BiClusO: A Novel Biclustering Approach and Its Application to Species-VOC Relational Data

Mohammad Bozlul Karim, Ming Huang, Naoaki Ono, Shigehiko Kanaya, and Md. Altaf-Ul-Amin

Abstract—In this paper, we propose a novel biclustering approach called BiClusO. Biclustering can be applied to various types of bipartite data such as gene-condition or gene-disease relations. For example, we applied BiClusO to bipartite relations between species and volatile organic compounds (VOCs). VOCs, which are emitted by different species, have huge environmental and ecological impacts. The biosynthesis of VOCs depends on different metabolic pathways which can be used to categorize the species. A previous study related to the KNApSAcK VOC database classified microorganisms based on their VOC profiles, which confirmed the consistency between VOC-based and pathogenicity-based classifications. However, due to limited data, classification of all species in terms of VOC profiles was not performed. In this study, we enriched our database with additional data collected from different online sources and journals. Then, by applying BiClusO to species-VOC relational data, we determined that VOC-based classification is consistent with taxonomy-based classification of the species. We also assessed the diversity of VOC pathways across different kingdoms of species.

Index Terms—Bicluster, bipartite graph, volatile organic compound, tanimoto coefficient, biclique

1 Introduction

Biclustering is an important data mining technique usually used to partition a sparse data matrix into a finite number of highly dense submatrices. In computational systems biology, two variable pairs such as gene-condition, gene-patient, or gene-disease are widely used with biclustering to identify the specific set of genes up-regulated and/or down-regulated by similar types of conditions, patients, or diseases. Also, some studies use biclustering to classify herbivorous species based on the plants they feed on to understand the evolutionary changes of the species in adapting to the availability of certain kinds of plants in their habitat [1]. Thus biclustering can be applied to various types of bipartite data in different fields.

In this paper, we introduce a new biclustering algorithm called BiClusO using the concept of the Tanimoto coefficient, relation number, and the DPClusO algorithm [2], [3], [4] and emphasize the construction of a simple graph by means of the first node set of a bipartite graph. In our approach, the edges of the simple graph are selected based on the common neighbors (in the second set) of the nodes of the first set. Thus, we create data folding and apply the DPClusO algorithm to generate a cluster set. After generating a cluster set, we unfold the data by assigning the members of the second node set to an individual cluster using the probability of their attachment to cluster nodes. Our algorithm can produce overlapping biclusters. Overlapping biclustering means a node of any side of a bipartite graph may belong to more than one bicluster. The problem of finding a minimum bicluster set which is either mutually exclusive or overlapping and covers all data elements from a bipartite graph has been proven to be NP-hard [5]. So, we select a single node set, i.e, the most significant node set, rather than both node sets to convert the biclustering problem to a simple graph clustering problem, for which there are polynomial-time heuristic algorithms.

As an example of our biclustering algorithm, we applied it to a database of Volatile Organic Compounds (VOCs). Under environmental temperature and pressure, VOCs can easily become vapor. Most VOCs contain carbon, along with hydrogen, oxygen, chlorine, fluorine, bromine, sulfur, or nitrogen. In this study, we mainly focused on VOCs emitted by living organisms. These are called biogenic VOCs. They have huge environmental and ecological impacts because they are the medium of mutual interactions between species. By facilitating the survival of species, VOCs play important roles in controlling the ecosystem with individual and combined effects. VOC emission is also substantially affected by the impacts of climate changes and ecosystem redistribution. The evolutionary development of species partially depends on changes in the environment and mutual interactions of species using VOCs. Backtracking VOC generation as a product of intricate metabolic pathways can identify complex cellular processes, which can then be used to categorize the biosynthesis mechanisms of these metabolites in different species. Having successful discovery of some VOC pathways, scientist are now trying to control the floral VOCs of plants through metabolic engineering such as up-regulating or down-regulating of biochemical steps and modification of existing pathways in an

- *The authors are with the Nara Institute of Science and Technology, Nara 630-0192, Japan. E-mail: hira9505040@gmail.com, {alex-mhuang, nono, amin-m}@is.naist.jp, skanaya@gtc.naist.jp.*

Manuscript received 7 Aug. 2018; revised 13 Mar. 2019; accepted 24 Apr. 2019. Date of publication 14 May 2019; date of current version 8 Dec. 2020.
(Corresponding author: Md. Altaf-Ul-Amin.)
Digital Object Identifier no. 10.1109/TCBB.2019.2914901

TABLE 1
Symbols and Their Meanings

Notation	Description
U, V	Two disjoint sets of nodes of a bipartite graph
E	Edge set between U and V
$N(u)$	Neighbor of u where $u \in U$ and $N(u) \subseteq V$
$N(v_ik)$	Neighbor of v_i in kth cluster $v_i \in V$, $N(v_i) \subseteq U$ and $\lvert N(v_ik)\rvert \leq \lvert N(v_i)\rvert$
R	Relation matrix with dimension $\lvert U\rvert \times \lvert U\rvert$ where $R_{i,j} = N(u_i) \cap N(u_j)$
T	Tanimoto cofficient matrix with dimension $\lvert U\rvert \times \lvert U\rvert$ where $T_{i,j} = (\lvert R_{i,j}\rvert / \lvert N(u_i) \cup N(u_j)\rvert)$
$Th_{rl}(\in R)$	Relation number threshold , usually any small element from R matrix
$Th_{tf}(\in T)$	Tanimoto cofficient threshold , usually any small element from T matrix
F	Boolean matrix of dimension $\lvert U\rvert \times \lvert U\rvert$ constructed by using R, T, Th_{rl}, Th_{tf} where $F_{i,j} \in \{0, 1\}$
Pvk	Probability set for finding second set of nodes in kth cluster. Where $vk = \{v_1, v_2, \ldots\ldots v_m\} \subseteq V$, the total m number of attached nodes in kth cluster
Pv_ik	Probability of node v_i to be included in kth cluster where $Pv_ik \in Pvk$ and $1 \leq i \leq \lvert Pvk\rvert$
$Pvth$	Threshold value of finding second set of nodes in a cluster, expressed as probability.
VOC_{Ck}	VOC set of kth cluster.

attempt to increase pollination and defense mechanisms [6]. Classification of species based on common emitting VOCs can lead to understanding the symbiosis of organisms in terms of VOCs. Also, clustering of VOCs based on chemical structure similarities can help to predict their pathways. In this study, we applied our algorithm to cluster the species-VOC relational data from the KNApSAcK database [7], [8], [9] and determined that the VOC-based classification of species was consistent with the taxonomy-based classification. We also assessed the diversity of VOC pathways across different kingdoms of species.

2 Proposed Biclustering Algorithm

2.1 The Concept of Biclustering

A bipartite graph is a graph that consists of two disjoint sets of nodes, U and V, such that each edge connects a node in U with a node in V, i.e., U and V are independent sets. Mathematically, a bipartite graph is denoted as $G = (U, V, E)$ where U and V are two disjoint partitions in G. An edge matrix is represented by a weight function $w : U \times V \rightarrow \{0, 1\}$ such that $w((u, v)) = 1$ for $(u, v) \in E$ and $w((u, v)) = 0$ for $(u, v) \notin E$.

A bicluster represents a pair of subsets $U' \subseteq U$ and $V' \subseteq V$ with $w : U' \times V' \rightarrow \{0, 1\}$ with a higher density of '1'. A biclustering algorithm finds a region bounded by $I = |U'|$ rows and $J = |V'|$ columns which maximizes the density and dimensions of the submatrix. In simple terms, a bipartite graph can be represented by a binary matrix. The concept of biclustering can also be extended for a general matrix. There have been several different approaches invented to seek maximal region biclusters [5], [10], [11], [12]. For example, Cheng and Church (2000) applied the greedy approach-based biclustering method to gene expression data. Amos, Roded, and Shamir developed Statistical-Algorithmic Method for Bicluster Analysis (SAMBA) using statistical data modeling to calculate vertex pair weighting, a hashing technique to find the heaviest bicliques, and a local score improvement procedure for addition or deletion of a single vertex. Factor Analysis for Bicluster Acquisition (FABIA) uses linear dependencies between samples and feature patterns. It is based on a multiplicative model which captures realistic non-Gaussian data distributions with heavy tails as observed in bipartite relationships in gene expression data. BiMax, on the other hand, recursively uses the divide and conquer method to enumerate all the maximal biclusters in a binary data matrix. Another statistically significant biclustering for a binary data matrix was defined by Koyutürk [13] which is referred to as Cmnk [14]. However, most of these approaches do not support clustering by controlling overlapping. Cmnk and BiMax, which are mainly designed for binary data sets, cannot produce good results when there are large number of zeros [14] in the input binary matrix. In this work, we propose a heuristic algorithm for finding biclusters in binary matrix data. Our approach can generate overlapping biclusters by utilizing the overlapping property of DPClusO.

Algorithm 1. Constructing R and T Matrices

```
Input: Bipartite graph $G = (U, V, E)$ represented by a matrix
       where the rows and columns correspond to the nodes
       of set $U$ and $V$ respectively.
Output: Two square matrices $R$ and $T$ containing relation number
        and Tanimoto coefficient between each pair of
        nodes of $U$.
   Pseudo Code:
1: Find node pairs $(u_i, u_j) \in U$ where $(i < j)$
2: Find neighbors set of $u_i$ and $u_j$ from $V$ as $N(u_i), N(u_j)$
3: Calculate the total number of the common neighbor
   $|N(u_i) \cap N(u_j)|$ from $N(u_i)$ , $N(u_j)$ and assign it to relation
   matrix as $R_{i,j}$
4: Calculate the Tanimoto coefficient
   $(|N(u_i) \cap N(u_j)|/|N(u_i) \cup N(u_j)|)$ between $u_i$ and $u_j$ and
   assign it to Tanimoto matrix as $T_{i,j}$
5: Output the $R, T$ matrix
```

2.2 The Biclustering Algorithm BiClusO

Table 1 summarizes the notations and their meanings that are used in the description of the algorithm. In this paper, we use data folding mechanism to create a simple graph G_s from the bipartite graph $G = (U, V, E)$ involving the nodes of set U. We discuss the proposed biclustering algorithm in terms of two separate algorithms. Algorithm 1 creates the relation matrix and the Tanimoto cofficient matrix. Based on these two matrices the second algorithm constructs G_s, applies DPClusO to generate clusters, and then unfolds the data by assigning the members of set V to individual clusters and thus completes the biclusters. The R and T matrices are diagonally symmetrical square matrices, and we only need to calculate the lower triangular parts of them. Condition $(i < j)$ in step 1 of Algorithm 1 only considers the lower triangular portions of R and T. If we have a total number of d elements in set V, then the degree of the nodes in set U may vary between 1 and d, considering the fact that no isolated node exists in U and V. This degree distribution makes Algorithm 1 output a relation matrix with values ranging between $(0, d)$. Here 0 means there is no common neighbor between two nodes. Step 1 to Step 4 loop through the nodes of set U. Calculating Step 3 and 4 only needs to determine the union and intersection operation on neighbors of nodes of set U from set V.

Algorithm 2 constructs a simple graph G_s using user defined threshold values Th_{rl} and Th_{tf} and applies the DPClusO clustering algorithm to find the cluster sets in G_s. The relation number threshold Th_{rl} and Tanimoto coefficient threshold Th_{tf} are used to select the edges of G_s. The values for these thresholds can be calculated based on the characteristics of the input bipartite graph.

Algorithm 2. Generate Biclusters

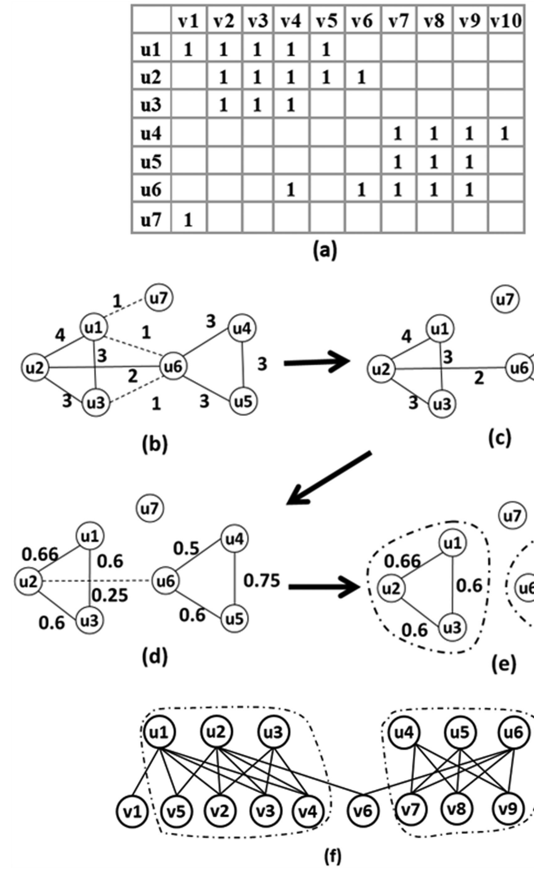
```
Input: R and T matrices from Algorithm 1, Relation number
       threshold $Th_{rl}$, Tanimoto coefficient thereshold $Th_{tf}$,
       Cluster Property CP, Cluster Density CD, Overlapping
       Coefficient OV, Pvth.
Output: Bicluster set
    Pseudo Code:
1: Construct boolean matrix $F$ such that $F_{i,j} = 1$ when
   $R_{i,j} \geq Th_{rl}$ and $T_{i,j} \geq Th_{tf}$, otherwise $F_{i,j} = 0$
2: For all $i, j$ Construct $G_s$ by adding an edge between $u_i$ and
   $u_j$ where $F_{ij} = 1$
3: Apply DPClusO to $G_s$ using parameters CP, CD, OV and
   generate cluster set $C_1, C_2, \ldots C_r$ where $C_i \subseteq U$.
4: Find out the neighbor nodes in set $V$ for each cluster from
   step 3
5: Calculate the probability set $Pvk$ of all neighbor nodes $v \in V$
   of $k$th cluster.
6: Select neighbor nodes of $k$th cluster by using condition
   $Pv_ik \geq Pvth$
7: Construct bicluster by attaching selected neighbor nodes to
   $k$th cluster
8: Repeat step 5 to 7 for $1 \leq k \leq r$
```

The DPClusO algorithm is mainly used to cluster a simple graph. The DPClusO algorithm takes three input parameters, which are the cluster density (CD), cluster property(CP), and the overlapping coefficient (OV). The details of these parameters can be found in [2]. The DPClusO algorithm generates overlapping clusters with nodes of set U where each cluster has links to nodes from set V. Let the kth cluster C_k have n nodes, $C_k = \{u_1, u_2, u_3 \ldots\ldots\ldots\ldots u_n\}$ where $u_i \in U$. The neighbor of C_k , that is $N(C_k)$ can be written as $N(C_k) = \{N(u_1) \cup N(u_2) \cup N(u_3) \ldots\ldots\ldots\ldots\ldots \cup N(u_n)\}$ where $N(C_k) \subseteq V$. Let the total number of distinct nodes in $N(C_k)$ be m, such that $N(C_k) = vk = \{v_1, v_2, v_3 \ldots\ldots\ldots\ldots v_m\}$ Step 5 in Algorithm 2 finds the probability set Pvk of all $v_i \in V$ nodes in the kth cluster. We calculate the probability of inclusion of v_i in cluster C_k by using the formula

$$Pv_ik = \frac{|N(v_ik)|}{|C_k|}. \tag{1}$$

Step 6 and 7 in Algorithm 2 filter out some of the $v \in V$ nodes by comparing this probability with the threshold and attaching the remaining nodes to the cluster. The algorithm can generate bicliques if they exists by setting $Pvth = 1$.

For clear understanding, here we explain our method by an example, shown in Fig. 1. First, let a simple bipartite graph $G = (U, V, E)$ be represented by the matrix of Fig. 1a. Here $|U| = 7$, $|V| = 10$, $u1$ has the neighbor set $\{v1, v2, v3, v4, v5\}$ and $u2$ has the neighbor set $\{v2, v3, v4, v5, v6\}$. $|N(u1) \cap N(u2)| = 4$ which is the relation number, i.e., the common neighbors between $u1$ and $u2$. The Tanimoto coefficient between these two nodes is $|N(u1) \cap N(u2)|/|N(u1) \cup N(u2)| = 0.66$. The simple graph in Fig. 1b was constructed by considering the elements of set U as nodes and by placing an edge between any node pair when the relation number between them is greater than 0. Next, we get the graph of Fig. 1c from the graph of Fig. 1b by filtering out the edges corresponding to relation number < 2. Fig. 1d then shows the Tanimoto coefficients corresponding to each edge. Then we get the graph of Fig. 1e by filtering out the edges where the Tanimoto coefficient is ≤0.25. Thus, filtering both removes the less important edges and improves the possibility of finding good clusters of densely connected subgraphs. By applying DPClusO algorithm to the graph of Fig. 1e, we get two clusters $\{u1, u2, u3\}$ and $\{u4, u5, u6\}$. To complete the biclusters, we add nodes from set V to these clusters using $Pvth > 0.5$. Finally, as shown in Fig. 1f, we get two biclusters $\{(u1, u2, u3), (v2, v3, v4, v5)\}$ and $\{(u4, u5, u6), (v7, v8, v9)\}$.



	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10
u1	1	1	1	1	1					
u2		1	1	1	1	1				
u3		1	1	1						
u4							1	1	1	1
u5							1	1	1	
u6				1		1	1	1	1	
u7	1									

Fig. 1. An example demonstrating the algorithm of BiClusO. (a) Matrix representation of a bipartite graph. (b) Graph construction by relation number. (c) Filtering by relation number threshold. (d) Generating Tanimoto coefficient of edges. (e) Filtering by Tanimoto threshold and applying DPClusO. (f) Second node attachment to each cluster.

3 Applying BiClusO to Species-VOC Data

To demonstrate the BiClusO algorithm, we applied it to species-VOC relationship data. This section describes that process and the results.

TABLE 2
VOC Data Count According to Different Taxonomy

Kingdom	Phylum	Class	Phylum	Order	Phylum	Family
Bacteria	11	24	43	60	420	773
Eubacteria	1	1	1	1	1	1
Euryarchaeota	1	1	1	1	2	5
Fungi	4	16	22	34	148	598
Plantae	3	5	7	16	139	793

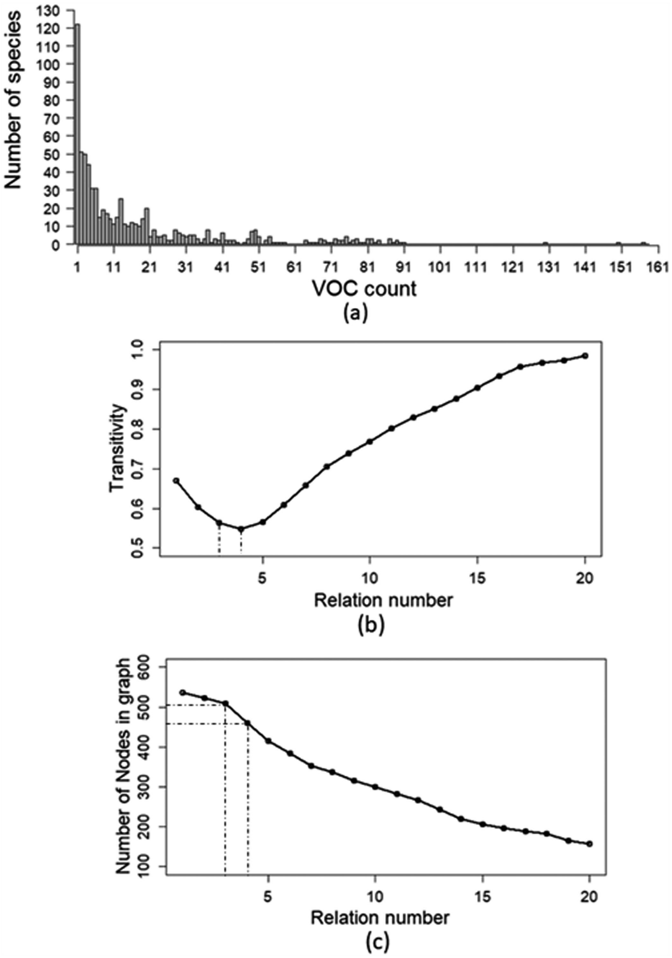


Fig. 2. (a) Frequency of species versus VOC count. (b) Transitivity versus relation number. (c) Node count versus relation number.

3.1 Data Collection and Preprocessing

We collected species-VOC relationship data from the KNApSAcK metabolite ecology section of the KNApSAcK family databases [9]. We also collected additional data from different papers and journals using Google Scholar and other publication sites [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38], [39], [40], [41], [42], [43], [44], [45], [46], [47], [48], [49], [50], [51], [52], [53], [54]. The final data consists of 12,410 species-VOC relations including 710 species and 1,740 different VOCs. We applied our proposed biclustering algorithm to the species-VOC relationship data.

As shown in Table 2, taxonomically, the species are from five different kingdoms. Most of the species belong to the plant, fungi, or bacteria kingdoms.

In most of the literatures, VOCs are described by chemical names. Sometimes a VOC is identified by different names according to different chemical naming conventions. For example, the sweet scent compound α-humulene has four different names; *alpha*-humulene; humulene; *alpha*-caryophyllene; and 3,7,10-humulatriene. For our study, we selected only one name for each compound. In total, the 710 different species with 1,740 distinct VOCs have many-to-many relationships which forms a big bipartite graph. The number of reported VOCs emitted by a species varied between 1 and 157. Fig. 2a shows the frequency of species with respect to the reported number of emitted VOCs based on our database. The uneven distribution of VOCs is due to the fact that some research emphasizes retention time or percentage volume of a small number of emitted VOCs over the complete set of emitted VOCs. Some experiments emphasize an important set of VOCs which have specific activities associated with characteristics such as plant growth, pollinator attraction, resisting enemies, or disease biomarkers. Only some literature reports the complete VOC profiles of species. From Fig. 2a, we can see that 122 species are associated with only one VOC, 51 species with only two VOCs, 50 species with only three VOCs, and so on. For our purposes, we completely discarded those species which are associated with only one or two VOCs. Finally, we produced an input bipartite graph G with dimension 540 x 1710.

3.2 Threshold Selection and Biclustering

In step three of algorithm two, the DPClusO algorithm generates a cluster set using three parameters. Based on our previous experience we set these DPClusO related thresholds as $CD = 0.5$, $CP = 0.5$ and $OV = 0.05$ [20], [21].

Two other important thresholds for the BiClusO algorithm are Th_{rl} and Th_{tf}, which are used to transform the bipartite graph to a simple graph. These thresholds are utilized for the two-step screening of the edges of the simple graph. In this work, we converted the species-VOC bipartite graph to a simple graph where species are the nodes. A relation number is the number of common VOCs between two species. Therefore, a lower value of Th_{rl} allows many edges in the simple graph which do not represent strong relations. Similarly, we can calculate Tanimoto similarity between any two species in the context of the associated VOCs, and a lower Th_{tf} will allow many noisy edges in the simple graph. At the same time, higher values of Th_{rl} and Th_{tf} will make many species as isolated nodes in the graph, and thus exclude them from the analysis. Therefore, we need to handle the trade-off between allowing meaningful edges and keeping a substantial number of species as non-isolated nodes in the simple graph.

In this work, we determined these thresholds based on the characteristics of the input data, i.e., the species-VOC relational data. The first step of screening was done based on Th_{rl}. A simple graph can be generated using different values of Th_{rl}. For example, Fig. 2b shows the plot of Th_{rl} versus the clustering co-efficient, and Fig. 2c shows the plot of Th_{rl} versus the non-isolated species, in the context of the generated simple graphs. In Fig. 2b, initially the clustering coefficient decreases mainly because of the removal of non-important edges. Also, from Fig. 2c, we can see that $Th_{rl} = 3$ allows about 70 more species in the network compared to $Th_{rl} = 4$. In this work, we empirically selected $Th_{rl} = 3$. We then performed the second step of screening based on Th_{tf}. We assume that the relationship between two species is strong if

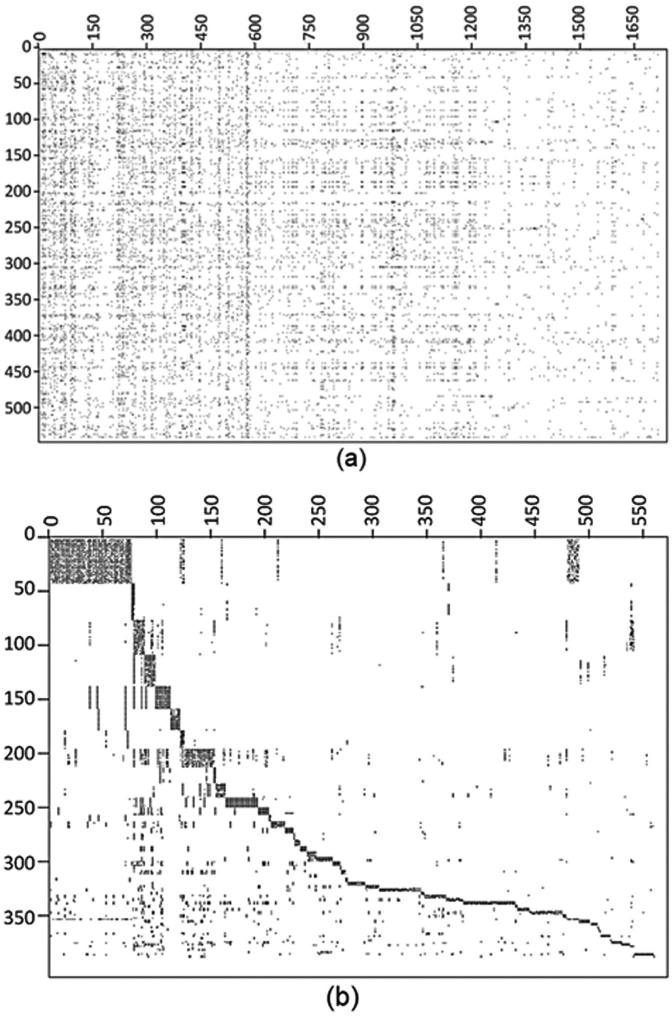


Fig. 3. Graphical presentation of matrix data (a) before and (b) after clustering. X axis represents VOCs and y axis represents species.

at least 50 percent of the VOCs of a species match those of the other species. For such a relationship, Th_{tf} should be greater than 0.33 as we prove in the following theorem.

Theorem 1. *The Tanimoto coefficient between two binary (0,1) vectors is greater than or equal 0.33 if the common number of 1 s between them is greater than or equal 50 percent of the 1 s of any of them.*

Proof. Let the number of 1 s in two binary vectors are x and y where c is the common number of 1 s between them. If c has to be more than or equal to 50 percent of each of x and y then we can write $c \geq x/2$ and $c \geq y/2$. From these conditions, we can write

$$4c \geq x + y,$$

or

$$3c \geq x + y - c.$$

Which is equivalent to

$$\frac{c}{x+y-c} \geq \frac{c}{3c},$$

so, the Tanimoto coefficient $\geq \frac{1}{3}$. □

Adding some margin to 0.33, we empirically selected $Th_{tf} \geq 0.4$ for this study. These empirical thresholds produced good results, as explained in the next section.

After observing different test data, we come to the conclusion that a minimum and reasonable value to start with the Th_{tf} is 0.33. As a simple way, the Th_{rl} can be decided by observing the sparseness/density of the input bipartite graph. For higher density graphs higher Th_{rl} is recommended. For example, the density of the species-VOC bipartite graph we utilized in the present work is roughly 1 percent and we used the Th_{rl} as 3. Usually, most practical graphs are sparse (density less than 5 percent) and using Th_{rl} between 2 to 5 is recommended.

4 Results and Discussion

In this section, we discuss the results produced by applying the proposed biclustering algorithm to the species-VOC relational data. We examine the properties of the clusters identified by this process, similarities between some of the clusters and some of the VOC groups, and the VOC diversity in terms of kingdoms of species identified by these results.

After applying the proposed biclustering algorithm to the species-VOC relational data in this study, we obtained a total of 57 clusters. Fig. 3a shows the matrix representing the species-VOC relational data before clustering, while Fig. 3b shows the same matrix after clustering. Visual comparison of Figs. 3a and 3b indicates that our algorithm efficiently separated the data into clusters. The dimensions of the two matrices in Fig. 3 are not the same because our algorithm filtered out species and VOCs for which there is not enough reported information, which failed to become part of a meaningful bicluster. The VOC set corresponding to each cluster was determined using $Pvth \geq 0.6$, which can be considered as characteristic VOCs of the corresponding taxonomic group.

4.1 Properties of Clusters

In the present case, a bicluster consists of one set of species and one set of VOCs. First, we assessed the richness of the species of similar taxonomic groups in each cluster. In each cluster, the richness of species belonging to similar hierarchical level, mostly at the family level, was determined by Fisher's exact test.

In Supplementary file 1, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2019.2914901, we summarized all the clusters, referred to as $C1$ to $C57$, with their classification, p-value, species, and VOCs determined with $Pvth \geq 0.6$. Out of 57 clusters, 36 clusters have p-values between 3.19E-52 and 9.46E-05, 12 clusters have values between 0.000134 and 0.006641, and 6 clusters have values between 0.019159 and 0.040969. Only 3 clusters have p-values > 0.05. Fig. 4 shows the distribution of p-values. The low p-values for 54 out of 57 clusters are significant, which indicates that the taxonomic classification is consistent with the VOC-based classification.

In the following, we briefly discuss the top 10 clusters based on the p-value and the most common VOCs corresponding to the respective clusters obtained by setting $Pvth = 1$ in most cases, i.e., based on bicliques.

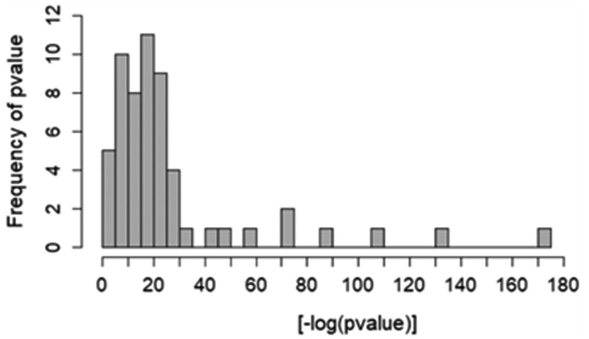


Fig. 4. Frequency of *p*-value of the clusters.

In $C1$, there are 45 species of genus salvia under the lamiaceae family (*p*-value 3.19E-52). The most common volatile compounds for this cluster are cedr-8-en-15-ol; humulene epoxideII; *p*-cymen-8-ol; and *trans*-caryophyllene which form a biclique. The essential oil composition of the salvia species contains medicinal and aromatic compounds. These species are commonly referred to as members of the mint family.

$C2$ is comprised of mostly bacteria in the family enterobacteriaceae (*p*-value 2.38E-32), with the most common VOCs being 1-decanol and 1-dodecanol. All of these species are gram negative bacteria mostly live in animal intestine.

For $C3$, we could not find any bi-cliques, but setting the coefficient $Pvth \geq 0.8$, we identified four common VOCs, 1-heptanol; 1-undecene; 3-methylbutan-1-ol; and hexan-1-ol. $C3$ consists mostly of bacteria from the burkholderiaceae family (*p*-value 5.87E-23). Species in this cluster are mostly pathogenic for humans or animals.

$C4$ (*p*-value 1.16E-40) are mostly streptomycetaceae family bacteria which are commercially used to produce antibiotics, antibacterial, antifungal, and antiparasitic metabolites by their secondary metabolism. The common VOC for this cluster is dimethyl disulfide.

$C5$ (*p*-value 3.87E-23) are mostly bacteria of the leuconostocaceae family. The most common VOCs are 2-methylpropan-1-ol; 2-phenylethanol; 2-phenylpropan-2-ol; 3-(Methylsulfanyl)-1-propanol; 3-methylbutan-2-ol; 3-methylbutanol; benzaldehyde; benzyl alcohol; butanoic acid; decanoic acid; dodecanoic acid; heptanoic acid; n-hexaneic acid; octanoic acid; tetradecanoic acid; and valeric acid. These gram positive bacteria can ferment glucose in the heterofermentative way to produce lactic acid.

$C6$ (*p*-value 1.65E-18) are mostly bacteria of the family prevotellaceae. Most of these bacteria are indigenous to the human and animal gastrointestinal tract and oral cavity. The most common VOCs are 12-methyltetradecanoic acid; 13-methyltetradecanoic acid; 3-hydroxy-15-methylhexadecanoic acid; 3-hydroxyhexadecanoic acid; hexadecanoic acid; and tetradecanoic acid.

$C7$ (*p*-value 1.05E-15) are mostly plants of the cannabaceae family. All of these plants are members of the eudicots class, and the most common VOC is *beta*-caryophyllene.

$C8$ (*p*-value 1.92E-27) are mostly fungi of the hypocreaceae family. These species are characterized as opportunistic avirulent plant symbionts. The most common VOCs are 3-methylbutanal; decanal; ethyl acetate; and nonanal.

$C10$ (*p*-value 3.82E-10) are mostly bacteria in the cyanobacteria phylum with the common characteristic of obtaining energy by photosynthesis. The most common VOC is *beta*-ionone-5,6-epoxide.

$C11$ (*p*-value 1.47E-13) are mostly fungi of the trichocomaceae family. These fungi are found in soil and cause disease in corpses. The most common VOCs are (Z)-2-penten-1-ol; 1-heptanol; 1-octanol; 1-pentanol; 1-penten-3-ol; 2(E)-octenal; 2-amylfuran; 2-heptenal; 2-hexanol; 2-n-butylfuran; 2-nonanone; 2-octanol; 2-octen-1-ol ;2-pentanol; 3-nonen-1-ol,(Z); 3-octanol; 3-octanone; 3-pentanol; 5-octen-1-ol,(Z)-; 5-octen-2-ol; 6-undecanone; chalcogran,(Z); conophthorin; cyclopentanone; heptan-2-ol; heptan-2-one; hexan-2-one; hexan-4-olide; hexanal; hexylformate; hexan-1-ol; nonalactone; octan-2-one; octan-3-ol; pentylhexanoate; propan-1-ol; and (Z)-3-hexen-1-ol.

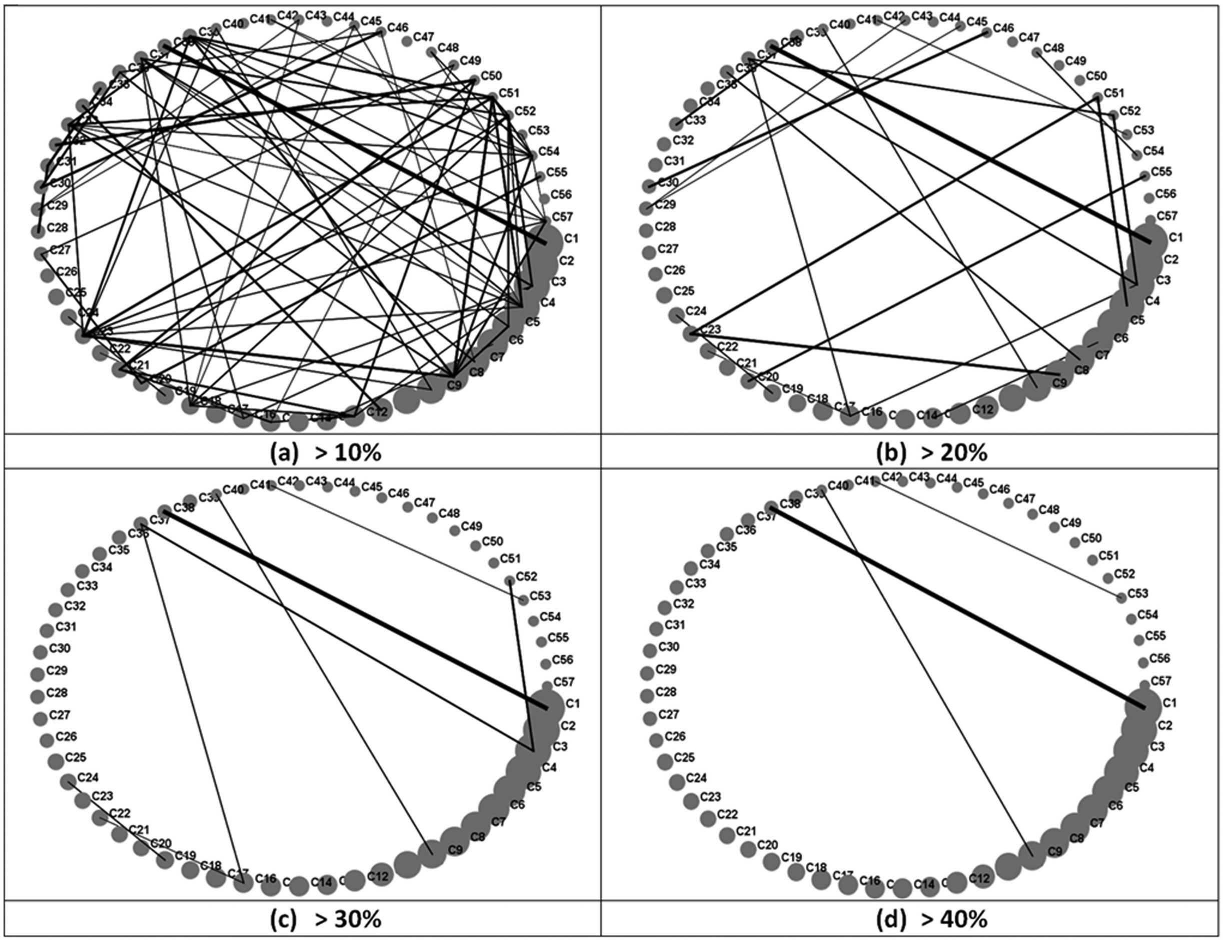
4.2 Relationship Between Clusters

We also assessed the similarity between clusters in terms of shared VOCs. Let a and b be the VOC sets of two clusters, then the percentage of $a \cap b$ in the context of $a \cup b$ is a measure of the Common VOC-based Similarity (*CVSim*) between those two clusters. In Fig. 5, we show the graphs where the nodes indicate the clusters, and the edges indicate a certain minimum *CVSim* between two clusters. In this figure, the size of a node is proportional to the number of species in it. In Fig. 5a, with minimum *CVSim* of 10 percent, we find many edges, implying that many cluster pairs have common VOCs. But, as we increase the *CVSim* threshold, the number of edges gradually decreases. At $CVSim >= 40\%$, there are only three edges linking 3 pair of clusters: $\{C1, C38\}$, $\{C9, C40\}$, and $\{C42, C53\}$. Cluster $C1$ with 41 species and cluster $C38$ with three species belong to the plant kingdom of family lamiaceae. Cluster $C42$ with three species and cluster $C53$ with three species belong to the family enterobacteriaceae in the kingdom of bacteria.

Finally, cluster $C9$ with 15 species and cluster $C40$ with two species belong to the bacteria kingdom in the families bacteroidaceae and veillonellaceae. With *CVSim* more than 35 percent, the cluster $C3$ with 32 species and cluster $C37$ with three species belong to the bacteria kingdom of family burkholderiaceae. Also, in examining these pairs of clusters, we noticed that $|VOC_{C38}| = 74$ and 61 of those are included in VOC_{C1}, $|VOC_{C42}| = 3$ and 2 of them are included in VOC_{C53}, and $|VOC_{C37}| = 8$ and 5 of them are included in VOC_{C3}. Each of these three pairs should be merged into a single cluster because the species belong to the same taxonomical group and the VOCs of one cluster are subsets of another cluster. However, they are divided into 2 clusters mainly because of the lack of enough reported data. Based on such results, we can predict some new species-VOC relations. For example, we can assume that VOCs reported for species of $C1$ are likely to be reported for species of $C38$ and vice versa, and similarly for the other pairs of clusters.

4.3 Structurally Similar VOC Groups (SSVGs)

In this section, we focus on relations between clusters of taxonomically similar species with Structurally Similar VOC Groups. This is important because it can be assumed that structurally similar metabolites may belong to the same or related metabolic pathways. Even though all metabolic pathways in living cells have not been determined, scientists are trying to discover pathways by matching metabolites in



(a) > 10%

(b) > 20%

(c) > 30%

(d) > 40%

Fig. 5. Relations among species clusters in terms of common VOC sets.

terms of their chemical and physical properties [15], [16]. VOC biosynthesis depends on transcriptional regulation by which different genes get involved in VOC emissions. Also, post transcriptional regulation may play important roles which yet to be discovered.

As the largest producers of VOCs, plants play important roles in the natural ecosystem. In plants, most of the VOCs are generated by using four major pathways: mevalonic acid (MVA), methylerythritol phosphate (MEP), Shikimate or phenylalanine, and lipoxygenase (LOX)[6], [17].

A metabolic pathway is generally defined as some of the consecutive steps of biochemical reactions, catalyzed by enzymes that occur inside a cell. In our study, there were a total of 505 VOCs included in the 57 biclusters generated from the species-VOC relational data. We generated PubChem IDs from the names of the VOCs and then downloaded SDF files for those VOCs and converted them to atom pair fingerprints (APFP) by ChemmineR [22] using the functions "sdf2ap" and "desc2fp" with default parameters. They compute the top 1,024 out of 4,096 most common atom pairs observed in the compound collection from DrugBank. To determine the structurally similar VOC pairs, we applied the "Tversky" similarity function to those APFP with $cut\ off > 0.85$, *alpha* = 1 and *beta* = 1. These coefficient values are recommended as the best to measure similarity among compounds [18], [19]. Thus we constructed a network of VOCs based on the chemical structure similarities between them. Then we applied the DPClusO tools [20] to create clusters from this network with $CD = 0.7$, $CP = 0.5$ and $OV = 0.05$. Doing this clustered the total 505 VOCs into 256 SSVGs, of which 62 are of size 2 or more and the rest are single VOCs. Supplementary Fig. 1a, available online, shows the cluster result (SSVGs) with at least two VOCs. Supplementary Fig. 1b, available online, shows the chemical 2d structure of a candidate element in each of the SSVGs with size of 3 or more, where we can visually see that the structures corresponding to different SSVGs are different.

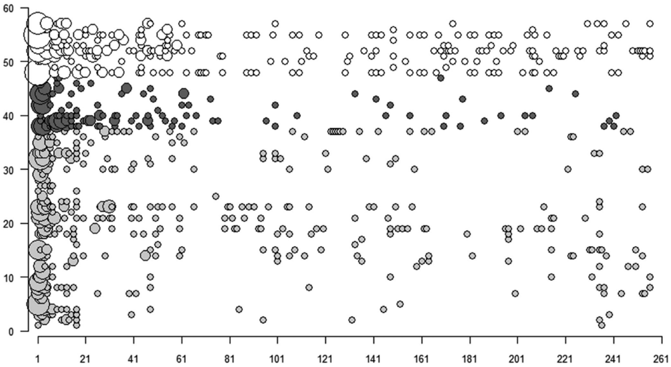


Fig. 6. Scatter plot of SSVGs for plant (White), fungi (Black), and bacteria (Gray). X axis represents SSVGs and y axis represents species clusters.

Fig. 6 shows the mapping of the VOCs belonging to 57 clusters to 256 SSVGs. When one or more VOCs belonging to

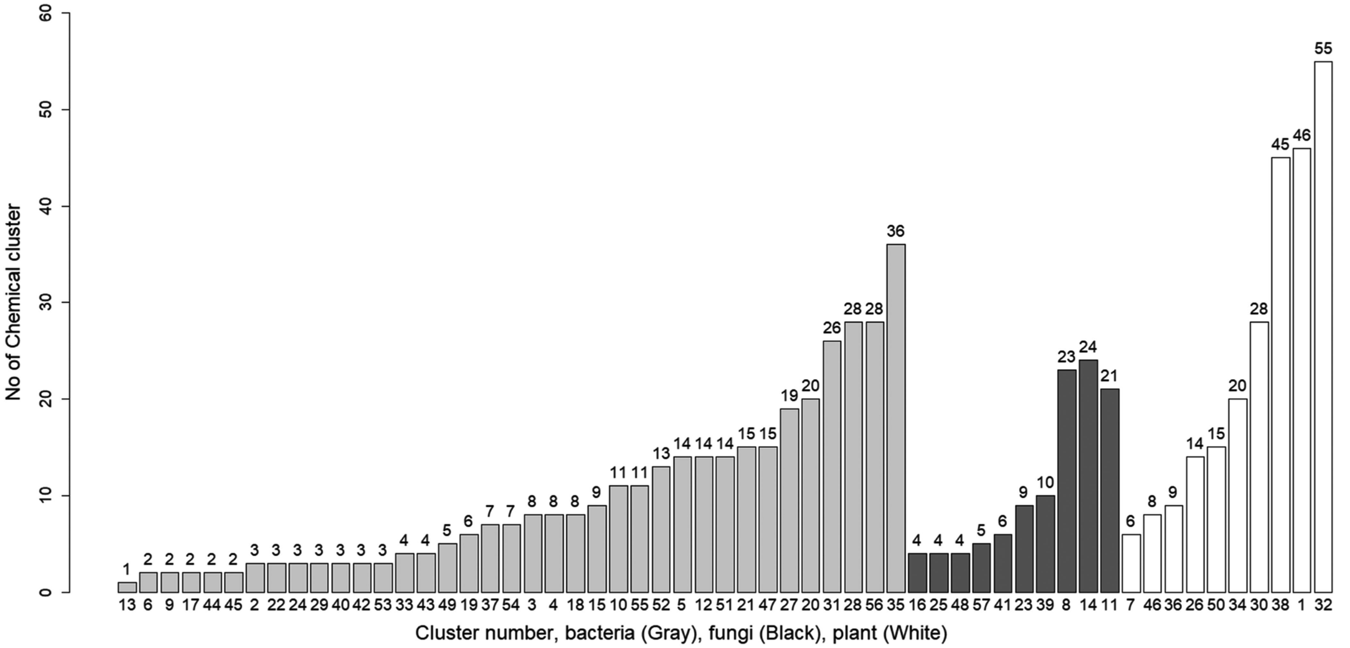


Fig. 7. Number of associated SSVGs in different species groups.

a cluster matched the VOCs of a SSVG, we placed a circle at the corresponding location on the map. The size of a circle on the map in Fig. 6 is proportional to the number of shared VOCs between a cluster and a SSVG. The color of a circle is white, black, or gray depending on whether the corresponding cluster belongs to the kingdom of plants, fungi, or bacteria. From the mapping, cluster $C1$ is linked to 46 SSVGs of which the first 9 groups are $SSVG1$, $SSVG8$, $SSVG11$, $SSVG18$, $SSVG21$, $SSVG27$, $SSVG26$, $SSVG28$ and $SSVG32$. In Supplementary Fig. s(b), available online, the representative compounds for these 9 groups are displayed with labels 1, 8, 11, 18, 21, 27, 26, 28, and 32. The sparsity of the map in Fig. 6 implies that the species of different taxonomical groups produce different types of VOCs in terms of chemical structure.

It can be hypothesized that structurally different VOCs are produced by different metabolic pathways. Thus, it may be suggested that the characteristic VOC pathways in different taxonomical groups are different. Such pathways evolved depending on the needs for survival and adaptation in the environment for certain classes of organisms. Further investigation into species-specific VOCs can provide more clues to the interaction of the species with the ecosystem.

4.4 VOC Diversity Across Three Kingdoms

Using the first taxonomical hierarchy of species, the kingdom, we classified the biclusters into three groups, i.e., plants, fungi, and bacteria. First, we compared plants, fungi, and bacteria in terms of the abundance of associated characteristic SSVGs. Out of 57 clusters, 10 clusters belong to plants, 10 belong to fungi, and 37 belong to bacteria. The histogram in Fig. 7 shows the number of associated SSVGs with different clusters in the three kingdoms.

To determine the differences between these three kingdoms in terms of the number of associated SSVGs to individual clusters, we performed a t-test. Based on the t-test, the p-values are: plants-fungi, 0.04934; plants-bacteria, 0.0295; and bacteria-fungi, 0.06914. From the results of the t-test, we can say that the groups of plants have more diversified VOCs compared to individual groups of bacteria or fungi. This is consistent with the fact that plants are a more advanced species compared to bacteria and fungi, and thus require more diversified VOCs for survival. Also, plants have different organs such as roots, leaves, stems, flowers, and fruits which can produce different amounts and types of VOCs at different times, e.g., different climate seasons, flowering seasons, diurnal and nocturnal times, for different purposes. Also, we conducted a t-test with p-value 0.04028 between eukaryotes and prokaryotes, i.e., bacteria versus combined clusters related to plant and fungi. The p-value indicates that individual groups of eukaryotes produce more diversified VOCs than individual groups of prokaryotes.

We also assessed the diversity of VOCs across the three kingdoms in terms of different chemical structures. Generally, each of the SSVGs can be considered as a group of chemical compounds related to similar or related metabolic pathways. In our dataset, most clusters were generated involving bacteria because we found more data related to bacteria in published literature. Separately, plants are associated with 136 SSVGs, fungi with 58 SSVGs, and bacteria with 148 SSVGs. Although individual bacterial groups are associated with smaller numbers of SSVGs, the total number of SSVGs related to bacterial clusters is the highest. This is because diversified types of bacteria live under diversified environments requiring diversified types of VOCs. Also, plants are associated with more diversified types of VOCs compared to fungi because plants are a more advanced species than fungi.

The Venn diagram in Fig. 8 shows the sharing of SSVGs across the kingdoms of plants, fungi, and bacteria. The evolutionary hierarchy of the tree of life reveals the gradual decline or uptake of biochemical traits by descendant species from their ancestors. The tree of life generated using DNA data from 3,000 species [23] reveals that bacteria are a predecessor to both fungi and plants, and that plants are a predecessor to fungi. DNA, converted to mRNA, creates different types of

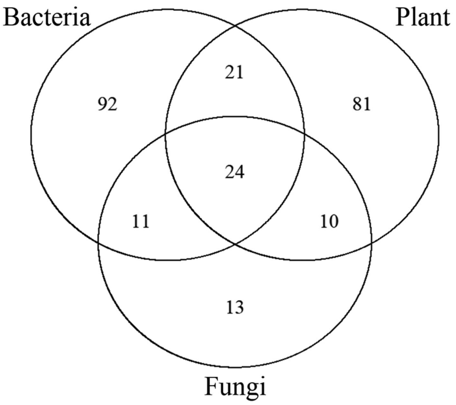


Fig. 8. Venn diagram showing the overlap of different SSVGs among three kingdoms.

proteins, and proteins engaged in different biochemical reactions eventually generate VOCs. Fig. 8 shows that many VOC pathways are conserved in all three kingdoms, or between any two kingdoms. The number of pathways conserved between plants and bacteria are more than those conserved between bacteria and fungi or plants and fungi. Plants and fungi probably acquired different VOC pathways from different bacterial species at different times through different interactions by horizontal gene transfers. Fig. 8 shows plants are related to 81 unique SSVGs, fungi to 13, and bacteria to 92. Different bacteria species living in many different types of microenviroments developed many unique types of VOCs. Plants being more advanced (need to handle different biological processes) and sessile (face different challenges of environmental stress) organisms developed many unique VOCs for their survival. Fungi are less advanced organisms and different types of fungi live in very similar environments, so they can survive with a somewhat smaller number of VOCs.

5 Conclusion

In this study, we developed a novel biclustering algorithm called BiClusO and applied it to species-VOC bipartite relational data to understand the diversity of VOCs and VOC pathways across species in different kingdoms. We developed BiClusO based on a data folding mechanism and the DPClusO algorithm previously developed by our group. By controlling different coefficients, our algorithm can identify high density biclusters as well as bicliques. We applied the BiClusO algorithm to species-VOC relationship data to identify groups of species in terms of common VOCs. Based on Fishers exact test we showed that VOC-based classification of species is consistent with their taxonomical classification. Inter-cluster relationships identified in terms of VOCs can help to predict the VOCs of new species in a similar taxonomical group. Furthermore, along with the algorithm, we used DPClusO and ChemmineR to identify structurally similar VOC groups. The relationships of those SSVGs to different groups of species imply that common VOC pathways in different taxonomical groups are different. Based on the results of the *t*-test, we showed that individual groups of plants are associated to more VOCs compared to groups of bacteria. Further analysis of the relations of these groups of species with SSVGs revealed the diversity of VOCs across the kingdoms of plants, fungi, and bacteria. Pairwise overlapping of SSVGs among three kingdoms reveals the conservation of evolutionary hierarchy in terms of biochemical traits.

Acknowledgments

This work was supported by the Ministry of Education, Culture, Sports, Science, and Technology of Japan (16K07223 and 17K00406) and NAIST Big Data Project and was partially supported by Platform Project for Supporting Drug Discovery and Life Science Research funded by the Japan Agency for Medical Research (18am0101111) and Development and the National Bioscience Database Center in Japan. The authors heartily thank Professor Michael Dean Barker of NAIST for his cordial cooperation.

References

[1] A. Muto-Fujita, K. Takemoto, S. Kanaya, T. Nakazato, T. Tokimatsu, N. Matsumoto, M. Kono, Y. Chubachi, K. Ozaki, and M. Kotera, “Data integration aids understanding of butterfly-host plant networks,” *Sci. Rep.*, vol. 7, 6 Mar. 2017, Art. no. 43368. doi: 10.1038/srep43368

[2] M. Altaf-Ul-Amin, et al., “Development and implementation of an algorithm for detection of protein complexes in large interaction networks,” *BMC Bioinf.*, vol. 7, no. 1, 2006, Art. no. 207.

[3] M. Altaf-Ul-Amin, M. Wada, and S. Kanaya, “Partitioning a PPI network into overlapping modules constrained by high-density and periphery tracking,” *Int. Scholarly Res. Netw.*, vol. 2012, 2012, Art. no. 726429.

[4] 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 Chemistry*, vol. 7, pp. 150–156, 2006.

[5] Y. Cheng and G. M. Church, “Biclustering of expression data,” in *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 2000, pp. 93–103.

[6] N. Dudareva, et al., “Biosynthesis, function and metabolic engineering of plant volatile organic compounds,” *New Phytologist*, vol. 198, no. 1, pp. 16–32, 2013.

[7] Y. Nakamura, et al., “KNApSAcK metabolite activity database for retrieving the relationships between metabolites and biological activities,” *Plant Cell Physiology*, vol. 55, no. 1, pp. e7–e7, 2014.

[8] F. M. Afendi, N. Ono, Y. Nakamura, K. Nakamura, L. K. Darusman, N. Kibinge, A. H. Morita, et al., “Data mining methods for omics and knowledge of crude medicinal plants toward big data biology,” *Comput. Structural Biotechnol. J.*, vol. 4, no. 5, 2013, Art. no. e201301010.

[9] A. A. Abdullah, M. Altaf-Ul-Amin, N. Ono, T. Sato, T. Sugiura, A. H. Morita, T. Katsuragi, A. Muto, T. Nishioka, and S. Kanaya, “Development and mining of a volatile organic compound database,” *BioMed Res. Int.*, vol. 2015, 2015, Art. no. 139254.

[10] A. Tanay, R. Sharan, and R. Shamir, “Discovering statistically significant biclusters in gene expression data,” *Bioinf.*, vol. 18, no. suppl_1, pp. S136–S144, 1 Jul. 2002.

[11] S. Hochreiter, U. Bodenhofer, M. Heusel, A. Mayr, A. Mitterecker, A. Kasim, T. Khamiakova, S. Van Sanden, D. Lin, W. Talloen, L. Bijnens, H. W. H. Göhlmann, Z. Shkedy, and D.-A. Clevert, “FABIA: Factor analysis for bicluster acquisition,” *Bioinf.*, vol. 26, no. 12, pp. 1520–1527, 15 Jun. 2010.

[12] A. Preli, S. Bleuler, P. Zimmermann, A. Wille, P. Bhlmann, W. Gruissem, L. Hennig, L. Thiele, and E. Zitzler, “A systematic comparison and evaluation of biclustering methods for gene expression data,” *Bioinf.*, vol. 22, no. 9, pp. 1122–1129, 1 May 2006.

[13] M. Koyuturk, W. Szpankowski, and A. Grama, “Bi-clustering gene-feature matrices for statistically significant dense patterns,” in *Proc. IEEE Comput. Syst. Bioinf. Conf.*, 2004, pp. 480–484.

[14] M. van Uitert, W. Meuleman, and L. Wessels, “Bi-clustering sparse binary genomic data,” *J. Comput. Biol.*, vol. 15, no. 10, pp. 1329–1345, 2008.

[15] Y. D. Cai, et al., “Prediction of compounds biological function (metabolic pathways) based on functional group composition,” *Mol. Diversity*, vol. 12, no. 2, pp. 131–137, 2008.

[16] M. A. Hamdalla, et al., “Metabolic pathway predictions for metabolomics: A molecular structure matching approach,” *J. Chemical Inf. Model.*, vol. 55, no. 3, pp. 709–718, 2015.

[17] C. Sanz and A. G. Prez, "Plant metabolic pathways and flavor biosynthesis," in *Handbook of Fruit and Vegetable Flavors*. Hoboken, NJ, USA: Wiley, 2010, pp. 129–155.

[18] M. Dunkel, S. Gnther, J. Ahmed, B. Wittig, and R. Preissner, "SuperPred: Drug classification and target prediction," *Nucleic Acids Res.*, vol. 36, no. Web Server issue, pp. W55–W59, 2008, doi: 10.1093/nar/gkn307.

[19] H. Matter, "Selecting optimally diverse compounds from structure databases: A validation study of two-dimensional and three-dimensional molecular descriptors," *J. Medicinal Chemistry*, vol. 40, no. 8, pp. 1219–1229, 11 Apr. 1997.

[20] M. B. Karim, N. Wakamatsu, and M. Altaf-Ul-Amin, "[Dedicated to Prof. T. Okada and Prof. T. Nishioka: Data science in chemistry] DPClusOST: A software tool for general purpose graph clustering," *J. Comput. Aided Chemistry*, vol. 18, pp. 76–93, 2017.

[21] R. Eguchi, M. B. Karim, P. Hu, T. Sato, N. Ono, S. Kanaya, and M. Altaf-Ul-Amin, "An integrative network-based approach to identify novel disease genes and pathways: A case study in the context of inflammatory bowel disease," *BMC Bioinf.*, vol. 19, no. 1, 2018, Art. no. 264.

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

[23] E. Pennisi, "Modernizing the tree of life," *Sci.*, vol. 300, pp. 1692–1697, 2003.

[24] S. D. Hatipoglu, et al., "Determination of volatile organic compounds in fourty five Salvia species by thermal desorption-GC-MS technique," *Rec. Natural Products*, vol. 10, no. 6, 2016, Art. no. 659.

[25] S. Lee, et al., "Volatile organic compounds emitted by Trichoderma species mediate plant growth," *Fungal Biol. Biotechnol.*, vol. 3, no. 1, 2016, Art. no. 7.

[26] V. R. Hinge, H. B. Patil, and A. B. Nadaf, "Aroma volatile analyses and 2AP characterization at various developmental stages in Basmati and Non-Basmati scented rice (Oryza Sativa L.) cultivars," *Rice*, vol. 9, no. 1, 2016, Art. no. 38.

[27] R. Ascrizzi, et al., "Patterns in volatile emission of different aerial parts of caper (Capparis spinosa L.)," *Chemistry Biodiversity*, vol. 13, no. 7, pp. 904–912, 2016.

[28] Y. Kong, et al., "Floral scents produced by Lilium and Cardiocrinum species native to China," *Biochemical Systematics Ecology*, vol. 70, pp. 222–229, 2017.

[29] N. Wiebelhaus, N. M. Kreitals, and J. R. Almirall, "Differentiation of marijuana headspace volatiles from other plants and hemp products using capillary microextraction of volatiles (CMV) coupled to gas-chromatography–mass spectrometry (GC–MS)," *Forensic Chemistry*, vol. 2, pp. 1–8, 2016.

[30] M. Kusano, et al., "Unbiased profiling of volatile organic compounds in the headspace of Allium plants using an in-tube extraction device," *BMC Res. Notes*, vol. 9, no. 1, 2016, Art. no. 133.

[31] H. A. Yamani, et al., "Antimicrobial activity of Tulsi (Ocimum tenuiflorum) essential oil and their major constituents against three species of bacteria," *Frontiers Microbiol.*, vol. 7, 2016, Art. no. 681.

[32] Y. C. Hoe, et al., "Flowering mechanisms, pollination strategies and floral scent analyses of syntopically coflowering Homalomena spp. (Araceae) on Borneo," *Plant Biol.*, vol. 18, no. 4, pp. 563–576, 2016.

[33] S. V. Zhigzhitzhapova, et al., "Chemical composition of volatile organic compounds of Artemisia vulgaris L.(Asteraceae) from the Qinghai–Tibet Plateau," *Ind. Crops Products*, vol. 83, pp. 462–469, 2016.

[34] D. Domik, et al., "A Terpene synthase is involved in the synthesis of the volatile organic compound Sodorifen of Serratia plymuthica 4Rx13," *Frontiers Microbiol.*, vol. 7, 2016, Art. no. 737.

[35] S. Erbas and H. Baydar, "Variation in scent compounds of oil-bearing rose (Rosa damascena Mill.) produced by headspace solid phase microextraction, hydrodistillation and solvent extraction," *Rec. Natural Products*, vol. 10, no. 5, 2016, Art. no. 555.

[36] A. B. Santos, et al., "Biogeneration of volatile organic compounds produced by Phormidium autumnale in heterotrophic bioreactor," *J. Appl. Phycology*, vol. 28, no. 3, pp. 1561–1570, 2016.

[37] S. Torbati, A. Movafeghi, and D. J. Djozan, "Identification of volatile organic compounds released from the leaves and flowers of Artemisia austriaca using the modified pencil lead as a fibre of solid phase microextraction," *J. Essential Oil Bearing Plants*, vol. 19, no. 5, pp. 1224–1233, 2016.

[38] A. Karmakar, A. Mukherjee, and A. Barik, "Floral volatiles with colour cues from two cucurbitaceous plants causing attraction of Aulacophora foveicollis," *Entomologia Experimentalis et Applicata*, vol. 158, no. 2, pp. 133–141, 2016.

[39] R. Cozzolino, et al., "Determination of volatile organic compounds in the dried leaves of Salvia species by solid-phase microextraction coupled to gas chromatography mass spectrometry," *Natural Product Res.*, vol. 30, no. 7, pp. 841–848, 2016.

[40] M. De Vrieze, et al., "Volatile organic compounds from native potato-associated Pseudomonas as potential anti-oomycete agents," *Frontiers Microbiol.*, vol. 6, 2015, Art. no. 1295.

[41] N. Killiny and S. E. Jones, "Profiling of volatile organic compounds released from individual intact juvenile and mature citrus leaves," *J. Plant Physiology*, vol. 208, pp. 47–51, 2017.

[42] D. M. Jaeger, J. B. Runyon, and B. A. Richardson, "Signals of speciation: Volatile organic compounds resolve closely related sage brush taxa, suggesting their importance in evolution," *New Phytologist*, vol. 211, no. 4, pp. 1393–1401, 2016.

[43] O. Liarzi, et al., "Use of the endophytic fungus Daldinia cf. concentrica and its volatiles as bio-control agents," *PloS One*, vol. 11, no. 12, 2016, Art. no. e0168242.

[44] U. Groenhagen, et al., "Coupled biosynthesis of volatiles and salinosporamide A in Salinispora tropica," *ChemBioChem*, vol. 17, no. 20, pp. 1978–1985, 2016.

[45] Y. Li, et al., "Volatile organic compounds emissions from Luculia pinceana flower and its changes at different stages of flower development," *Molecules*, vol. 21, no. 4, 2016, Art. no. 531.

[46] P. R. R. Mesquita, et al., "Discrimination of Eugenia uniflora L. bio-types based on volatile compounds in leaves using HS-SPME/GCMS and chemometric analysis," *Microchemical J.*, vol. 130, pp. 79–87, 2017.

[47] B. L. SánchezOrtiz, et al., "Antifungal, anti-oomycete and phytotoxic effects of volatile organic compounds from the endophytic fungus Xylaria sp. strain PB3f3 isolated from Haematoxylon brasiletto," *J. Appl. Microbiol.*, vol. 120, no. 5, pp. 1313–1325, 2016.

[48] E. A. Estrella-Parra, et al., "Volatile organic compounds from Pachyrhizus ferrugineus and Pachyrhizus erosus (Fabaceae) leaves," *Boletn Latinoamericano y del Caribe de Plantas Medicinales y Aromáticas*, vol. 15, no. 3, pp. 175–181, 2016.

[49] C. Thongpoon and T. Machan, "Aroma volatile composition of millingtonia hortensis Linn. F. Flower growing in Chiang Rai, Thailand," *Asian J. Chemistry*, vol. 28, no. 2, 2016, Art. no. 329.

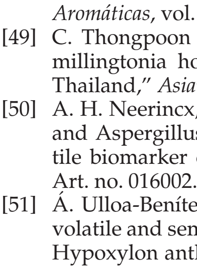
[50] A. H. Neerincx, et al., "Identification of Pseudomonas aeruginosa and Aspergillus fumigatus mono and co-cultures based on volatile biomarker combinations," *J. Breath Res.*, vol. 10, no. 1, 2016, Art. no. 016002.

[51] Á. Ulloa-Benítez, et al., "Phytotoxic and antimicrobial activity of volatile and semi-volatile organic compounds from the endophyte Hypoxylon anthochroum strain Blaci isolated from Bursera lancifolia (Burseraceae)," *J. Appl. Microbiol.*, vol. 121, no. 2, pp. 380–400, 2016.

[52] F. J. Cuevas, et al., "Effect of management (organic vs conventional) on volatile profiles of six plum cultivars (Prunus salicina Lindl.). A chemometric approach for varietal classification and determination of potential markers," *Food Chemistry*, vol. 199, pp. 479–484, 2016.

[53] M. Kupska and H. H. Jeleń, "In-tube extraction for the determination of the main volatile compounds in Physalis peruviana L," *J. Separation Sci.*, vol. 40, no. 2, pp. 532–541, 2017.

[54] W. Raza, et al., "Response of tomato wilt pathogen Ralstonia solanacearum to the volatile organic compounds produced by a biocontrol strain Bacillus amyloliquefaciens SQR-9," *Sci. Rep.*, vol. 6, 2016, Art. no. 24856.



Mohammad Bozlul Karim received the MEng degree in information and communication technology from the Bangladesh University of Engineering and Technology (BUET), Dhaka, Bangladesh, in 2010. He is currently working toward the PhD degree in graph clustering and its application to different biological network analyses in the Nara Institute of Science and Technology (NAIST), Nara, Japan.

Ming Huang (M'12) received the PhD degree in biomedical engineering from the University of Aizu, Aizu, Japan, in 2012. He is currently an assistant professor with the Nara Institute of Science and Technology and a visiting scholar with the Biomedical Engineering Department, University of California, Davis. His research interests include health informatics based on bioinformatics for health promotion and disease management. He is a member of the IEEE.

Naoaki Ono received the PhD degree in complex systems from the University of Tokyo, Tokyo, Japan, in 2001. He was a postdoctoral researcher with the Biophysics Department, Kyoto University, Kyoto, Japan, in 2001. He worked as a research scientist with the Advanced Telecommunication Research Institute, Kyoto, Japan, from 2002 to 2005. He was a research assistant with the Department of Information Science, Osaka University, Osaka, Japan, from 2006 to 2012. He moved to the Department of Information Science, Nara Institute of Science and Technology (NAIST), Nara, Japan, as an assistant professor, in 2012. He is currently an associate professor with the Data Science Center, NAIST. He coauthored three books in complex systems, network analysis, and bioinformatics. He has published about 100 journal and conference papers. His current interest include application of statistical learning methods for bioinformatics and systems biology.

Shigehiko Kanaya received the BSc degree in bio-science from the Science University of Tokyo, Japan, in 1985, and the PhD degree from the Toyohashi University of Technology, Japan, in 1990. He served as an assistant professor in information engineering with Yamagata University, in 1990, guest associate professor with the National Institute of Genetics, in 1996, associate professor with Electronic and Information Engineering, in 1999, associate professor with the Applied Bio System Engineering, Yamagata University, in 2000, guest researcher with Bio Radical Institute (Yamagata Prefecture), in 2000, associate professor with the Research and Education Center for Genetic Information, NAIST, in 2001, associate professor with the Graduate School of Information Science, NAIST, in 2002, professor with the Graduate School of Information Science, NAIST, in 2004, and is currently working as a professor with the Nara Institute of Science and Technology, Japan.

Md. Altaf-Ul-Amin received the BSc degree in electrical and electronic engineering from the Bangladesh University of Engineering and Technology (BUET), Dhaka, the MSc degree in electrical, electronic and systems engineering from the University Keban Gsaan Malaysia (UKM), and the PhD degree from the Nara Institute of Science and Technology (NAIST), Japan. He previously worked in several universities in Bangladesh, Malaysia, and Japan. Currently, he is working as an associate professor with the Computational Systems Biology Lab, NAIST. He is conducting research on network biology, systems biology, cheminformatics, and biological databases.