Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 139254, 13 pages
http://dx.doi.org/10.1155/2015/139254

Research Article

Development and Mining of a Volatile Organic Compound Database

Azian Azamimi Abdullah,[1,2] Md. Altaf-Ul-Amin,[1] Naoaki Ono,[1] Tetsuo Sato,[1] Tadao Sugiura,[1] Aki Hirai Morita,[1] Tetsuo Katsuragi,[3] Ai Muto,[4] Takaaki Nishioka,[1] and Shigehiko Kanaya[1]

[1]*Graduate School of Information Science, Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma, Nara 630-0192, Japan*
[2]*School of Mechatronic Engineering, Universiti Malaysia Perlis (UniMAP), Ulu Pauh, 02600 Arau, Perlis, Malaysia*
[3]*Department of Computer Science and Engineering, Toyohashi University of Technology, 1-1 Hibarigaoka, Tempaku, Toyohashi, Aichi 441-8580, Japan*
[4]*Graduate School of Biological Science, Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma, Nara 630-0192, Japan*

Correspondence should be addressed to Md. Altaf-Ul-Amin; amin-m@is.naist.jp and Shigehiko Kanaya; skanaya@gtc.naist.jp

Received 14 April 2015; Accepted 14 June 2015

Academic Editor: Zhirong Sun

Volatile organic compounds (VOCs) are small molecules that exhibit high vapor pressure under ambient conditions and have low boiling points. Although VOCs contribute only a small proportion of the total metabolites produced by living organisms, they play an important role in chemical ecology specifically in the biological interactions between organisms and ecosystems. VOCs are also important in the health care field as they are presently used as a biomarker to detect various human diseases. Information on VOCs is scattered in the literature until now; however, there is still no available database describing VOCs and their biological activities. To attain this purpose, we have developed KNApSAcK Metabolite Ecology Database, which contains the information on the relationships between VOCs and their emitting organisms. The KNApSAcK Metabolite Ecology is also linked with the KNApSAcK Core and KNApSAcK Metabolite Activity Database to provide further information on the metabolites and their biological activities. The VOC database can be accessed online.

1. Introduction

Recently big data has become an important topic that has significant roles to play in versatile disciplines of scientific research. Big data biology is a data-intensive science which has emerged because of the rapidly increasing volume of molecular biological data in omics fields such as genomics, transcriptomics, proteomics, and metabolomics [1–3]. With the explosively growing data scale, the development of biological databases incorporating different species has become a very important theme in big data biology. To address this need, we have developed KNApSAcK Family Databases (DBs), which have been utilized in a number of studies in metabolomics. The KNApSAcK Family Database systems previously have been used to understand the medicinal usage of plants based on traditional and modern knowledge [4, 5]. To facilitate a comprehensive understanding of the interactions between the metabolites of organisms and the chemical-level contribution of metabolites to human health, a Metabolite Activity DB known as the KNApSAcK Metabolite Activity DB has been constructed [6] and a network-based approach has been proposed to analyze the relationships between 3D structure and biological activities of the metabolites [7].

Metabolomics is the scientific study of quantification of profiles and analysis of chemical processes involving metabolites in a comprehensive fashion. In general, metabolites can be divided into two groups: primary and secondary metabolites. Primary metabolites are directly involved in the normal growth, development, and reproduction. On the other hand, secondary metabolites are not directly involved in these

processes but usually have important ecological function, such as inter- or intraspecies communication, antifungal and antimicrobial activities, and also defense against pests and pathogens. Large portions of these defense compounds are volatile organic compounds (VOCs) that are involved in different ways of defense: direct defense and indirect defense. VOCs constitute only a small proportion of the total number of secondary metabolites produced by living organisms; however because of their important roles in chemical ecology specifically in the biological interactions between organisms and ecosystems, revealing and analyzing the roles of these VOCs is essential for understanding the interdependence of organisms [8].

VOCs originate from major pathways of secondary metabolisms of many living organisms, including human, animals, microorganisms, and plants. In plant kingdom, VOCs are responsible for internal and external communication between plants and herbivores, pathogens, pollinators, and parasitoids such as defense and attractant [9]. Microbial volatiles are widely used as biomarkers to detect human diseases [10]. This is because bacteria have a recognizable metabolism that produces bacteria-specific VOCs, which might be used for noninvasive diagnostic purposes [11]. For example, a volatile organic compound called methyl nicotine produced by *Mycobacterium tuberculosis* bacteria can be used as a noninvasive and rapid diagnostic marker for detection of Tuberculosis (TB) diseases [12]. Human also produces VOCs. Hundreds of volatiles are emitted from the human body in breath, blood, skin, and urine. These compounds reflect the different metabolic conditions of an individual [13]. Therefore, differences between the volatile profiles of individual humans can be used as an indicator to evaluate and monitor "disease" or "health" status. A review of breath analysis in disease diagnosis using volatile profiles is presented by Lourenço and Turner [14]. Breath analysis can be used as a biomarker to identify patients related to breast cancer [15], colorectal cancer [16], pulmonary tuberculosis [17], and lung cancer [18].

Advancements in analytical methods such as gas chromatography-mass spectrometry (GCMS), proton transfer reaction mass spectrometry (PTR-MS), and selected ion flow tube mass spectrometry (SIFT-MS) have provided an opportunity to identify the volatile metabolites of living organisms in research laboratories. These analytical approaches generate a large amount of data and require specialized mathematical, statistical, and bioinformatics tools to analyze such data. Despite the advances in sampling and detection by these analytical methods, only few databases have been developed to handle these large and complex datasets. There are few volatile organic compound databases which can be accessed freely; however their applicability is often limited by several elements. Most of these databases only focus on volatiles which are emitted by certain living organisms and have limited applications. For example, the Superscent database [19] only provides structure information of flavors and scents, and the mVOC database [20] provides information of microbial volatiles only. Flavornet [21] features compounds identified in experiments employing gas chromatography olfactometry (GC-O) analysis, and Pherobase [22] is focused on insect pheromones and semiochemicals. The vocBinBase [23] is a mass spectral database for volatiles which can allow for tracking and identification of volatile compounds in complex mixtures. None of these databases provide information on biological activities of VOCs and species-species interaction based on volatiles. Information on volatiles emission from microorganisms, plants, and other organisms is scattered in the literature until now, but there is no public and up-to-date database that accumulated comprehensive information of volatiles and their biological activities.

In the present study, we have developed a VOC database of microorganisms, fungi, and plants as well as human being, which comprises the relation between emitting species, the volatiles, and their biological activities. We have deposited the VOC data into KNApSAcK Metabolite Ecology Database and this database is currently available at http://kanaya.naist.jp/MetaboliteEcology/top.jsp. Apart from the database development, we also analyzed the VOC data using hierarchical clustering and network clustering based on DPClus [24, 25]. In addition, we also performed the heatmap clustering based on Tanimoto coefficient as the similarity index of the chemical structure to cluster all VOCs emitted by various biological species to understand the relationships between chemical structures of VOCs and their biological activities.

2. Methods

2.1. Data Collection and Database Development. The data were collected by an extensive literature search up to January 2015 on PubMed (http://www.ncbi.nlm.nih.gov/pubmed) and Google Scholar (http://scholar.google.co.jp/). The PubMed search provided 60 articles based on the keywords "volatile organic compounds" and "metabolites." The information on VOCs, emitting species, target species, and their biological activities was extracted and deposited into KNApSAcK Metabolite Ecology Database. The KNApSAcK Metabolite Ecology is also linked to the KNApSAcK Core and KNApSAcK Metabolite Activity Database to provide further information on the metabolites and their biological activities. Data were divided into two types: (1) microorganisms species-VOC binary relations; (2) emitting species-VOC-target species triplet relations.

2.2. Clustering of Species Based on VOC Similarity. Clustering is an unsupervised learning method, which is the task of grouping a set of objects into the same group (cluster) based on similarity or distance measures. This technique is important for knowledge discovery and has been applied in many applications such as machine learning, pattern recognition, image analysis, and bioinformatics [26–28]. In this study, we utilized hierarchical clustering and graph clustering methods for classifying the VOC emitting species. Both methods are discussed separately in the following.

2.2.1. Hierarchical Clustering. We used hierarchical agglomerative clustering method, which starts out by putting each observation into its own separate cluster. It then examines all the distances between all the observations and pairs together the two closest ones to form a new cluster. The process

continues until all the observations are included in a single cluster. The result of clustering is usually represented by a dendrogram. In our case, we used a species versus VOC matrix. Let this matrix be called M and $M_{ik} = 1$ if the species i is related to the kth VOC or otherwise $M_{ik} = 0$. Hierarchical methods require a distance matrix and hence we determined the Euclidean distances between species. Euclidean distance d between species i and species j can be calculated as follows:

$$d(i, j) = \sqrt{\sum_{k=1}^{n} \left(M_{ik} - M_{jk}\right)^2}. \quad (1)$$

Here, n is the number of VOCs and there are 1088 VOCs in our data. Based on Euclidean distance, we perform Ward's hierarchical clustering analysis using R, an open source programming language.

2.2.2. Graph Clustering Based on DPClus. DPClus is a graph clustering software [24], which has been developed based on a graph clustering algorithm that can extract densely connected nodes as a cluster [25]. This algorithm can be applied to an undirected simple graph $G = (N, E)$ that consists of a finite set of nodes N and a finite set of edges E. Two important parameters are used in this algorithm, which are density d_k and cluster property cp_{nk}. 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

$$cp_{nk} = \frac{E_{nk}}{d_k \times N_k} \quad (2)$$

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. In this study, we apply the DPClus algorithm to identify certain groups of microorganism species based on VOC similarity. A network is constructed where a node represents a microorganism species and an edge indicates high VOC similarity between the corresponding species pair. We selected 5% of the organism pairs based on lower Euclidean distance between them. We used the nonoverlapping mode with the following DPClus settings: Cluster property cp_{nk} was set to 0.5, density value d_k was set to 0.6, and minimum cluster size was set to 2.

2.3. Clustering of VOCs Based on Chemical Structure Similarity. We also performed a classification of VOCs based on their chemical structure similarity. In order to determine the similarity between two chemical compounds, we used Tanimoto coefficient as similarity measure. The Tanimoto coefficient is defined as (3), which is the proportion of the features shared between two compounds divided by their union [29]

$$\text{Tanimoto}_{A,B} = \frac{AB}{A + B - AB}. \quad (3)$$

The variable AB is the number of features (or on-bits in binary fingerprint) common in both compounds, while A and B are the number of features that are related to individual compounds, respectively. The Tanimoto coefficient has a range from 0 to 1 with higher values indicating greater similarity than lower ones. Additionally, a Tanimoto coefficient value larger than 0.85 indicates that the compared compounds may have similar biological activity [30]. For the purpose of calculating Tanimoto coefficient, it is obligatory to assign fingerprints to the compounds. ChemMine package in R was used to generate atom pair fingerprints and calculation of Tanimoto coefficient [31, 32]. 2D compound structures in the generic structure definition file (SDF) format were obtained from PubChem database (https://pubchem.ncbi.nlm.nih.gov) and then were imported into ChemmineR package in one batch file. The atom pair descriptors are calculated during the SDF import and stored in a searchable descriptor database as a list object.

Based on Tanimoto similarity measure between chemical structures, heatmap clustering was performed for classifying the VOCs. We also determined the p values of the clusters based on hypergeometric distribution using

$$p \text{ value} = 1 - \sum_{i=0}^{K-1} \frac{\binom{V}{i}\binom{N-V}{C-i}}{\binom{N}{C}}. \quad (4)$$

Here N is the total number of VOCs, C is the size of a cluster, and V and K, respectively, are the number of VOCs of a certain category in the whole data and in the cluster. The hypergeometric distribution is used to calculate the statistical significance of having drawn specific K successes (out of N total draws) from the whole population. The test is often used to identify which subpopulations are over- or underrepresented in a sample. The calculated p value implies the probability of getting K or more VOCs of a particular category in a cluster when the cluster is formed by random selection. Lower p value indicates that the statistical significance is high.

Our purpose is to relate a structure group to a biological activity if and only if the structure group is overrepresented by VOCs associated with that biological activity.

3. Results and Discussion

3.1. KNApSAcK VOC Database. At present, we have accumulated 1088 VOCs emitted by 517 microorganisms species and 341 VOCs emitted by other biological species including plants, animals, and human. These VOC data have been deposited into KNApSAcK Metabolite Ecology Database, which allows users to search information on VOCs using the KNApSAcK compound ID and metabolite name. This KNApSAcK Metabolite Ecology Database is also linked to the KNApSAcK Core and KNApSAcK Metabolite Activity Database to provide further information on the volatile metabolites and their biological activities. The VOC database can be accessed online at http://kanaya.naist.jp/MetaboliteEcology/top.jsp. Figure 1 shows the main window of the KNApSAcK Metabolite Ecology Database, which shows the search type and search condition. For search type, users can choose either partial or exact string matching searches by clicking the corresponding button, that is,

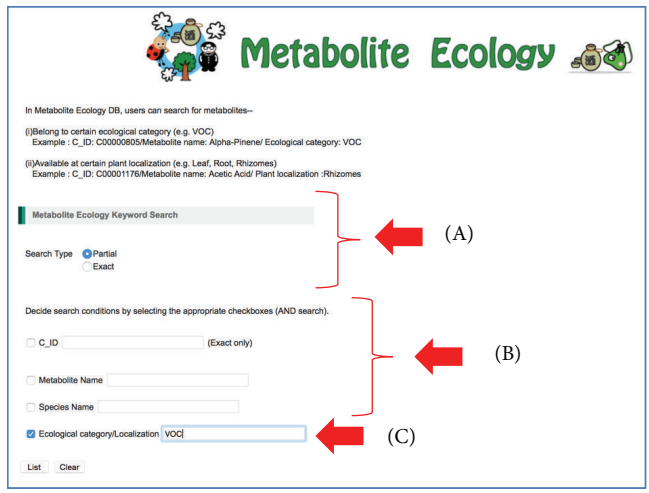


Figure 1: The main window of the KNApSAcK Metabolite Ecology Database. (A) Section used to select the search type. (B) Section used to select the search conditions and to input keywords. (C) Users can input "VOC" in the text box for the ecological/localization category to search VOC data.

Input word = [match type: partial, ecological category/localization: VOC]

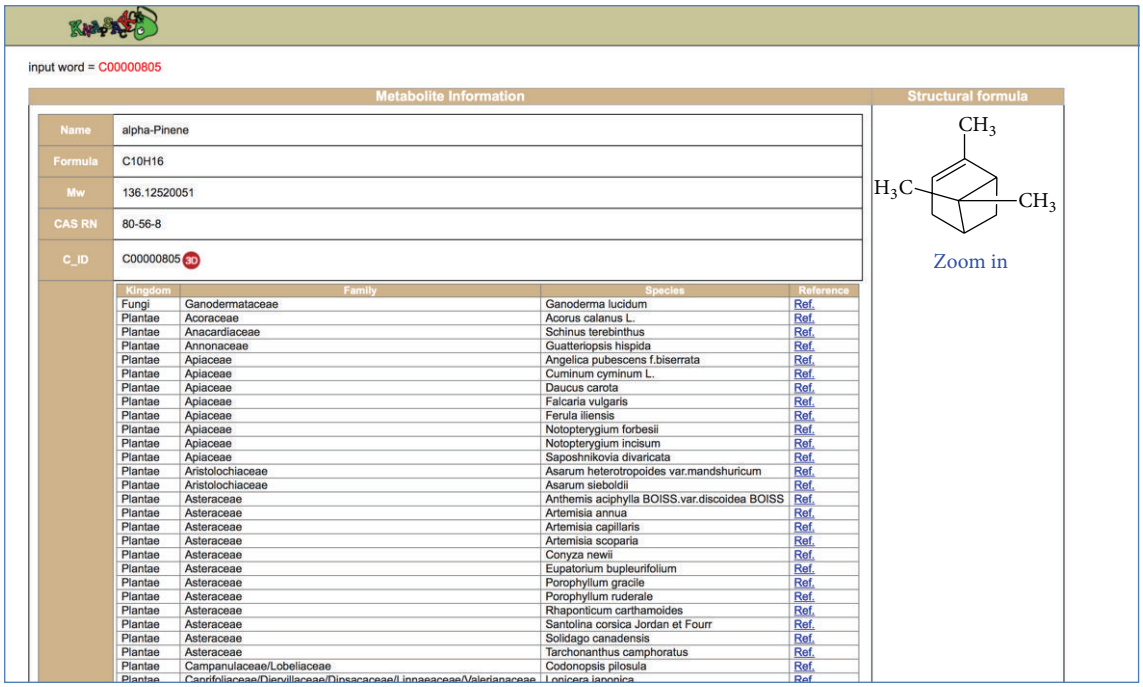
C_ID	Metabolite Name	Species Name	Ecological category/ Localization	Reference
-	(+)-2-Carene	Solanum lycopersicum	VOC	J Chem Ecol (2012) 38:1376_1386
-	(+)-4-Carene	Human (Urine)	VOC	Talanta 89 (2012) 360_368
-	(+)-Limonene	Solanum lycopersicum	VOC	J Chem Ecol (2012) 38:1376_1386
-	(+)-Sativene	Cladosporium cladosporioides	VOC	Sensors 2013, 13, 13969-13981
-	(+)-Sativene	Cochliobolus sativus	VOC	Fiers M, Lognay G, Fauconnier M-L, Jijakli MH (2013) PLoS ONE 8(6): e66805
-	(-)-Caryophyllene oxide	Solanum lycopersicum	VOC	J Chem Ecol (2012) 38:1376_1386
-	(2E,6E)-2,6-Nonadienoic acid	Human (Skin)	VOC	Wound Rep Reg (2010) 18 391_400
-	(E)-2-Hexenal	Solanum lycopersicum	VOC	Zebelo et al. BMC Plant Biology 2014,14:140
-	(E)-alpha-Bergamotene	Polygonum minus	VOC	Molecules 2014, 19, 19220-19242
-	(Z)-Myrtanol	Polygonum minus	VOC	Molecules 2014, 19, 19220-19242
-	-(Z)-Myrtanol	Polygonum minus	VOC	Molecules 2014, 19, 19220-19242
-	1,2,3,4,4a,5,6,8aoctahydro-4a,8-dimethyl-2-(1-methylethenyl)-Naphthalene	Polygonum minus	VOC	Molecules 2014, 19, 19220-19242
-	1,2,4-Trimethylbenzene	Human (Urine)	VOC	Talanta 89 (2012) 360_368
-	1,2-Benzenedicarboxylic acid	Polygonum minus	VOC	Molecules 2014, 19, 19220-19242
C00000805	alpha-pinene A	Cladosporium cladosporioides	VOC	Sensors 2013, 13, 13969-13977
C00000806	(-)-beta-Pinene	Solanum lycopersicum	VOC	Zebelo et al. BMC Plant Biology 2014,14:140
C00000816	beta-Pinene A	Proteaceae	VOC	J Chem Ecol (2013) 39:438_446
C00000823	Limonene	Polygonum minus	VOC	Molecules 2014, 19, 19220-19242
C00000823	Limonene	Proteaceae	VOC	J Chem Ecol (2013) 39:438_446
C00000853	Myrcene A	Proteaceae	VOC	J Chem Ecol (2013) 39:438_446
C00000861	Terpinolene A	Solanum lycopersicum	VOC	J Chem Ecol (2012) 38:1376_1386
C00001176	Acetic acid A	Burkholderia tropica	VOC	Bioengineered 4:4, 236_243 July/August 2013
C00001189	3-Methylbutanoic acid A	Staphylococcus aureus	VOC	PLOS Pathogens, May 2013, Volume 9, Issue 5

(A)

(B)

Figure 2: The results retrieved for VOCs search in the KNApSAcK Metabolite Ecology Database. (A) The VOC, which does not have a KNApSAcK compound ID. (B) The VOC, which has a KNApSAcK compound ID. Users can click at the C_ID to find more information on the metabolite and at the button "A" to retrieve its biological activity.

partial or exact (Figure 1(A)). Other check boxes can also be selected to specify different search conditions (Figure 1(B)) such as KNApSAcK compound ID (C_ID), metabolite name, species name, and ecological category or localization. To search VOC data, users can input "VOC" in the text box for the ecological category/localization category, select the corresponding check box, and then click the List button (Figure 1(C)). Part of the results retrieved by entering "VOC" in the text box is shown in Figure 2. The attributes in the list are C_ID, which corresponds to the KNApSAcK compound ID, metabolite name, species name (VOCs emitting species), ecological category/localization (VOC), and references (the source of the VOCs information), from left to right. During the literature search, it turned out that many VOCs do not have a KNApSAcK compound ID but might be biologically relevant. Therefore, those VOCs were also included into the database. For example, in the first line, the VOC with the name (+)-2-Carene (Figure 2(A)) does not have a KNApSAcK compound ID; however it was produced by *Solanum lycopersicum*. In the future, we will find more information on these VOCs and assign the KNApSAcK compound ID to these metabolites. On the other hand, information related to the VOCs that have KNApSAcK compound ID can be obtained by clicking the C_ID as in Figure 2(B). Figure 3 shows the search results obtained by clicking the C_ID, C00000805, which were retrieved from the KNApSAcK Core Database. Users can retrieve further knowledge of this metabolite, such as molecular formula, molecular weight, CAS RN, 3D structure, and other species information, which also produce the corresponding metabolite. To understand the relationships between VOCs and their biological activities, we also integrate


input word = C00000805

Metabolite Information

Name	alpha-Pinene
Formula	C10H16
Mw	136.12520051
CAS RN	80-56-8
C_ID	C00000805

Structural formula

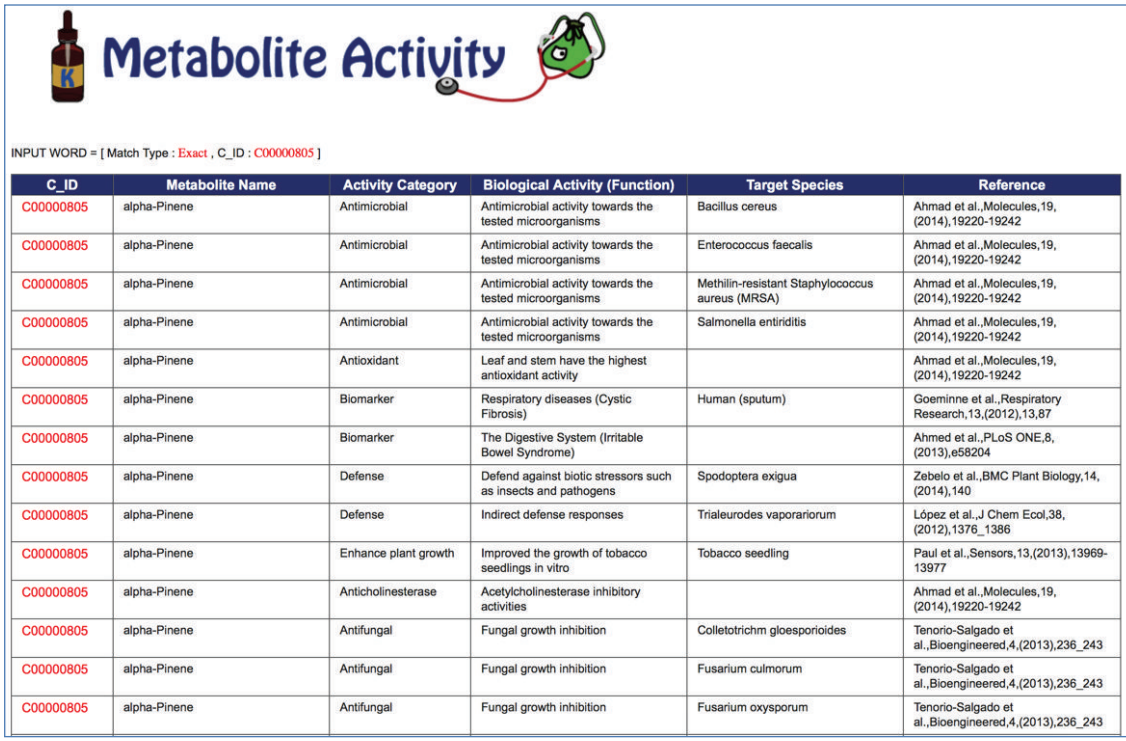
Kingdom	Family	Species	Reference
Fungi	Ganodermataceae	Ganoderma lucidum	Ref.
Plantae	Acoraceae	Acorus calanus L.	Ref.
Plantae	Anacardiaceae	Schinus terebinthus	Ref.
Plantae	Annonaceae	Guatteriopsis hispida	Ref.
Plantae	Apiaceae	Angelica pubescens f.biserrata	Ref.
Plantae	Apiaceae	Cuminum cyminum L.	Ref.
Plantae	Apiaceae	Daucus carota	Ref.
Plantae	Apiaceae	Falcaria vulgaris	Ref.
Plantae	Apiaceae	Ferula iliensis	Ref.
Plantae	Apiaceae	Notopterygium forbesii	Ref.
Plantae	Apiaceae	Notopterygium incisum	Ref.
Plantae	Apiaceae	Saposhnikovia divaricata	Ref.
Plantae	Aristolochiaceae	Asarum heterotropoides var.mandshuricum	Ref.
Plantae	Aristolochiaceae	Asarum sieboldii	Ref.
Plantae	Asteraceae	Anthemis aciphylla BOISS.var.discoidea BOISS	Ref.
Plantae	Asteraceae	Artemisia annua	Ref.
Plantae	Asteraceae	Artemisia capillaris	Ref.
Plantae	Asteraceae	Artemisia scoparia	Ref.
Plantae	Asteraceae	Conyza newii	Ref.
Plantae	Asteraceae	Eupatorium bupleurifolium	Ref.
Plantae	Asteraceae	Porophyllum gracile	Ref.
Plantae	Asteraceae	Porophyllum ruderale	Ref.
Plantae	Asteraceae	Rhaponticum carthamoides	Ref.
Plantae	Asteraceae	Santolina corsica Jordan et Fourr	Ref.
Plantae	Asteraceae	Solidago canadensis	Ref.
Plantae	Asteraceae	Tarchonanthus camphoratus	Ref.
Plantae	Campanulaceae/Lobeliaceae	Codonopsis pilosula	Ref.
Plantae	Caprifoliaceae/Dipsacaceae/Linnaeaceae/Valerianaceae	Lonicera japonica	Ref.

Figure 3: An example of the search results obtained by clicking the C_ID, C00000805, which were retrieved from the KNApSAcK Core Database.

the KNApSAcK Metabolite Ecology Database with KNApSAcK Metabolite Activity Database. Information on VOCs related to biological activity can be obtained by clicking the "A" button as in Figure 2(B). Figure 4 shows the search result of biological activity related to C_ID C00000805, which were retrieved from the KNApSAcK Metabolite Activity Database. The attributes in the list are C_ID, metabolite name, activity category, biological activity, target species, and references, from left to right. Here, the metabolite known as *alpha-pinene* (C_ID C00000805) has few biological activity categories such as antimicrobial, antioxidant, biomarker, defense, plant growth enhancement, anticholinesterase, antifungal, dermatitic, irritant, psychotomimetic, and toxic.

3.2. Clustering of Microorganisms Based on VOC Similarity. Initially, the accumulated VOC data were divided into two types: (1) microorganisms species-VOC binary relations; (2) emitting species-VOC-target species triplet relations. This section focuses on the clustering analysis result of the first type of data, which is the relationship between microorganism species and their emitting VOCs. Until now, we have accumulated 1088 compounds produced by 517 microorganisms. Figure 5 shows the log-log relation between the number of VOCs, M, and the frequency of species, N. The pattern roughly follows power law [33]. Figure 5 shows that there are 92 species that emit only one type of VOC (Point x). Highest 50 types of VOCs are emitted by an individual species and there are 14 such species in our present data (Point y). From this statistical analysis, we can say that most microorganism species emit a few VOCs, which can act as their odor fingerprint. The information of emitting species and compounds has been converted into a 517×1088 binary matrix ("1" indicates presence while "0" indicates absence). The binary matrix then was used to calculate the Euclidean distance between species. From the Euclidean distance, hierarchical clustering of species was performed. Figure 6 shows a hierarchical dendrogram plot of microorganism species based on VOC presence. Here, we cut the dendrogram tree to 50 clusters and the threshold height for this clustering is 7. Supplementary Table 1 (in Supplementary Material available online at http://dx.doi.org/10.1155/2015/139254) shows the species name with their corresponding clusters and the pathogenicity of the microorganism species. Interestingly, 77 species from 517 species are known as pathogenic bacteria and are classified into six clusters, which are clusters 6, 27, 35, 40, 47, and 48. Out of these six clusters, three clusters, that is, clusters 35, 40, and 47, contain 100% pathogenic bacterial species such as *P. aeruginosa*, *K. pneumoniae,* and *E. coli*. The other three clusters contain both pathogenic and nonpathogenic species. For example, cluster 6 consists of 11 (7.2%) pathogenic bacterial species while cluster 27 comprises only one (7.7%) pathogen species. Cluster 48 contains 4 (16%) pathogenic bacterial species. Out of all 50 clusters, the rest of 44 clusters contain nonpathogenic species. These results imply that VOCs emitted by some pathogenic bacteria are different from those emitted by nonpathogenic bacteria. These results show consistency between VOC and pathogenicity based classification of microorganisms.

In order to extract different and more information, we constructed a network by inserting edges between species for which the Euclidean distance is less than a threshold. The threshold was decided to include the lowest 5% distances as edges in the network. We then determined the high-density clusters in that network by applying the graph clustering



INPUT WORD = [Match Type : Exact , C_ID : C00000805]

C_ID	Metabolite Name	Activity Category	Biological Activity (Function)	Target Species	Reference
C00000805	alpha-Pinene	Antimicrobial	Antimicrobial activity towards the tested microorganisms	Bacillus cereus	Ahmad et al.,Molecules,19, (2014),19220-19242
C00000805	alpha-Pinene	Antimicrobial	Antimicrobial activity towards the tested microorganisms	Enterococcus faecalis	Ahmad et al.,Molecules,19, (2014),19220-19242
C00000805	alpha-Pinene	Antimicrobial	Antimicrobial activity towards the tested microorganisms	Methilin-resistant Staphylococcus aureus (MRSA)	Ahmad et al.,Molecules,19, (2014),19220-19242
C00000805	alpha-Pinene	Antimicrobial	Antimicrobial activity towards the tested microorganisms	Salmonella entiriditis	Ahmad et al.,Molecules,19, (2014),19220-19242
C00000805	alpha-Pinene	Antioxidant	Leaf and stem have the highest antioxidant activity		Ahmad et al.,Molecules,19, (2014),19220-19242
C00000805	alpha-Pinene	Biomarker	Respiratory diseases (Cystic Fibrosis)	Human (sputum)	Goeminne et al.,Respiratory Research,13,(2012),13,87
C00000805	alpha-Pinene	Biomarker	The Digestive System (Irritable Bowel Syndrome)		Ahmed et al.,PLoS ONE,8, (2013),e58204
C00000805	alpha-Pinene	Defense	Defend against biotic stressors such as insects and pathogens	Spodoptera exigua	Zebelo et al.,BMC Plant Biology,14, (2014),140
C00000805	alpha-Pinene	Defense	Indirect defense responses	Trialeurodes vaporariorum	López et al.,J Chem Ecol,38, (2012),1376_1386
C00000805	alpha-Pinene	Enhance plant growth	Improved the growth of tobacco seedlings in vitro	Tobacco seedling	Paul et al.,Sensors,13,(2013),13969-13977
C00000805	alpha-Pinene	Anticholinesterase	Acetylcholinesterase inhibitory activities		Ahmad et al.,Molecules,19, (2014),19220-19242
C00000805	alpha-Pinene	Antifungal	Fungal growth inhibition	Colletotrichm gloesporioides	Tenorio-Salgado et al.,Bioengineered,4,(2013),236_243
C00000805	alpha-Pinene	Antifungal	Fungal growth inhibition	Fusarium culmorum	Tenorio-Salgado et al.,Bioengineered,4,(2013),236_243
C00000805	alpha-Pinene	Antifungal	Fungal growth inhibition	Fusarium oxysporum	Tenorio-Salgado et al.,Bioengineered,4,(2013),236_243

Figure 4: An example of the search result of biological activity related to C_ID C00000805, which were retrieved from the KNApSAcK Metabolite Activity Database.

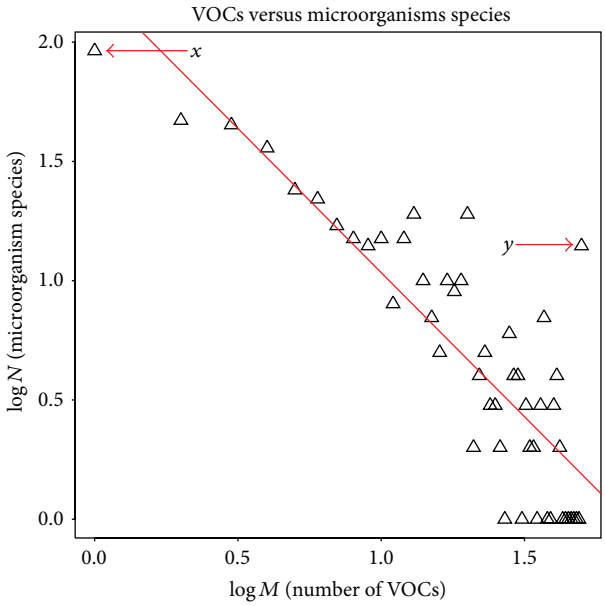


Figure 5: The log-log relation between the number of VOCs and the number of related microorganisms' species.

algorithm DPClus. Supplementary Figure 1 shows the overall network, which displays all the generated clusters in such a way that intracluster edges are green and intercluster edges are red. Figure 7(a) shows the hierarchical connected graph of the clustering result, where the green nodes represent clusters of microorganism species and the red edges represent the interaction between clusters. The radius of a green node in the hierarchical graph in Figure 7 is proportional to the logarithm of the number of nodes in the cluster it represents. The width of a red edge in the hierarchical graph between a pair of clusters is proportional to the number of edges between those clusters in the original graph. Figure 7(b) shows the independent nodes of the hierarchical graph, which indicates that these clusters do not interact with other clusters.

Overall, DPClus generated 50 clusters where 20 clusters are connected nodes to each other while the remaining 30 clusters are independent nodes. Only cluster 1 contains both pathogenic and nonpathogenic microorganisms. Clusters 2, 7, 14, 21, 26, and 40 consist of only pathogenic bacteria while the other clusters are consisting of only nonpathogenic bacteria. These results imply that pathogenicity of microorganisms can be linked to characteristic combinations of identical VOCs emitted by them. Some of the pathogenic members of cluster 1 such as *Klebsiella pneumoniae*, *Escherichia coli*, *Staphylococcus aureus,* and *Pseudomonas aeruginosa* are very highly connected to other pathogenic clusters, for example, clusters 2 and 7. Figure 7(a) shows that clusters 2, 7, 14, 21, 26, and 40 are connected by red edges, which reflect VOC similarity between pathogenic microorganisms. Also, there is VOC based similarity between nonpathogenic species of cluster 1 and clusters 10, 13, 16, 18, 19, 23, 24, 33, and 36. The red edges between clusters 4 and 8 and between clusters 9 and 15 are also because of VOC similarity between nonpathogenic species of those clusters. Here it is noteworthy that the rest of 30 clusters consisting of nonpathogenic species are independent clusters,

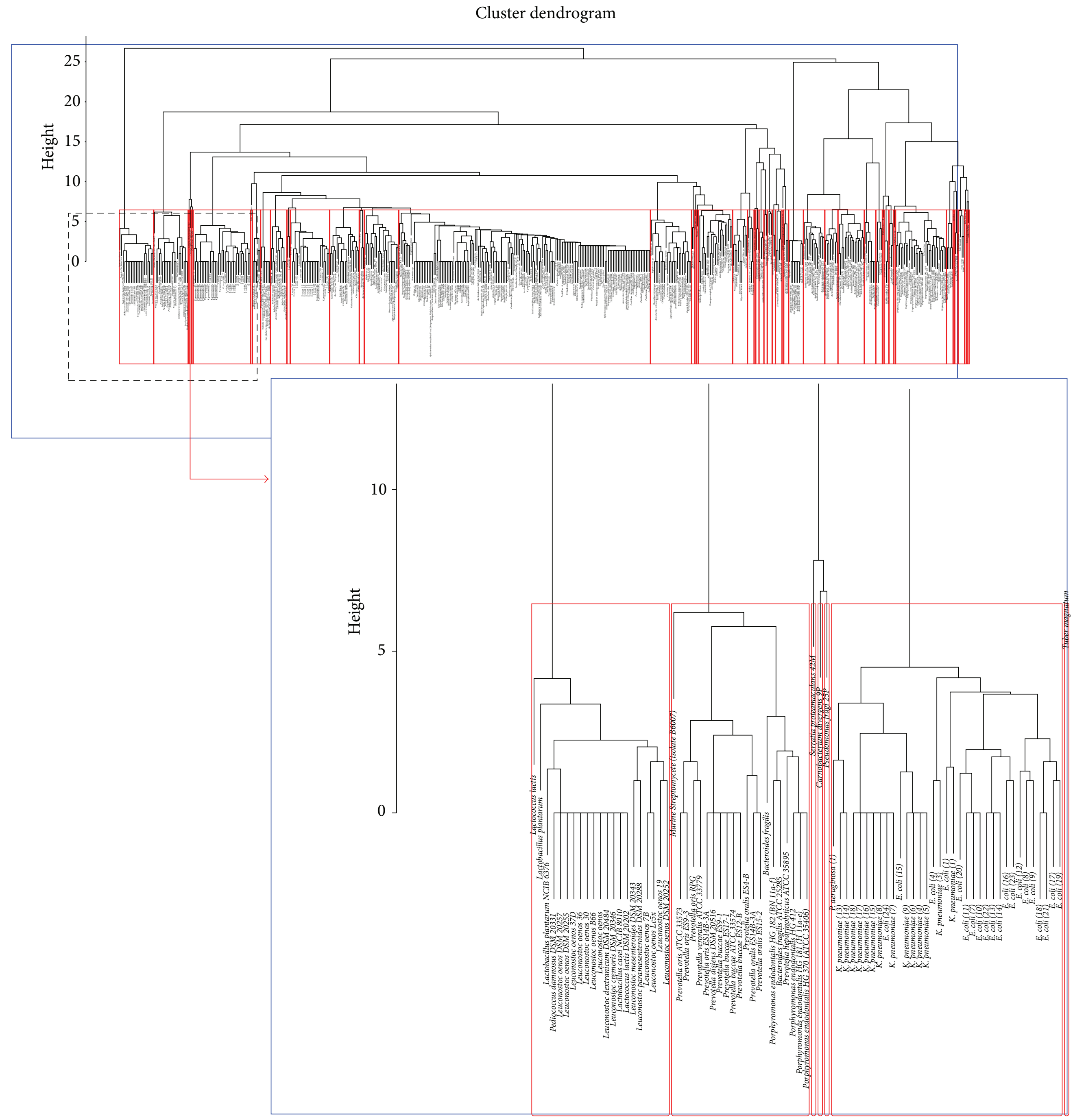


Figure 6: Hierarchical dendrogram plot of microorganism species based on VOC presence.

which implies that many nonpathogenic groups of species emit quite unique types of VOCs as shown in Figure 7(b). Supplementary Figure 2 shows the microorganism species belong to cluster 1 (pathogenic and nonpathogenic), cluster 7 (pathogenic only), and cluster 10 (nonpathogenic species only), respectively. Here the internal nodes of a cluster are shown connected by green edges and its neighboring clusters are shown connected by red edges. To evaluate the stability of graph clustering results by DPClus, we also clustered the networks generated by several random samplings of 80% or more edges of the original network. We found that DPClus can still cluster the microorganisms species based on pathogenicity.

The results of network clustering and hierarchical clustering are similar in the sense that both results indicated that VOC based classification of microorganisms is consistent with their classification based on pathogenicity. However, clustering by DPClus further revealed existence and nonexistence of relations between different pathogenic and nonpathogenic groups of microorganisms.

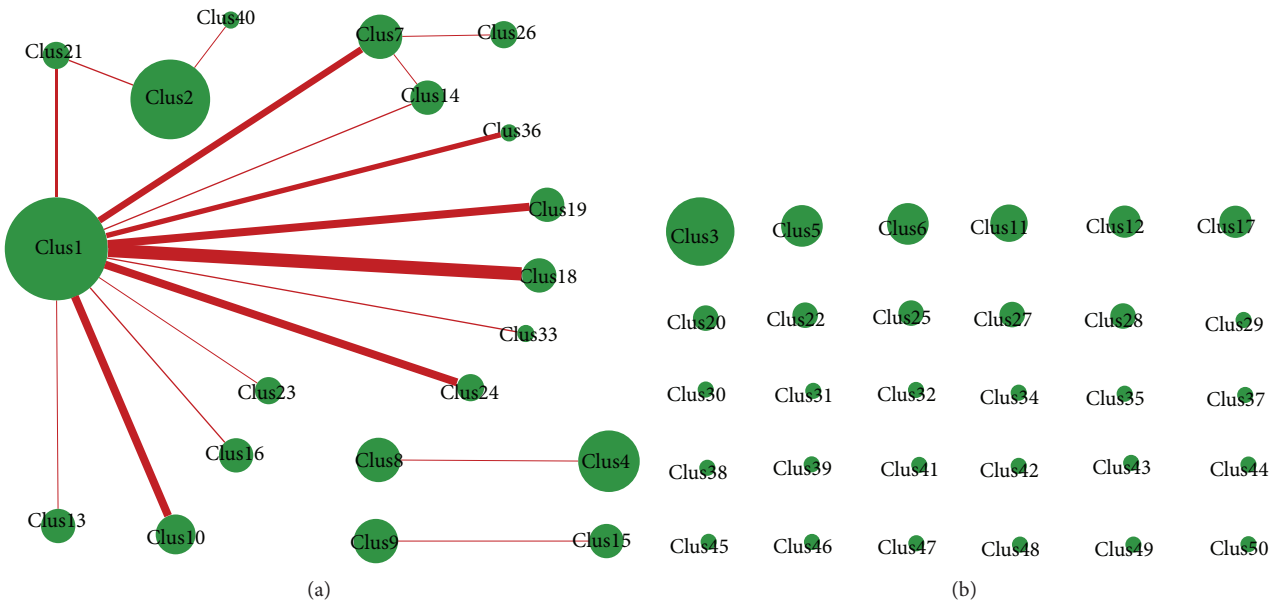


FIGURE 7: Hierarchical graph of DPClus clustering result. (a) Connected nodes. (b) Independent nodes.

3.3. Clustering Analysis of VOCs Based on Chemical Structure Similarity. The first type of data focused on microorganism species only but the second type of data includes VOCs emitted by other biological species such as plants, animals, and humans. The second data type that we have accumulated until now is 1044 species-species interactions via 341 VOCs associated with 11 groups of biological activities. The biological activities of VOCs are classified into two types: (i) chemical ecology related activities, in which most VOCs are involved in interaction between species for survival of organisms such as defense and antimicrobial, and (ii) human health care related activities, in which many VOCs are widely used as disease biomarker and odor. From our accumulated data, 57.3% of the activities belong to chemical ecology such as antifungal, antimicrobial, attractant, defense, plant growth enhancement, root growth inhibition, and repellent activities and 42.7% are human health related activities such as disease biomarker, odor, anticholinesterase, and antioxidant as shown in Figure 8. There are many VOCs, which have several biological activities. Thus, it is important to investigate the relationships between VOCs and their biological activities statistically. Initially, we determined pairwise chemical structural similarity between VOCs based on Tanimoto coefficient. 2D compound structures in the generic structure definition file (SDF) format of all 341 VOCs were obtained from PubChem database (https://pubchem.ncbi.nlm.nih.gov) and then were imported into ChemmineR package in one batch file. We calculated the chemical structure similarity using Tanimoto coefficient. Then, we converted the Tanimoto similarity matrix into distance matrix by subtracting each of the similarity values from 1. Based on distance matrix, we performed heatmap clustering and the result is shown in Figure 9. White and red colours indicate the extreme distance values of 0 and 1, respectively, and the intermediate distance values are indicated by the intensity of the red colour.

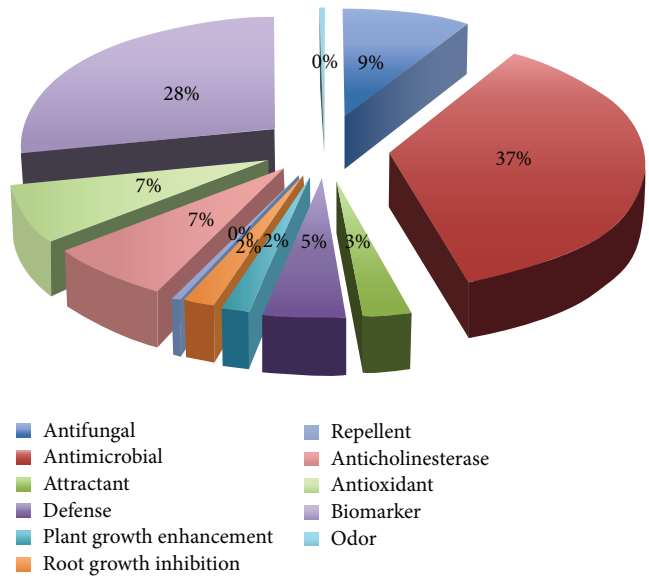


FIGURE 8: Pie chart showing the relative frequencies VOCs belonging to 11 biological activities.

From the heatmap plot, we tentatively outlined 11 clusters of VOCs. The count of VOCs belonging to each activity group in each cluster is shown in Table 1. To assess the richness of VOCs of similar activity in individual clusters, we determined their p values based on hypergeometric distribution which are also shown in Table 1. The major types of chemical compounds belonging to each cluster and their corresponding biological activities are mentioned in Table 2. The chemical structures of the VOCs belonging to all clusters (cluster 1 to cluster 11) are shown in Supplementary Figure 2.

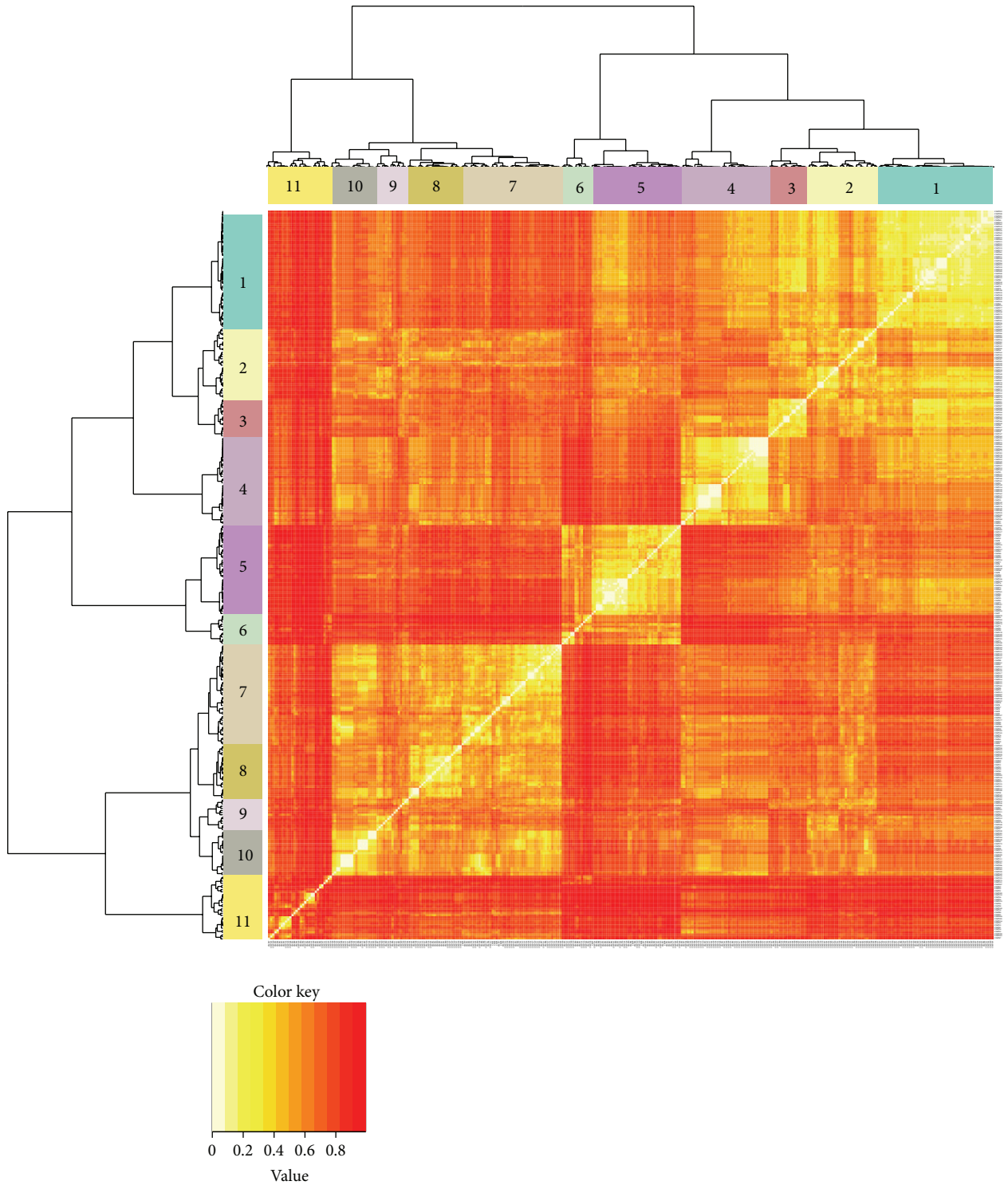


Figure 9: Heatmap clustering of VOCs based on chemical structure similarity determined by Tanimoto coefficient.

From this result, we can see that there are 55 VOCs belonging to cluster 1 and mainly involved with anticholinesterase, antimicrobial, antioxidant, and defense activities, for example, beta-caryophyllene, isocaryophyllene, and caryophyllene. All compounds in cluster 1 are terpenoids, of which 15 VOCs are monoterpenoids (10 carbon units) and 40 VOCs are sesquiterpenoids (15 carbon units). There are 33 VOCs in cluster 2 and the p values corresponding to anticholinesterase, antimicrobial, and antioxidant are 4.849×10^{-4}, 9.696×10^{-4}, and 4.849×10^{-4}, respectively. Some of the VOCs that are classified into cluster 2 are monoterpenoids and sesquiterpenoids such as beta-linalool, terpinen-4-ol,

Table 1: The count of VOCs belonging to each activity group in each cluster and their p value based on hypergeometric distribution.

Biological activity	Cluster ID (Count)	Cluster 1 (55)	Cluster 2 (33)	Cluster 3 (41)	Cluster 4 (18)	Cluster 5 (21)	Cluster 6 (25)	Cluster 7 (47)	Cluster 8 (15)	Cluster 9 (42)	Cluster 10 (14)	Cluster 11 (30)
Anticholinesterase	p value (count)	5.28×10^{-7} (26)	4.849×10^{-4} (15)	0.6181 (8)	0.9109 (2)	1.274×10^{-2} (9)	0.9147 (3)	1 (0)	0.3596 (4)	0.9994 (2)	1 (0)	1 (0)
Antifungal	p value (count)	0.9128 (2)	0.9115 (1)	0.5399 (3)	2.561×10^{-2} (4)	1 (0)	0.5176 (2)	0.6403 (3)	0.6571 (1)	0.3099 (4)	1 (0)	0.3291 (3)
Antimicrobial	p value (count)	2.10×10^{-6} (26)	9.696×10^{-4} (15)	0.6898 (8)	0.9281 (2)	1.871×10^{-2} (9)	0.8246 (4)	0.9999 (1)	0.4049 (4)	0.9997 (2)	1 (0)	1 (0)
Antioxidant	p value (count)	5.28×10^{-7} (26)	4.849×10^{-4} (15)	0.6181 (8)	0.9109 (2)	1.274×10^{-2} (9)	0.9147 (3)	1 (0)	0.3596 (4)	0.9994 (2)	1 (0)	1 (0)
Attractant	p value (count)	0.9708 (2)	0.8144 (2)	1 (0)	0.4831 (2)	1 (0)	0.1661 (4)	3.829×10^{-2} (8)	0.1356 (3)	1.983×10^{-2} (8)	1 (0)	1 (0)
Biomarker	p value (count)	1 (8)	0.9999 (11)	1.835×10^{-3} (34)	0.7944 (10)	1.444×10^{-2} (18)	0.9821 (11)	4.42×10^{-5} (41)	0.9948 (5)	6.071×10^{-2} (31)	1.036×10^{-2} (13)	1.963×10^{-3} (26)
Defense	p value (count)	3.35×10^{-9} (22)	9.258×10^{-2} (7)	0.9758 (2)	0.6764 (2)	0.7594 (2)	0.8418 (2)	0.9987 (1)	0.8668 (1)	0.9787 (2)	1 (0)	1 (0)
Plant growth enhancement	p value (count)	6.01×10^{-2} (4)	1 (0)	1 (0)	1 (0)	1 (0)	3.531×10^{-3} (4)	0.7778 (1)	1 (0)	1 (0)	1 (0)	0.6069 (1)
Root growth inhibition	p value (count)	0.1749 (5)	0.8632 (1)	0.9183 (1)	1 (0)	0.7111 (1)	4.111×10^{-2} (4)	0.7672 (2)	0.5847 (1)	0.7062 (2)	0.5591 (1)	0.8347 (1)
Odor	p value (count)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	2.29×10^{-5} (4)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)
Repellent	p value (count)	1 (0)	7.551×10^{-2} (2)	1 (0)	1 (0)	1.871×10^{-3} (3)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)

Table 2: Summary of clustering result and its descriptions related to chemical structures and biological activities.

Cluster ID (count)	Description on chemical structures	Related biological activities
Cluster 1 (55 VOCs)	All compounds are terpenoids. 15 VOCs are monoterpenoids (10 carbon units) and 40 VOCs are sesquiterpenoids (15 carbon units).	Anticholinesterase, antimicrobial, antioxidant, and defense.
Cluster 2 (33 VOCs)	17 VOCs are alcohol, aldehyde, ketone, epoxide, and ester of terpenoids. The other VOCs are alcohol, aldehyde, carboxylic acid, ester, and ketone of straight-chain alkenes.	Anticholinesterase, antimicrobial, and antioxidant.
Cluster 3 (41 VOCs)	Alkanes.	Biomarker.
Cluster 4 (18 VOCs)	Alkenes.	Antifungal.
Cluster 5 (21 VOCs)	Aldehyde, ester, carboxylic acid, and ketone of C8–C18 alkanes.	Anticholinesterase, antimicrobial, antioxidant, biomarker, and repellent.
Cluster 6 (25 VOCs)	21 VOCs are alcohol and ether of C3–C8 alkanes.	Plant growth enhancement and root growth inhibition and odor.
Cluster 7 (47 VOCs)	45 VOCs are ester, carboxylic acid, ketone, and aldehyde of noncyclic C2–C9 alkanes.	Attractant and biomarker.
Cluster 8 (15 VOCs)	VOCs consist of epoxide, ethers, esters, and alcohols.	—
Cluster 9 (42 VOCs)	24 VOCs are aromatic alcohols, carboxylic acids, esters, ketones, and ethers. 16 VOCs are aromatic compounds consisting of C and H atoms. One VOC consists of C, H, and Br atoms. One VOC is an alkane ester.	Attractant.
Cluster 10 (14 VOCs)	Aromatic compounds. 12 VOCs are heteroaromatic compounds that consist of one or more sulfur, nitrogen, or oxygen atoms.	Biomarker.
Cluster 11 (30 VOCs)	VOCs are quite diverse in chemical elements, C0–C6 small molecules.	Biomarker.

p-menth-1-en-8-ol, drimenol, and nerolidol. 17 VOCs are alcohol, aldehyde, ketone, epoxide, and ester of terpenoids. The other VOCs are alcohol, aldehyde, carboxylic acid, ester, and ketone of straight-chain alkenes.

For cluster 3, there are 41 compounds and the main biological activity involved is biomarker for various diseases such as colorectal cancer and asthma. We obtained small p value (1.835×10^{-3}) for biomarker activity of cluster 3. All compounds are alkanes; most of them are emitted in human breath such as octane, isobutane, 2-methylpentane, methylcyclohexane, hexane, and cyclohexane.

There are 18 compounds in cluster 4 and all of them are alkenes such as beta-farnesene, alpha-caryophyllene, ocimene, and beta-ocimene. These compounds are mainly associated with chemical ecology activity, which is antifungal and the p value for this activity is 2.561×10^{-2}. For cluster 5, there are 21 VOCs which are aldehyde, ester, carboxylic acid, and ketone of C8–C18 alkanes. Cluster 5 is significantly related with multiple biological activities, which are anticholinesterase, antimicrobial, antioxidant, biomarker, and repellent activities.

There are 25 VOCs in clusters 6 and 21, of them are alcohol and ether of C3–C8 alkanes. We also obtained small p value for plant growth enhancement activity (3.531×10^{-3}), root growth inhibition activity (4.111×10^{-2}), and odor activity (2.29×10^{-5}) for cluster 6. An example of VOCs involved in plant growth enhancement activity is 2,3-butanediol and there are many reports that this compound released by soil microorganisms had improved plant growth and increased pathogen resistance [34, 35]. For odor activity, compounds involved are in alcohol sulfanylalkanols chemical class group such as 2-methyl-3-sulfanylbutan-1-ol and 3-methyl-3-sulfanylhexan-1-ol. These compounds have a pungent sweat/kitchen odor, also reminiscent of onions with some fruity connotations which are transformed into the volatile substances by bacterial enzymes present only in corynebacteria.

There are 47 VOCs in clusters 7 and 45, of them are ester, carboxylic acid, ketone, and aldehyde of noncyclic C2–C9 alkanes. Cluster 7 is significantly related with multiple biological activities, which are attractant (p value = 3.829×10^{-2}) and biomarker for various diseases (p value = 4.42×10^{-5}). Aldehydes belong to cluster 7 such that acetaldehyde, propanal, hexanal, 2-methyl-butanal, pentanal, heptanal, and 3-methyl-butanal are mostly used as biomarker for various diseases including cancers and irritable bowel syndrome. In cluster 8, there are 15 VOCs belonging to this cluster, which consist of epoxide, ethers, esters, and alcohols. In cluster 9, there are 42 VOCs and the main biological activity is attractant (p value = 1.983×10^{-2}). All VOCs belonging to cluster 9 are aromatic compounds, in which 24 VOCs are aromatic alcohols, carboxylic acids, esters, ketones, and ethers. 16 VOCs are aromatic compounds consisting of C and H atoms. One VOC consists of C, H, and Br atoms. One VOC is an alkane ester. Also, there are 14 VOCs in cluster 10 which are aromatic compounds. 12 VOCs are heteroaromatic compounds that consist of one or more sulfur, nitrogen, or oxygen atoms. On the other hand, 30 VOCs belonging

to cluster 11 are of diverse types of C0–C6 small molecules with low molecular weight, ranging from hydrogen cyanide (27.02534) to tetrachloroethyene (165.8334). The main biological activity for cluster 10 and cluster 11 is biomarker for various diseases. The p values for biomarker activity for cluster 10 and cluster 11 are 1.036×10^{-2} and 1.963×10^{-3}, respectively. The major VOCs involved in this activity are isoxazole, 2,3-dimethyl-pyrazine, and 2-methyl-pyrazine which are mostly produced in human urine and can be used as biomarker for autism spectrum disorders.

The heatmap clustering shows that there are strong links between chemical structure of VOCs and their biological activities. Comparative activity relationships between chemical ecology and human health care activity will lead to systematization of metabolomics combined with human and ecological metabolic pathways.

4. Conclusion

In the present study, we have developed a database of VOCs emitted by various living organisms including microorganisms, plants, animals, and human, which can be accessed at KNApSAcK Metabolite Ecology Database. Apart from VOC biological activities related to human health care, more than half of the biological activities are associated with chemical ecology. Hierarchical clustering and graph clustering by DPClus algorithm were utilized to extract specific microorganism species clusters based on VOC similarity. We found consistency between VOC and pathogenicity based classification of microorganisms. Additionally, heatmap clustering based on Tanimoto similarity measure was used to cluster the VOCs emitted by various species. We found that similar chemical structures of VOCs indicate possibilities of exhibiting similar biological activities. In future work, we will accumulate more data and perform comprehensive analysis of the VOCs in the context of human health care and chemical ecology. The KNApSAcK Metabolite Ecology Database may be useful for the discovery of novel agriculture tools and also for the noninvasive identification of biomarkers in medical diagnostic field as well as systematic research in various omics fields, especially metabolomics integrated with ecosystems.

Conflict of Interests

The authors declare that there is no financial interest or conflict of interests regarding the publication of this paper.

Acknowledgments

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. J. Patterson, J. Cooper, P. M. Kirk, R. L. Pyle, and D. P. Remsen, "Names are key to the big new biology," *Trends in Ecology and Evolution*, vol. 25, no. 12, pp. 686–691, 2010.

[2] T. Hey, S. Tansley, and M. Tolle, *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Microsoft Research, Washington, DC, USA, 2009.

[3] S. Kelling, W. M. Hochachka, D. Fink et al., "Data-intensive science: a new paradigm for biodiversity studies," *BioScience*, vol. 59, no. 7, pp. 613–620, 2009.

[4] F. M. Afendi, T. Okada, M. Yamazaki et al., "KNApSAcK family databases: integrated metabolite-plant species databases for multifaceted plant research," *Plant & Cell Physiology*, vol. 53, no. 2, article e1, 2012.

[5] S. H. Wijaya, H. Husnawati, F. M. Afendi et al., "Supervised clustering based on DPClusO: prediction of plant-disease relations using Jamu formulas of KNApSAcK database," *BioMed Research International*, vol. 2014, Article ID 831751, 15 pages, 2014.

[6] Y. Nakamura, F. M. Afendi, A. K. Parvin et al., "KNApSAcK metabolite activity database for retrieving the relationships between metabolites and biological activities," *Plant and Cell Physiology*, vol. 55, no. 1, article e7, 2014.

[7] Y. Ohtana, A. A. Abdullah, M. Altaf-Ul-Amin et al., "Clustering of 3D-structure similarity based network of secondary metabolites reveals their relationships with biological activities," *Molecular Informatics*, vol. 33, no. 11-12, pp. 790–801, 2014.

[8] D. D. Rowan, "Volatile metabolites," *Metabolites*, vol. 1, no. 1, pp. 41–63, 2011.

[9] Y. Iijima, "Recent advances in the application of metabolomics to studies of biogenic volatile organic compounds (BVOC) produced by plant," *Metabolites*, vol. 4, no. 3, pp. 699–721, 2014.

[10] L. D. J. Bos, P. J. Sterk, and M. J. Schultz, "Volatile metabolites of pathogens: a systematic review," *PLoS Pathogens*, vol. 9, no. 5, Article ID e1003311, 8 pages, 2013.

[11] N. Yusuf, A. Zakaria, M. I. Omar et al., "In-vitro diagnosis of single and poly microbial species targeted for diabetic foot infection using e-nose technology," *BMC Bioinformatics*, vol. 16, article 158, 2015.

[12] M. Syhre and S. T. Chambers, "The scent of *Mycobacterium tuberculosis*," *Tuberculosis*, vol. 88, no. 4, pp. 317–323, 2008.

[13] M. Shirasu and K. Touhara, "The scent of disease: volatile organic compounds of the human body related to disease and disorder," *Journal of Biochemistry*, vol. 150, no. 3, pp. 257–266, 2011.

[14] C. Lourenço and C. Turner, "Breath analysis in disease diagnosis: methodological considerations and applications," *Metabolites*, vol. 4, no. 2, pp. 465–498, 2014.

[15] M. Phillips, R. N. Cataneo, C. Saunders, P. Hope, P. Schmitt, and J. Wai, "Volatile biomarkers in the breath of women with breast cancer," *Journal of Breath Research*, vol. 4, no. 2, Article ID 026003, 8 pages, 2010.

[16] D. F. Altomare, M. Di Lena, F. Porcelli et al., "Exhaled volatile organic compounds identify patients with colorectal cancer," *British Journal of Surgery*, vol. 100, no. 1, pp. 144–150, 2013.

[17] M. Phillips, R. N. Cataneo, R. Condos et al., "Volatile biomarkers of pulmonary tuberculosis in the breath," *Tuberculosis*, vol. 87, no. 1, pp. 44–52, 2007.

[18] M. Hakim, Y. Y. Broza, O. Barash et al., "Volatile organic compounds of lung cancer and possible biochemical pathways," *Chemical Reviews*, vol. 112, no. 11, pp. 5949–5966, 2012.

[19] M. Dunkel, U. Schmidt, S. Struck et al., "SuperScent—a database of flavors and scents," *Nucleic Acids Research*, vol. 37, no. 1, pp. D291–D294, 2009.

[20] M. C. Lemfack, J. Nickel, M. Dunkel, R. Preissner, and B. Piechulla, "mVOC: a database of microbial volatiles," *Nucleic Acids Research*, vol. 42, no. 1, pp. D744–D748, 2014.

[21] T. Acree and H. Arn, "Flavornet and human odor space," http://www.flavornet.org/index.html.

[22] A. M. El-Sayaed, "The Pherobase: Database of Insect Pheromones and Semiochemicals," http://www.pherobase.com.

[23] K. Skogerson, G. Wohlgemuth, D. K. Barupal, and O. Fiehn, "The volatile compound BinBase mass spectral database," *BMC Bioinformatics*, vol. 12, article 321, 2011.

[24] 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," *Journal of Computer Aided Chemistry*, vol. 7, pp. 150–156, 2006.

[25] 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, article 207, 2006.

[26] R. Shen, A. B. Olshen, and M. Ladanyi, "Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis," *Bioinformatics*, vol. 25, no. 22, pp. 2906–2912, 2009.

[27] T. Hwang, G. Atluri, M. Xie et al., "Co-clustering phenome-genome for phenotype classification and disease gene discovery," *Nucleic Acids Research*, vol. 40, no. 19, article e146, 2012.

[28] Y. Liu, Q. Gu, J. P. Hou, J. Han, and J. Ma, "A network-assisted co-clustering algorithm to discover cancer subtypes based on gene expression," *BMC Bioinformatics*, vol. 15, article 37, 2014.

[29] P. Willett, J. M. Barnard, and G. M. Downs, "Chemical similarity searching," *Journal of Chemical Information and Computer Sciences*, vol. 38, no. 6, pp. 983–996, 1998.

[30] D. E. Patterson, R. D. Cramer, A. M. Ferguson, R. D. Clark, and L. E. Weinberger, "Neighborhood behavior: a useful concept for validation of 'molecular diversity' descriptors," *Journal of Medicinal Chemistry*, vol. 39, no. 16, pp. 3049–3059, 1996.

[31] Y. Cao, A. Charisi, L.-C. Cheng, T. Jiang, and T. Girke, "ChemmineR: a compound mining framework for R," *Bioinformatics*, vol. 24, no. 15, pp. 1733–1734, 2008.

[32] T. W. H. Backman, Y. Cao, and T. Girke, "ChemMine tools: an online service for analyzing and clustering small molecules," *Nucleic Acids Research*, vol. 39, no. 2, pp. W486–W491, 2011.

[33] H. Jeong, S. P. Mason, A.-L. Barabási, and Z. N. Oltvai, "Lethality and centrality in protein networks," *Nature*, vol. 411, no. 6833, pp. 41–42, 2001.

[34] C. M. Ryu, M. A. Farag, C. H. Hu et al., "Bacterial volatiles promote growth in *Arabidopsis*," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 8, pp. 4927–4932, 2003.

[35] M. D'Alessandro, M. Erb, J. Ton et al., "Volatiles produced by soil-borne endophytic bacteria increase plant pathogen resistance and affect tritrophic interactions," *Plant, Cell and Environment*, vol. 37, no. 4, pp. 813–826, 2014.