportional to the number of microorganism species in the cluster while the thickness of red edges are proportional to the number of similarity-based relation.

![Figure 1](image_url)

**Figure 1.** Graph clustering based on DPClus (a) Connected node (b) Independent node.

From the clustering of DPClus, there are 50 clusters are generated where 20 cluster are connected nodes to each other (Figure 2(a)) while the other 30 clusters are independent clusters (Figure 2(b)). From this result we can conclude that the pathogenicity of microorganism can be connected by the similarity of VOCs emitted. Pathogenic microorganism such as Escherichia coli, Klebsiella pneumonia and Staphylococcus aureus that comes from cluster 1 are connected to the other pathogenic cluster for example cluster 2. The results of this graph-clustering and hierarchical clustering are similar because both results show that the VOCs based classification of microorganism is consistent with their classification based on pathogenicity. However, through DPClus we can see the relation between each microorganism to VOCs emitted by their species since DPClus can show the interaction being generated between clusters.

### 3.2. Classification of species using supervised machine learning methods

The classification was done by using supervised machine learning approach by separating the pathogens based on binary VOC data extracted. The binary VOC data set represents the input while pathogenicity status represents the desired output. The extracted data contains the 1088 VOCs data and their predicted pathogenicity for 517 microorganisms acts as target. The data are being split into testing and training with 30% and 70% respectively.

From the result obtained, SVM provide highest accuracy with 99% and 100% and 95% for sensitivity and precision respectively. From Table 1, it is predicted that 131 species is tested non-
pathogenic while 22 species are correctly predicted as pathogenic. However, one species is falsely predicted as pathogenic. This incorrect classification may result from the presence of volatile compound. In training confusion matrix, there are 3 species was incorrectly classified as pathogenic while another 308 species are non-pathogenic. However, 53 species are correctly classified as pathogenic species.

**Table 1. Confusion Matrix for SVM**

| Prediction       | Reference       | Testing | Training |
|------------------|-----------------|---------|----------|
|                  | Non-pathogenic |         |          |
| Non-pathogenic   | 131             | 308     | 2        |
| Pathogenic       | 0               | 22      | 53       |

For the comparison between classifier, RF was used beside SVM to choose the most reliable classifier among those two. The number of trees (ntree) selected for the random forest algorithm were set to 100. The accuracy obtained by using RF is high with 95%. It also has 100% sensitivity but only 70% specificity. From both testing and training, we can confirm that 65 microorganisms are pathogenic while 13 species which being incorrectly predicted as non-pathogenic but turn up being pathogenic.

**Table 2. Confusion Matrix for RF**

| Prediction       | Reference       | Testing | Training |
|------------------|-----------------|---------|----------|
|                  | Non-pathogenic |         |          |
| Non-pathogenic   | 131             | 308     | 6        |
| Pathogenic       | 0               | 16      | 49       |

Figure 2 shows the ROC curve for RF classifier. The rate of true positive is high similar with the rate of false positive. From both classifier of supervised machine learning being tested, SVM shows the highest accuracy with 99% compared to RF (95%). This result confirmed that SVM are more reliable compared to RF even though the accuracy difference is small.

**Figure 2.** ROC curve for the rate accuracy between species and VOCs.
4. Conclusion
In worldwide healthcare, the development of technologies for medical purposes is growing in fast rate to prevent infectious diseases that endanger human population. A non-invasive monitoring tool is preferable for the detection of volatile metabolite. Through this study, VOCs can act as a potential biomarker to detect the presence of volatile microorganisms. From the study involving 1088 VOCs emitted by 517 microorganisms proved that the classification of microorganism species can be done based on volatile metabolite contents similarity. Based on DPClus, we can see the relation between each microorganism to VOCs emitted by their species since DPClus can show the interaction being generated between clusters. For the supervised learning method, SVM shows the highest accuracy with 99% compared to RF for the classification of microorganisms species based on VOC similarity.

References
[1] G. D. Ehrlich, N. L. Hiller, and F. Hu, “What makes pathogens pathogenic,” Genome Biol., vol. 9, no. 6, p. 225, 2008.
[2] K. A. Bliven and A. T. Maurelli, “Antivirulence genes: Insights into pathogen evolution through gene loss,” Infect. Immun., vol. 80, no. 12, pp. 4061–4070, 2012.
[3] Q. Wu, M. Chen, H. Zhou, X. Zhou, and Y. Wang, “Metabolite profiles of Populus in response to pathogen stress,” Biochem. Biophys. Res. Commun., vol. 465, no. 3, pp. 421–426, 2015.
[4] D. D. Rowan, “Volatile Metabolites,” Metabolites, vol. 1, no. 1, pp. 41–63, 2011.
[5] E. Pichersky, J. P. Noel, N. Dudareva, and A. Arbor, “Biosynthesis of Plant Volatiles: Nature ’ s Diversity and Ingenuity PV Biosynthetic Pathways Branch Off from Primary Metabolism,” vol. 311, no. 5762, pp. 808–811, 2006.
[6] M. Heil and J. Ton, “Long-distance signalling in plant defence,” Trends Plant Sci., vol. 13, no. 6, pp. 264–272, 2008.
[7] N. Yusuf, A. Zakaria, M. I. Omar, A. Y. M. Shakaff, M. J. Masnan, L. M. Kamarudin, N. Abdul Rahim, N. Z. I. Zakaria, A. A. Abdullah, A. Othman, and M. S. Yasin, “In-vitro diagnosis of single and poly microbial species targeted for diabetic foot infection using e-nose technology,” BMC Bioinformatics, vol. 16, p. 158, 2015.
[8] Yusuf, N., M. I. Omar, A. Zakaria, A. J. Jeffree, R. Thiumani, A. A. Abdullah, A. Y. M. Shakaff et al. "Evaluation of E-nose technology for detection of the causative bacteria in different culture media on diabetic foot infection." In 2014 IEEE Conference on Biomedical Engineering and Sciences (ICBES), pp. 67-70. IEEE, 2014.
[9] K. Liu, A. A. Abdullah, M. Huang, and T. Nishioka, “Novel Approach to Classify Plants Based on Metabolite Content Similarity,” Biomed Res. Int., vol. 2017.
[10] Ohtana, Y., Abdullah, A.A., Altaf-Ul-Amin, M., Huang, M., Ono, N., Sato, T., Sugiura, T., Horai, H., Nakamura, Y., Morita, A. and Lange, K.W., 2014. Clustering of 3D-Structure Similarity Based Network of Secondary Metabolites Reveals Their Relationships with Biological Activities. Molecular informatics, 33(11-12), pp.790-801.
[11] A. A. Abdullah et al., “Development and Mining of a Volatile Organic Compound Database,” Biomed Res. Int., vol. 2015, 2015.
[12] F. M. Afendi et al., “KNApSAcK family databases: Integrated metabolite-plant species databases for multifaceted plant research,” Plant Cell Physiol., vol. 53, no. 2, pp. 1–12, 2012.
[13] M. C. Lemfack, J. Nickel, M. Dunkel, R. Preissner, and B. Piechulla, “MVOC: A database of microbial volatiles,” Nucleic Acids Res., vol. 42, no. D1, pp. 1–90, 2014.
[14] M. C. Lemfack, B. Gohlke, S. M. T. Toguem, S. Preissner, B. Piechulla, and R. Preissner, “mVOC 2.0: a database of microbial volatiles,” no. November, pp. 1-5, 2017.
[15] A. A. Abdullah, T. Nishioka, T. Katsuragi, and N. Ono, “Analysis of Volatile Metabolites Emitted by Various Species to Reveal Their Roles in Chemical Ecology and Healthcare,” no. March, pp. 30–31, 2015.
[16] M. Altaf-Ul-Amin, H. Tsuji, K. Kurokawa, H. Asahi, Y. Shinbo, and S. Kanaya, “DPClus : A density-periphery based graph clustering software mainly focused on detection of protein complexes in interaction networks,” J. Comput. Aided Chem., vol. 7, pp. 150–156, 2006.
[17] M. Altaf-Ul-Amin, Y. Shinbo, K. Mihara, K. Kurokawa, and S. Kanaya, “Development and implementation of an algorithm for detection of protein complexes in large interaction networks,” BMC Bioinformatics, vol. 7, p. 207, 2006.

[18] D. K. Srivastava and L. Bhambhu, “Data classification using support vector machine,” J. Theor. Appl. Inf. Technol., vol. 12, pp. 1–7, 2009.

[19] B. López, F. Torrent-Fontbona, R. Viñas, and J. M. Fernández-Real, “Single Nucleotide Polymorphism relevance learning with Random Forests for Type 2 diabetes risk prediction,” Artif. Intell. Med., pp. 3–9, 2017.