NAIST-IS-DD1461205

Doctoral Dissertation

Data-Intensive Science of Relationships Among Species, Volatile Organic Compounds and Biological Activities

Azian Azamimi binti Abdullah

March 16, 2017

Department of Applied Informatics
Graduate School of Information Science
Nara Institute of Science and Technology
Japan

A Doctoral Dissertation
submitted to the Graduate School of Information Science,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
Doctor of Engineering

Thesis Committee:

Professor Shigehiko Kanaya	(Supervisor)
Professor Keiichi Yasumoto	(Co-supervisor)
Associate Professor Md. Altaf-Ul-Amin	(Co-supervisor)
Assistant Professor Naoaki Ono	(Co-supervisor)

Data-Intensive Science of Relationships Among Species, Volatile Organic Compounds and Biological Activities[1]

Azian Azamimi binti Abdullah

Abstract

Volatile organic compounds (VOCs) are small molecules with low molecular weight that exhibit high vapor pressure under ambient conditions. In this study, we have developed a VOC database of microorganisms, fungi, plants as well as human being, which comprises the relation between emitting species, VOC, and their biological activities. We have deposited the VOC data into KNApSAcK Metabolite Ecology Database and this database is currently available online. Accumulated data are divided into two types: (1) microorganisms species-VOC relations, and (2) emitting species-VOC-biological activity relations. Initially, we performed hierarchical clustering and graph clustering by DPClus algorithm to extract clusters of microorganisms based on VOC similarity. Both clustering results indicated that VOC based classification of microorganisms is consistent with their classification based on pathogenicity. For the second data, we performed heatmap clustering utilizing Tanimoto coefficient as the similarity index between chemical structures to cluster all VOCs. We further accessed the statistical significance of the clusters using hypergeometric *p*-values to understand the relationships between chemical structures of VOCs and their biological activities. Additionally, we also compared several types of hierarchical clustering methods with DPClus clustering to classify VOCs using fingerprint-based similarity measure between chemical structures. Our research indicates that similar chemical structures of VOCs indicate possibilities of exhibiting similar biological activities. We extended our findings by using supervised machine learning methods to predict biological activities of VOCs based on chemical structures. We have developed 72 classification models for the prediction of biological activities of VOCs by 9 types of fingerprints and trained by Deep Neural Network (DNN), Gradient Boosting Machine (GBM), Random Forest

(RF) and Generalized Linear Model (GLM). Based on our computational results, PubChem fingerprints trained with GBM method are suggested to be used as the input for the prediction compared to other fingerprints and machine learning methods. Generally, GBM method can outperform DNN in term of classifying VOCs. GBM method has advantage in term of computational speed and requires less parameter for optimization. Hence, we highly recommend using GBM method for the prediction of biological activities of VOCs based on chemical structures.

Keywords: Volatile organic compounds, species, biological activity, clustering

[1] Doctoral Dissertation, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DD1461205, March 16, 2017.

Acknowledgements

First and foremost, I thank Allah, the Lord of the Worlds, the Beneficent, and the Most Merciful, for giving me the light and for enabling me to complete this dissertation. There were many obstacles during this journey, but Alhamdulillah with his help and will, I manage to go through it.

I owe special thanks to my supervisor, Professor Shigehiko Kanaya, for his supervision and continuous support. He always taught me on how to be a good researcher and a scientist.

I also want to express my gratitude for Professor Keiichi Yasumoto for taking his time to review my thesis and for his insightful recommendations.

My thanks go to Associate Professor Md. Altaf-Ul-Amin, Assistant Professor Naoaki Ono and Assistant Professor Tetsuo Sato for all their support, valuable comments and suggestions. I also want to extend particular thanks to Professor Takaaki Nishioka for his knowledge sharing in biochemistry, Associate Professor Tadao Sugiura, and Assistant Professor Ming Huang. I am also grateful to Mrs. Minako Ohashi for her help in administrative matter and Mrs. Aki Hirai Morita for her support in developing KNApSAcK Family Databases.

I wish to thank all my fellow lab members in Computational Systems Biology Laboratory especially Dr. Sony Hortono Wijaya, Dr. Kibinge Nelson Kipchirchir, Dr. Tetsuo Katsuragi and Ms. Lidwina Ayu Andarini for their support in my studies. I am really grateful for having such a good place for conducting research.

I also would like to express my appreciation to all staff in NAIST International Student Affairs Section for their contribution throughout my studies. I also owe tremendous debt of gratitude to all my friends in Japan and Malaysia for their continuous support and motivation.

I also wish to thank to Universiti Malaysia Perlis and Ministry of Education

Malaysia for their financial support during my doctoral course in Japan. The scholarship has offered me a wide knowledge and new experience.

Last but not least, my special and highest appreciation to all my family members, especially for my parents, Mr. Abdullah Ibrahim and Mrs. Che Mah Yahya for their unconditional love, support and prayers for my success. I also wish to express my gratitude to my late husband, Mr. Ahmad Faiz Ahmad Najah for his last blessing and love. He is my main motivation to undertake and complete this doctoral study. May Allah bless his soul, forgive him of his sins and grant him the highest levels of paradise. Ameen.

March 16, 2017
Azian Azamimi binti Abdullah

List of Abbreviations

VOC	Volatile organic compound
SAR	Systemic acquired resistance
GCMS	Gas chromatography mass spectrometry
PTR-MS	Proton transfer reaction mass spectrometry
SIFT-MS	Selected ion flow tube mass spectrometry
GC-O	Gas chromatography olfactometry
DB	Database
DNN	Deep Neural Network
GBM	Gradient Boosting Machine
RF	Random Forest
GLM	Generalized Linear Model
MRSA	Methilin-resistant Staphylococcus aureus
PGPR	Plant growth-promoting rhizobacteria
mVOC	Microbial volatile organic compounds
QSAR	Quantitative structure–activity relationships
QSPR	Quantitative structure–property relationships
e-nose	Electronic nose
1D	One-dimensional
2D	Two-dimensional
3D	Three-dimensional
SDF	Structure definition file
MSE	Mean squared error
InChIKey	IUPAC International Chemical Identifier

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This dissertation summarizes the author's research experience in developing a novel volatile organic compound (VOC) database and analyzing the relationships among species, volatile organic compounds and biological activities. This study also attempts to predict the biological activity of volatile organic compound based on their chemical structures by using various machine-learning methods. This chapter describes the general background, the research problem, and research objectives and also explains what are to be expected from the rest of the dissertation.

1.1. Background

Metabolomics is the scientific study of quantification of low mass compounds 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 functions, such as inter- or intra-species communication, antifungal, antimicrobial activities and also as a defense against pests and pathogens (Agostini-costa et al. 2012). Secondary metabolites are often colored, fragrant, or flavorful compounds and largely fall into three classes of compounds: alkaloids, terpenoids and phenolics. Small proportions produced by these secondary metabolites are volatile organic compounds (VOCs) that play important roles in chemical ecology and human healthcare.

VOCs can be defined as small compounds ranging in between C5 to C20 carbon count with a molecular weight in the range of 50 to 200 Daltons (Rowan

2011). They comprise of a diverse chemical group of organic compounds with various biological functions and have high vapor pressures under ambient conditions. Their high vapor pressure results from a low boiling point, which causes large numbers of molecules to evaporate from the liquid or solid form of the compound and enter the surrounding air, a trait known as volatility. Living organisms including human, animals, microorganisms and plants produce VOCs naturally. The naturally produced VOCs play important roles in communication between plants and they also serve as signaling molecules by passing information between organisms. For human and other animals, VOCs are important as scents and flavor of food. Recently, an increased number of researchers are utilizing VOCs as a biomarker to identify various kinds of diseases. Here, we elaborate further details of the importance of VOCs for living organisms specifically in chemical ecology, agriculture and human healthcare.

(a) Chemical ecology

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. The total amount of VOCs emitted globally to the atmosphere is estimated to exceed 1 Pg per year, and these VOCs include mainly plant-produced VOCs, isoprene, monoterpenes and other oxygenated carbon compounds, such as herbivore-induced volatiles and green leaf volatiles (Iijima 2014). Many studies have been performed that showed the emission of VOCs from plants occur as significant cues, signals, or defense responses to wounding, herbivore infestation, pathogen infection, and pollination. The emitted VOCs are responsible for internal and external communication between plants and herbivores, pathogens, pollinators, and parasitoids as shown in Fig. 1.1. Plants emit VOCs from their roots, leaves, fruits and flowers and use these compounds internally as defensive and signaling systems to induce levels of systemic acquired resistance (SAR) to pests and diseases. Some VOCs, such as methyl jasmonate α-pinene, camphene, and 1,8-cineol may inhibit the growth of other plants. VOCs produced by plant organs such as fruits and flowers also can act as external signaling molecules or semiochemicals by attracting pollinators and seed dispersers (Delory et al. 2016). They also contribute to the attraction of pest insects and beneficial insect predators in tritrophic interactions. Apart from plants, VOCs also act as a major communication among insects and other arthropods. Female insects use specific VOCs as sex pheromones to attract mates (Reddy & Guerrero 2004). Insects also use VOCs to mark pathways between nest and food and for defense (de Bruyne & Baker 2008).

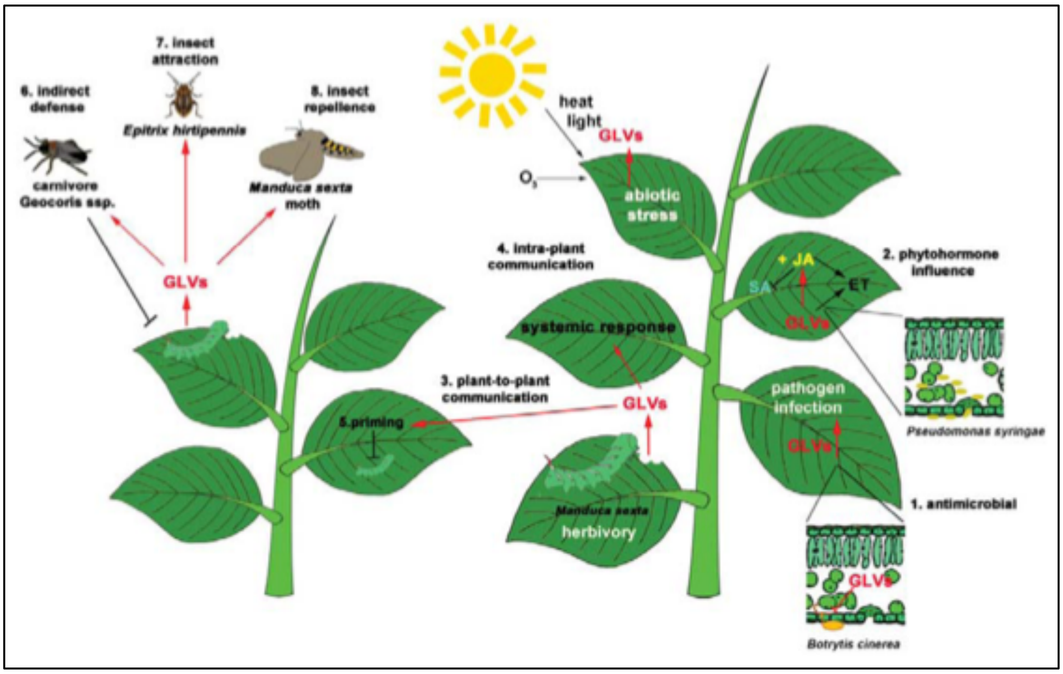


Figure 1.1. *Internal and external communication between plants, herbivores, pathogens, pollinators, and parasitoids* (Scala et al. 2013).

(b) Agriculture

Another important application field of VOCs is agriculture. Conventional agricultural industry relies on a wide use of chemical pesticides and fertilizers. However, increased demand for organic products shows that consumers prefer reduced chemical use. Therefore, a novel sustainable agriculture needs to be developed for crop protection and prevention from using harmful chemicals. VOCs emitted by bacteria and fungi might have the potential as an alternative to the use of chemical pesticides to protect plants from pests and pathogens (Kanchiswamy et al. 2015a). It is because VOCs released by some plant growth-promoting rhizobacteria (PGPR) can enhance plant growth as well as inhibit the growth of other microorganisms, as shown in Fig. 1.2. For example, acetoin and 2,3-butanediol released by rhizobacteria were found to promote the growth of Arabidopsis thaliana seedlings (Kai et al. 2016). A number of frequently emitted VOCs such as hexanal and 2-E-hexenal show antifungal activity and have been developed as an alternative to synthetic chemicals (Ayseli & Ipek 2015). Chemical ecologists also consider microbial VOCs as potential signaling molecules or semiochemicals that function as attractants and repellents to insects and other invertebrates. Pheromone traps are VOC based equipment for controlling pests without using harmful pesticides (Beck & Vannette 2016). In this strategy, pest insects may be diverted away from high-value crops using attractants, while simultaneously being repelled from high-value crops with repellents. Furthermore, natural enemies of insect pests, which are predators and parasitoids, may be simultaneously attracted making the use of semiochemicals a much more viable integrated management strategy than broad-spectrum chemical insecticides. For agriculture scientists, microbial VOCs are seen as biocontrol agents to control various phytopathogens and as biofertilizers for plant growth promotion (Kanchiswamy et al. 2015b). Fig. 1.3 shows the combination of two distinct approaches, which are the identification of very early biomarkers using a knowledge base of translational genomic information on host and pathogen responses and also the development of novel sensors that capture biomarkers for disease detection (Dandekar et al. 2010; Cheung et al. 2015). The authors claimed that it should be possible to identify and defend the crop by interdicting pathogen spread prior to the rapid expansion phase of the disease. These examples indicate that the VOCs might have a potential impact on crop welfare and sustainable agriculture.

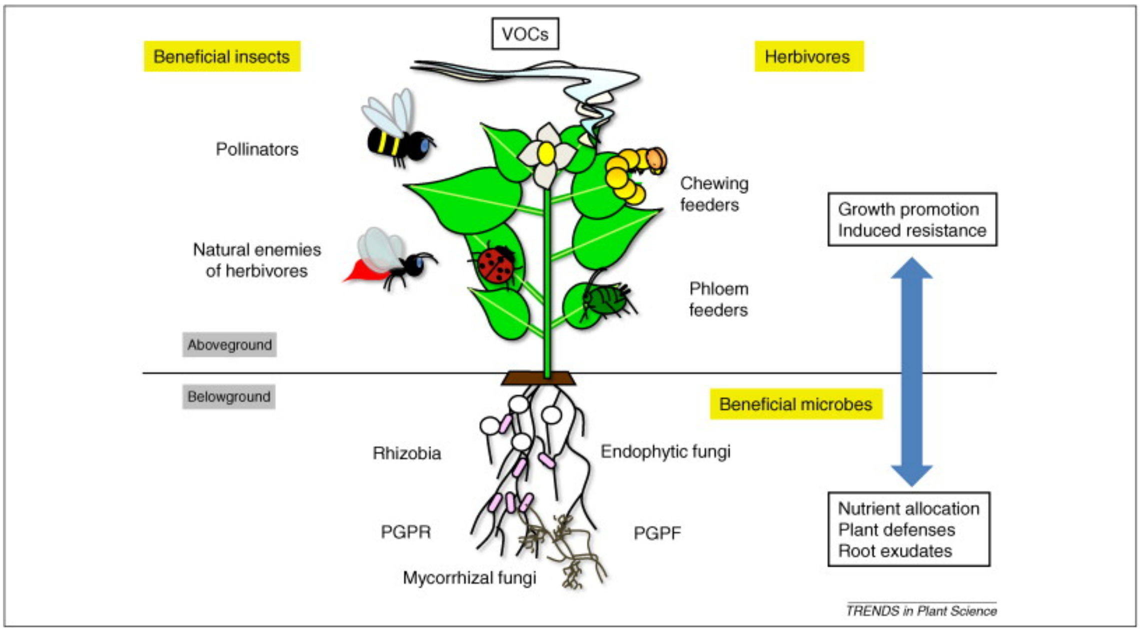


Figure 1.2. *The beneficial microbes such as PGPR can enhance plant growth and induce resistance in aerial plant tissues* (Pineda et al. 2010).

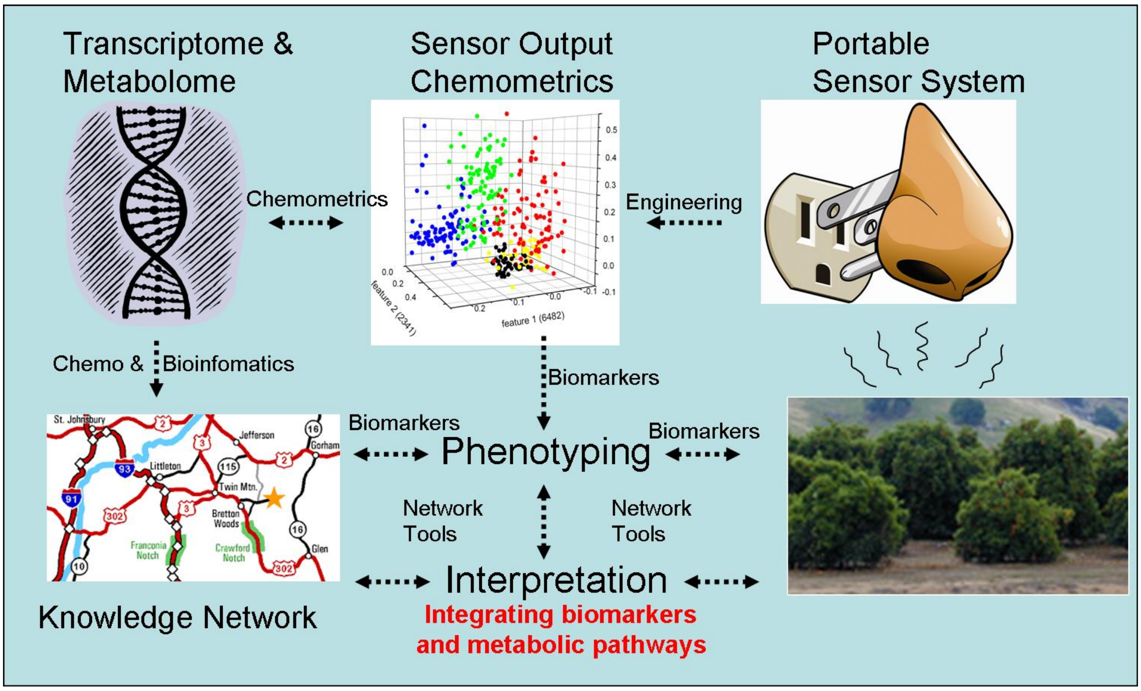


Figure 1.3. *Combination of portable sensor system and bioinformatics knowledge for the management of vector-borne diseases of specialty crops* (Dandekar et al. 2010).

(c) Human healthcare

Recently, there are many research works in applying VOCs as human disease biomarker. This is because the volatiles produced by humans reflect the different metabolic phenotypes (metabotypes) of individuals and may be useful as non-invasive biomarkers to evaluate and monitor disease or health status (Holmes et al. 2008). Hundreds of volatiles are emitted through the human body in breath, blood, skin, fecal and urine as shown in Fig. 1.4 (Shirasu & Touhara 2011; Buljubasic & Buchbauer 2015). A comprehensive review of breath analysis in disease diagnosis using volatile profiles has been presented by (Lourenço & Turner 2014). Breath analysis can be used as a biomarker to identify patients related to breast cancer (Wang, Sun, et al. 2014; Phillips et al. 2010; Phillips et al. 2003), colorectal cancer (Altomare et al. 2013; Amal et al. 2016), pulmonary tuberculosis (Syhre & Chambers 2008; Phillips et al. 2007), and lung cancer

(Hakim et al. 2012; Wang, Dong, et al. 2014; Capuano et al. 2015). Some recent developments of electronic-nose (e-nose) technologies, particularly involving breath analysis, with the potential for providing many new diagnostic applications for the detection of specific human diseases associated with different organs in the body has been summarized by (Wilson 2015). This is also supported by (Fitzgerald et al. 2016), where they stated that the e-nose technology can contribute to personalized medicine approach and have potential to develop early detection for stress-related disorders through analysis of VOCs from exhaled breath. Besides the exhaled breath, fecal and urine headspace VOCs also can be used to diagnose gastrointestinal illness (Garner et al. 2007; Probert et al. 2009; Arasaradnam et al. 2014; Chan et al. 2016).

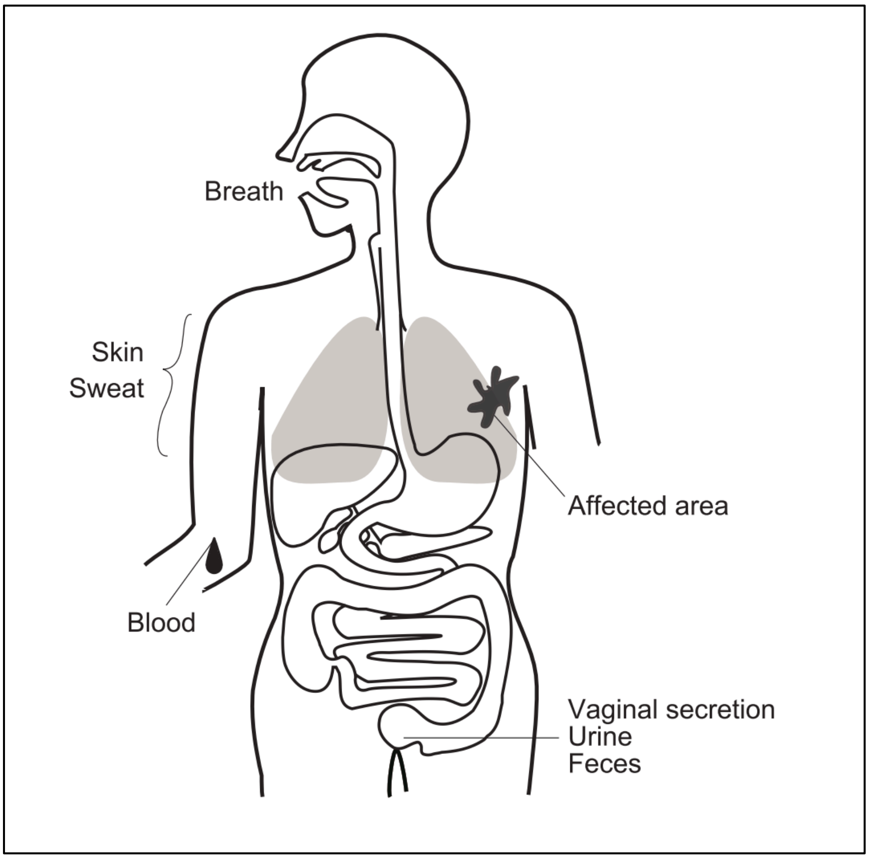


Figure 1.4. *Hundreds of VOCs are emitted through the human body* (Shirasu & Touhara 2011).

Microbial volatiles are also widely used as biomarkers to detect human diseases. This is because bacteria have a recognizable metabolism that produces bacteria-specific VOCs, which might be used for non-invasive diagnostic purposes (Bos et al. 2013). For example, an electronic nose has been used to determine the causative bacteria responsible for diabetic foot infection by recognizing its volatiles (Yusuf et al. 2015). Recently, some authors reported that skin microbiota may play a major role in human attractiveness to blood-sucking insects (Verhulst et al. 2010; Dormont et al. 2013). These insects are responsible for transmission of widespread and sometimes deadly infectious diseases, including malaria (Wong et al. 2012; De Moraes et al. 2014), dengue (Paixão et al. 2014) and zika virus (Didier Musso 2016). These examples indicate that disease-specific VOCs have potential as diagnostic olfactory biomarkers of infectious diseases, metabolic diseases, genetic disorders and other kinds of diseases.

1.2. Research Problem and Objectives

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 tool 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. For example, the Superscent database (Dunkel et al. 2009) only provides structure information of flavors and scents, and the mVOC database (Lemfack et al. 2014) provides information of microbial volatiles only. Flavornet (Arn & Acree 1998) features compounds identified in experiments employing gas chromatography olfactometry (GC-O) analysis, and Pherobase (El-Sayed 2014) is focused on insect pheromones and semiochemicals. The vocBinBase (Skogerson et al. 2011) 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, but there is no public and up-to-date database that accumulated comprehensive information of volatiles and their biological activities. To meet this purpose, we attempt to develop a novel VOC database that accumulates information of emitting species, VOC and biological activities. The main objective of this study is to explore and identify the diversity roles of volatile organic compounds emitted by various species such as plants, microorganisms and human and also to develop a novel database of VOCs extracted from the literature. Second objective is to analyze the relationships between VOCs and microorganisms species based on VOC similarity by using clustering methods. The third objective is to analyze the relationships between other species, VOCs and biological activities based on chemical structural similarity by using unsupervised and supervised machine learning methods.

1.3. Dissertation outline

This dissertation outline is organized as follows. In Chapter 2, the development of VOC database and information on accumulated data is explained. We explain on how potential user can utilize this database for systematic studies in metabolomics. Chapter 3 describes hierarchical clustering and network clustering based on DPClus algorithm for classifying the microorganism species based on VOC metabolites content similarity. In Chapter 4, we discussed on heatmap clustering based on Tanimoto coefficient as the similarity index between chemical structures to cluster all VOCs emitted by other biological species. The resulted clusters were then further accessed by *p*-value based on hypergeometric distribution to understand the relationships between chemical structures of VOCs and their biological activities. Then, we also compared several different clustering methods to determine the degree of cluster overlap and how well it classified chemical structures of VOCs into clusters. Additionally, we extended our analysis by implementing supervised machine learning methods such as Deep Neural Network (DNN), Gradient Boosting Machine (GBM), Random Forest (RF) and Generalized Linear Model (GLM) using different type of molecular fingerprints as classification models for predicting the biological activities of VOCs based on their chemical structures. Finally, Chapter 5 gives conclusing remarks of this dissertation.

Chapter 2

Development of a VOC Database

2.1. Background

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 (Hey et al. 2009; Kelling et al. 2009; Patterson et al. 2010). 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 (Afendi et al. 2012; Afendi et al. 2013; Wijaya et al. 2014). A review of the KNApSAcK DB utilization in scientific work is presented by (Ikeda et al. 2013). Data also has been accumulated in the KNApSAcK DB in order to facilitate the comprehensive understanding of healthy cuisine ingredients, as well as metabolomics (Katsuragi et al. 2013). 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 (Nakamura et al. 2013; Nakamura et al. 2014) and a network-based approach has been proposed to analyze the relationships between 3D structure and biological activities of the metabolites (Ohtana et al. 2014).

In this study, we have developed a VOC database of microorganisms, fungi, and plants as well as human being, which comprises the relation between emitting species, VOC and their biological activities (Abdullah et al. 2015). We have deposited the VOC data into KNApSAcK Metabolite Ecology Database, a part of KNApSAcK family databases and this database is currently available at http://kanaya.naist.jp/MetaboliteEcology/top.jsp. In this chapter, we describe the development of a VOC database and explain how potential users can utilize this database for metabolomics studies.

2.2. Methods

The data were collected by an extensive literature search on PubMed (http://www.ncbi.nlm.nih.gov/pubmed) and Google Scholar. The PubMed search provided more than 100 articles based on the keywords "volatile organic compounds" and "metabolites". The information on VOCs, emitting species, target species and their biological activities were extracted and deposited into KNApSAcK Metabolite Ecology Database. The KNApSAcK Metabolite Ecology is also linked to the KNApSAcK Core and KNApSAcK Metabolite Activity databases 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 – biological activities triplet relations.

2.3. Results and Discussion

At present, we have accumulated 1088 VOCs emitted by 517 microorganisms species and 341 VOCs emitted by other biological species including plants, fungi, animals and human with their related biological activities. 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. The main window of KNApSAcK Family Databases is shown in Fig. 2.1 and user can access the KNApSAcK Metabolite Ecology Database by clicking the corresponding button.

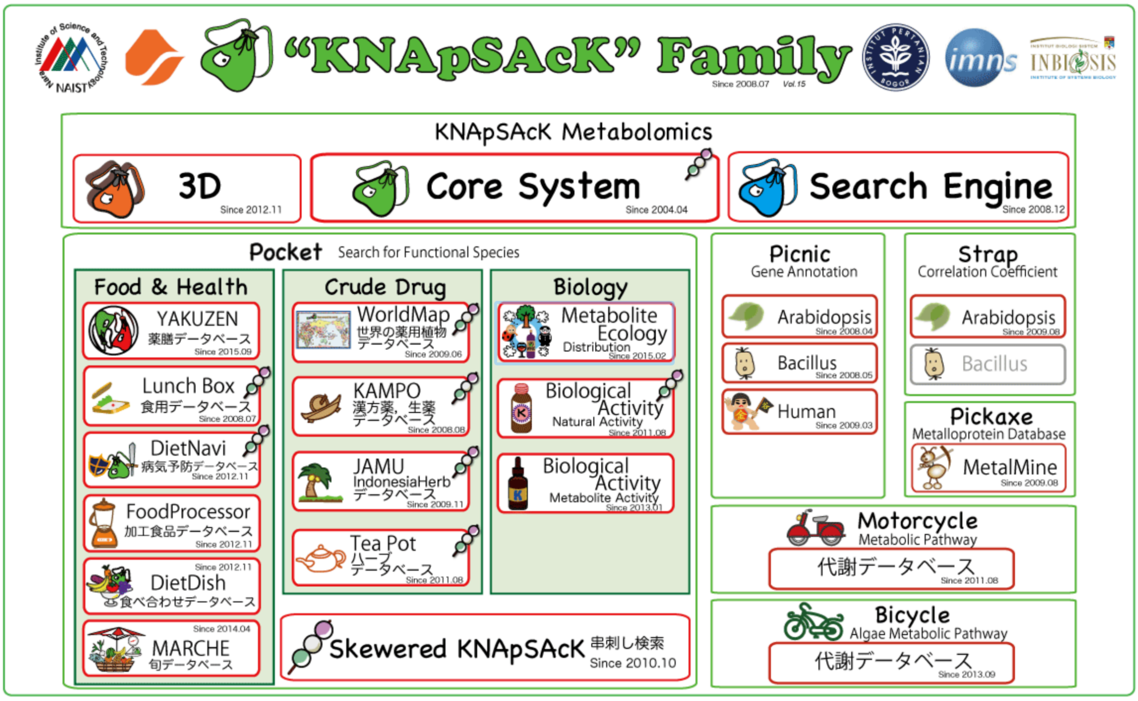
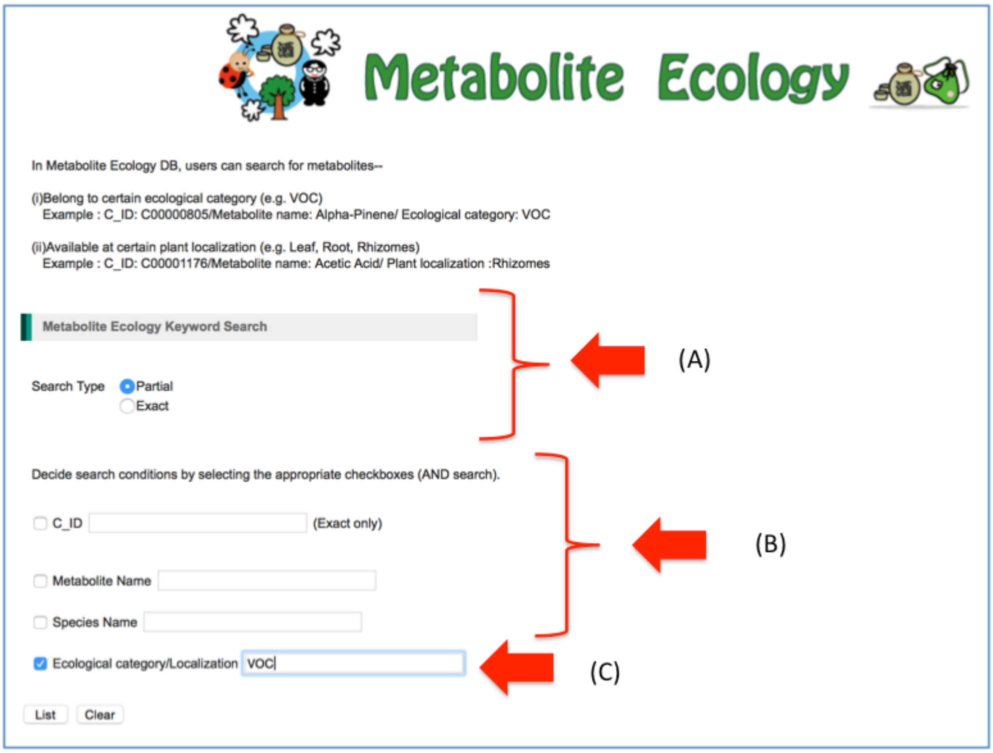

NAIST Nara Institute of Science and Technology
"KNApSAcK" Family
Since 2008.07 Vol.15
imns
INBIOSIS
KNApSAcK Metabolomics
3D Since 2012.11
Core System Since 2004.04
Search Engine Since 2008.12
Pocket Search for Functional Species
Food & Health
YAKUZEN 薬膳データベース Since 2015.09
Lunch Box 食用データベース Since 2008.07
DietNavi 病気予防データベース Since 2012.11
FoodProcessor 加工食品データベース Since 2012.11
Since 2012.11 DietDish 食べ合わせデータベース
Since 2014.04 MARCHE 旬データベース
Crude Drug
WorldMap 世界の薬用植物データベース Since 2009.06
KAMPO 漢方薬，生薬データベース Since 2008.08
JAMU IndonesiaHerb データベース Since 2009.11
Tea Pot ハーブデータベース Since 2011.08
Biology
Metabolite Ecology Distribution Since 2015.02
Biological Activity Natural Activity Since 2011.08
Biological Activity Metabolite Activity Since 2013.01
Skewered KNApSAcK 串刺し検索 Since 2010.10
Picnic Gene Annotation
Arabidopsis Since 2008.04
Bacillus Since 2008.05
Human Since 2009.03
Strap Correlation Coefficient
Arabidopsis Since 2009.08
Bacillus
Pickaxe Metalloprotein Database
MetalMine Since 2009.08
Motorcycle Metabolic Pathway
代謝データベース Since 2011.08
Bicycle Algae Metabolic Pathway
代謝データベース Since 2013.09

Figure 2.1. *The main window of the KNApSAcK Family Databases. (http://kanaya.naist.jp/KNApSAcK_Family/)*

Figure 2.2 shows the main window of the KNApSAcK Metabolite Ecology Database, which shows the search types and search conditions. For search type, users can choose either partial or exact string matching searches by clicking the corresponding button, i.e. Partial or Exact (Fig. 2.2A). Other check boxes can also be selected to specify different search conditions (Fig. 12.2B) 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 checkbox and then click the List button (Fig. 2.2C).



In Metabolite Ecology DB, users can search for metabolites--

(i)Belong to certain ecological category (e.g. VOC)
Example : C_ID: C00000805/Metabolite name: Alpha-Pinene/ Ecological category: VOC

(ii)Available at certain plant localization (e.g. Leaf, Root, Rhizomes)
Example : C_ID: C00001176/Metabolite name: Acetic Acid/ Plant localization :Rhizomes

Metabolite Ecology Keyword Search

Search Type ● Partial ○ Exact

Decide search conditions by selecting the appropriate checkboxes (AND search).

□ C_ID (Exact only)

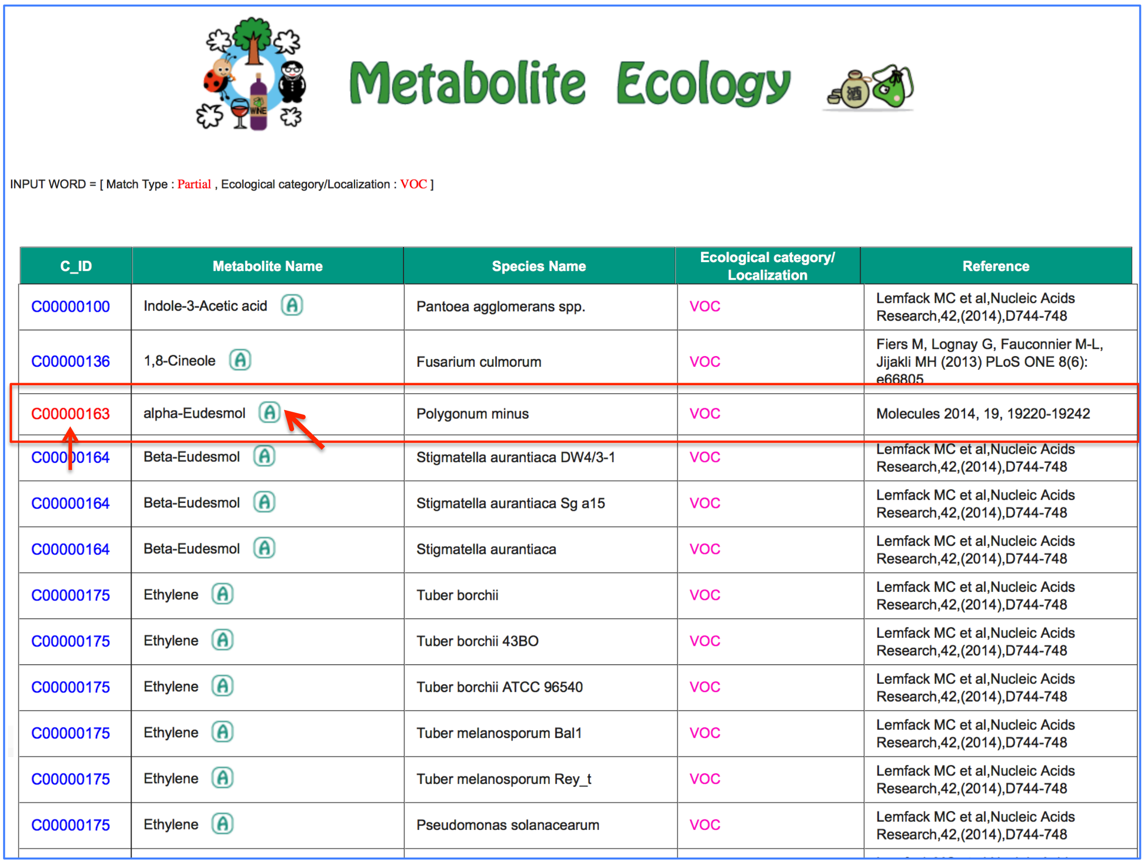
□ Metabolite Name

□ Species Name

☑ Ecological category/Localization VOC

Figure 2.2. *The main window of the KNApSAcK Metabolite Ecology Database.*

Part of the result retrieved by entering 'VOC' in the text box is shown in Fig. 2.3. 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 VOC's information), from left to right. As an example, the metabolite known as alpha-Eudesmol (C_ID C00000163) is emitted by *Polygonum minus*, generally known as 'kesum' in Malaysia. This particular plant is among the most commonly used food additive, flavoring agent and traditionally used to treat stomach and body aches (Christapher et al. 2014; Vikram et al. 2014). Information related to the VOCs that have KNApSAcK compound ID can be obtained by clicking the C_ID as in Fig. 2.3. Figure 2.4 shows the search results obtained by clicking the C_ID, C00000163, 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, InChlKey and other species information, which also produce the corresponding metabolite.

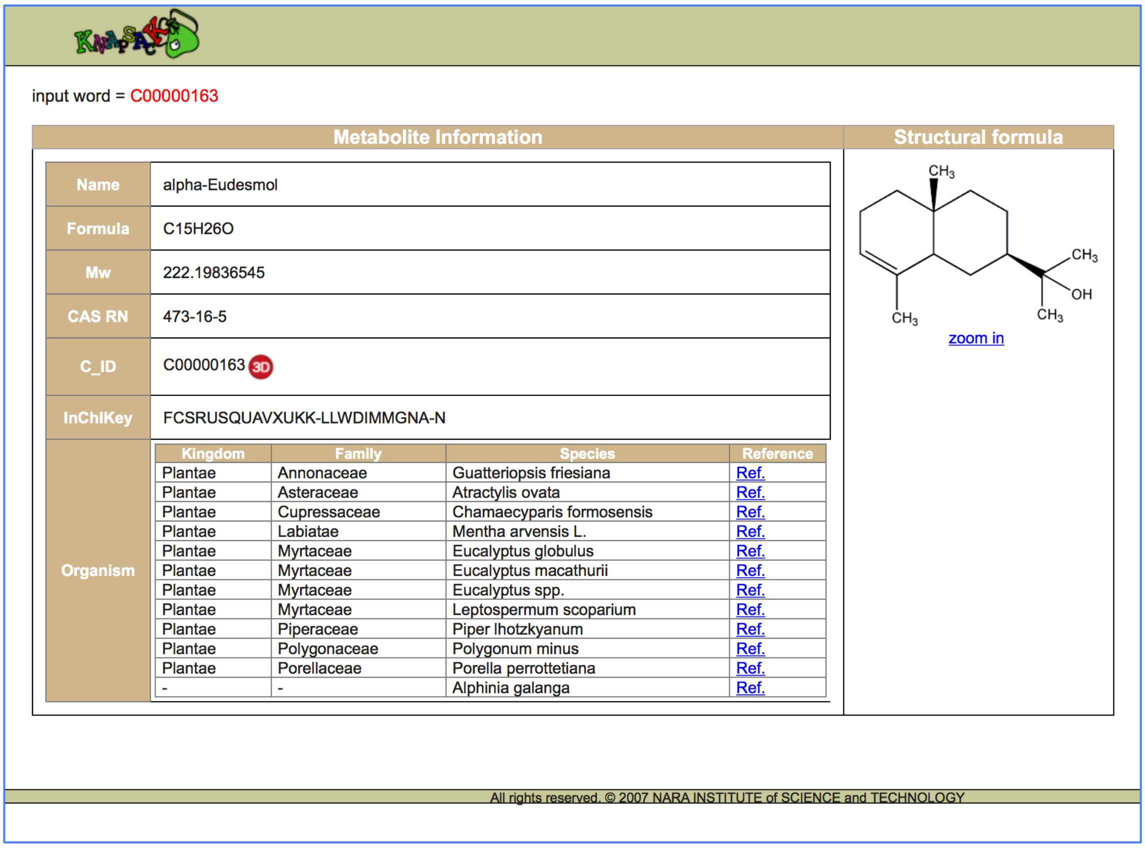


Metabolite Ecology

INPUT WORD = [Match Type : Partial , Ecological category/Localization : VOC]

C_ID	Metabolite Name	Species Name	Ecological category/ Localization	Reference
C00000100	Indole-3-Acetic acid Ⓐ	Pantoea agglomerans spp.	VOC	Lemfack MC et al,Nucleic Acids Research,42,(2014),D744-748
C00000136	1,8-Cineole Ⓐ	Fusarium culmorum	VOC	Fiers M, Lognay G, Fauconnier M-L, Jijakli MH (2013) PLoS ONE 8(6): e66805
C00000163	alpha-Eudesmol Ⓐ	Polygonum minus	VOC	Molecules 2014, 19, 19220-19242
C00000164	Beta-Eudesmol Ⓐ	Stigmatella aurantiaca DW4/3-1	VOC	Lemfack MC et al,Nucleic Acids Research,42,(2014),D744-748
C00000164	Beta-Eudesmol Ⓐ	Stigmatella aurantiaca Sg a15	VOC	Lemfack MC et al,Nucleic Acids Research,42,(2014),D744-748
C00000164	Beta-Eudesmol Ⓐ	Stigmatella aurantiaca	VOC	Lemfack MC et al,Nucleic Acids Research,42,(2014),D744-748
C00000175	Ethylene Ⓐ	Tuber borchii	VOC	Lemfack MC et al,Nucleic Acids Research,42,(2014),D744-748
C00000175	Ethylene Ⓐ	Tuber borchii 43BO	VOC	Lemfack MC et al,Nucleic Acids Research,42,(2014),D744-748
C00000175	Ethylene Ⓐ	Tuber borchii ATCC 96540	VOC	Lemfack MC et al,Nucleic Acids Research,42,(2014),D744-748
C00000175	Ethylene Ⓐ	Tuber melanosporum Bal1	VOC	Lemfack MC et al,Nucleic Acids Research,42,(2014),D744-748
C00000175	Ethylene Ⓐ	Tuber melanosporum Rey_t	VOC	Lemfack MC et al,Nucleic Acids Research,42,(2014),D744-748
C00000175	Ethylene Ⓐ	Pseudomonas solanacearum	VOC	Lemfack MC et al,Nucleic Acids Research,42,(2014),D744-748

Figure 2.3*. The results retrieved for VOC's search in the KNApSAcK Metabolite Ecology Database.*



input word = C00000163

Metabolite Information

Structural formula

Name	alpha-Eudesmol
Formula	C15H26O
Mw	222.19836545
CAS RN	473-16-5
C_ID	C00000163 3D
InChIKey	FCSRUSQUAVXUKK-LLWDIMMGNA-N

Organism

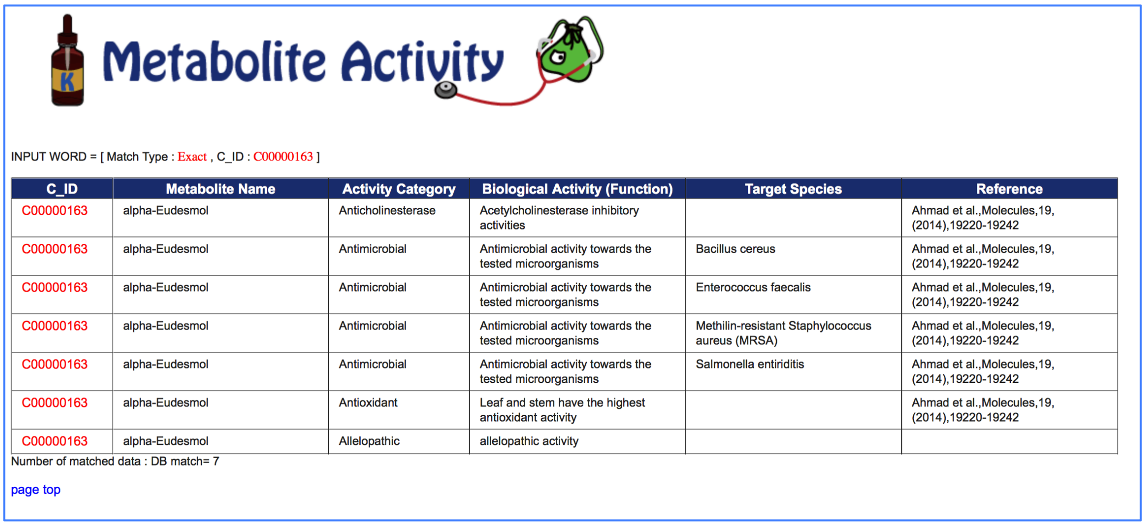
Kingdom	Family	Species	Reference
Plantae	Annonaceae	Guatteriopsis friesiana	Ref.
Plantae	Asteraceae	Atractylis ovata	Ref.
Plantae	Cupressaceae	Chamaecyparis formosensis	Ref.
Plantae	Labiatae	Mentha arvensis L.	Ref.
Plantae	Myrtaceae	Eucalyptus globulus	Ref.
Plantae	Myrtaceae	Eucalyptus macathurii	Ref.
Plantae	Myrtaceae	Eucalyptus spp.	Ref.
Plantae	Myrtaceae	Leptospermum scoparium	Ref.
Plantae	Piperaceae	Piper lhotzkyanum	Ref.
Plantae	Polygonaceae	Polygonum minus	Ref.
Plantae	Porellaceae	Porella perrottetiana	Ref.
-	-	Alphinia galanga	Ref.

All rights reserved. © 2007 NARA INSTITUTE of SCIENCE and TECHNOLOGY

Figure 2.4. An example of the search results obtained by clicking the C_ID, C00000163, which were retrieved from the KNApSAcK Core Database. User can find out other organisms, which also emit this particular metabolite.

To understand the relationships between VOCs and their biological activities, we also integrate the KNApSAcK Metabolite Ecology Database with KNApSAcK Metabolite Activity Database. Information on biological activities of VOCs can be obtained by clicking the 'A' button in Fig. 2.3. Figure 2.5 shows the search result of biological activities related to C_ID C00000163, which was retrieved from the KNApSAcK Metabolite Activity Database. The attributes in the list are C_ID, metabolite name, activity category, biological activity (function), target species and references, from left to right. From the database, user can find out that this particular VOC, known as alpha-Eudesmol has been emitted by

Polygonum minus and have several biological activities such as anticholinesterase, antimicrobial, antioxidant and allelopathic against target species (*Bacillus cereus*, *Enterococcus faecalis*, *Methilin-resistant Staphylococcus aureus* (MRSA), *Salmonella entiriditis*). This information might be useful for the discovery of novel agriculture tools, as well as the development for market of pharmaceutical agents in the future.



INPUT WORD = [Match Type : Exact , C_ID : C00000163]

C_ID	Metabolite Name	Activity Category	Biological Activity (Function)	Target Species	Reference
C00000163	alpha-Eudesmol	Anticholinesterase	Acetylcholinesterase inhibitory activities		Ahmad et al.,Molecules,19,(2014),19220-19242
C00000163	alpha-Eudesmol	Antimicrobial	Antimicrobial activity towards the tested microorganisms	Bacillus cereus	Ahmad et al.,Molecules,19,(2014),19220-19242
C00000163	alpha-Eudesmol	Antimicrobial	Antimicrobial activity towards the tested microorganisms	Enterococcus faecalis	Ahmad et al.,Molecules,19,(2014),19220-19242
C00000163	alpha-Eudesmol	Antimicrobial	Antimicrobial activity towards the tested microorganisms	Methilin-resistant Staphylococcus aureus (MRSA)	Ahmad et al.,Molecules,19,(2014),19220-19242
C00000163	alpha-Eudesmol	Antimicrobial	Antimicrobial activity towards the tested microorganisms	Salmonella entiriditis	Ahmad et al.,Molecules,19,(2014),19220-19242
C00000163	alpha-Eudesmol	Antioxidant	Leaf and stem have the highest antioxidant activity		Ahmad et al.,Molecules,19,(2014),19220-19242
C00000163	alpha-Eudesmol	Allelopathic	allelopathic activity		

Number of matched data : DB match= 7

page top

Figure 2.5*. An example of the search result of biological activities related to C_ID C00000163, which was retrieved from the KNApSAcK Metabolite Activity Database.*

2.4. Summary

In this chapter, we described on the development of a VOC database and explained how to utilize this database for metabolomics studies. Initially, data were accumulated by an extensive literature search through PubMed and Google Scholar. Information on VOCs, emitting species, target species and related biological activity were identified and extracted into an excel format. The data then, were deposited into KNApSAcK Metabolite Ecology Database. Until now, we have accumulated about 1088 VOCs emitted by microorganisms species and 341 VOCs by other biological species such as plants, insects as well as human with the corresponding biological activities. Apart from VOC biological activities related to human healthcare, more than half of the biological activities are associated with chemical ecology. The KNApSAcK Metabolite Ecology Database may be useful for the discovery of novel agricultural tools by focusing on the identification of plant growth promoting rhizobacteria and also the discovery of signature volatiles of plant pathogenic species. This database also can be utilized for the non-invasive identification of biomarkers in the medical diagnostic field as well as a systematic research in various omics fields, especially metabolomics integrated with ecosystems.

Chapter 3

Clustering of Microorganisms Species Based on VOC Similarity

In previous chapter, we described the development of a novel VOC database, which is known as KNApSAcK Metabolite Ecology Database. There were two types of accumulated data: 1) Microorganisms species – VOC relations, 2) Emitting species – VOC – biological activity relations. This chapter focuses on the clustering analysis result of the first type of data, which are the relationships between microorganism species and their emitting VOCs. Until now, we have accumulated 1088 VOCs emitted by 517 microorganisms species (Abdullah et al. 2015).

3.1. Background

A microorganism or microbe is a microscopic living organism, which may be single-celled (Madigan 2012) or multicellular. They are universal in the biosphere and are often found in large quantities and diverse compositions (microbiome). Microorganisms are very diverse and include all bacteria, archaea and most protozoa. This group also contains some species of fungi, algae, and certain microscopic animals, such as rotifers. Microorganisms are also exploited in biotechnology, both in traditional food and beverage preparation, and in modern technologies based on genetic engineering. A small proportion of microorganisms are pathogenic, causing disease and even death in plants and animals (Alberts et al. 2002).

It is well known that microbes produce a diversity of natural compounds, e.g. antibiotics. Many of these small molecules (<300 Da) exhibit high-vapour pressures and low boiling points, and, together with a lipophilic character, these features support volatility, which are known as VOCs. In the past decade, studies on microbial volatile organic compounds (mVOC) attracted many researchers (Korpi et al. 2009; Piechulla & Degenhardt 2014; Lemfack et al. 2014). The aromas of wines, cheese and other milk product, which are usually recognized as pleasant by human are resulted from volatiles produced by microorganisms. On the other hand, microorganisms also produce the unpleasant malodorous smells during the process of putrefaction such as amines, indole, sulphur compounds and ammonia. The earthy and muddy smell of wet forest soils is due to the emission of the volatile geosmin released by Streptomyces species (Gerber 1967; Cane et al. 2006). The human microbial flora at any given anatomical site is relatively specifically accompanied by a typical volatile organic compound (VOC) profile such as gases released by the gut, foot odour and sweat smell. The VOC mixture of breath originates from more than one source within the respiratory system and respiratory disorders can result in odorous gases being expelled into the air, which can be useful for diagnostic purposes (Cheepsattayakorn & Cheepsattayakorn 2013). For example, a compound known as methyl nicotine can be a promising biomarker to be used as a non-invasive diagnostic tool for detection of Mycobacterium tuberculosis (Syhre & Chambers 2008). The emission of 2-nonanone of Pseudomonas aeruginosa VOCs may be used as in vivo marker to detect lung infections (Carroll et al. 2005). A group of researchers has investigated the performance of electronic nose (e-nose) technique performing direct measurement of static headspace with algorithm and data interpretations which was validated by GC-MS, to determine the causative bacteria responsible for diabetic foot infection based on their volatiles (Yusuf et al. 2015). Other than diagnostic tool, some volatile compounds produced by microorganisms such as higher alcohols (2-methyl-1-butanol, 3-methyl-1-butanol and isobutanol) can be used as biofuels (Blombach & Eikmanns 2011). Metabolic engineering can be used to improve the production of natural microbial alcohols for the bio renewable fuels

(Ingram et al. 2010; Nozzi et al. 2014; Lee et al. 2015). More than 10 000 microbial species are described and at least a million are expected to exist on earth, the VOC profiles of a surprisingly small number of microorganisms were investigated so far. Considering the importance and the central roles of VOCs in our biosphere, our first objective was to accumulate the data related to VOCs and their emitting microorganisms species. Using this accumulated data, we performed hierarchical clustering and graph clustering for classifying the VOC emitting species based on volatile metabolite content similarity.

3.2. Datasets

We used the first type of data, which are 1088 VOCs emitted by 517 microorganisms species, as mentioned in Chapter 2. The information of emitting species and volatile compounds has been converted into a 517×1088 binary matrix ("1" indicates presence while "0" indicates absence), where rows represent as microorganism species and columns represent VOCs emitted by the corresponding species as shown in Table 2.1. The binary matrix then, was used to calculate the distance between species and to perform the clustering.

Table 2.1. *Representation of microorganism species and volatile organic compounds as a two-dimensional binary matrix.*

Species	VOCs					
	VOC$_1$	VOC$_2$	VOC$_3$	VOC$_4$	…	VOC$_M$
S$_1$	1	0	1	1	…	0
S$_2$	1	1	0	0	…	0
S$_3$	0	1	0	1	…	0
…	…	…	…	…	…	…
S$_N$	1	0	0	0	…	1

3.3. Methods

Clustering is an unsupervised learning method, which is the task of grouping a set of objects into groups (clusters) based on similarity or distance measures (Jain et al. 1999). This technique is important for knowledge discovery and has been applied in many applications such as machine learning, pattern recognition, image analysis and bioinformatics (Thalamuthu et al. 2006; Diao et al. 2011; Clifford et al. 2011; Richard et al. 2013). The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. There are many clustering methods based on different algorithms. Typical cluster models include:

(1) Connectivity models: for example, hierarchical clustering builds models based on distance connectivity.

(2) Centroid models: for example, the k-means algorithm represents each cluster by a single mean vector.

(3) Distribution models: clusters are modeled using statistical distributions, such as multivariate normal distributions used by the Expectation-maximization algorithm.

(4) Density models: for example, DBSCAN and OPTICS defines clusters as connected dense regions in the data space.

(5) Subspace models: in Biclustering, clusters are modeled with both cluster members and relevant attributes.

(6) Group models: some algorithms do not provide a refined model for their results and just provide the grouping information.

(7) Graph-based models: a subset of nodes in a graph such that every two nodes in the subset are connected by an edge can be considered as a prototypical form of cluster.

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:

3.3.1. Hierarchical clustering

We used hierarchical agglomerative clustering method, which starts out by putting each observation into its own separate cluster (Johnson 1967; Murtagh & Contreras 2011; Murtagh & Contreras 2012). The result of clustering is usually represented by a dendrogram. The reason why we choose hierarchical clustering is that, it is easy to use and our objective is to find a specific group for microorganism species. In our case, we used a Species vs. VOC matrix. Let this matrix be called *M* and M_{ik}=1 if the species *i* is related to the k_{th} VOC or otherwise M_{ik}=0. 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 equation (3.1):

$$d(i,j) = \sqrt{\sum_{k=1}^{n}(M_{ik} - M_{jk})^2} \tag{3.1}$$

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

3.3.2. Graph clustering based on DPClus

DPClus is a graph clustering software (Md. Altaf-Ul-Amin et al. 2006), which has been developed based on a graph-clustering algorithm that can extract densely connected nodes as a cluster (Md Altaf-Ul-Amin et al. 2006). Initially, the algorithm was purposely developed to detect and visualize clusters of proteins in interaction networks which mostly represent molecular biological functional units. We explore the possibility of this algorithm to other applications as well and here, we apply the DPClus algorithm to find a cohesive group for our accumulated microorganism data. 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: 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 $(|E|_{max})$. Equation (3.2) represents the cluster property of node *n* with respect to cluster *k*:

$$cp_{nk} = \frac{E_{nk}}{d_k \times N_k} \tag{3.2}$$

N_k is the number of nodes in cluster *k*. E_{nk} is the total number of edges between the node *n* and the nodes of cluster *k*. In this study, we applied the DPClus algorithm to identify certain groups of microorganism species, based on VOC similarity. A network was 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 the lower Euclidean distance between them. We used the non-overlapping mode with the following DPClus settings: Cluster property cp_{nk} was set to 0.5, density value d_k was set to value in between 0.6 and 0.9, and minimum cluster size was set to 2, as recommended by (Md Altaf-Ul-Amin et al. 2006).

3.4. Results and discussion

3.4.1. Hierarchical clustering result

Fig. 3.1 shows the log-log relation between the number of VOCs, *M* and the frequency of species, *N*. The pattern roughly follows power-law (Jeong et al. 2001). From this figure, we can see 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. Fig. 3.2 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. We also enlarged the clusters that consisting 100% pathogenic microorganisms species, which are clusters 35, 40 and 47.

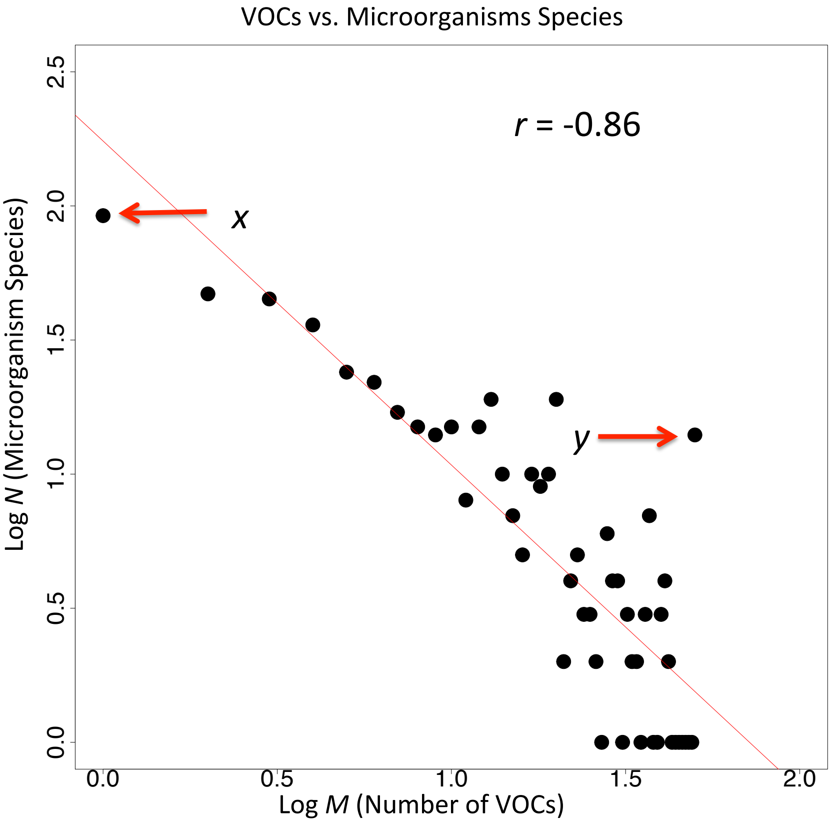


Figure 3.1. *The log-log relation between the number of VOCs and the number of related microorganisms species.*

Table 1 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 i.e. clusters 35, 40 and 47 (Fig. 3.2) contain 100% pathogenic bacterial species such as *Pseudomonas aeruginosa*, *Klebsiella pneumoniae* and *Escherichia coli*. The other three clusters contain both pathogenic and non-pathogenic species. For example, cluster 6 consists of 11 (7.2%) pathogenic bacterial species while cluster 27 comprises of 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 non-pathogenic species. These results imply that VOCs emitted by some pathogenic bacteria are different from those emitted by non-pathogenic bacteria. These results show consistency between VOC and pathogenicity-based classification of microorganisms.

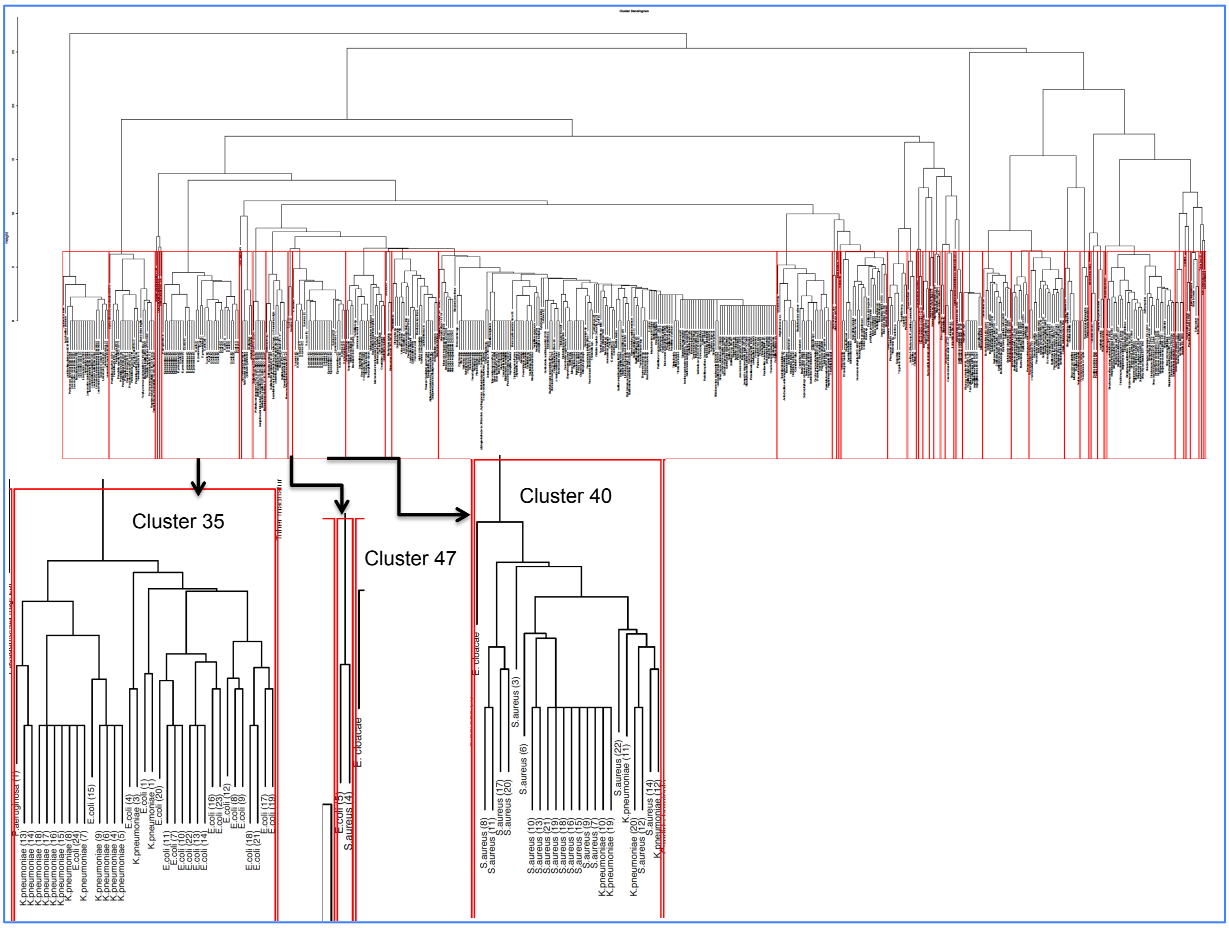


Figure 3.2. *Hierarchical dendrogram plot of microorganism species based on VOC similarity. Cluster 35, 40 and 47 contain 100% pathogenic species and the detail description on each cluster with the related microorganisms species is given in Table 3.1.*

***Table 3.1**. Microorganisms species name corresponding to the clusters and their pathogenicity.*

Species No	Species Name	Cluster No	Pathogenicity
1	Chondromyces crocatus	1	Non-pathogenic
2	Nannocystis exedens	2	Non-pathogenic
3	Nannocystis exedens Na eB37	2	Non-pathogenic
4	Nannocystis exedens subsp. cinnabarina Na c29	2	Non-pathogenic
5	Tuber magnatum	3	Non-pathogenic
6	Phoma sp.	4	Non-pathogenic
7	Tuber melanosporum	5	Non-pathogenic
8	Penicillium roqueforti (IBT 16404)	6	Non-pathogenic
9	Nannocystis exedens Na e485	6	Non-pathogenic
10	Streptomyces citreus	6	Non-pathogenic
11	Acremonium furcatum BAFC 51375	6	Non-pathogenic
12	Bacillus strains	6	Non-pathogenic
13	Fistulina hepatica (Schaeffer: Fr.) Fr	6	Non-pathogenic
14	Octadecabacter sp.	6	Non-pathogenic
15	Octadecabacter sp. ARK10255b	6	Non-pathogenic
16	Halomonas venusta	6	Non-pathogenic
17	Planococcus citreus	6	Non-pathogenic
18	Enterobacter agglomerans	6	Non-pathogenic
19	Aspergillus versicolor	6	Non-pathogenic
20	Streptomyces griseus	6	Non-pathogenic
21	Escherichia sp.	6	Non-pathogenic
22	Bacillus pumilus (BSH-4)	6	Non-pathogenic
23	Burkholderia andropogonis LMG 2129	6	Non-pathogenic
24	Burkholderia sordidicola LMG 22029	6	Non-pathogenic
25	Limnobacter thiooxidans LMG 19593	6	Non-pathogenic
26	Stenotrophomonas rhizophilla ep10-p69	6	Non-pathogenic
27	Staphylococcus aureus (5)	6	Pathogenic
28	Agaricus bisporus	6	Non-pathogenic

29	Aspergillus candidus	6	Non-pathogenic
30	Emericella nidulans	6	Non-pathogenic
31	Mycoleptodonoides aitchisonii TUFC10099	6	Non-pathogenic
32	Pseudomonas aeruginosa(2)	6	Pathogenic
33	Pseudomonas fluorescens L13-6-12	6	Non-pathogenic
34	Pseudomonas sp.	6	Non-pathogenic
35	Pseudomonas trivialis 3Re2-7	6	Non-pathogenic
36	Shewanella spp.	6	Non-pathogenic
37	Muscodor fengyangensis (ZJLQ374)	6	Non-pathogenic
38	Penicillium expansum	6	Non-pathogenic
39	Muscodor fengyangensis (ZJLQ070)	6	Non-pathogenic
40	Muscodor fengyangensis (ZJLQ151)	6	Non-pathogenic
41	Muscodor albus I-41. 3s	6	Non-pathogenic
42	Bacillus pumilus (ZB13)	6	Non-pathogenic
43	Bacillus subtilis(BL02)	6	Non-pathogenic
44	Carnobacterium maltaromaticum	6	Non-pathogenic
45	Penicillium crustosum	6	Non-pathogenic
46	Escherichia coli(3)	6	Pathogenic
47	Mycobacterium tuberculosis	6	Pathogenic
48	Psedomonas taetroleus	6	Non-pathogenic
49	Penicillium cyclopium	6	Non-pathogenic
50	Paecilomyces variotii	6	Non-pathogenic
51	Jannaschia helgolandensis strain HEL-26	6	Non-pathogenic
52	Klebsiella pneumoniae	6	Pathogenic
53	Tuber oligospermum	6	Non-pathogenic
54	Thermoactinomyces spp.	6	Non-pathogenic
55	Actinobacteria	6	Non-pathogenic
56	Oscillatoria chalybea	6	Non-pathogenic
57	Oscillatoria sp.	6	Non-pathogenic
58	Streptomyces lavendulae	6	Non-pathogenic
59	Cytophaga strains	6	Non-pathogenic
60	a marine Arctic bacterium	6	Non-pathogenic

61	Burkholderia sordidicola LMG 22029	6	Non-pathogenic
62	Roseobacter clade	6	Non-pathogenic
63	Serratia plymuthica 3Re4-18	6	Non-pathogenic
64	Staphylococcus epidermidis 2P3-18	6	Non-pathogenic
65	Stenotrophomonas rhizophila P69	6	Non-pathogenic
66	Arthrobacter globiformis	6	Non-pathogenic
67	Streptomycete sp.	6	Non-pathogenic
68	Penicillium palitans (commune) (IBT 15899)	6	Non-pathogenic
69	Staphylococcus aureus (1)	6	Pathogenic
70	Agaricus campestris	6	Non-pathogenic
71	Trichoderma aureoviride IMI 91968	6	Non-pathogenic
72	Aspergillus fumigatus	6	Non-pathogenic
73	Sulfitobacter sp. Bio-007	6	Non-pathogenic
74	Cenococcum geophilum	6	Non-pathogenic
75	Wolinella curva CCUG 13146 (35224)	6	Non-pathogenic
76	aerobic Gram-negative bacteria	6	Non-pathogenic
77	Klebsiella pneumoniae (2)	6	Pathogenic
78	Staphylococcus aureus (2)	6	Pathogenic
79	Trichoderma pseudokoningii (T64)	6	Non-pathogenic
80	Streptomyces àntibioticus CBS 659.68	6	Non-pathogenic
81	psychrotrophic bacteria	6	Non-pathogenic
82	Trichoderma viride (T60)	6	Non-pathogenic
83	Thermomonospora fusca	6	Non-pathogenic
84	Alternaria Alternata	6	Non-pathogenic
85	Paenibacillus polymyxa (BMP-11)	6	Non-pathogenic
86	Puccinia graminis var. tritici	6	Non-pathogenic
87	Trichoderma sp.	6	Non-pathogenic
88	Escherichia coli(2)	6	Pathogenic
89	Bacillus amyloliquefaciens IN937a	6	Non-pathogenic
90	Bacillus subtilis GB03	6	Non-pathogenic
91	Penicillium digitatum	6	Non-pathogenic
92	Bacillus spp.	6	Non-pathogenic

93	Citrobacter freundii	6	Non-pathogenic
94	Micrococcus luteus	6	Non-pathogenic
95	Staphylococcus aureus	6	Pathogenic
96	Daedalea juniperina	6	Non-pathogenic
97	Muscodor fengyangensis (ZJLQ023)	6	Non-pathogenic
98	Muscodor fengyangensis (ZJLQ024)	6	Non-pathogenic
99	Oscillatoria perornata	6	Non-pathogenic
100	Spirulina platensis	6	Non-pathogenic
101	Penicillium clavigerum	6	Non-pathogenic
102	Xanthomonas campestris pv campestris	6	Non-pathogenic
103	Roseovarius spp.	6	Non-pathogenic
104	benthic cyanobacteria (Calothrix, Plectonema)	6	Non-pathogenic
105	Cyanobacterial biofilms	6	Non-pathogenic
106	Aerobasidium pullulans	6	Non-pathogenic
107	Escherichia coli(6)	6	Pathogenic
108	Dipodascus aggregatus	6	Non-pathogenic
109	Pseudomonas solanacearum	6	Non-pathogenic
110	Aspergillus clavatus	6	Non-pathogenic
111	Blastomyces dermatitidis	6	Non-pathogenic
112	Ceratocystis fimbriata	6	Non-pathogenic
113	Mucor hiemalis	6	Non-pathogenic
114	Tuber borchii 43BO	6	Non-pathogenic
115	Tuber borchii ATCC 96540	6	Non-pathogenic
116	Tuber melanosporum Bal1	6	Non-pathogenic
117	Tuber melanosporum Rey_t	6	Non-pathogenic
118	Bifidobacterium adolescentis DPC6044	6	Non-pathogenic
119	Lactobacillus brevis DPC6108	6	Non-pathogenic
120	Anabaena	6	Non-pathogenic
121	Fossombronia pusilla	6	Non-pathogenic
122	Lyngbya	6	Non-pathogenic
123	Sigmatella aurantiaca	6	Non-pathogenic
124	Streptomyces sulfureus	6	Non-pathogenic

125	Streptomyces UC5319	6	Non-pathogenic
126	Pseudomonas fluorescens AN5	6	Non-pathogenic
127	Fomes annosus	6	Non-pathogenic
128	Flavobacteria sp.	6	Non-pathogenic
129	Chromobacterium sp.	6	Non-pathogenic
130	Clitocybe geotropa	6	Non-pathogenic
131	Fomes scutellatus	6	Non-pathogenic
132	Marasmius oreacles	6	Non-pathogenic
133	Pholiota aurea	6	Non-pathogenic
134	Pantoea agglomerans spp.	6	Non-pathogenic
135	Pseudonocardia sp.	6	Non-pathogenic
136	Saccharomonospora sp.	6	Non-pathogenic
137	Thermomonospora sp.	6	Non-pathogenic
138	Lactobacillus fermentum	6	Non-pathogenic
139	Fomes pomaceus	6	Non-pathogenic
140	Alphaproteobacteria (Rhizobium , Sphingomonas , Methylobacterium , Roseovarius)	6	Non-pathogenic
141	Betaproteobacteria (Variovorax , Zogloea)	6	Non-pathogenic
142	Deleya spp.	6	Non-pathogenic
143	Photobacterium spp.	6	Non-pathogenic
144	Plantibacter spp.	6	Non-pathogenic
145	Pseudoalteromonas spp.	6	Non-pathogenic
146	Rhizobium ssp.	6	Non-pathogenic
147	Rhodococcus spp.	6	Non-pathogenic
148	Sphingomonas spp.	6	Non-pathogenic
149	Variovorax spp.	6	Non-pathogenic
150	Vibrio spp.	6	Non-pathogenic
151	Zogloea ssp.	6	Non-pathogenic
152	Bacillus popillae	6	Non-pathogenic
153	Penicillium chrysogenum (IBT 15921)	6	Non-pathogenic
154	Penicillium chrysogenum (IBT 15996)	6	Non-pathogenic
155	Azoarcus evansii	6	Non-pathogenic

156	Acinetobacter calcoaceticus	6	Non-pathogenic
157	Tilletia caries	6	Non-pathogenic
158	Tilletia controversa	6	Non-pathogenic
159	Tilletia foetida	6	Non-pathogenic
160	Bacillus thuringensis	6	Non-pathogenic
161	Chondromyces crocatus Cm c2	7	Non-pathogenic
162	Chondromyces crocatus Cm c5	7	Non-pathogenic
163	Myxobacterium spp.	8	Non-pathogenic
164	Myxococcus xanthus	8	Non-pathogenic
165	Stigmatella aurantiaca	9	Non-pathogenic
166	Stigmatella aurantiaca DW4/3-1	9	Non-pathogenic
167	Stigmatella aurantiaca Sg a15	9	Non-pathogenic
168	Streptomyces caviscabies	10	Non-pathogenic
169	Streptomyces sp. GWS-BW-H5.	10	Non-pathogenic
170	Streptomyces coelicolor	11	Non-pathogenic
171	Streptomyces albidoflavus	11	Non-pathogenic
172	Streptomyces albidoflavus AMI 246	11	Non-pathogenic
173	Streptomyces albus	11	Non-pathogenic
174	Streptomyces albus IFO 13014	11	Non-pathogenic
175	Streptomyces antibioticus	11	Non-pathogenic
176	Streptomyces antibioticus ETH 22014	11	Non-pathogenic
177	Streptomyces aureofaciens ETH 13387	11	Non-pathogenic
178	Streptomyces coelicolor ATCC 21666	11	Non-pathogenic
179	Streptomyces coelicolor DSM 40233	11	Non-pathogenic
180	Streptomyces diastatochromogenes IFO 13814	11	Non-pathogenic
181	Streptomyces griseus ATCC 23345	11	Non-pathogenic
182	Streptomyces griseus IFO 13849	11	Non-pathogenic
183	Streptomyces hirsutus ATCC 19773	11	Non-pathogenic
184	Streptomyces hirsutus ETH 1666	11	Non-pathogenic
185	Streptomyces hygroscopicus ATCC 27438	11	Non-pathogenic
186	Streptomyces murinus DSM 40091	11	Non-pathogenic
187	Streptomyces murinus NRRL 8171	11	Non-pathogenic

188	Streptomyces olivaceus ETH 6445	11	Non-pathogenic
189	Streptomyces olivaceus ETH 7437	11	Non-pathogenic
190	Streptomyces rishiriensis AMI 224	11	Non-pathogenic
191	Streptomyces spp. AMI 240	11	Non-pathogenic
192	Streptomyces spp. AMI 243	11	Non-pathogenic
193	Streptomyces thermoviolaceus CBS 111.62	11	Non-pathogenic
194	Actinomycetes	11	Non-pathogenic
195	Streptomyces albus subsp. pathocidicus IFO 13812	11	Non-pathogenic
196	Streptomyces antibioticus CBS 659.68	11	Non-pathogenic
197	Streptomyces aureofaciens ETH 28832	11	Non-pathogenic
198	Streptomyces diastatochromogenes ETH 18822	11	Non-pathogenic
199	Streptomyces hygroscopicus IFO 13255	11	Non-pathogenic
200	Streptomyces thermoviolaceus IFO 12382	11	Non-pathogenic
201	Streptomyces spp.	12	Non-pathogenic
202	Bacillus	13	Non-pathogenic
203	Tuber borchii	14	Non-pathogenic
204	Tuber indicum	15	Non-pathogenic
205	marine Streptomycete (isolate B6007)	16	Non-pathogenic
206	Prevotella buccae ATCC 33574	16	Non-pathogenic
207	Prevotella buccae ES12-B	16	Non-pathogenic
208	Prevotella buccae ES17-1	16	Non-pathogenic
209	Prevotella buccae ES9-1	16	Non-pathogenic
210	Prevotella disiens DSM 20516	16	Non-pathogenic
211	Prevotella heparinolyticus ATCC 35895	16	Non-pathogenic
212	Prevotella oris ATCC 33573	16	Non-pathogenic
213	Prevotella oris ES14B-3A	16	Non-pathogenic
214	Prevotella oris ES9-3	16	Non-pathogenic
215	Prevotella oris RPG	16	Non-pathogenic
216	Prevotella veroralis ATCC 33779	16	Non-pathogenic
217	Porphyromonas endodontalis HG 181 (H 11a-e)	16	Non-pathogenic
218	Porphyromonas endodontalis HG 182 (BN 11a-f)	16	Non-pathogenic
219	Porphyromonas endodontalis HG 370 (ATCC 35406)	16	Non-pathogenic

220	Porphyromonas endodontalis HG 412	16	Non-pathogenic
221	Prevotella oralis ES4-B	16	Non-pathogenic
222	Bacteroides fragilis	16	Non-pathogenic
223	Bacteroides fragilis ATCC 25285	16	Non-pathogenic
224	Prevotella oralis ES14B-3A	16	Non-pathogenic
225	Prevotella oralis ES15-2	16	Non-pathogenic
226	ARK10063	17	Non-pathogenic
227	Bjerkandera adusta	17	Non-pathogenic
228	Bjerkandera adusta CBS 595.78	17	Non-pathogenic
229	Armillaria mellea	18	Non-pathogenic
230	Pholiota squarrosa	18	Non-pathogenic
231	Stropharia rugosoannulata	18	Non-pathogenic
232	Verticillium longisporum	19	Non-pathogenic
233	Candida tropicalis	19	Non-pathogenic
234	Salmonella enterica	19	Non-pathogenic
235	Shigella flexneri	19	Non-pathogenic
236	Tuber panniferum	19	Non-pathogenic
237	Tuber excavatum	19	Non-pathogenic
238	Penicillium aurantiogriseum	19	Non-pathogenic
239	Ascocoryne sarcoides NRRL 50072	19	Non-pathogenic
240	Aspergillus ornatus	19	Non-pathogenic
241	Neurospora sitophila ATCC 46892	19	Non-pathogenic
242	Neurospora sp.	19	Non-pathogenic
243	Penicillium chrysogenum	19	Non-pathogenic
244	penicillium paneum (Conidia)	19	Non-pathogenic
245	Tuber uncinatum	19	Non-pathogenic
246	Ceratocystis sp.	19	Non-pathogenic
247	Thielaviopsis basicola	19	Non-pathogenic
248	Mycobacterium bovis	19	Non-pathogenic
249	Muscodor albus CZ-620	19	Non-pathogenic
250	Muscodor crispans	19	Non-pathogenic
251	Boletus variegatus	19	Non-pathogenic

252	Fomes sp.	19	Non-pathogenic
253	Dinoroseobacter shibae	20	Non-pathogenic
254	Dinoroseobacter shibae strain DFL-27	20	Non-pathogenic
255	Loktanella sp.	20	Non-pathogenic
256	Loktanella sp. Bio-204	20	Non-pathogenic
257	Carnobacterium divergens 9P	21	Non-pathogenic
258	Dinoroseobacter sp.	22	Non-pathogenic
259	Stigmatella sp.	22	Non-pathogenic
260	Calothrix	23	Non-pathogenic
261	Phormidium sp.	23	Non-pathogenic
262	Plectonema	23	Non-pathogenic
263	Calothrix parietina	23	Non-pathogenic
264	Plectonema notatum	23	Non-pathogenic
265	Plectonema sp.	23	Non-pathogenic
266	Tolypothrix	23	Non-pathogenic
267	Tolypothrix distorta	23	Non-pathogenic
268	Calothrix sp.	23	Non-pathogenic
269	Burkholderia ambifaria LMG 19467	24	Non-pathogenic
270	Burkholderia ambifaria LMG 17828	24	Non-pathogenic
271	Burkholderia ambifaria LMG 19182	24	Non-pathogenic
272	Alcaligenes	25	Non-pathogenic
273	Alcaligenes faecalis	25	Non-pathogenic
274	Arthrobacter nitroguajacolius	25	Non-pathogenic
275	Lysobacter gummosus	25	Non-pathogenic
276	Sporosarcina ginsengisoli	25	Non-pathogenic
277	Stenotrophomonas maltophilia	26	Non-pathogenic
278	Serratia marcescens	26	Non-pathogenic
279	Bacillus simplex	26	Non-pathogenic
280	Bacillus subtilis	26	Non-pathogenic
281	Bacillus weihenstephanensis	26	Non-pathogenic
282	Microbacterium oxydans	26	Non-pathogenic
283	Streptomyces lateritius	26	Non-pathogenic

284	Escherichia coli	27	Pathogenic
285	Burkholderia anthina LMG 20980	27	Non-pathogenic
286	Burkholderia gladioli LMG 2216	27	Non-pathogenic
287	Burkholderia glumae LMG 2196	27	Non-pathogenic
288	Burkholderia caledonica LMG 19076	27	Non-pathogenic
289	Burkholderia caribensis LMG 18531	27	Non-pathogenic
290	Burkholderia caryophylli LMG 2155	27	Non-pathogenic
291	Burkholderia fungorum LMG 16225	27	Non-pathogenic
292	Burkholderia glathei LMG 14190	27	Non-pathogenic
293	Burkholderia lata LMG 22485	27	Non-pathogenic
294	Serratia plymuthica IC14	27	Non-pathogenic
295	Burkholderia graminis LMG 18924	27	Non-pathogenic
296	Cellulomonas uda	27	Non-pathogenic
297	Paenibacillus polymyxa	28	Non-pathogenic
298	Paenibacillus polymyxa E681	28	Non-pathogenic
299	Trichoderma viride	29	Non-pathogenic
300	Tuber aestivum	30	Non-pathogenic
301	Tuber brumale	31	Non-pathogenic
302	Tuber mesentericum	31	Non-pathogenic
303	Tuber rufum	31	Non-pathogenic
304	Tuber simonea	31	Non-pathogenic
305	Pseudomonas fragi 25P	32	Non-pathogenic
306	Burkholderia lata LMG 6993	33	Non-pathogenic
307	Burkholderia phenazinium LMG 2247	33	Non-pathogenic
308	Burkholderia phytofirmans LMG 22487	33	Non-pathogenic
309	Burkholderia pyrrocinia LMG 21822	33	Non-pathogenic
310	Burkholderia terricola LMG 20594	33	Non-pathogenic
311	Chromobacterium violaceum	33	Non-pathogenic
312	Chromobacterium violaceum CV0	33	Non-pathogenic
313	Pseudomonas putida	33	Non-pathogenic
314	Pseudomonas putida ISOf	33	Non-pathogenic
315	Serratia marcescens MG1	33	Non-pathogenic

316	Serratia plymuthica HRO-C48	33	Non-pathogenic
317	Burkholderia sacchari LMG 19450	33	Non-pathogenic
318	Burkholderia thailandensis LMG 20219	33	Non-pathogenic
319	Pseudomonas fluorescens WCS 417r	33	Non-pathogenic
320	Serratia entomophilia A1MO2	33	Non-pathogenic
321	Serratia proteamaculans B5a	33	Non-pathogenic
322	Burkholderia tropica LMG 22274	34	Non-pathogenic
323	Burkholderia cepacia LMG 1222 358	34	Non-pathogenic
324	Burkholderia hospita LMG 20598	34	Non-pathogenic
325	Burkholderia kururiensis LMG 19447	34	Non-pathogenic
326	Burkholderia phenoliruptrix LMG 22037	34	Non-pathogenic
327	Burkholderia xenovorans LMG 21463	34	Non-pathogenic
328	Serratia sp.	34	Non-pathogenic
329	Pandoraea norimbergensis LMG 18379	34	Non-pathogenic
330	**Escherichia coli(1)**	**35**	**Pathogenic**
331	**Escherichia coli(4)**	**35**	**Pathogenic**
332	**Klebsiella pneumoniae (1)**	**35**	**Pathogenic**
333	**Klebsiella pneumoniae (3)**	**35**	**Pathogenic**
334	**Escherichia coli(7)**	**35**	**Pathogenic**
335	**Escherichia coli(8)**	**35**	**Pathogenic**
336	**Escherichia coli(9)**	**35**	**Pathogenic**
337	**Escherichia coli(10)**	**35**	**Pathogenic**
338	**Escherichia coli(11)**	**35**	**Pathogenic**
339	**Escherichia coli(12)**	**35**	**Pathogenic**
340	**Escherichia coli(13)**	**35**	**Pathogenic**
341	**Escherichia coli(14)**	**35**	**Pathogenic**
342	**Escherichia coli(15)**	**35**	**Pathogenic**
343	**Escherichia coli(16)**	**35**	**Pathogenic**
344	**Escherichia coli(17)**	**35**	**Pathogenic**
345	**Escherichia coli(18)**	**35**	**Pathogenic**
346	**Escherichia coli(19)**	**35**	**Pathogenic**
347	**Escherichia coli(20)**	**35**	**Pathogenic**

348	**Escherichia coli(21)**	**35**	**Pathogenic**
349	**Escherichia coli(22)**	**35**	**Pathogenic**
350	**Escherichia coli(23)**	**35**	**Pathogenic**
351	**Escherichia coli(24)**	**35**	**Pathogenic**
352	**Klebsiella pneumoniae (4)**	**35**	**Pathogenic**
353	**Klebsiella pneumoniae (5)**	**35**	**Pathogenic**
354	**Klebsiella pneumoniae (6)**	**35**	**Pathogenic**
355	**Klebsiella pneumoniae (7)**	**35**	**Pathogenic**
356	**Klebsiella pneumoniae (8)**	**35**	**Pathogenic**
357	**Klebsiella pneumoniae (9)**	**35**	**Pathogenic**
358	**Klebsiella pneumoniae (15)**	**35**	**Pathogenic**
359	**Klebsiella pneumoniae (16)**	**35**	**Pathogenic**
360	**Klebsiella pneumoniae (17)**	**35**	**Pathogenic**
361	**Klebsiella pneumoniae (18)**	**35**	**Pathogenic**
362	**Pseudomonas aeruginosa(1)**	**35**	**Pathogenic**
363	**Klebsiella pneumoniae (13)**	**35**	**Pathogenic**
364	**Klebsiella pneumoniae (14)**	**35**	**Pathogenic**
365	Pseudomonas aurantiaca	36	Non-pathogenic
366	Pseudomonas chlororaphis	36	Non-pathogenic
367	Pseudomonas corrugate	36	Non-pathogenic
368	Pseudomonas fluorescens	36	Non-pathogenic
369	Cupriavidus necator LMG 1199	37	Non-pathogenic
370	Klebsiella sp.	37	Non-pathogenic
371	Pseudomonas aeruginosa PUPa3	37	Non-pathogenic
372	Citrobacter sp.	37	Non-pathogenic
373	Enterobacter spp.	37	Non-pathogenic
374	Lactobacillus brevis	37	Non-pathogenic
375	Lactobacillus hilgardii	37	Non-pathogenic
376	Oenococcus oeni	37	Non-pathogenic
377	Lactobacillus lactis	37	Non-pathogenic
378	Alpha proteobacteria	37	Non-pathogenic
379	Gamma proteobacteria	37	Non-pathogenic

380	Klebsiella oxytoca	37	Non-pathogenic
381	Lactobacillus sp.	37	Non-pathogenic
382	Lactococcus sp.	37	Non-pathogenic
383	Schizophyllum commune	37	Non-pathogenic
384	Alphaproteobacteria (e.g. Roseobacter sp.)	37	Non-pathogenic
385	Betaproteobacteria (Alcaligenes faecalis)	37	Non-pathogenic
386	Desulfovibrio acrylicus	37	Non-pathogenic
387	Parasporobacterium paucivorans	37	Non-pathogenic
388	Treponema denticola	37	Non-pathogenic
389	Brevibacterium linens	37	Non-pathogenic
390	Aspergillus flavus	38	Non-pathogenic
391	Aspergillus flavus NRRL 18543	38	Non-pathogenic
392	Aspergillus flavus NRRL 25347	38	Non-pathogenic
393	Aspergillus niger	38	Non-pathogenic
394	Aspergillus niger NRRL 326	38	Non-pathogenic
395	Aspergillus parasiticus NRRL 5862	38	Non-pathogenic
396	Penicillium glabrum NRRL 766	38	Non-pathogenic
397	Rhizopus stolonifer	38	Non-pathogenic
398	Rhizopus stolonifer NRRL 54667	38	Non-pathogenic
399	Serratia proteamaculans 42M	39	Non-pathogenic
400	**E. cloacae**	**40**	**Pathogenic**
401	**Staphylococcus aureus (3)**	**40**	**Pathogenic**
402	**Klebsiella pneumoniae (11)**	**40**	**Pathogenic**
403	**Staphylococcus aureus (6)**	**40**	**Pathogenic**
404	**Staphylococcus aureus (8)**	**40**	**Pathogenic**
405	**Staphylococcus aureus (10)**	**40**	**Pathogenic**
406	**Staphylococcus aureus (11)**	**40**	**Pathogenic**
407	**Staphylococcus aureus (13)**	**40**	**Pathogenic**
408	**Staphylococcus aureus (14)**	**40**	**Pathogenic**
409	**Staphylococcus aureus (17)**	**40**	**Pathogenic**
410	**Staphylococcus aureus (20)**	**40**	**Pathogenic**
411	**Klebsiella pneumoniae (10)**	**40**	**Pathogenic**

412	**Klebsiella pneumoniae (12)**	**40**	**Pathogenic**
413	**Klebsiella pneumoniae (19)**	**40**	**Pathogenic**
414	**Klebsiella pneumoniae (20)**	**40**	**Pathogenic**
415	**Staphylococcus aureus (7)**	**40**	**Pathogenic**
416	**Staphylococcus aureus (9)**	**40**	**Pathogenic**
417	**Staphylococcus aureus (12)**	**40**	**Pathogenic**
418	**Staphylococcus aureus (15)**	**40**	**Pathogenic**
419	**Staphylococcus aureus (16)**	**40**	**Pathogenic**
420	**Staphylococcus aureus (18)**	**40**	**Pathogenic**
421	**Staphylococcus aureus (19)**	**40**	**Pathogenic**
422	**Staphylococcus aureus (21)**	**40**	**Pathogenic**
423	**Staphylococcus aureus (22)**	**40**	**Pathogenic**
424	biofilms A (Rivularia sp./C. parietina community)	41	Non-pathogenic
425	C. parietina	41	Non-pathogenic
426	Cyanobacteria	41	Non-pathogenic
427	Rivularia sp.	41	Non-pathogenic
428	Laccaria bicolor	42	Non-pathogenic
429	Paxillus involutus MAJ	42	Non-pathogenic
430	Paxillus involutus NAU	42	Non-pathogenic
431	Penicillium sp.	43	Non-pathogenic
432	Desulfovibrio gigas	43	Non-pathogenic
433	Methanobacterium formicicum	43	Non-pathogenic
434	Methanobacterium thermoautotrophicum	43	Non-pathogenic
435	Methanosarcina barkeri	43	Non-pathogenic
436	Aeromonas veronii	43	Non-pathogenic
437	Geobacillus stearothermophilus	43	Non-pathogenic
438	Clostridium collagenovorans	43	Non-pathogenic
439	Desulfovibrio vulgaris	43	Non-pathogenic
440	Enterobacter cloacae	43	Non-pathogenic
441	Rhodobacter spaeroides	43	Non-pathogenic
442	Rhodocyclus tenuis	43	Non-pathogenic
443	Rhodospirillum rubrum	43	Non-pathogenic

444	Aspergillus sp.	43	Non-pathogenic
445	Candida humicola	43	Non-pathogenic
446	Scopulariopsis brevicaulis	43	Non-pathogenic
447	Methanobacterium sp.	43	Non-pathogenic
448	Bacterium from CFB group	43	Non-pathogenic
449	Saccharomyces cerevisiae Y1001	44	Non-pathogenic
450	Serratia spp. B2675	44	Non-pathogenic
451	Serratia spp. B675	44	Non-pathogenic
452	Saccharomyces cerevisiae	44	Non-pathogenic
453	Xanthomonas campestris pv. vesicatoria 85-10	45	Non-pathogenic
454	Sulfitobacter pontiacus	46	Non-pathogenic
455	Sulfitobacter pontiacus BIO-007	46	Non-pathogenic
456	Sulfitobacter sp.	46	Non-pathogenic
457	Loktanella hongkongensis strain Bio-204	46	Non-pathogenic
458	Sulfitobacter dubius BIO-205	46	Non-pathogenic
459	bacterial strains from the North Sea, the Arctic Ocean, or of terrestrial origin	46	Non-pathogenic
460	Oceanibulbus indolifex HEL-45	46	Non-pathogenic
461	Roseobacter gallaeciensis strain PIC-68	46	Non-pathogenic
462	Stappia marina strain DFL-11	46	Non-pathogenic
463	Sulfitobacter sp. PIC-70	46	Non-pathogenic
464	**Escherichia coli(5)**	**47**	**Pathogenic**
465	**Staphylococcus aureus (4)**	**47**	**Pathogenic**
466	Staphylococcus sp.	48	Pathogenic
467	Staphylococcus xylosus	48	Pathogenic
468	Clostridium sp.	48	Pathogenic
469	Bacteroides distasonis	48	Non-pathogenic
470	Bacteroides ovatus	48	Non-pathogenic
471	Bacteroides thetaiotamicron	48	Non-pathogenic
472	Bacteroides vulgatus	48	Non-pathogenic
473	Capnocytophaga ochracea ATCC 33596	48	Non-pathogenic
474	Clostridium bifermentans	48	Non-pathogenic

475	Clostridium sporogenes	48	Non-pathogenic
476	Fusobacterium nucleatum	48	Non-pathogenic
477	Porphyromonas gingivalis	48	Non-pathogenic
478	Porphyromonas gingivalis FDC381	48	Non-pathogenic
479	Porphyromonas gingivalis W83	48	Pathogenic
480	Prevotella intermedia ATCC 25261	48	Non-pathogenic
481	Prevotella loescheii	48	Non-pathogenic
482	Prevotella loescheii ATCC 15930	48	Non-pathogenic
483	Veillonella spp.	48	Non-pathogenic
484	Actinobacillus actinomycetemcomitans Y4	48	Non-pathogenic
485	Bacteroides bivius	48	Non-pathogenic
486	Clostridium butyricum	48	Non-pathogenic
487	Clostridium cadaverum	48	Non-pathogenic
488	Clostridium fallax	48	Non-pathogenic
489	Clostridium histolyticum	48	Non-pathogenic
490	Clostridium tertium	48	Non-pathogenic
491	Lactobacillus casei NCIB 8010	49	Non-pathogenic
492	Lactobacillus plantarum	49	Non-pathogenic
493	Lactobacillus plantarum NCIB 6376	49	Non-pathogenic
494	Lactococcus lactis	49	Non-pathogenic
495	Lactococcus lactis DSM 20202	49	Non-pathogenic
496	Leuconostoc cremoris DSM 20346	49	Non-pathogenic
497	Leuconostoc dextranicum DSM 20484	49	Non-pathogenic
498	Leuconostoc mesenteroides DSM 20343	49	Non-pathogenic
499	Leuconostoc oenos	49	Non-pathogenic
500	Leuconostoc oenos B66	49	Non-pathogenic
501	Leuconostoc oenos 19	49	Non-pathogenic
502	Leuconostoc oenos 30	49	Non-pathogenic
503	Leuconostoc oenos 36	49	Non-pathogenic
504	Leuconostoc oenos 37D	49	Non-pathogenic
505	Leuconostoc oenos 7B	49	Non-pathogenic
506	Leuconostoc oenos DSM 20252	49	Non-pathogenic

507	Leuconostoc oenos DSM 20255	49	Non-pathogenic
508	Leuconostoc oenos DSM 20257	49	Non-pathogenic
509	Leuconostoc oenos Lc5x	49	Non-pathogenic
510	Leuconostoc paramesenteroides DSM 20288	49	Non-pathogenic
511	Pediococcus damnosus DSM 20331	49	Non-pathogenic
512	Bacteroides gracilis CCUG 13143 (ATCC 33236)	50	Non-pathogenic
513	Bacteroides ureolyticus CCUG 7319 (ATCC 33387)	50	Non-pathogenic
514	Campylobacter fetus subsp. venerealis CCUG 538 (ATCC 19438)	50	Non-pathogenic
515	Wolinella recta FDC 371 (ATCC 33238)	50	Non-pathogenic
516	Wolinella succinogenes CCUG 12550 (ATCC 29543)	50	Non-pathogenic
517	Wolinella curva CCUG 13146 (ATCC 35224)	50	Non-pathogenic

3.4.2. Graph-clustering based on DPClus result

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 algorithm DPClus. Fig. 3.3 shows the overall network, which displays all the generated clusters in such a way that intra cluster edges are green and inter cluster edges are red. Fig. 3.4 (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 Fig. 3.4 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. Nodes enclosed by dotted rectangle are consisting of only pathogenic bacteria, the only node enclosed by the dotted circle is consisting of both pathogenic and non-pathogenic bacteria and the rest nodes are consisting of only non-pathogenic bacteria. Fig. 3.4 (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 rest 30 clusters are independent nodes. Only cluster 1 contains both pathogenic and non-pathogenic microorganisms. Clusters 2, 7, 14, 21, 26 and 40 consist of only pathogenic bacteria while the other clusters are consisting of only non-pathogenic 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 e.g. cluster 2 and 7.

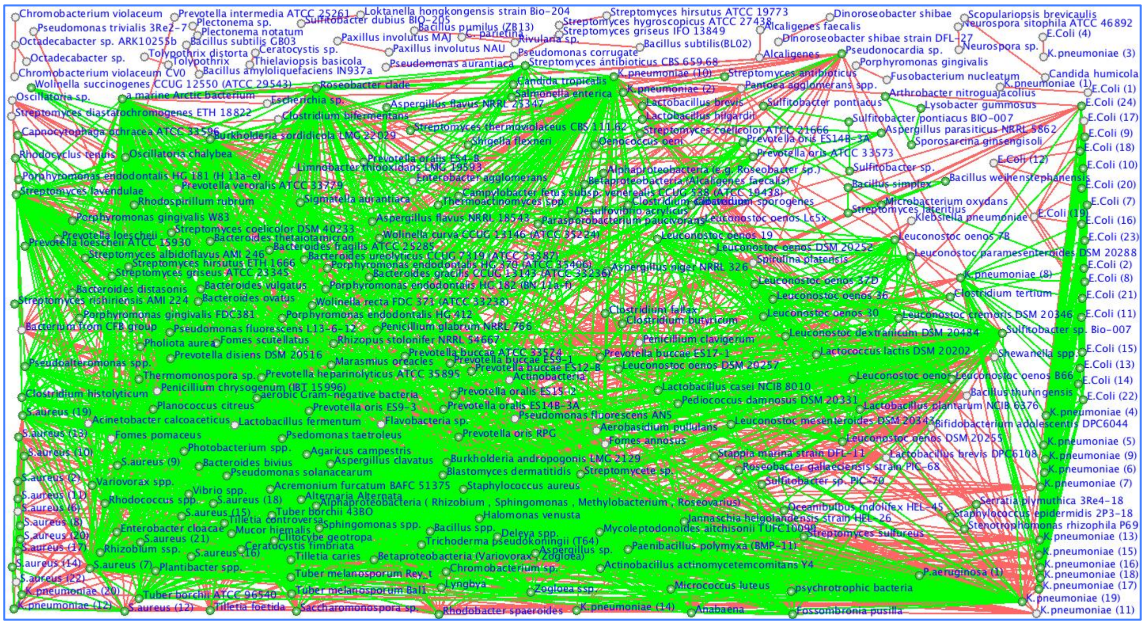


Figure 3.3. *Overall DPClus network, which displays all the generated clusters.*

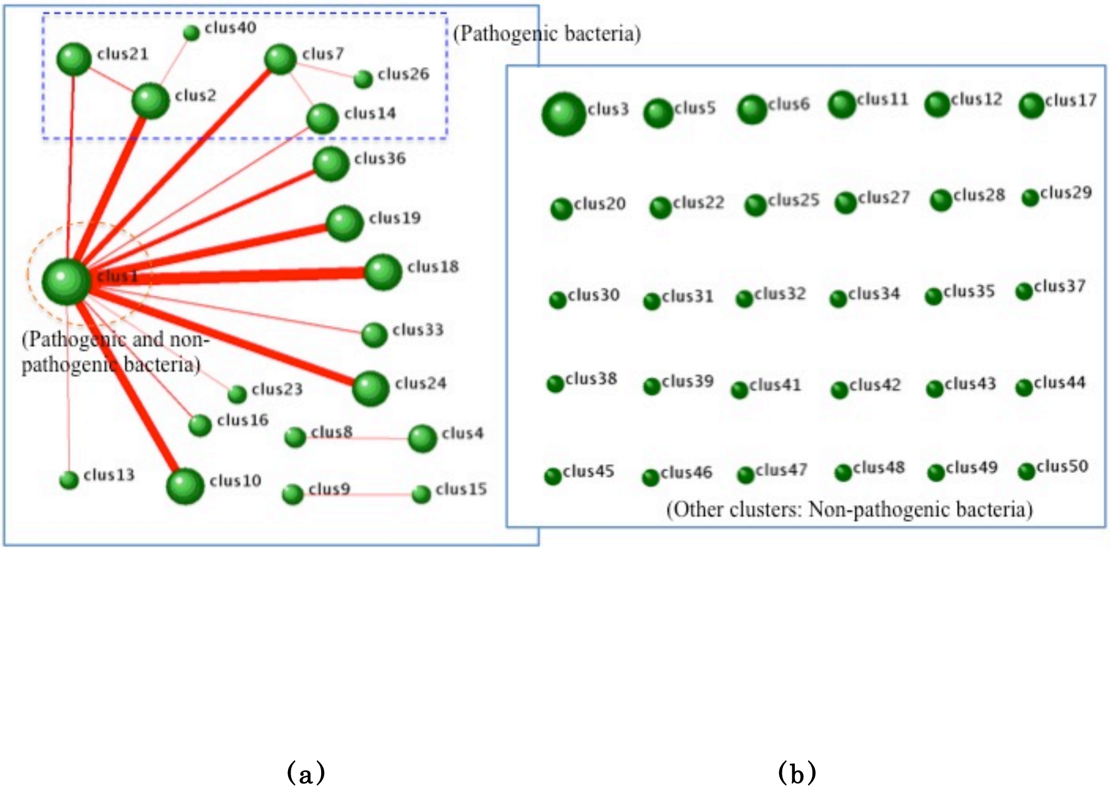


Figure 3.4. *Hierarchical graph of DPClus clustering result in case of* cp_{in} = *0.5 and* d_{in} = *0.6. (a) Connected nodes. (b) Independent nodes.*

Fig. 3.4 (a) shows that cluster 2, 7, 14, 21, 26 and 40 are connected by red edges, which reflect VOC similarity between pathogenic microorganisms. Also, there are VOC based similarity between non-pathogenic species of cluster 1 and clusters 10, 13, 16, 18, 19, 23, 24, 33 and 36. The red edges between cluster 4 and 8 and between cluster 9 and 15 are also because of VOC similarity between non-pathogenic species of those clusters. Here it is noteworthy that the rest of 30 clusters consisting of non-pathogenic species are independent clusters, which implies that many non-pathogenic groups of species emit quite unique types of VOCs as shown in Fig. 3.4 (b).

Fig. 3.5 shows the microorganism species belong to cluster 1 (pathogenic and non-pathogenic), cluster 7 (pathogenic only) and cluster 10 (non-pathogenic 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.

Here, we also examined different values of density, d_{in} to the clustering result. We used cp_{in}= 0.5 and d_{in} = 0.6 for the experiments discussed in Fig. 3.4 and Fig. 3.5. However the variation of density value, d_{in} can also affect the outcome of the clustering. Fig. 3.6, Fig. 3.7 and Fig. 3.8 show the hierarchical graph of DPClus clustering for d_{in} = 0.7, d_{in} = 0.8 and d_{in} = 0.9, respectively. If high value is used for d_{in}, the generated clusters are of high density but smaller in size and hence relatively more in number. 51 clusters were generated for d_{in} = 0.7, 52 clusters were generated for d_{in} = 0.8, and 55 clusters were generated for d_{in} = 0.9. However, many such clusters are consisting of only two, three or four pathogenic microorganisms. The highest number of clusters containing pathogenic clusters ≥ 4 is obtained in case of cp_{in} = 0.5 and d_{in} = 0.6. Hence, the classification results between pathogenic and non-pathogenic microorganism species are best obtained in case of cp_{in} = 0.5 and d_{in} = 0.6. In general, from the periphery tracking point of view, we consider that a reasonable and balanced value for cp_{in} is 0.5 and d_{in} = 0.6 because it is in the middle of the parameter space. However it can be said that the larger the value of cp_{in} and d_{in} the more spherical the structure of the generated complexes (Md Altaf-Ul-Amin et al. 2006).

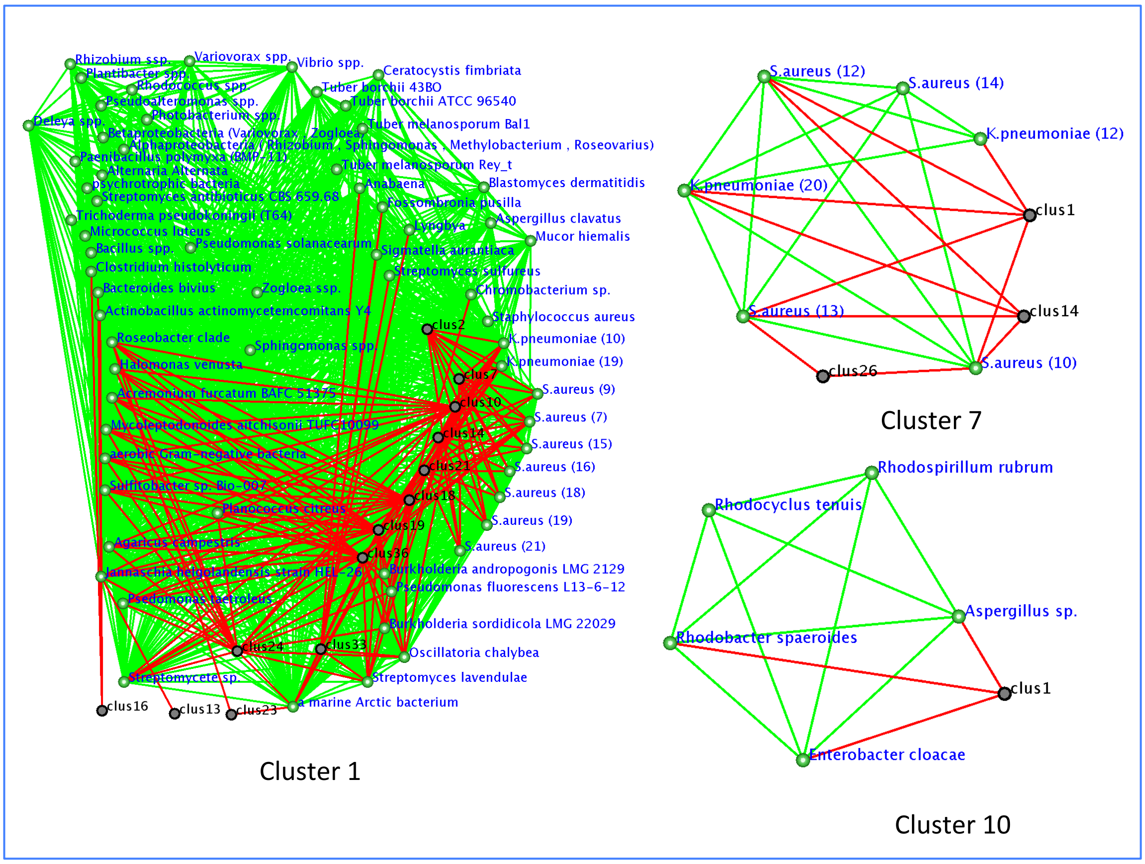


Figure 3.5. *The three example clusters of microorganism species that classify the microorganism species according to their pathogenicity.*

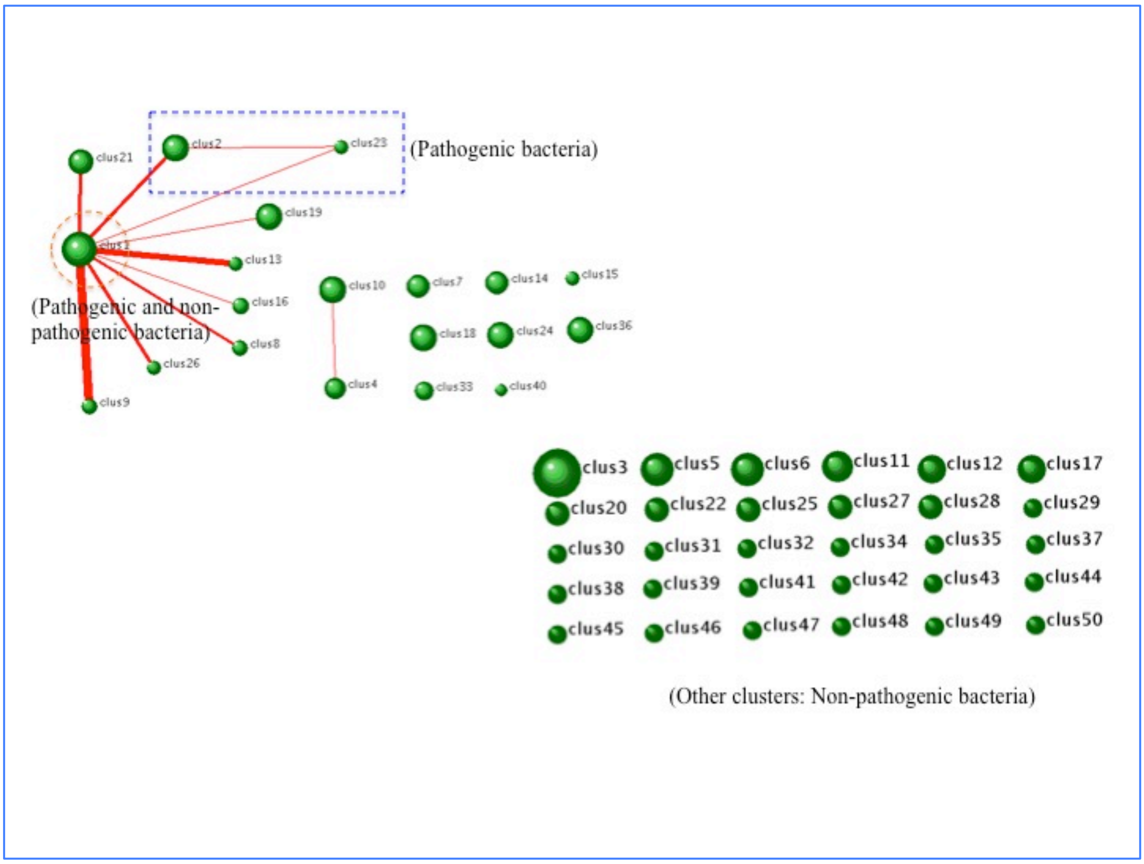


Figure 3.6. *Hierarchical graph of DPClus clustering result in case of cp_{in}= 0.5 and d_{in} = 0.7.*

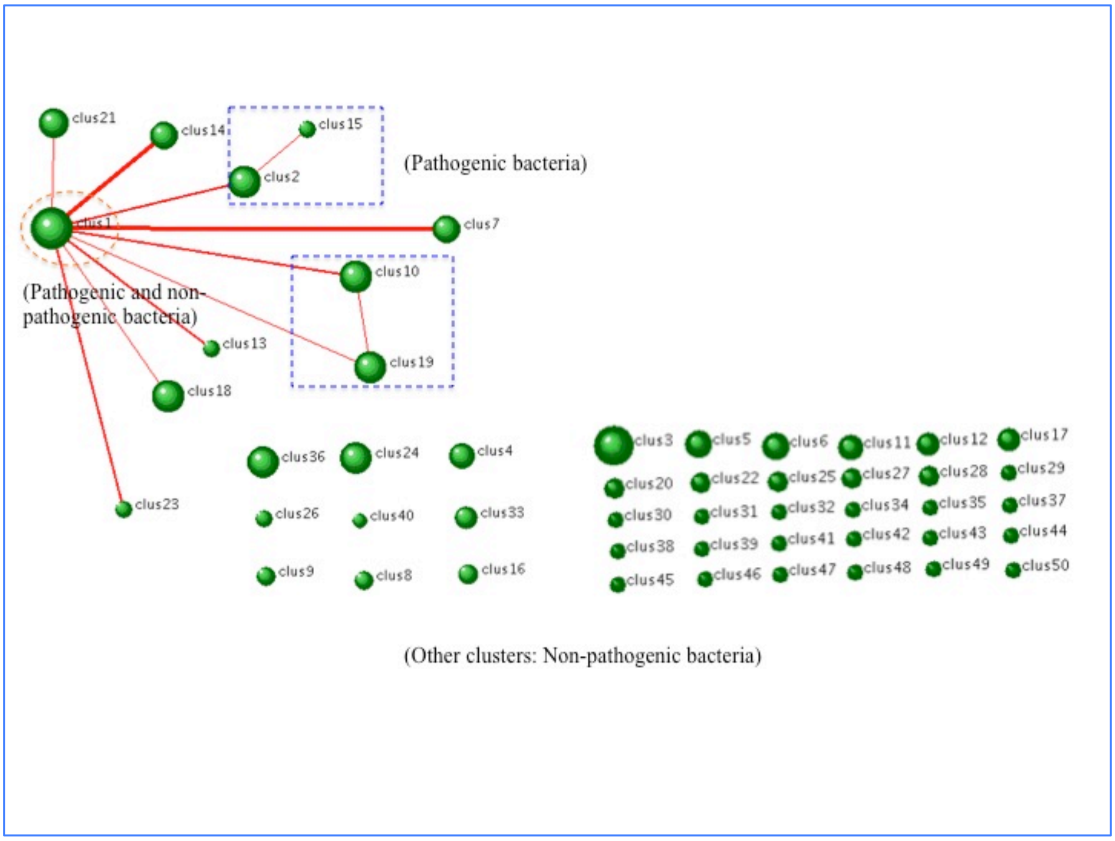


Figure 3.7. *Hierarchical graph of DPClus clustering result in case of* cp_{in}= *0.5 and* d_{in} = *0.8.*

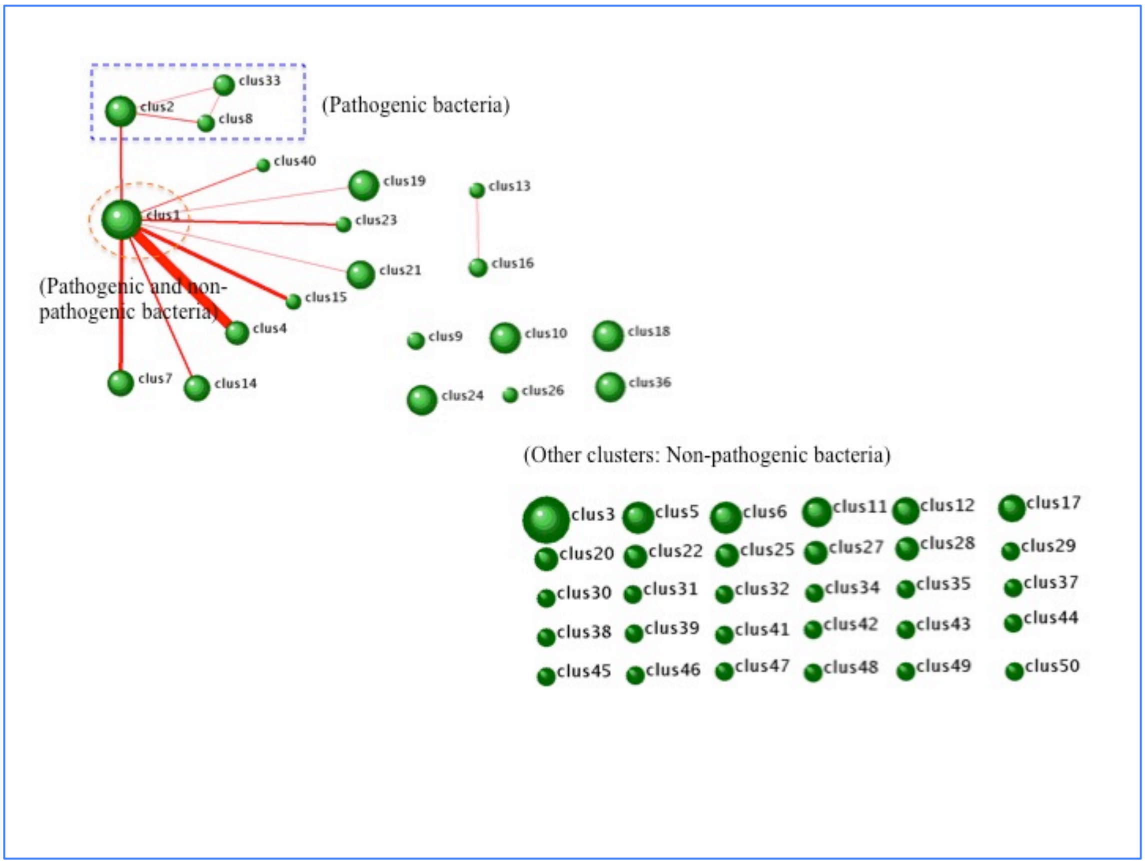


Figure 3.8. *Hierarchical graph of DPClus clustering result in case of cp_{in}= 0.5 and d_{in} = 0.9.*

3.5. Summary

The results of hierarchical clustering and graph clustering based on DPClus algorithm 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 non-existence of relations between different pathogenic and non-pathogenic groups of microorganisms. The variation of input density value, d_{in} can affect the outcome of the clustering. It is important to choose the d_{in} value for DPClus clustering. The classification results between pathogenic and non-pathogenic microorganism species are best obtained in case of cp_{in} = 0.5 and d_{in} = 0.6. It is because the highest number of clusters containing pathogenic clusters (6 clusters consisting pathogenic species) is obtained in case of cp_{in} = 0.5 and d_{in} = 0.6. The classification achieved by DPClus is better in a sense it produced more clusters with 100% membership of either pathogenic or non-pathogenic microorganisms.

Chapter 4

Classification of VOCs Based On Chemical Structural Similarity

In previous chapter, we described the clustering analysis methods to cluster the microorganism species based on VOC metabolite contents similarity. In this chapter, we focus on the second type of data that we have accumulated in our database; VOCs emitted by other organisms such as plants, animals and humans with their related biological activities. For the second data, we performed heatmap clustering utilizing Tanimoto coefficient as the similarity index between chemical structures to cluster all VOCs. We further accessed the statistical significance of the clusters using hypergeometric *p*-values to understand the relationships between chemical structures of VOCs and their biological activities. We also compared several types of hierarchical clustering methods (single, complete, average, centroid, median linkage and Ward's method) with DPClus algorithm to cluster the chemical structures of VOCs using Tanimoto coefficient as a similarity measure. Additionally, we extended our analysis by implementing supervised machine learning methods such as Deep Neural Network (DNN), Gradient Boosting Machine (GBM), Random Forest (RF) and Generalized Linear Model (GLM) as classification models for predicting the biological activities of VOCs based on their chemical structures.

4.1. Background

Chemical similarity or molecular similarity refers to the similarity of chemical elements, molecules or chemical compounds with respect to either structural or functional qualities, i.e. the effect that the chemical compound has on reaction partners in inorganic or biological settings. Biological effects and thus also similarity of effects are usually quantified using the biological activity of a compound. In pharmacology, biological activity describes the beneficial or adverse effects of a drug on living matter (Miller-Keane 1993). When a drug is a complex chemical mixture, this activity is exerted by the substance's active ingredient or pharmacophore but can be modified by the other constituents. Among the various properties of chemical compounds, biological activity plays a crucial role since it suggests uses of the compounds in the medical applications. However, chemical compounds may show some adverse and toxic effects which may prevent their use in medical practice. The notion of chemical similarity is one of the most important concepts in chemoinformatics (Nikolova & Jaworska 2003; P. 2014; Maggiora et al. 2014; Cereto-Massagué et al. 2015). It plays an important role in modern approaches to predict the properties of chemical compounds and also in conducting drug design studies by screening large databases containing structures of available chemicals.

The importance of structural similarity derives in large part from the *Similar Property Principle*, which states that molecules that are structurally similar are likely to have similar properties (Maggiora & Shanmugasundaram 2004). This relationship underlies a range of chemoinformatics techniques such as similarity searching, molecular diversity analysis, clustering and a range of quantitative structure activity relationships (QSAR) or quantitative structure–property relationships (QSPR) methods.

A similarity measure has three components: the representation or descriptor that is used to characterize the two molecules that are being compared; the weighting scheme that is used to reflect the relative importance of different parts of the representation; and the similarity coefficient that is used to quantify the degree of resemblance between two appropriately weighted structural representations. A comprehensive review by (Willett 2009) provides the detail explanations on the three components. The book by (Todeschini & Consonni 2000; Todeschini et al. 2012; Consonni & Todeschini 2012) is the standard work on ways of describing chemical structures. Many of the descriptors reported by these authors have been used in studies of molecular similarity. They are commonly divided into three classes: whole molecule or one-dimensional (1D) descriptors; descriptors that can be calculated from two-dimensional (2D) representations of molecules; and descriptors that can be calculated from three-dimensional (3D) representations. The role of weighting schemes in similarity measures has attracted much less attention to date than have the roles of the representation and of the similarity coefficient. In a weighted fingerprint, each fragment has a weight assigned to it reflecting its relative degree of importance, so that a high-weighted fragment occurring in both the reference structure and a database structure would make a greater contribution to the overall structural similarity than would a low-weighted fragment. The most obvious form of weighting is the number of times that a fragment occurs in a molecule, so that a fingerprint encodes fragment occurrences, rather than the fragment incidences encoded in a binary fingerprint. Similarity coefficients have been developed for use in many different application domains, and there is hence a wide range available that can be used for the measurement of structural similarity. For example, a recent study by (Todeschini et al. 2012; Consonni & Todeschini 2012) discussed no less than 51 different coefficients that can be used to compute the similarity between binary fingerprints, and there have hence been many comparative studies to identify the most appropriate for chemical applications. One of the earliest such studies, (P. Willett et al. 1998) showed that the Tanimoto coefficient, an example of the class of coefficients known as association coefficients, provided an effective way of comparing 2D fingerprints. The studies conducted by (Bajusz et al. 2015) also proved that Tanimoto coefficient, along with Dice index, Cosine coefficient and Soergel distance were identified to be the best metrics for similarity calculations. This is one of the reasons why we choose Tanimoto coefficient as similarity measure to calculate similarity between two volatile organic compounds and perform clustering.

Here, we investigate the relationships between chemical structures of VOCs and biological activities by applying unsupervised (clustering) and supervised machine learning methods (Deep Neural Network, Gradient Boosting Machine, Random Forest and Generalized Linear Model) as classification models for predicting the biological activities of VOCs based on their chemical structures.

4.2. Datasets

The second data type that we have accumulated until now are 1044 species-species interactions via 341 VOCs associated with 11 groups of biological activities. The biological activities of VOCs are classified into two types: (1) chemical ecology related activities, in which most VOCs involved in interaction between species for survival of organisms such as defense and antimicrobial, (2) 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, enhance plant growth, inhibit root growth and repellent activities and 42.7% are human health related activities such as disease biomarker, odor, anti-cholinesterase and antioxidant as shown in Fig. 4.1. The detail explanations related to these 11 biological activities are described in Table 4.1.

Table 4.1. *The description on each of biological activities.*

Biological activities	Description on biological activities
Antifungal	Limits or prevents the growth of yeasts and other fungal organisms.
Antimicrobial	Kills or inhibits the growth of microorganisms.
Attractant	A substance, known as pheromone that attracts

	animals, such as insects.
Defense	Adaptation that promotes the survivability of an organism by protecting it from its natural enemies.
Enhance plant growth	Increase or promote plant growth.
Inhibit root growth	Decrease root growth of plants.
Repellent	A substance that deters insects or other pests from approaching or settling.
Anti-cholinesterase	An agent that inhibits acetylcholinesterase, the enzyme that breaks down acetylcholine at junctions of cholinergic nerve endings and effector organs or postsynaptic neurons.
Antioxidant	A molecule that inhibits the oxidation of other molecules.
Biomarker	A molecule, by which a particular pathological process or disease can be identified.
Odor	The property of a substance that activates the sense of smell.

There are many VOCs, which have several biological activities. Fig. 4.2 shows the relative frequencies of VOCs, which have several biological activities. There are 239 VOCs (about 70%), which have only one specific biological activity. 28 VOCs have 2 biological activities, 52 VOCs have 3 biological activities, 17 VOCs have 4 biological activities, 3 VOCs have 5 biological activities and only 2 VOCs have 6 biological activities. For simplicity, we empirically select the most relevant biological activity to each particular compound and the resultant distribution of the compounds with refer to biological activities is shown in Fig. 4.3. It facilitates to investigate the relationships between VOCs and their biological activities.

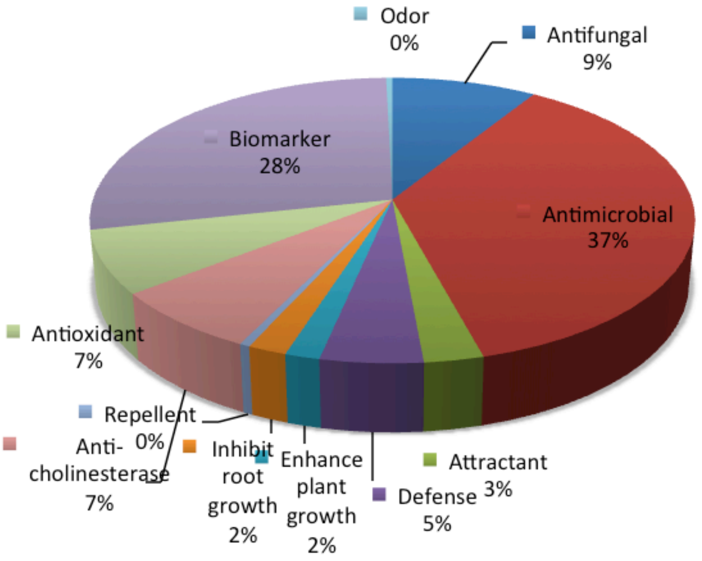


Figure 4.1. *Pie chart showing the relative frequencies VOCs belonging to 11 biological activities.*

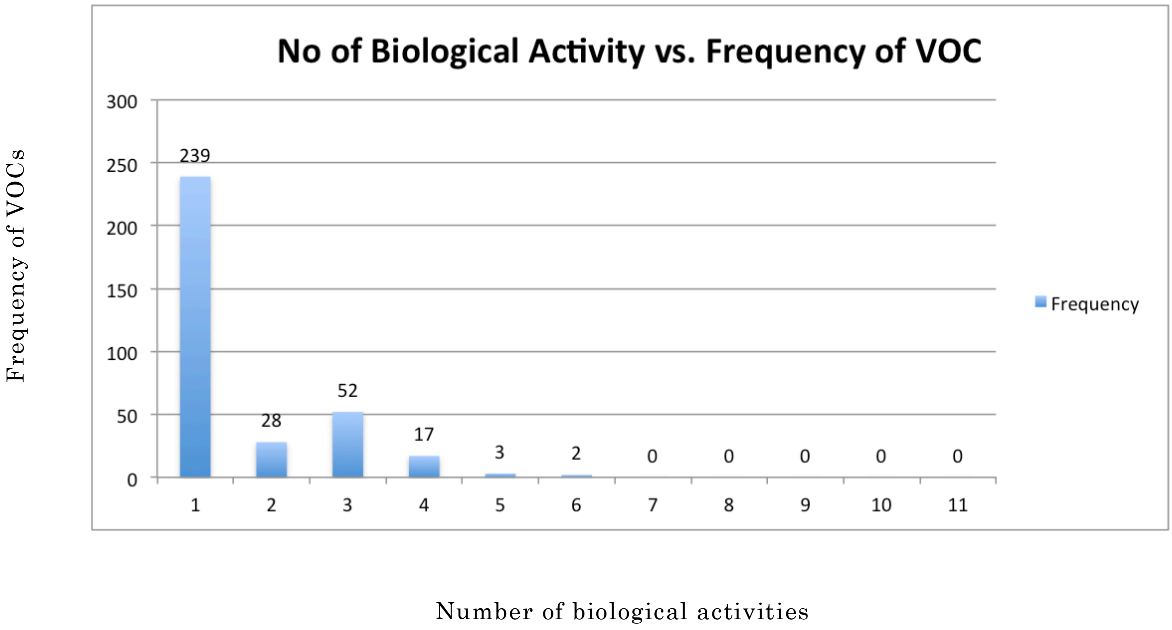


Figure 4.2. *The relative frequencies of VOCs, which have several biological activities.*

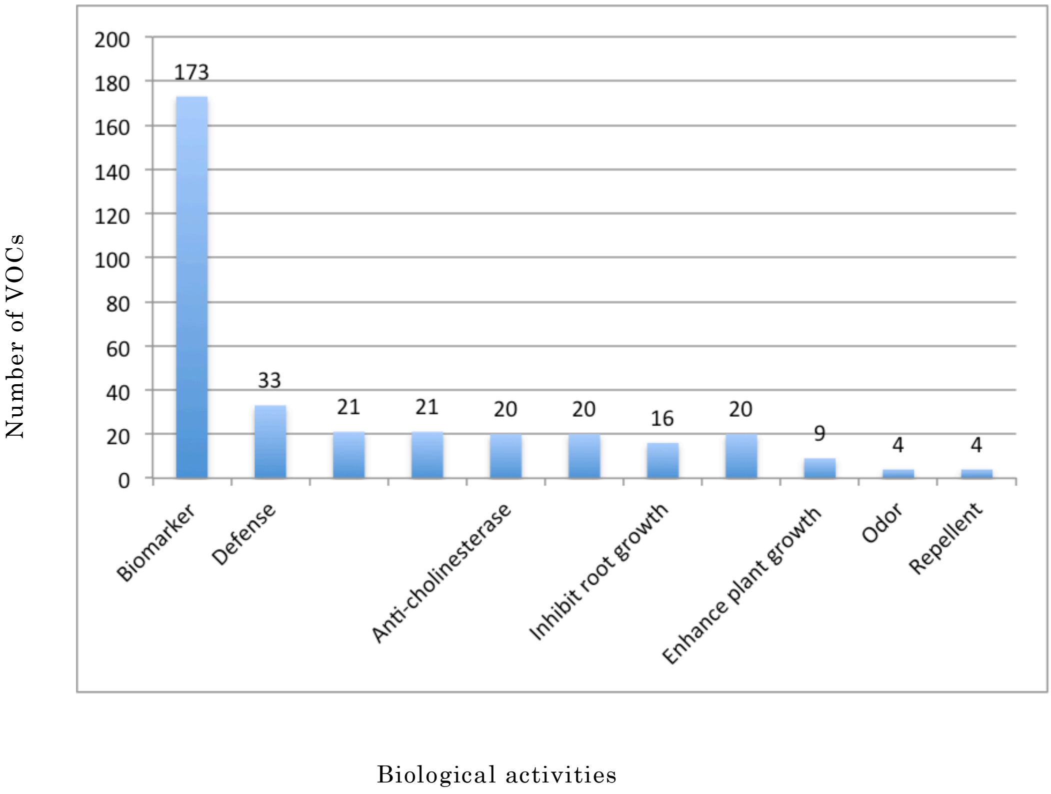


Figure 4.3. *The most relevant biological activity for each of VOCs.*

4.3. Methods

4.3.1. Heatmap clustering and hypergeometric distribution

We performed 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 application of Tanimoto coefficient in cheminformatics has been reported in (Butina 1999; Godden et al. 2000; Cha et al. 2009; Rojas-Cherto et al. 2012; Dimitrov et al. 2014). Recently, (Liu et al. 2013) has used Tanimoto coefficient as a novel approach to classify plants based on metabolite content similarity. The Tanimoto coefficient is defined as equation (4.1), which is the proportion of the features shared between two compounds divided by their union (Peter Willett et al. 1998).

$$Tanimoto_{A,B} = \frac{AB}{A+B-AB} \tag{4.1}$$

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 (Patterson et al. 1996). For the purpose of calculating Tanimoto coefficient, it is obligatory to assign fingerprints to the compounds. ChemMine package in R was used to generate binary fingerprints and calculation of Tanimoto coefficient (Cao et al. 2008; Cao et al. 2014). 2-D 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 binary PubChem fingerprints are calculated during the SDF import and stored in a searchable descriptor database as a list object. The detail description of PubChem fingerprint can be referred in Appendix A.

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 equation (4.2).

$$p - value = 1 - \sum_{i=0}^{K-1} \frac{\binom{V}{i}\binom{N-V}{C-i}}{\binom{N}{C}} \tag{4.2}$$

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 a specific K successes (out of N total draws) from the whole population. The test is often used to identify which sub-populations are over- or under-represented 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.

4.3.2. Comparison of clustering methods

Based on Tanimoto similarity measure, we applied two different clustering methods to classify the VOCs, which are DPClus clustering and hierarchical clustering. Both methods were described in Chapter 3. The reason why we apply both methods is that, we want to determine the degree of cluster overlap and how well it classified chemical structures of VOCs into clusters. Additionally, we also point out the advantages and limitations of both clustering methods.

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

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

4.3.3. Molecular fingerprints

To extend our findings, we also developed classification models to predict biological activities of VOCs based on their chemical structures by several machine learning methods such as Deep Neural Network (LeCun et al. 2015; Ma et al. 2015; Chandra & Sharma 2016), Gradient Boosting Machine (Friedman 2001; Friedman 2002; Natekin & Knoll 2013; Chen & Guestrin 2016), Random Forest (Breiman 2001) and Generalized Linear Model (Cook 1998). Eight types of molecular fingerprints are used to represent the molecules, as following:

(1) PubChem (PubChem, 881 bits),

(2) CDK (CDK, 1024 bits),

(3) Extended CDK (Extended, 1024bits),

(4) MACCS (MACCS, 166 bits),

(5) Klekota-Roth (KR, 4860 bits) (Klekota & Roth 2008),

(6) Substructure (SubFP, 307 bits),

(7) Estate (Estate, 79 bits),

(8) Atom pairs (AP, 780 bits) (Carhart et al. 1985).

The detail description on substructures and each fingerprint method can be referred to this website (http://www.scbdd.com/chemdes/list-fingerprints/). We also proposed a new type of fingerprint, by combining all features and substructures obtained by these fingerprints (Combine, 9121 bits). The reason why we use many types of fingerprints, is that we want to investigate which fingerprint method can generate the best prediction model. We converted the SDF files of all 341 VOCs into binary fingerprints using ChemDes software (Dong et al. 2015). After we obtained the binary matrix of fingerprints, we performed the data-processing method by removing all columns that contain "0". This is because it might be not relevant for the classification of VOCs based on substructures. The features or substructures displayed in binary matrix, was used as input to the classification models. There are 11 classes of biological activities, which have been used as outputs for the classification model. The VOC-Substructure-Biological activities relations can be represented as a matrix, shown in Table 4.2 where rows represent VOCs and columns represent substructures of molecular fingerprints. We added one additional column to represent biological activities for each of VOCs.

Table 4.2. *Representation of VOCs, substructures and biological activities as a two-dimensional matrix.*

VOCs	Substructures						Biological
	S_1	S_2	S_3	S_4	…	S_M	Activities
VOC_1	1	0	1	1	…	0	Antimicrobial
VOC_2	1	1	0	0	…	0	Biomarker
VOC_3	0	1	0	1	…	0	Defense
…	…	…	…	…	…	…	
VOC_N	1	0	0	0	…	1	Odor

Machine learning algorithms are generally developed in computer science or adjacent disciplines and find their way into chemical modeling by a process of diffusion. Recently, machine learning methods are popular in chemoinformatics and quantitative structure–activity relationships (QSAR), which usually predicting the unknown property values of a test set of molecules based on the known values for a training set. An example of existing machine learning algorithms is given in Fig. 4.4. We implemented four types of supervised machine learning methods for predicting biological activities of VOCs, which are Deep Neural Network (DNN), Gradient Boosting Machine (GBM), Random Forest (RF) and Generalized Linear Model (GLM) using H2O package in R program (Intelligence 2015; Anqi et al. 2015).

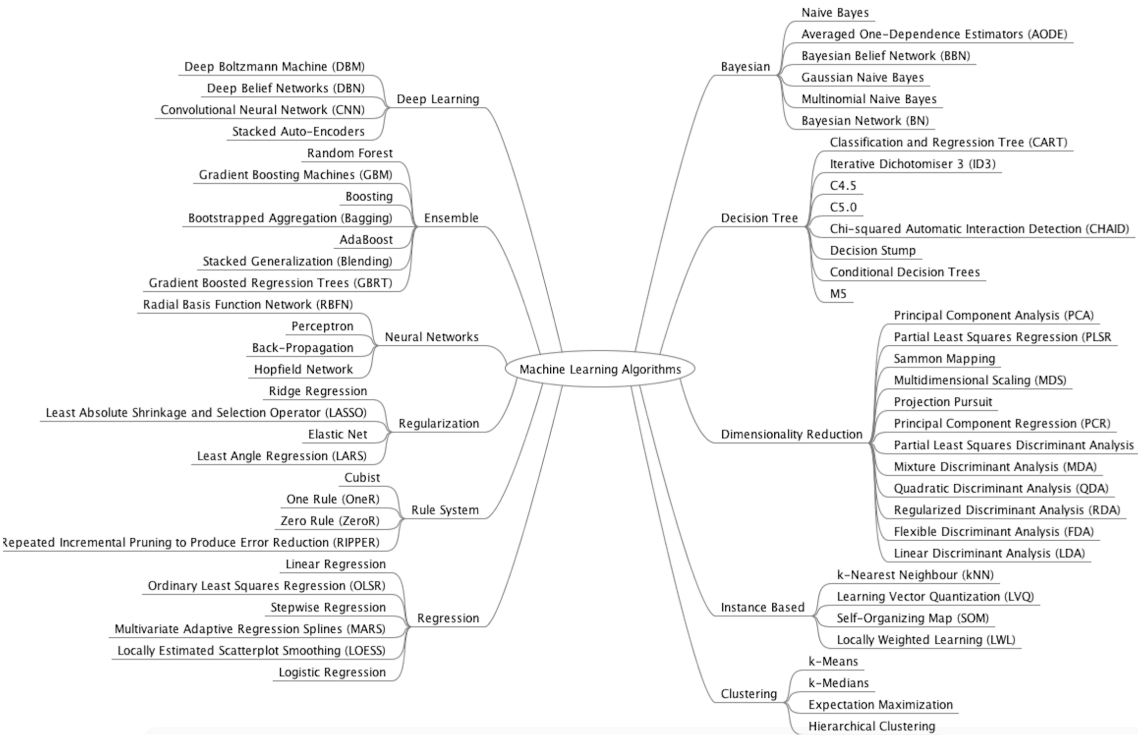

Machine Learning Algorithms
Deep Learning
Deep Boltzmann Machine (DBM)
Deep Belief Networks (DBN)
Convolutional Neural Network (CNN)
Stacked Auto-Encoders
Ensemble
Random Forest
Gradient Boosting Machines (GBM)
Boosting
Bootstrapped Aggregation (Bagging)
AdaBoost
Stacked Generalization (Blending)
Gradient Boosted Regression Trees (GBRT)
Neural Networks
Radial Basis Function Network (RBFN)
Perceptron
Back-Propagation
Hopfield Network
Regularization
Ridge Regression
Least Absolute Shrinkage and Selection Operator (LASSO)
Elastic Net
Least Angle Regression (LARS)
Rule System
Cubist
One Rule (OneR)
Zero Rule (ZeroR)
Repeated Incremental Pruning to Produce Error Reduction (RIPPER)
Regression
Linear Regression
Ordinary Least Squares Regression (OLSR)
Stepwise Regression
Multivariate Adaptive Regression Splines (MARS)
Locally Estimated Scatterplot Smoothing (LOESS)
Logistic Regression
Bayesian
Naive Bayes
Averaged One-Dependence Estimators (AODE)
Bayesian Belief Network (BBN)
Gaussian Naive Bayes
Multinomial Naive Bayes
Bayesian Network (BN)
Decision Tree
Classification and Regression Tree (CART)
Iterative Dichotomiser 3 (ID3)
C4.5
C5.0
Chi-squared Automatic Interaction Detection (CHAID)
Decision Stump
Conditional Decision Trees
M5
Dimensionality Reduction
Principal Component Analysis (PCA)
Partial Least Squares Regression (PLSR
Sammon Mapping
Multidimensional Scaling (MDS)
Projection Pursuit
Principal Component Regression (PCR)
Partial Least Squares Discriminant Analysis
Mixture Discriminant Analysis (MDA)
Quadratic Discriminant Analysis (QDA)
Regularized Discriminant Analysis (RDA)
Flexible Discriminant Analysis (FDA)
Linear Discriminant Analysis (LDA)
Instance Based
k-Nearest Neighbour (kNN)
Learning Vector Quantization (LVQ)
Self-Organizing Map (SOM)
Locally Weighted Learning (LWL)
Clustering
k-Means
k-Medians
Expectation Maximization
Hierarchical Clustering

Figure 4.4. *Example of existing machine learning algorithms.*

4.3.4. Deep Neural Network (DNN)

A neural network is network composed of simulated "neurons". Each neuron has multiple inputs and one output. Each input arrow is associated with a weight, *wi*. The neuron is also associated with a function, *f(z)*, called the activation function, and a default bias term *b*, as shown in Fig. 4.5.

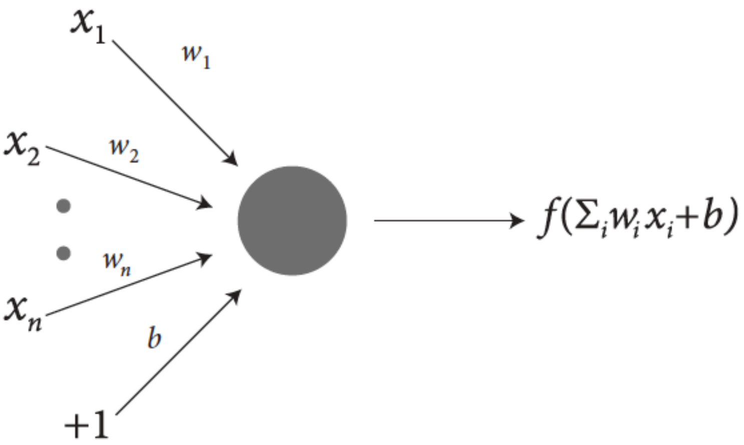


Figure 4.5. *The basic unit of a neuron.*

A row of neurons form a layer of the neural network, and a DNN is built from several layers of neurons (Fig. 4.6).

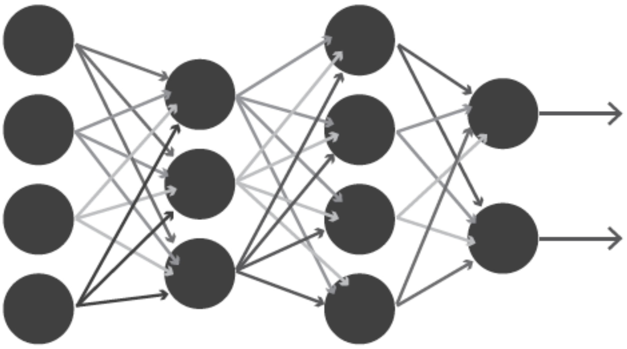


Figure 4.6. *The deep neural network model.*

Normally, there are three types of layers in a DNN: (1) the input layer, where the fingerprint of a molecule is entered (2) the output layer where predictions are generated (3) the hidden (middle) layers; the word "deep" in deep neural nets implies more than one hidden layer. In this study, we used 3 types of activation functions in the hidden layers: (1) the tanh function, (2) the rectified linear unit (ReLU) function (Maas et al. 2013), and (3) the maxout function (Goodfellow et al. 2013; Zhang et al. 2014). The output layer can have one or more neurons, and each output neuron generates prediction for biological activities. The layout of a DNN, including the number of layers and the number of neurons in each layer, needs to be defined, along with the choice of the activation function in each neuron. Dropout is a method, which remove some of neurons in the input and hidden layer to avoid over fitting (Srivastava et al. 2014; Baldi & Sadowski 2014). Since there are a lot of parameters that can impact model accuracy, we implemented hyper-parameter tuning for the network optimization. Multi-dimensional hyper-parameter optimization (more than 4 parameters) can be more efficient with random parameter search. For a random parameter search, we did a loop over models with parameters drawn uniformly from a given range, and then we chose the best parameter for our DNN network.

4.3.5. Gradient boosting machine (GBM)

Gradient boosting machines are a family of powerful machine-learning techniques for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees (Natekin & Knoll 2013). It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function. Gradient boosting involve three elements; (1) A loss function to be optimized, (2) A weak learner to make predictions and (3)

An additive model to add weak learners to minimize the loss function. It is clear that these elements would greatly affect the GBM model properties. The GBM framework provides the practitioner with such design flexibility.

4.3.6. Random forest (RF)

RF is an ensemble method that consists of many decision trees (Breiman 2001) for classification and regression tasks. It operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes output by individual trees. For prediction a new sample is pushed down the tree. It is assigned the label of the training sample in the terminal node it ends up in. This procedure is iterated over all trees in the ensemble and the mode vote of all trees is reported as the random forest prediction.

4.3.7. Generalized linear model (GLM)

In statistics, the generalized linear model (GLM) is a flexible generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution. The GLM generalizes linear regression by allowing the linear model to be related to the response variable via a link function and by allowing the magnitude of the variance of each measurement to be a function of its predicted value. Generalized linear models were formulated by (Nelder & Wedderburn 1972) as a way of unifying various other statistical models, including linear regression, logistic regression and Poisson regression. They proposed an iteratively reweighted least squares method for maximum likelihood estimation of the model parameters. Maximum-likelihood estimation remains popular and is the default method on many statistical computing packages. Other approaches, including Bayesian approaches and least squares fits to variance stabilized responses, have been developed.

4.3.8. Evaluation of model performance

The performance of multi-classification models were measured by mean squared error (MSE) value. We conducted two experiments; (1) Using all datasets as training, (2) Using 10-fold cross-validation technique. In this technique, the compounds were randomly divided into ten parts, where nine parts were used for training and remaining part for testing. This process is carried out ten times in such a way that each part was used once for testing.

4.4. Results and discussion

4.4.1. Heatmap clustering and hypergeometric distribution

Initially, we determined pairwise chemical structural similarity between VOCs based on Tanimoto coefficient. 2-D 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 (Cao et al. 2014) 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 Fig. 4.7. White and red colors indicate the extreme distance values of 0 and 1 respectively and the intermediate distance values are indicated by the intensity of the red color. 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 4.3. 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 4.3. The major types of chemical compounds belonging to each cluster and their corresponding biological activities are mentioned in Table 4.4. The chemical structures of the VOCs belonging to all clusters (Cluster 1 to Cluster 11) are shown in Appendix B.

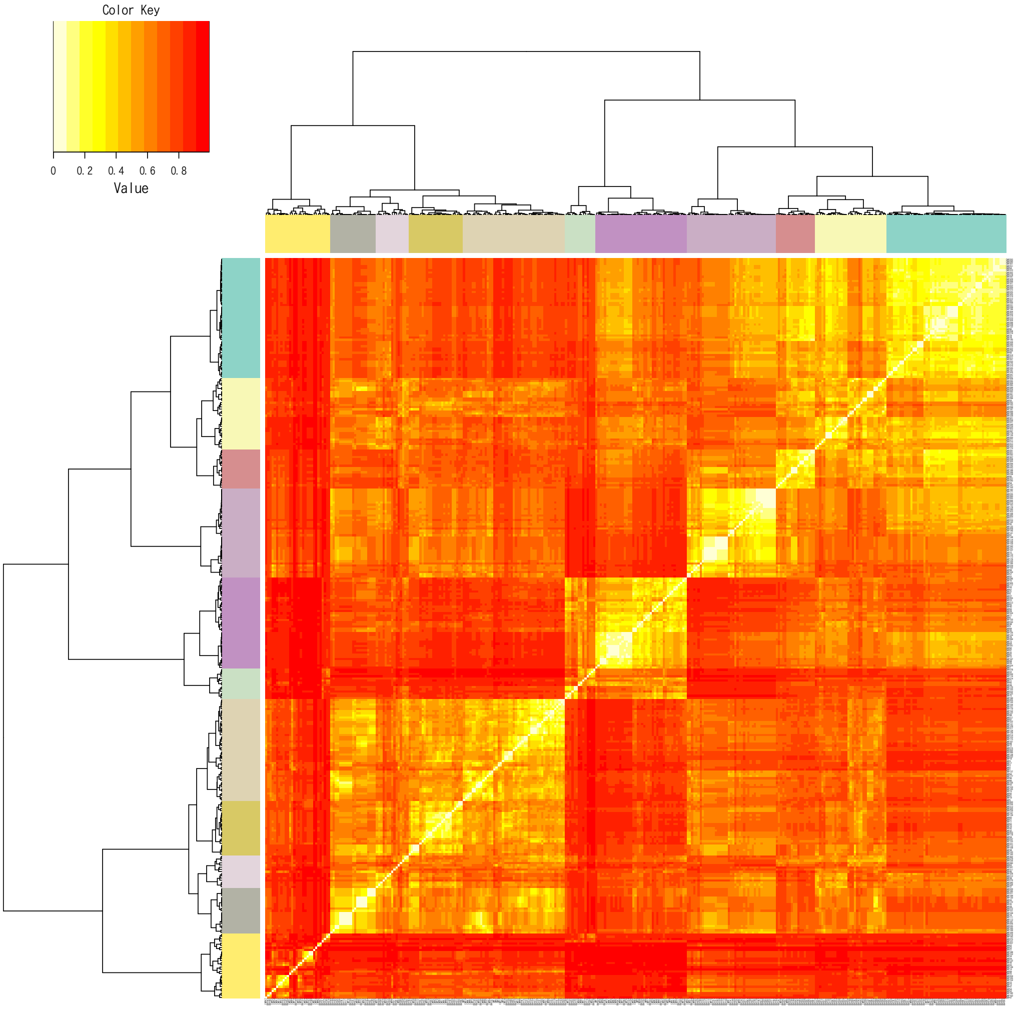


Figure 4.7. *Heatmap clustering of VOCs based on chemical structure similarity determined by Tanimoto coefficient.*

Table 4.3. *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)
Anti-cholinesterase	*p*-value	5.28x10^{-7}	4.849x10^{-4}	0.6181	0.9109	1.274x10^{-2}	0.9147	1	0.3596	0.9994	1	1
	(Count)	(26)	(15)	(8)	(2)	(9)	(3)	(0)	(4)	(2)	(0)	(0)
Antifungal	*p*-value	0.9128	0.9115	0.5399	2.561x10^{-2}	1	0.5176	0.6403	0.6571	0.3099	1	0.3291
	(Count)	(2)	(1)	(3)	(4)	(0)	(2)	(3)	(1)	(4)	(0)	(3)
Antimicrobial	*p*-value	2.10x10^{-6}	9.696x10^{-4}	0.6898	0.9281	1.871x10^{-2}	0.8246	0.9999	0.4049	0.9997	1	1
	(Count)	(26)	(15)	(8)	(2)	(9)	(4)	(1)	(4)	(2)	(0)	(0)
Antioxidant	*p*-value	5.28x10^{-7}	4.849x10^{-4}	0.6181	0.9109	1.274x10^{-2}	0.9147	1	0.3596	0.9994	1	1
	(Count)	(26)	(15)	(8)	(2)	(9)	(3)	(0)	(4)	(2)	(0)	(0)
Attractant	*p*-value	0.9708	0.8144	1	0.4831	1	0.1661	3.829x10^{-2}	0.1356	1.983x10^{-2}	1	1
	(Count)	(2)	(2)	(0)	(2)	(0)	(4)	(8)	(3)	(8)	(0)	(0)
Biomarker	*p*-value	1	0.9999	1.835x10^{-3}	0.7944	1.444x10^{-2}	0.9821	4.42x10^{-5}	0.9948	6.071x10^{-2}	1.036x10^{-2}	1.963x10^{-3}
	(Count)	(8)	(11)	(34)	(10)	(18)	(11)	(41)	(5)	(31)	(13)	(26)
Defense	*p*-value	3.35x10^{-9}	9.258x10^{-2}	0.9758	0.6764	0.7594	0.8418	0.9987	0.8668	0.9787	1	1
	(Count)	(22)	(7)	(2)	(2)	(2)	(2)	(1)	(1)	(2)	(0)	(0)
Enhance Plant Growth	*p*-value	6.01x10^{-2}	1	1	1	1	3.531x10^{-3}	0.7778	1	1	1	0.6069
	(Count)	(4)	(0)	(0)	(0)	(0)	(4)	(1)	(0)	(0)	(0)	(1)
Inhibit Root growth	*p*-value	0.1749	0.8632	0.9183	1	0.7111	4.111x10^{-2}	0.7672	0.5847	0.7062	0.5591	0.8347
	(Count)	(5)	(1)	(1)	(0)	(1)	(4)	(2)	(1)	(2)	(1)	(1)
Odor	*p*-value	1	1	1	1	1	2.29x10^{-5}	1	1	1	1	1
	(Count)	(0)	(0)	(0)	(0)	(0)	(4)	(0)	(0)	(0)	(0)	(0)
Repellent	*p*-value	1	7.551x10^{-2}	1	1	1.871x10^{-3}	1	1	1	1	1	1
	(Count)	(0)	(2)	(0)	(0)	(3)	(0)	(0)	(0)	(0)	(0)	(0)

Table 4.4. *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) and 40 VOCs are sesquiterpenoids (15 carbon).	Anti-cholinesterase, antimicrobial, antioxidant, 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.	Anti-cholinesterase, antimicrobial, 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.	Anti-cholinesterase, antimicrobial, antioxidant, biomarker, repellent.
Cluster 6 (25 VOCs)	21 VOCs are alcohol and ether of C3-C8 alkanes.	Enhance plant growth, inhibit root growth, odor.
Cluster 7 (47 VOCs)	45 VOCs are ester, carboxylic acid, ketone and aldehyde of non-cyclic C2-C9 alkanes.	Attractant, 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 hetero-aromatic 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.

From this result, we can see that there are 55 VOCs belong to Cluster 1 and mainly involved with anti-cholinesterase, 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 anti-cholinesterase, antimicrobial and antioxidant are $4.849 \text{x} 10^{-4}$, $9.696 \text{x} 10^{-4}$ and $4.849 \text{x} 10^{-4}$ respectively. Some of the VOCs that are classified into Cluster 2 are monoterpenoids and sesquiterpenoids such as beta-linalool, terpinen-4-ol, 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 activities involved is biomarker for various diseases such as colorectal cancer and asthma. We obtained small *p*-value ($1.835 \text{x} 10^{-3}$) for biomarker activity of Cluster 3. All compounds are alkanes, of which most of them are emitted in human breath such as octane, isobutane, 2-methylpentane, methylcyclohexane, hexane and cyclohexane.

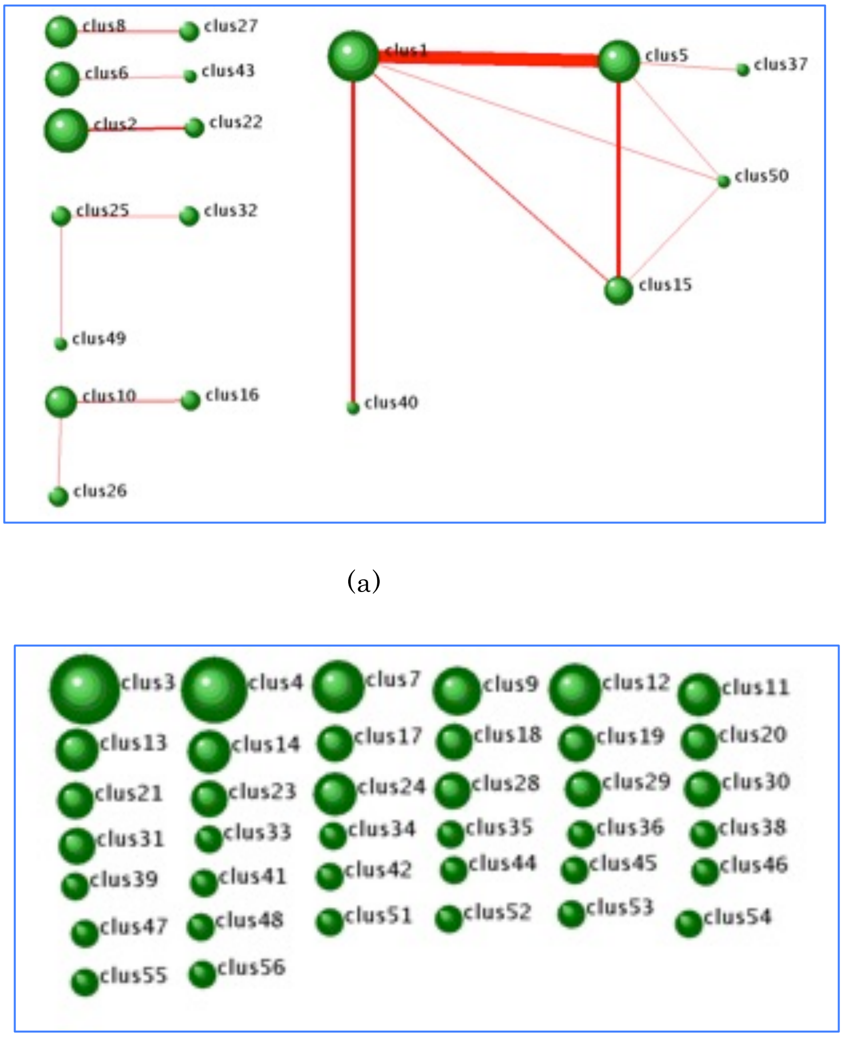
There are 18 compounds in Cluster 4 and all of them are alkenes such as beta-farnesene, alpha-caryophyllene, ocimene and beta-ocimene. These compounds are mainly associated with chemical ecology activity, which is antifungal and the *p*-value for this activity is $2.561 \text{x} 10^{-2}$. For Cluster 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, that are anti-cholinesterase, antimicrobial, antioxidant, biomarker and repellent activities. There are 25 VOCs in Cluster 6 and 21 of them are alcohol and ether of C3-C8 alkanes. We also obtained small *p*-value for enhance plant growth activity ($3.531 \text{x} 10^{-3}$), inhibit root growth ($4.111 \text{x} 10^{-2}$) and odor activity ($2.29 \text{x} 10^{-5}$) for cluster 6. An example of VOCs involved in enhance plant growth 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 (Ryu et al. 2003; D'Alessandro et al. 2014). 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 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 Cluster 7 and 45 of them are ester, carboxylic acid, ketone and aldehyde of non-cyclic C2-C9 alkanes. Cluster 7 is significantly related with multiple biological activities, which are attractant (*p*-value = 3.829x10^{-2}) and biomarker for various diseases (*p*-value = 4.42x10^{-5}). Aldehydes belong to Cluster 7 such as acetaldehyde, propanal, hexanal, 2-methyl-butanal, pentanal, heptanal and 3-methyl-butanal are mostly used as a biomarker for various diseases including cancer and irritable bowel syndrome. In Cluster 8, there are 15 VOCs belong 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 = 1.983x10^{-2}). All VOCs belong 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 hetero-aromatic compounds that consist of one or more sulfur, nitrogen or oxygen atoms. In Cluster 11, which consists of 30 VOCs, but their VOCs are quite diverse in chemical elements and have C0-C6 small molecules, ranging from hydrogen cyanide (27.02534 g/mol) to tetrachloroethyene (165.8334 g/mol). The main biological activity for Cluster 10 and Cluster 11 are biomarker for various diseases. The *p*-values for biomarker activity for Cluster 10 and Cluster 11 are 1.036x10^{-2} and 1.963x10^{-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 (Cozzolino et al. 2014; Dieme et al. 2015).

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.

4.4.2. Comparison of clustering methods

To extend our studies, we compared two different clustering methods (DPClus and hierarchical clustering) to cluster all 341 VOCs that we accumulated in our database. In case of DPClus algorithm, we used 0.6 as input density d_{in} and 0.5 as input cluster property cp_{in}. DPClus generated 56 clusters. Figure 4.8 (a) shows the interacted clusters while Fig. 4.8 (b) shows the independent clusters of DPClus.



(a)

(b)

Figure 4.8. *(a) Interacted clusters of DPClus. (b) Independent clusters of DPClus.*

To be consistent, we extracted 50 clusters based on hierarchical clustering. The size of the biggest cluster generated by DPClus is 18 while in case of hierarchical clustering it is 98 (centroid's method) as shown in Fig. 4.9. It is also observed that in hierarchical clustering, there is some imbalance in the size of generated clusters. On the other hand, the clusters generated by DPClus algorithm are in balanced size (Abdullah et al. 2016).

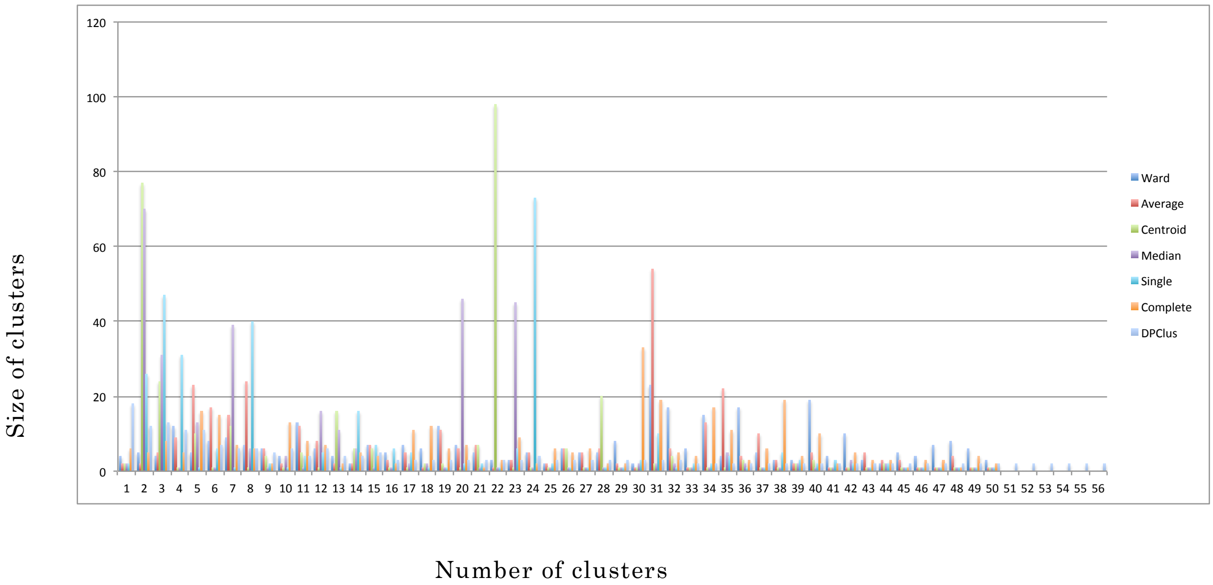


Figure 4.9. *Distribution of cluster with refer to size generated by DPClus and each linkage method of hierarchical clustering.*

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

$$m = \frac{i^2}{a \times b}$$

Here, *a* is the size of a cluster generated by DPClus, *b* is the size of a clusters generated by hierarchical method and *i* is the size of the intersection set of *a* and *b*. The calculated matching score is ranged between 0 and 1, where value of 1 shows the maximum overlapping score between two generated clusters. The distribution of matching score between DPClus and each method of hierarchical clustering is shown as Fig. 4.10.

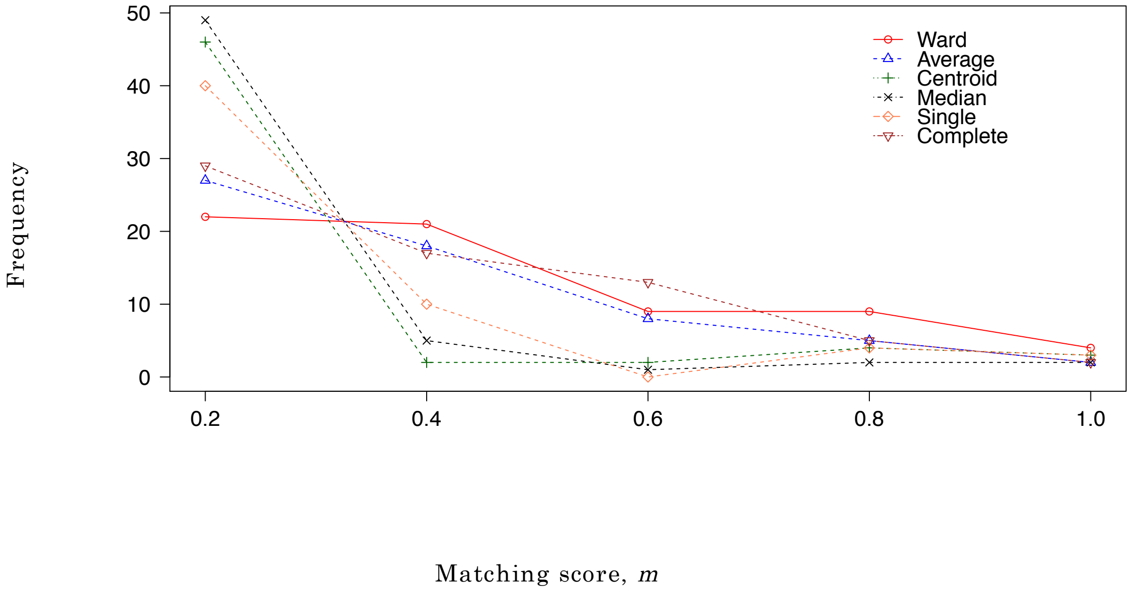


Figure 4.10. *Distribution of clusters with refer to matching score for DPClus and each linkage method of hierarchical clustering.*

From this figure, we can observe that the Ward's method of hierarchical clustering has the most matching clusters with DPClus algorithm while median has the least matching clusters with DPClus. Both DPClus and hierarchical methods generated clusters of VOCs with high structural similarity and similar biological activity. For example, cluster 1 generated by DPClus algorithm contains 18 VOCs, which are terpenoids and their biological activities are anti-cholinesterase, antimicrobial and defense activities. These results somehow aligned with our previous results, where we have shown that structurally similar group of VOCs generated by hierarchical clustering correspond to similar biological functions by conducting statistical analysis involving hypergeometric distribution based *p*-values (Abdullah et al. 2015).

Another comparison measure is Rand index. The Rand index or Rand measure, is a measure of the similarity between two data clustering methods (Rand 1971). It is calculated by using the following equation,

$$Rand = \frac{a+d}{a+b+c+d}$$

where a is number of point pairs in the same cluster A and B, b is number of point pairs in the same cluster A not B, c is number of point pairs in the same cluster B not A and d is number of point pairs in different cluster A and B. Table 4.5 shows the Rand index value for the comparison between DPClus and each hierarchical clustering methods. From this table, we can see that the Complete method has high similarity with Ward and Average method. DPClus has high similarity with Average and Ward's method, but has low similarity with Median method. This result somehow is aligned with the result displayed in Fig. 4.10, where we have shown that DPClus has lowest similarity with Median method. It seems that DPClus clustering is quite different from each of hierarchical clustering method. This is because the different nature of the clustering algorithm itself. Hierarchical clustering builds models based on distance connectivity while DPClus is a graph clustering method, based on a subset of nodes in a graph such that every two nodes in the subset are connected by an edge which can be considered as a prototypical form of clustering.

Table 4.5. *Rand index value for comparison between two clustering methods.*

	Ward	Average	Centroid	Median	Single	Complete	DPClus
Ward							
Average	0.875						
Centroid	0.717	0.729					
Median	0.714	0.707	0.710				
Single	0.788	0.793	0.666	0.704			
Complete	0.883	0.884	0.718	0.727	0.799		
DPClus	0.701	0.723	0.650	0.642	0.675	0.718	

In this section, we discussed two different clustering methods, namely DPClus graph clustering and hierarchical clustering to cluster the chemical structures of volatile organic compounds (VOCs) using Tanimoto coefficient as chemical similarity measure. Additionally, we compared the performances of DPClus algorithm with 6 different methods of hierarchical clustering, which are single, complete, average, centroid, median and Ward's method. Based on matching score, we found that Ward's method has the most matching clusters with DPClus while median has the least matching clusters. Using Rand index, we found that Complete method has similarity with Average and Ward's method. Compared to hierarchical clustering, DPClus can give a better visualization of how generated clusters are interacted with each other and we found that VOCs belonging to the interacted clusters have similar chemical structure, which indicates possibilities of exhibiting similar biological activities. In conclusion, chemical similarity measure can be used to predict biological activities of a compound and this can be applied in the medical and agrotechnology fields.

4.4.3. Supervised machine-learning methods

In previous section, we compared the performance of two clustering methods; DPClus graph clustering and hierarchical clustering to classify volatile organic compounds (VOCs) using fingerprint-based similarity measure between chemical structures. Using the same datasets as mentioned earlier, we extended the studies by implementing supervised machine learning methods to classify the VOCs based on chemical structures. The main difference between unsupervised (clustering) method and supervised machine learning is that, supervised machine learning methods need an output class variable. In supervised learning, each example is a pair consisting of an input object and a desired output.

In this study, we have developed 72 classification models to predict biological activities of VOCs by four types of supervised machine-learning methods, which are DNN, GBM, RF and GLM. Eight types of molecular fingerprints are used to represent the molecules, which are PubChem (PubChem, 881 bits), CDK (CDK, 1024 bits), Extended CDK (Extended, 1024bits), MACCS (MACCS, 166 bits), Klekota-Roth (KR, 4860 bits), Substructure (Sub, 307 bits), Estate (Estate, 79 bits) and atom pairs (AP, 780 bits). We also proposed a new type of fingerprint, called Combine (Combine, 9121 bits) by combining all features or substructures obtained by these eight fingerprints. After removing all "0" columns from the binary matrix, we input as classification models to the machine learning methods for prediction of biological activities.

It is difficult and time-consuming to find the best parameters for DNN due to the large number of adjustable parameters. Hence we took the approach by choosing the best parameter by using the multi-dimensional hyper-parameter optimization method. We selected the best parameter and then, compared with the default parameter. Table 4.6 shows the DNN parameter used in this study.

Table 4.6. *List of DNN parameters used in this study.*

Parameter list	DNN 1 (default)	DNN 2	DNN 3	DNN 4	DNN 5
Activation function	Rectifier	Tanh	Maxout	Rectifier with Dropout	Maxout
Input dropout ratio				20%	
Hidden dropout ratio				20%, 20%, 20%, 20%, 20%	
Hidden layer 1	200	200	200	200	200
Hidden layer 2	200	200	200	200	200
Hidden layer 3				200	200
Hidden layer 4				200	200
Hidden layer 5				200	200
Epoch	10	10	10	10000	10000

We used the default setting for DNN 1; Rectifier activation function, 200 neurons in both hidden layer 1 and hidden layer 2 and epochs was set to 10. We varied the parameter for DNN 2 and DNN 3 by using the Tanh and Maxout activation function. For DNN 4, we selected the best parameter based on multi-dimensional hyper-parameter optimization method; Rectifier activation function with dropout, 5 hidden layers, 200 neurons in every hidden layer, 20% dropout rate in input layer and each of hidden layer and the epoch was set to 10000. For DNN 5, we used the Maxout activation function, 5 hidden layers, 200 neurons in every hidden layer and the epoch was set to 10000. Other than DNN, we also compared the classification performance of GBM, RF and GLM methods. Table 4.7 shows the list of classification models using different fingerprints and machine learning methods.

Table 4.7. *List of 72 classification models using different fingerprints and machine learning methods.*

Model No	Fingerprint + Machine Learning Method
1	Combine + DNN1 (default)
2	Combine + DNN2
3	Combine + DNN3
4	Combine + DNN4
5	Combine + DNN5
6	Combine + RF
7	Combine + GBM
8	Combine + GLM
9	KR + DNN1 (default)
10	KR + DNN2
11	KR + DNN3
12	KR + DNN4
13	KR + DNN5
14	KR + RF
15	KR + GBM
16	KR + GLM
17	PubChem + DNN1 (default)
18	PubChem + DNN2
19	PubChem + DNN3
20	PubChem + DNN4
21	PubChem + DNN5
22	PubChem + RF
23	PubChem + GBM
24	PubChem + GLM
25	CDK + DNN1 (default)
26	CDK + DNN2
27	CDK + DNN3
28	CDK + DNN4
29	CDK + DNN5

30	CDK + RF
31	CDK + GBM
32	CDK + GLM
33	Extended + DNN1
34	Extended + DNN2
35	Extended + DNN3
36	Extended + DNN4
37	Extended + DNN5
38	Extended + RF
39	Extended + GBM
40	Extended + GLM
41	AP + DNN1 (default)
42	AP + DNN2
43	AP + DNN3
44	AP + DNN4
45	AP + DNN5
46	AP + RF
47	AP + GBM
48	AP + GLM
49	Sub + DNN1 (default)
50	Sub + DNN2
51	Sub + DNN3
52	Sub + DNN4
53	Sub + DNN5
54	Sub + RF
55	Sub + GBM
56	Sub + GLM
57	Estate + DNN1 (default)
58	Estate + DNN2
59	Estate + DNN3
60	Estate + DNN4
61	Estate + DNN5

62	Estate + RF
63	Estate + GBM
64	Estate + GLM
65	MACCS + DNN1 (default)
66	MACCS + DNN2
67	MACCS + DNN3
68	MACCS + DNN4
69	MACCS + DNN5
70	MACCS + RF
71	MACCS + GBM
72	MACCS + GLM

We conducted two types of experiments; (1) Using all datasets as training and (2) Using 10-fold cross validation technique. A full list of classification results for both experiments (in terms of MSE value and accuracy) is available in the Appendix C.

For the first experiment, by using all datasets as training, the best classification model was developed by Klekota-Roth fingerprint trained with Deep Neural Network 4 (DNN 4) method, with MSE value 0.05420784. Second best classification model was developed by PubChem fingerprint with MSE value 0.05871162, followed by MACCS fingerprint with MSE value 0.07807859. Both fingerprints were also trained with Deep Neural Network 4 (DNN 4). The best parameter for deep learning was obtained by using rectifier activation function with dropout rate at 20%. Number of hidden layer was set to 5 and 200 neurons for each of hidden layer. Estate and atom pair fingerprint did not perform well in the classification model. This is because the length of the Estate fingerprint is only 79 bits, which is too short to characterize molecules. Too much information loss led to the bad prediction.

For the second experiment, we adopted the 10-fold cross-validation technique to evaluate the performance of our models. The lowest MSE error was obtained by using PubChem fingerprint trained by GBM method at 0.39318013, followed by Combine fingerprint also trained by GBM method. The obtained MSE error was 0.39837325. MACCS fingerprint trained by GBM method also gave good MSE value at 0.39979038 compared to other models. The worst performance was obtained using Extended fingerprints trained with Deep Neural Network 4 (DNN 4) and Estate fingerprint trained with Deep Neural Network 3 (DNN 3).

Fig. 4.11 shows the performance of 72 classification models (MSE value) by using all datasets as training and 10-fold cross validation technique. It seems that all data are distributed randomly and there is no correlation between the performance obtained by using all datasets as training and 10-fold cross validation technique.

1

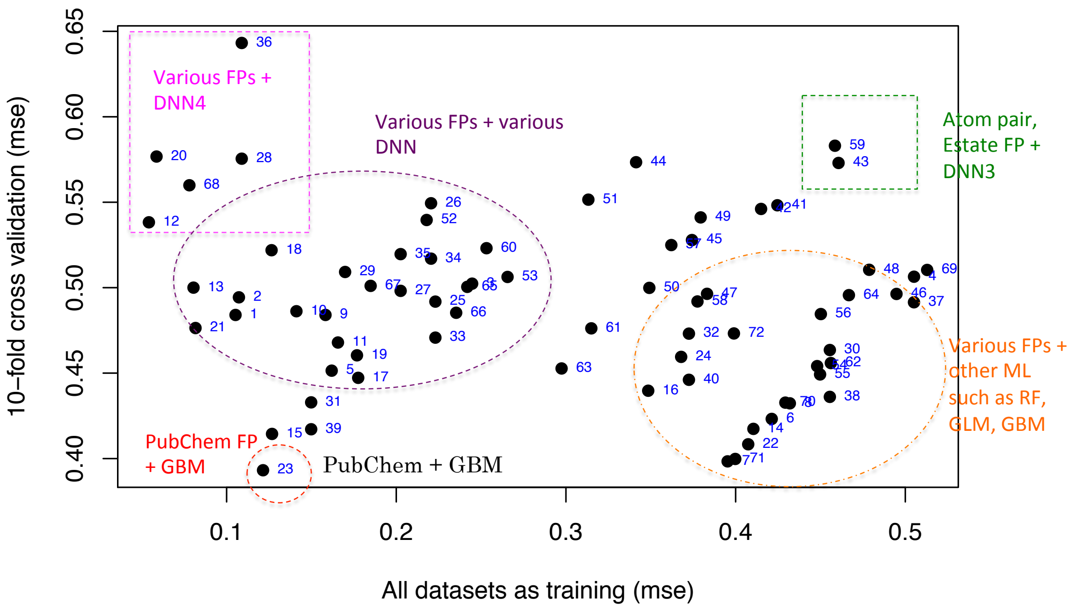


Figure 4.11. *Performance of 72 classification models by using all datasets as training and 10-fold cross validation technique (MSE value).*

We observed that there are two types of models: 1) the left side is affected by over-fitting problem, and 2) the right side is not changed for both experiments. The left side points, which most of the combination of fingerprint types and DNN methods suffered from over-fitting problems due to the many parameters of DNN. The performance of DNN is good when using all datasets as training, however it becomes worst when we used 10-fold cross validation technique, such as model No 12 (Klekota-Roth fingerprint trained with DNN4 method) and model No 36 (Extended fingerprint trained with DNN4 method). This might be caused by the small number of our sample data and many parameters of DNN. DNN always requires a large amount of data to be trained, usually more than 50,000 samples. In our study, we only have 341 VOC data for the classification task. In theory, over-fitting is a major problem for DNN and we have proved this experimentally. Moreover, the Klekota-Roth and Extended fingerprints have many substructures or features (more than 1000), which need to be trained and as a result, they are suffering from over-fitting problems. The right side points did not change much for both experiments. For example, the classification model No 43 (atom pair fingerprint trained with DNN3 method) and model No 59 (Estate fingerprint trained with DNN3 method) performed poorly in both experiments. From this result, we can understand two things; 1) Atom pair and Estate fingerprint did not perform well in model building, 2) DNN3 is the worst, compared to other DNN models. Atom pair fingerprint are a structural descriptor type that is defined by the shortest paths among the non-hydrogen atoms in a molecule. Each path is described by the types of atoms in a pair, the length of their shortest bond path, the number of their pi electrons and the non-hydrogen atoms bonded to them. The number of atom pairs describing a molecule grows with its number of atoms. The fingerprints provided by PubChem are a binary representation of the presence and absence of a library of 881 substructure features. Compared to atom pairs, the PubChem fingerprints are a knowledge-based system that stores less information than the much more complex and unbiased atom pair concept. PubChem fingerprints are also less sensitive than atom pair descriptors. The length of the Estate fingerprint is only 79 bits, which is too short to characterize molecules and some of the information might be loss, which cause the bad prediction. It is also observed that hyper parameters of DNN can affect the overall performance. The reason why DNN3 performed poorly for both experiments, is because the Maxout activation function and a small number of epochs. Rectifier activation function is a better choice for this classification task.

Based on Fig. 4.10, we can observe that the classification model No 23 (PubChem fingerprint trained with GBM method) gives good results in both experiments. This model obtained MSE value = 0.1214795 when using all datasets as training and MSE value = 0.39318013 in case of 10-fold cross validation technique. The results show that GBM method is good at predicting biological activities of VOCs. GBM appears to be a very effective and efficient machine-learning method. It is efficient because it achieves these results with much less computational effort than either of those methods and produces much smaller models. This is also supported by (Sheridan et al. 2016), where they compared eXtreme Gradient Boosting (XGBoost) to random forest and single-task deep neural nets on 30 in-house data sets and found that XGBoost can make prediction better than those of random forest and almost as good as those of deep neural nets. Overall, GBM results somehow are contrary with DNN results.

We also evaluated the performance of all 72 models in term of classification accuracy. Classification accuracy is the ratio of correct predictions to total predictions made and often presented as a percentage by multiplying the result by 100. Fig. 4.12 shows the performance of 72 classification models in term of accuracy value (%) by using all datasets as training and 10-fold cross validation technique. Also, it can be seen that all data are distributed randomly and there is no correlation between the performance obtained by using all datasets as training and 10-fold cross validation technique. Similarly to MSE result, we observed that there are two types of models: 1) the right side is affected by over-fitting problem, and 2) the left side is not changed for both experiments. The right side models, such as model No 12 (Klekota-Roth fingerprint trained with DNN4 method), model No 20 (PubChem fingerprint trained with DNN4 method) and model No 36 (Extended fingerprint trained with DNN4 method) give good classification result when using all datasets as training, however it becomes worst when we used 10-fold cross validation technique. This might be caused by the small number of our sample data, many parameters of DNN and large number of features need to be trained, which we have explained previously.

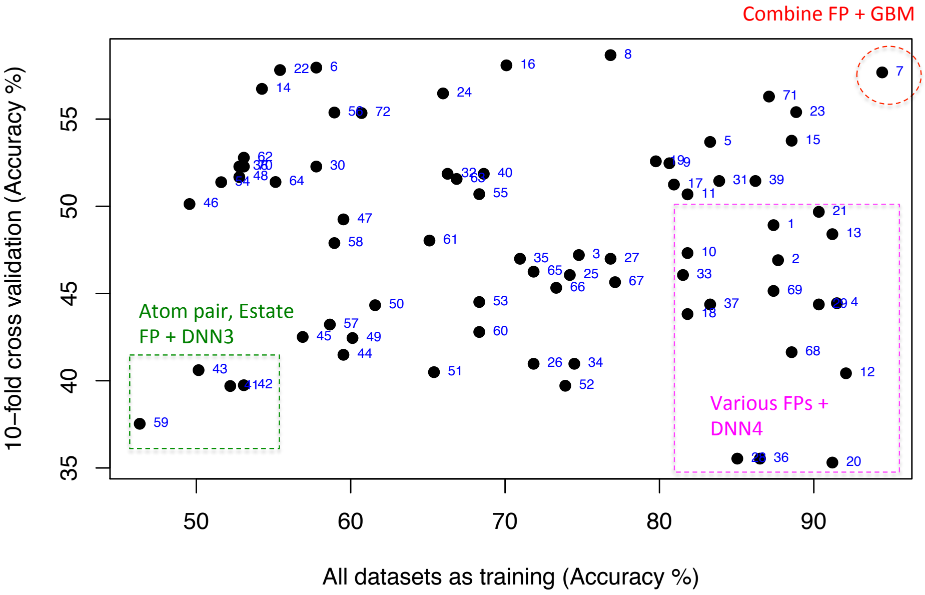


Figure 4.12. *Performance of 72 classification models by using all datasets as training and 10-fold cross validation technique (accuracy).*

Contrarily, there are few models which performed poorly in both experiments. The classification model No 43 (atom pair fingerprint trained with DNN3 method) and model No 59 (Estate fingerprint trained with DNN3 method) performed poorly in case of using all datasets as training and 10-fold cross validation technique. This is due to the small number of substructures for Estate fingerprint, which is too short to characterize molecules. The atom pair fingerprint is also known as a very sensitive fingerprint and this is the reason why it performed poorly in both experiments. Based on Fig. 4.11, we observed that the classification model No 7 (Combine fingerprint trained with GBM method) gives good results in both experiments. This model obtained accuracy value of 94.4% when using all datasets as training and 57.7% in case of 10-fold cross validation technique. The results show that GBM method is good at predicting biological activities of VOCs. This result somehow is aligned with our previous result shown in Fig. 4.10, where we proved that GBM appears to be a very effective and efficient algorithm, compared to other machine learning methods.

4.4.4. Identification of important substructures

In the Section 4.4.3, we explained the performance of DNN, GBM, RF and GLM machine learning methods to predict the biological activities of VOCs based on chemical structures. The best classification model was built by using GBM algorithm along with PubChem fingerprint. Hence, we identified the important substructures for PubChem fingerprint using GBM algorithm for the purpose of predicting biological activities. The H2O R package has implemented the method of Gedeon (Gedeon 1997) in order to find the variable importance in descending order of importance. GBM algorithm can automatically calculate variable importance, which include the absolute and relative predictive strength of each feature in the prediction task. The most important substructures (top 5) during classification were PubChemFP430, PubChemFP2, PubChemFP334, PubChemFP14 and PubChemFP839. The detail description for each of the substructures are given below (refer to Appendix A for detail).

1) PubchemFP430: C(-C)(-C)(=C)

2) PubchemFP2: >= 16 H

3) PubchemFP334: C(~C)(~C)(~C)(~C)

4) PubchemFP14: >= 1 N

5) PubchemFP839: CC1CC(C)CC1

It was observed that PubChem fingerprint number 430 ranked the highest in the context of classifying biological activities of VOCs. This fingerprint represents the detailed atom neighborhood. The second highest rank was PubChem fingerprint number 2, which represents the presence of 16 hydrogen atoms in a compound. Based on the frequency of these substructures presence in each of VOC, we identified the most relevant biological activity for each compound. For example, PubChem fingerprint number 430 has contributed most to the attractant and biomarker activities. Most of VOCs, which have attractant and biomarker activities consist only this particular fingerprint. PubChem fingerprint number 2, which represents the presence of 16 hydrogen atoms contribute to chemical ecology activities such as repellent, antimicrobial, antifungal and defense. PubChem fingerprint number 334 has occurred in most of VOCs, which have human healthcare activities such as anti-cholinesterase and antioxidant activities. PubChem fingerprint number 14, which represents the presence of one nitrogen atom in a compound contribute to biomarker and antifungal activities. This makes sense because the VOCs released in human breath are nitrogen containing such as dimethylamine and ammonia. Studies have shown elevated levels of inflammatory and oxidative stress biomarkers such as nitrogen oxides in patients with asthma, COPD, bronchiectasis and cystic fibrosis (Montuschi 2007; N.M. & M. 2008). It is also known that organonitrogen compounds containing an aliphatic nitrogen have significant antifungal properties (Mullen et al. 1989) and azole (a class of five-membered heterocyclic compounds containing a nitrogen atom and at least one other non-carbon atom as part of the ring) antifungal drugs are the most widely employed antifungal agents in clinical practice (Dodds-Ashley 2010). PubChem fingerprint number 839, which indicates the presence of complex SMARTS patterns, contribute to enhance plant growth and inhibit root growth.

These substructures or features have significant relationships with biological activities and are considered important for prediction of biological activities of VOCs.

4.4.5. Testing the recommended model on new datasets

It is important to show that the recommended fingerprint and machine learning method will also apply to other new datasets that have been not been part of the model building. Thus, 120 additional new VOC datasets were selected from other sources, which some of them were obtained from KNApSAcK Metabolite Activity DB and other VOC data were accumulated by literature search. Table 4.8 shows the new VOC data and predicted result by using our recommended model. We found that 73 VOC were predicted correctly, which give about 60.8% accuracy. The confusion matrix between actual and predicted biological activities is shown in Table 4.9. The low accuracy of prediction results is because of small sample of our training datasets, which is not sufficient enough for the model to learn and predict new data. Most of VOCs are predicted as biomarker, maybe because the training data was dominated by biomarkers. Also, 12 of the 14 odor VOCs were predicted as biomarkers. It is maybe because odor and biomarker VOCs are chemical structurally similar. It is recommended to increase the quantity of sample datasets so that the data for each of activity are well-balanced and this can increase the prediction accuracy.

Table 4.8. *List of new VOC datasets and predicted results.*

New VOC data	Actual Biological Activity	Predicted Activity
1,6-dioxacyclododecane-7,12-dione	Biomarker	Biomarker
Potassium lespedezate	Enhance plant growth	Enhance plant growth
Impericine	Anticholinesterase	Anticholinesterase
Aloe emodin;Rhabarberone	Repellent	Biomarker
Chrysophanol	Antioxidant	Biomarker
Annonalide	Inhibit root growth	Inhibit root growth
Limbatolide A	Anticholinesterase	Anticholinesterase
Limbatolide B	Anticholinesterase	Anticholinesterase

Potassium chelidonate	Enhance plant growth	Enhance plant growth
Kompasinol A;Maackoline	Antioxidant	Biomarker
Capillin	Antifungal	Biomarker
Carvacrol	Antifungal	Biomarker
L-Lactic acid;(S)-(+)-Lactic acid	Antioxidant	Biomarker
Rishitin	Defense	Defense
N-Isobutyroylbuxahyrcanine	Anticholinesterase	Antioxidant
Grandinol	Inhibit root growth	Inhibit root growth
1, 10-(1-butenylidene) bis benzene	Biomarker	Biomarker
Neoeriocitrin	Antimicrobial	Antimicrobial
Lappaconitine	Antioxidant	Antioxidant
Brassinolide	Enhance plant growth	Enhance plant growth
Diallyl sulfide	Odor	Biomarker
Cyclotetrasiloxane	Biomarker	Biomarker
Gentianose	Enhance plant growth	Biomarker
Tridecane	Biomarker	Biomarker
Tetradecane	Biomarker	Biomarker
Glabranin	Antimicrobial	Anti-cholinesterase
5-Hydroxy-1,4-naphthoquinone	Antifungal	Biomarker
Menthol	Defense	Biomarker
Castasterone	Enhance plant growth	Anti-cholinesterase
Dimethylsilanediol	Biomarker	Biomarker

Isouvaretin	Antimicrobial	Antimicrobial
Lunularic acid	Enhance plant growth	Biomarker
Momilactone A	Inhibit root growth	Inhibit root growth
Indole-3-butyric acid	Inhibit root growth	Inhibit root growth
Volicitin	Defense	Biomarker
Glucolimnanthin	Odor	Biomarker
(-)-N-Methylcytisine	Repellent	Biomarker
Potassium isolespedezate	Enhance plant growth	Enhance plant growth
Sinigrin	Attractant	Biomarker
Lepidimoide	Enhance plant growth	Enhance plant growth
Caffeine	Defense	Biomarker
(+)-Camphor;Camphor	Repellent	Anti-cholinesterase
(-)-Menthone	Odor	Biomarker
1,8-Cineole;Eucalyptol	Inhibit root growth	Inhibit root growth
Menthyl acetate	Odor	Biomarker
Tomatine	Repellent	Antioxidant
Momilactone B	Inhibit root growth	Inhibit root growth
4-heptanone	Biomarker	Biomarker
Myrcene	Attractant	Attractant
Emodin	Antifungal	Biomarker
Eugenol	Antioxidant	Biomarker
(+)-Marmesin;Marmesin	Anticholinesterase	Anticholinesterase

Medicarpin	Defense	Biomarker
Salicylic acid	Defense	Attractant
Gossypol	Defense	Biomarker
cis-trans-Nepetalactone	Repellent	Biomarker
Kurramine-2'-beta-N-oxide	Anticholinesterase	Anticholinesterase
(R)-(-)-Carvone	Odor	Biomarker
Dihydrozeatin	Enhance plant growth	Enhance plant growth
(+)-Coniine	Odor	Biomarker
Chalcogran	Attractant	Attractant
Flindersiachromone	Odor	Biomarker
Agnuside;Buddlejoside A	Attractant	Attractant
Dolichodial	Repellent	Anti-cholinesterase
(+)-Iridodial	Repellent	Biomarker
Naringin	Antimicrobial	Antimicrobial
Narirutin	Antimicrobial	Antimicrobial
Carvone oxide	Attractant	Attractant
(+)-Pulegone	Odor	Odor
Demissine	Repellent	Biomarker
Styraxin	Odor	Odor
24-Epibrassinolide	Inhibit root growth	Inhibit root growth
trans-Cinnamic acid	Enhance plant growth	Enhance plant growth
trans-Zeatin	Enhance plant growth	Enhance plant growth

Nordihydroguaiaretic acid	Antimicrobial	Biomarker
Luteolin	Antioxidant	Antioxidant
8R-Hydroxylinoleic acid	Antifungal	Antifungal
(-)-Jasmonic acid	Enhance plant growth	Enhance plant growth
trans-2-Hexenal	Antifungal	Antifungal
Myricetin	Antioxidant	Antioxidant
Integerrimine	Attractant	Attractant
Methyl jasmonate	Enhance plant growth	Enhance plant growth
2-hexyl-1-octanol	Biomarker	Biomarker
(+)-Ascorbic acid	Antioxidant	Antioxidant
28-Homocastasterone	Enhance plant growth	Enhance plant growth
4-pentadiene	Biomarker	Biomarker
Sucrose;(+)-Sucrose	Antioxidant	Antioxidant
Pyrethrins	Defense	Defense
Aniline	Biomarker	Biomarker
N6-Benzyladenine	Enhance plant growth	Enhance plant growth
Ethylene	Defense	Defense
Camalexin	Defense	Defense
p-Coumaric acid	Antifungal	Antifungal
Geraniol	Attractant	Attractant
Chlorophyll a	Odor	Biomarker
Gibberellin A1;GA1	Enhance plant growth	Enhance plant growth

Allicin	Odor	Biomarker
Tomatidine	Repellent	Antioxidant
Protoanemonin;Protoanemonene	Enhance plant growth	Enhance plant growth
Skatole	Odor	Biomarker
Pinocembrin	Antimicrobial	Antimicrobial
Isopimpinellin	Antifungal	Biomarker
Seselin;Amyrolin;Seseline	Antifungal	Biomarker
Actinidine	Attractant	Attractant
Caffeic acid	Antifungal	Biomarker
Thymol	Antifungal	Biomarker
(-)-Epicatechin	Anticholinesterase	Anticholinesterase
Hesperetin	Antimicrobial	Antimicrobial
Lycorine;(-)-Lycorine	Anticholinesterase	Anticholinesterase
Uvaretin	Antimicrobial	Antimicrobial
Crinamine	Anticholinesterase	Anticholinesterase
Citronellal	Repellent	Anticholinesterase
N-Methylfuntumine	Anticholinesterase	Anticholinesterase
Piperonal	Odor	Biomarker
Isoeugenol	Odor	Biomarker
Shikimic acid 3-phosphate;S3P	Enhance plant growth	Enhance plant growth
(+)-Catechin	Antioxidant	Antioxidant
N6-(delta2-Isopentenyl)adenine;2iP	Enhance plant growth	Enhance plant growth

Naringenin;(-)-Naringenin	Antimicrobial	Antimicrobial
Galanthamine	Anticholinesterase	Anticholinesterase

Table 4.9. *Confusion matrix between actual and predicted biological activities.*

		Predicted												
Actual		Biomarker	Enhance plant growth	Anticholinesterase	Repellent	Antioxidant	Inhibit root growth	Antifungal	Defense	Antimicrobial	Odor	Attractant	Error	Error (%)
Biomarker	10	10	0	0	0	0	0	0	0	0	0	0	0	0.00
Enhance plant growth	19	2	16	1	0	0	0	0	0	0	0	0	3	2.50
Anticholinesterase	11	0	0	10	0	1	0	0	0	0	0	0	1	0.83
Repellent	10	5	0	3	0	2	0	0	0	0	0	0	10	8.33
Antioxidant	10	4	0	0	0	6	0	0	0	0	0	0	4	3.33
Inhibit root growth	7	0	0	0	0	0	7	0	0	0	0	0	0	0.00
Antifungal	11	8	0	0	0	0	0	3	0	0	0	0	8	6.67
Defense	10	5	0	0	0	0	0	0	4	0	0	1	6	5.00
Antimicrobial	10	1	0	1	0	0	0	0	0	8	0	0	2	1.67
Odor	14	12	0	0	0	0	0	0	0	0	2	0	12	10.00
Attractant	8	1	0	0	0	0	0	0	0	0	0	7	1	0.83
TOTAL	120	48	16	15	0	9	7	3	4	8	2	8	47	39.17

4.5. Summary

In this chapter, we examined the classification of VOCs based on chemical structural similarity. The main reason is to find the relationships between chemical structures of VOCs and biological activities based on the *Similar Property Principle*, which states that molecules that are structurally similar are likely to have similar properties (Maggiora & Shanmugasundaram 2004). Based on heatmap clustering and hypergeometric distribution result, we found that there are strong links between chemical structure of VOCs and their biological activities. Additionally, we also compared several types of hierarchical clustering methods with DPClus clustering to classify VOCs using fingerprint-based similarity measure between chemical structures. We extended our findings by building models using supervised machine learning methods to predict biological activities of VOCs based on chemical structures. We found that molecular fingerprints can be used for predicting biological activities of volatile metabolites. It is recommended to use PubChem and Combine fingerprint trained with Gradient Boosting Machine (GBM) method in the context of classifying VOCs. GBM method has advantage in term of computational speed and require less parameters for optimization, compared to other machine learning methods.

Chapter 5

Conclusions

In order to obtain a better understanding of the relationships among species, VOC and biological activities, we utilized data-intensive science for discovering and identifying natural diversity of VOCs. This study has been started by accumulating VOC data from literature and also scientific reports. We found that many VOCs are produced by plants, microorganisms, insects and also human. Each of VOCs produced by different organisms are very unique and have its specific function or biological activities. This study is conducted in order to further investigate the relationships among organisms, emitting VOCs and their corresponding biological activities.

In this dissertation, we have discussed a database of VOCs emitted by various living organisms including microorganisms, fungi, plants, animals and humans, which can be accessed at KNApSAcK Metabolite Ecology Database. Apart from VOC biological activities related to human healthcare, 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 clusters of microorganism species based on VOC similarity. We found consistency between VOC and pathogenicity based classification of microorganisms. Additionally, we also compared several types of hierarchical clustering methods with DPClus clustering to classify VOCs using fingerprint-based similarity measure between chemical structures. Our research indicates that similar chemical structures of VOCs indicate possibilities of exhibiting similar biological activities. We extended our findings by using supervised machine learning methods to predict biological activities of VOCs based on chemical structures. We have developed 72 classification model for the prediction of biological activities of VOCs by 9 types of fingerprints and trained by Deep Neural Network, Gradient Boosting Machine, Random Forest and Generalized Linear Model. Based on the computational results, PubChem fingerprints was suggested to be used as the input for the prediction, compared to other fingerprints. Gradient boosting machine (GBM) method can outperform Deep Neural Network (DNN) in term of classifying VOCs. GBM method has advantage in term of computational speed and require less parameters for optimization. Hence, we highly recommend to use Gradient Boosting Machine for the prediction of biological activities of VOCs based on chemical structures.

In future, more VOCs can be accumulated, and comprehensive analysis can be performed in the context of human healthcare and chemical ecology. The KNApSAcK Metabolite Ecology Database may be useful for the discovery of novel agricultural tools and also for the non-invasive identification of biomarkers in the medical diagnostic field as well as a systematic research in various omics fields, especially metabolomics integrated with ecosystems. It is hoped that the KNApSAcK Metabolite Ecology Database can be as a reference tool for the users to find information on volatile metabolites with related biological activities for the application in agriculture, ecosystems and healthcare industry.

Bibliography

Abdullah, A.A. et al., 2016. Comparison of Clustering Methods in the Context of Chemical Structure Similarity Based Classification of VOCs. *Procedia Chemistry*, 20, pp.40–44.

Abdullah, A.A. et al., 2015. Development and Mining of a Volatile Organic Compound Database. *BioMed Research International*, 2015.

Afendi, F.M. et al., 2013. Data Mining Methods for Omics and Knowledge of Crude Medicinal Plants toward Big Data Biology Abstract : Molecular biological data has rapidly increased with the recent progress of the Omics fields , e . g ., genomics , transcriptomics , proteomics and me. *Computational and structural biotechnology journal*, 4(5).

Afendi, F.M. et al., 2012. KNApSAcK family databases: integrated metabolite-plant species databases for multifaceted plant research. *Plant & cell physiology*, 53(2), p.e1.

Agostini-costa, T.S. et al., 2012. Secondary Metabolites. *Chromatography and Its Applications*, 1, pp.131–164.

Alberts, B., Johnson, a & Lewis, J., 2002. Introduction to Pathogens. In *Molecular Biology of the Cell*. pp. 473–475.

Altaf-Ul-Amin, M. et al., 2006. Development and implementation of an algorithm for detection of protein complexes in large interaction networks. *BMC Bioinformatics*, 7(1), pp.1–13.

Altaf-Ul-Amin, M. et al., 2006. DPClus : A density-periphery based graph clustering software mainly focused on detection of protein complexes in interaction networks. *Journal of Computer Aided Chemistry*, 7, pp.150–156.

Altomare, D.F. et al., 2013. Exhaled volatile organic compounds identify patients with colorectal cancer. *British Journal of Surgery*, 100(1), pp.144–150.

Amal, H. et al., 2016. Breath testing as potential colorectal cancer screening tool. *International Journal of Cancer*, 138(1), pp.229–236.

Anqi, A., Aiello, S. & Rao, A., 2015. Package "h2o."

Arasaradnam, R.P. et al., 2014. Review article: next generation diagnostic modalities in gastroenterology--gas phase volatile compound biomarker detection. *Alimentary pharmacology & therapeutics*, 39(8), pp.780–9.

Arn, H. & Acree, T.E., 1998. Flavornet: A database of aroma compounds based on odor potency in natural products. *Developments in Food Science*, 40(C), p.27.

Ayseli, M.T. & Ipek, Y., 2015. Trends in Food Science & Technology Flavors of the future : Health benefits of flavor precursors and volatile compounds in plant foods. , 48, pp.2015–2017.

Bajusz, D., Rácz, A. & Héberger, K., 2015. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics*, 7(1), pp.1–13.

Baldi, P. & Sadowski, P., 2014. The dropout learning algorithm. *Artificial Intelligence*, 210(1), pp.78–122.

Beck, J.J. & Vannette, R.L., 2016. Harnessing Insect-Microbe Chemical Communications to Control Insect Pests of Agricultural Systems. *Journal of Agricultural and Food Chemistry*, p.acs.jafc.6b04298.

Blombach, B. & Eikmanns, B.J., 2011. Current knowledge on isobutanol production with Escherichia coli, Bacillus subtilis and Corynebacterium glutamicum. *Bioengineered bugs*, 2(6), pp.346–350.

Bos, L.D.J., Sterk, P.J. & Schultz, M.J., 2013. Volatile metabolites of pathogens: a systematic review. *PLoS pathogens*, 9(5), p.e1003311.

Breiman, L., 2001. Random forests. *Machine learning*, 45(1), pp.5–32.

de Bruyne, M. & Baker, T.C., 2008. Odor detection in insects: Volatile codes. *Journal of Chemical Ecology*, 34(7), pp.882–897.

Buljubasic, F. & Buchbauer, G., 2015. The scent of human diseases: A review on specific volatile organic compounds as diagnostic biomarkers. *Flavour and Fragrance Journal*, 30(1), pp.5–25.

Butina, D., 1999. Unsupervised data base clustering based on daylight's fingerprint and Tanimoto similarity: A fast and automated way to cluster small and large data sets. *Journal of Chemical Information and Computer Sciences*, 39(4), pp.747–750.

Cane, D.E. et al., 2006. Geosmin biosynthesis in Streptomyces avermitilis. Molecular cloning, expression, and mechanistic study of the germacradienol/geosmin synthase. *The Journal of antibiotics*, 59(8), pp.471–479.

Cao, Y. et al., 2014. ChemmineR : Cheminformatics Toolkit for R. , (2008), pp.1–46.

Cao, Y. et al., 2008. ChemmineR: A compound mining framework for R. *Bioinformatics*, 24(15), pp.1733–1734.

Capuano, R. et al., 2015. The lung cancer breath signature: a comparative analysis of exhaled breath and air sampled from inside the lungs. *Scientific reports*, 5(April), p.16491.

Carhart, R.E., Smith, D.H. & Venkataraghavan, R., 1985. Atom pairs as molecular features in structure-activity studies: definition and applications. *Journal of Chemical Information and Computer Science*, 25(2), pp.64–73.

Carroll, W. et al., 2005. Detection of volatile compounds emitted by Pseudomonas aeruginosa using selected ion flow tube mass spectrometry. *Pediatric pulmonology*, 39(5), pp.452–6.

Cereto-Massagué, A. et al., 2015. Molecular fingerprint similarity search in virtual screening. *Methods*, 71(C), pp.58–63.

Cha, S., Choi, S. & Tappert, C., 2009. Anomaly between Jaccard and Tanimoto coefficients. In *Proceedings of Student-Faculty Research Day, CSIS, Pace University*. pp. 1–8.

Chan, D.K., Leggett, C.L. & Wang, K.K., 2016. Diagnosing gastrointestinal illnesses using fecal headspace volatile organic compounds. *World J Gastroenterol*, 22(224), pp.1639–1649.

Chandra, B. & Sharma, R.K., 2016. Fast learning in Deep Neural Networks. *Neurocomputing*, 171, pp.1205–1215.

Cheepsattayakorn, A. & Cheepsattayakorn, R., 2013. Review Article Breath Tests in Respiratory and Critical Care Medicine : From Research to Practice in Current Perspectives. , 2013.

Chen, T. & Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System. In *KDD*. pp. 1–10.

Cheung, W.H.K. et al., 2015. Volatile organic compound (VOC) profiling of citrus tristeza virus infection in sweet orange citrus varietals using thermal desorption gas chromatography time of flight mass spectrometry (TD-GC/TOF-MS). *Metabolomics*, 11(6), pp.1514–1525.

Christapher, P.V. et al., 2014. Review on Polygonum minus. Huds, a commonly used food additive in Southeast Asia. *Pharmacognosy research*, 7(1), pp.1–6.

Clifford, H. et al., 2011. Comparison of clustering methods for investigation of genome-wide methylation array data. *Frontiers in Genetics*, 2(DEC), pp.1–11.

Consonni, V. & Todeschini, R., 2012. New similarity coefficients for binary data. *Match-Communications in Mathematical and Computer Chemistry*, 68, pp.581–592.

Cook, R.J., 1998. Generalized linear model. *Encyclopedia of Biostatistics*, 6(2), p.e16104.

Cozzolino, R. et al., 2014. Use of solid-phase microextraction coupled to gas chromatography-mass spectrometry for determination of urinary volatile organic compounds in autistic children compared with healthy controls. *Analytical and Bioanalytical Chemistry*, 406(19), pp.4649–4662.

D'Alessandro, M. et al., 2014. Volatiles produced by soil-borne endophytic bacteria increase plant pathogen resistance and affect tritrophic interactions. *Plant, Cell and Environment*, 37(4), pp.813–826.

Dandekar, A.M. et al., 2010. Analysis of early host responses for asymptomatic disease detection and management of specialty crops. *Critical reviews in immunology*, 30(3), pp.277–89.

Delory, B.M. et al., 2016. Root-emitted volatile organic compounds: can they mediate belowground plant-plant interactions? *Plant and Soil*, pp.1–26.

Diao, B. et al., 2011. A graph-clustering approach to search important molecular markers and pathways of Parkinson's disease. *African Journal of Biotechnology*, 10(69), pp.15656–15661.

Didier Musso, D.J.G., 2016. Zika Virus. *Nature*, 11(1), pp.10–20.

Dieme, B. et al., 2015. Metabolomics Study of Urine in Autism Spectrum Disorders Using a Multiplatform Analytical Methodology. *Journal of Proteome Research*, 14(12), pp.5273–5282.

Dimitrov, I. et al., 2014. AllergenFP: Allergenicity prediction by descriptor fingerprints. *Bioinformatics*, 30(6), pp.846–851.

Dodds-Ashley, E., 2010. Management of drug and food interactions with azole antifungal agents in transplant recipients. *Pharmacotherapy*, 30(8), pp.842–54.

Dong, J. et al., 2015. ChemDes: An integrated web-based platform for molecular descriptor and fingerprint computation. *Journal of Cheminformatics*, 7(1), pp.1–10.

Dormont, L., Bessière, J.-M. & Cohuet, A., 2013. Human skin volatiles: a review.

Journal of chemical ecology, 39(5), pp.569–78.

Dunkel, M. et al., 2009. SuperScent--a database of flavors and scents. *Nucleic acids research*, 37(Database issue), pp.D291-4.

El-Sayed, A.M., 2014. The Pherobase: Database of pheromones and semiochemicals. *The Pherobase*. Available at: http://www.pherobase.com.

Fitzgerald, J.E. et al., 2016. Artificial Nose Technology: Status and Prospects in Diagnostics. *Trends in Biotechnology*, xx, pp.1–10.

Friedman, J.H., 2001. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), pp.1189–1232.

Friedman, J.H., 2002. Stochastic gradient boosting. *Computational Statistics and Data Analysis*, 38(4), pp.367–378.

Garner, C.E. et al., 2007. Volatile organic compounds from feces and their potential for diagnosis of gastrointestinal disease. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, 21(8), pp.1675–88.

Gedeon, T.D., 1997. Data mining of inputs: analysing magnitude and functional measures. *International journal of neural systems*, 8(2), pp.209–218.

Gerber, N.N., 1967. Geosmin, an earthy smelling substance isolated from actinomycetes. *Biotechnology and Bioengineering*, 9(3), pp.321–327.

Godden, J.W., Xue, L. & Bajorath, J., 2000. Combinatorial preferences affect molecular similarity/diversity calculations using binary fingerprints and Tanimoto coefficients. *Journal of Chemical Information and Modeling*, 40(1), pp.163–166.

Goodfellow, I.J. et al., 2013. Maxout Networks. *Proceedings of the 30th International Conference on Machine Learning (ICML)*, 28, pp.1319–1327.

Hakim, M. et al., 2012. Volatile organic compounds of lung cancer and possible biochemical pathways. *Chemical reviews*, 112(11), pp.5949–66.

Hey, T., Tansley, S. & Tolle, K., 2009. *The Fourth Paradigm*, Available at: http://research.microsoft.com/en-us/collaboration/fourthparadigm/.

Holmes, E., Wilson, I.D. & Nicholson, J.K., 2008. Metabolic Phenotyping in Health and Disease. *Cell*, 134(5), pp.714–717.

Iijima, Y., 2014. Recent Advances in the Application of Metabolomics to Studies of Biogenic Volatile Organic Compounds (BVOC) Produced by Plant. *Metabolites*, 4(3), pp.699–721.

Ikeda, S. et al., 2013. Systematization of the protein sequence diversity in enzymes related to secondary metabolic pathways in plants, in the context of big data biology inspired by the KNApSAcK motorcycle database. *Plant and Cell Physiology*, 54(5), pp.711–727.

Ingram, L.O. et al., 2010. Metabolic engineering for production of biorenewable fuels and chemicals: Contributions of synthetic biology. *Journal of Biomedicine and Biotechnology*, 2010.

Intelligence, M., 2015. High Performance Machine Learning in R with H2O. , (October).

Jain, a. K., Murty, M.N. & Flynn, P.J., 1999. Data clustering: a review. *ACM Computing Surveys*, 31(3), pp.264–323.

Jeong, H. et al., 2001. Lethality and centrality in protein networks. *Nature*, 411(6833), pp.41–42.

Johnson, S.C., 1967. Hierarchical clustering schemes. *Psychometrika*, 32(3), pp.241–254.

Kai, M., Effmert, U. & Piechulla, B., 2016. Bacterial-plant-interactions: Approaches to unravel the biological function of bacterial volatiles in the rhizosphere. *Frontiers in Microbiology*, 7(FEB).

Kanchiswamy, C.N., Malnoy, M. & Maffei, M.E., 2015a. Bioprospecting bacterial and fungal volatiles for sustainable agriculture. *Trends in Plant Science*, 20(4), pp.206–211.

Kanchiswamy, C.N., Malnoy, M. & Maffei, M.E., 2015b. Chemical diversity of microbial volatiles and their potential for plant growth and productivity. *Frontiers in plant science*, 6(March), p.151.

Katsuragi, T. et al., 2013. Cuisine Omics : Fundamental Structures of Zouni and Retortable Pouched Pack of Curry Unveiled by Multivariate Analysis Based on Food Ingredients. *Foods & Food Ingredients J. Jpn., Vol.*, 218(1), pp.43–60.

Kelling, S. et al., 2009. Data-intensive science: a new paradigm for biodiversity studies. *Bioscience*, 59, pp.613–620.

Klekota, J. & Roth, F.P., 2008. Chemical substructures that enrich for biological activity. *Bioinformatics*, 24(21), pp.2518–2525.

Korpi, A., Järnberg, J. & Pasanen, A.-L., 2009. Microbial volatile organic compounds.

Critical reviews in toxicology, 39(2), pp.139–93.

LeCun, Y., Bengio, Y. & Hinton, G., 2015. Deep learning. *Nature*, 521(7553), pp.436–444. Available at: http://dx.doi.org/10.1038/nature14539.

Lee, S.Y., Kim, H.M. & Cheon, S., 2015. Metabolic engineering for the production of hydrocarbon fuels. *Current Opinion in Biotechnology*, 33, pp.15–22.

Lemfack, M.C. et al., 2014. mVOC: a database of microbial volatiles. *Nucleic acids research*, 42(Database issue), pp.D744-8.

Liu, K. et al., 2013. Novel Approach to Classify Plants Based on Metabolite Content Similarity.

Lourenço, C. & Turner, C., 2014. Breath Analysis in Disease Diagnosis: Methodological Considerations and Applications. *Metabolites*, 4(2), pp.465–498.

Ma, J. et al., 2015. Deep neural nets as a method for quantitative structure-activity relationships. *Journal of Chemical Information and Modeling*, 55(2), pp.263–274.

Maas, A.L., Hannun, A.Y. & Ng, A.Y., 2013. Rectifier Nonlinearities Improve Neural Network Acoustic Models. In *Proceedings of the 30 th International Conference on Machine Learning*. p. 6.

Madigan, M., 2012. Brock Biology of Microorganisms, 13th edn. *International Microbiology*, pp.550–551.

Maggiora, G. et al., 2014. Molecular similarity in medicinal chemistry. *Journal of Medicinal Chemistry*, 57(8), pp.3186–3204.

Maggiora, G.M. & Shanmugasundaram, V., 2004. Molecular similarity measures. *Methods in molecular biology (Clifton, N.J.)*, 275(2), pp.1–50.

Miller-Keane, 1993. Miller-Keane Encyclopedia and Dictionary of Medicine, Nursing, and Allied Health. *Gastroenterology Nursing*, 15, p.258.

Montuschi, P., 2007. Analysis of exhaled breath condensate in respiratory medicine: methodological aspects and potential clinical applications. *Therapeutic advances in respiratory disease*, 1(1), pp.5–23.

De Moraes, C.M. et al., 2014. Malaria-induced changes in host odors enhance mosquito attraction. *Proceedings of the National Academy of Sciences of the United States of America*, 111(30), pp.11079–84.

Mullen, G.B. et al., 1989. Studies on antifungal agents. 19. Effect of the C-5-aromatic substitution on the in vitro activity of novel 3,5-substituted isoxazolidines. *Chemotherapy*, 35(1), pp.39–42.

Murtagh, F. & Contreras, P., 2012. Algorithms for hierarchical clustering: An overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(1), pp.86–97.

Murtagh, F. & Contreras, P., 2011. Methods of Hierarchical Clustering. *Computer*, 38(2), pp.1–21.

N.M., G. & M., A., 2008. Biomarkers in exhaled breath condensate: A review of collection, processing and analysis. *Journal of Breath Research*, 2(3), p.no pagination.

Nakamura, K. et al., 2013. KNApSAcK-3D: a three-dimensional structure database of plant metabolites. *Plant & cell physiology*, 54(2), p.e4.

Nakamura, Y. et al., 2014. KNApSAcK Metabolite Activity Database for retrieving the relationships between metabolites and biological activities. *Plant & cell physiology*, 55(1), p.e7.

Natekin, A. & Knoll, A., 2013. Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, 7(DEC).

Nelder, J. & Wedderburn, R., 1972. Generalized Linear Models. *Journal of the Royal Statistical Society: Series A*, 135, pp.370–384.

Nikolova, N. & Jaworska, J., 2003. Approaches to Measure Chemical Similarity– a Review. *QSAR Combinatorial Science*, 22(910), pp.1006–1026.

Nozzi, N.E. et al., 2014. Metabolic engineering for higher alcohol production. *Metabolic Engineering*, 25, pp.174–182.

Ohtana, Y. et al., 2014. Clustering of 3D-Structure Similarity Based Network of Secondary Metabolites Reveals Their Relationships with Biological Activities. *Molecular Informatics*, p.n/a-n/a.

P., W., 2014. The calculation of molecular structural similarity: Principles and practice. *Molecular Informatics*, 33(6–7), pp.403–413.

Paixão, K. da S. et al., 2014. Volatile semiochemical-conditioned attraction of the male yellow fever mosquito, Aedes aegypti, to human hosts. *Journal of Vector Ecology*, 40(1), pp.1–6.

Patterson, D.E. et al., 1996. Neighborhood behavior: A useful concept for validation of “molecular diversity” descriptors. *Journal of Medicinal Chemistry*, 39(16), pp.3049–3059.

Patterson, D.J. et al., 2010. Names are key to the big new biology. *Trends in Ecology and Evolution*, 25(12), pp.686–691.

Phillips, M. et al., 2010. Volatile biomarkers in the breath of women with breast cancer. *Journal of breath research*, 4(2), p.26003.

Phillips, M. et al., 2007. Volatile biomarkers of pulmonary tuberculosis in the breath. *Tuberculosis (Edinburgh, Scotland)*, 87(1), pp.44–52.

Phillips, M. et al., 2003. Volatile markers of breast cancer in the breath. *The breast journal*, 9(3), pp.184–91.

Piechulla, B. & Degenhardt, J., 2014. The emerging importance of microbial volatile organic compounds. *Plant, cell & environment*, 37(4), pp.811–2.

Pineda, A. et al., 2010. Helping plants to deal with insects: The role of beneficial soil-borne microbes. *Trends in Plant Science*, 15(9), pp.507–514.

Probert, C.S.J. et al., 2009. Volatile organic compounds as diagnostic biomarkers in gastrointestinal and liver diseases. *Journal of gastrointestinal and liver diseases : JGLD*, 18(3), pp.337–43.

Rand, W.M., 1971. Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*, 66(336), pp.846–850.

Reddy, G.V.P. & Guerrero, A., 2004. Interactions of insect pheromones and plant semiochemicals. *Trends in Plant Science*, 9(5), pp.253–261.

Richard, S. et al., 2013. Bayesian Hierarchical Clustering for Studying Cancer Gene Expression Data with Unknown Statistics. , 8(10), pp.0–11.

Rojas-Cherto, M. et al., 2012. Metabolite identification using automated comparison of high-resolution multistage mass spectral trees. *Analytical chemistry*, 84(13), pp.5524–5534.

Rowan, D.D., 2011. Volatile Metabolites. *Metabolites*, 1(1), pp.41–63.

Ryu, C.-M. et al., 2003. Bacterial volatiles promote growth in Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America*, 100(8), pp.4927–32.

Scala, A. et al., 2013. Green leaf volatiles: A plant's multifunctional weapon against herbivores and pathogens. *International Journal of Molecular Sciences*, 14(9), pp.17781–17811.

Sheridan, R.P. et al., 2016. Extreme Gradient Boosting as a Method for Quantitative Structure-Activity Relationships. *Journal of Chemical Information and*

Modeling, p.acs.jcim.6b00591.

Shirasu, M. & Touhara, K., 2011. The scent of disease: Volatile organic compounds of the human body related to disease and disorder. *Journal of Biochemistry*, 150(3), pp.257–266.

Skogerson, K. et al., 2011. The volatile compound BinBase mass spectral database. *BMC bioinformatics*, 12(1), p.321.

Srivastava, N. et al., 2014. Dropout: prevent NN from overfitting. *Journal of Machine Learning Research*, 15, pp.1929–1958.

Syhre, M. & Chambers, S.T., 2008. The scent of Mycobacterium tuberculosis. *Tuberculosis (Edinburgh, Scotland)*, 88(4), pp.317–23.

Thalamuthu, A. et al., 2006. Evaluation and comparison of gene clustering methods in microarray analysis. *Bioinformatics*, 22(19), pp.2405–2412.

Todeschini, R. et al., 2012. Similarity coefficients for binary chemoinformatics data: Overview and extended comparison using simulated and real data sets. *Journal of Chemical Information and Modeling*, 52(11), pp.2884–2901.

Todeschini, R. & Consonni, V., 2000. Handbook of Molecular Descriptors. *New York*, 11, p.688.

Verhulst, N.O. et al., 2010. Chemical ecology of interactions between human skin microbiota and mosquitoes. *FEMS microbiology ecology*, 74(1), pp.1–9.

Vikram, P. et al., 2014. A recent review on phytochemical constituents and medicinal properties of kesum (Polygonum minus Huds.). *Asian Pacific journal of tropical biomedicine*, 4(6), pp.430–5.

Wang, C., Dong, R., et al., 2014. Exhaled volatile organic compounds as lung cancer biomarkers during one-lung ventilation. *Scientific Reports*, 4, p.7312.

Wang, C., Sun, B., et al., 2014. Volatile Organic Metabolites Identify Patients with Breast Cancer, Cyclomastopathy, and Mammary Gland Fibroma. *Scientific Reports*, 4, pp.1–6.

Wijaya, S.H. et al., 2014. Supervised clustering based on DPClusO: Prediction of plant-disease relations using Jamu formulas of KNApSAcK database. *BioMed Research International*, 2014.

Willett, P., 2009. Similarity methods in chemoinformatics. *Annual Review of Information Science and Technology*, 43, pp.1–117.

Willett, P., Barnard, J.M. & Downs, G.M., 1998. Chemical similarity searching.

Journal of Chemical Information and Modeling, 38(6), pp.983–996.

Willett, P., Barnard, J.M. & Downs, G.M., 1998. Chemical Similarity Searching. *Journal of Chemical Information and Computer Sciences*, 38(6), pp.983–996.

Wilson, A., 2015. Advances in Electronic-Nose Technologies for the Detection of Volatile Biomarker Metabolites in the Human Breath. *Metabolites*, 5(1), pp.140–163.

Wong, R.P.M., Flematti, G.R. & Davis, T.M.E., 2012. Investigation of volatile organic biomarkers derived from Plasmodium falciparum in vitro. *Malaria journal*, 11(1), p.314.

Yusuf, N. et al., 2015. In-vitro diagnosis of single and poly microbial species targeted for diabetic foot infection using e-nose technology. *BMC bioinformatics*, 16, p.158.

Zhang, X. et al., 2014. Improving Deep Neural Network Acoustic Models using Generalized Maxout Networks. *Icassp*, pp.2–6.

Achievements

Reviewed publications

1. Yuki Ohtana, Azian Azamimi Abdullah, Md. Altaf-Ul-Amin, Ming Huang, Naoaki Ono, Tetsuo Sato, Tadao Sugiura, Hisayuki Horai, Yukiko Nakamura, Aki Morita (Hirai), Klaus W. Lange, Nelson K. Kibinge, Tetsuo Katsuragi, Tsuyoshi Shirai, and Shigehiko Kanaya, 2014, Clustering of 3D-Structure Similarity Based Network of Secondary Metabolites Reveals Their Relationships with Biological Activities, Molecular Informatics, 2014, 33, 790 – 801 (in Chapter 2)
2. Azian Azamimi Abdullah, Md. Altaf-Ul-Amin, Naoaki Ono, Tetsuo Sato, Tadao Sugiura, Aki Hirai Morita, Tetsuo Katsuragi, Ai Muto, Takaaki Nishioka, and Shigehiko Kanaya, 2015, Development and Mining of a Volatile Organic Compound Database, BioMed Research International, Volume 2015 (2015), Article ID 139254 (in Chapter 2, 3 and 4).
3. Nurlisa Yusuf, Ammar Zakaria, Mohammad Iqbal Omar, Ali Yeon Md Shakaff, Maz Jamilah Masnan, Latifah Munirah Kamarudin, Norasmadi Abdul Rahim, Nur Zawatil Isqi Zakaria, Azian Azamimi Abdullah, Amizah Othman, Mohd Sadek Yasin, "In-vitro diagnosis of single and poly microbial species targeted for diabetic foot infection using e-nose technology", BMC Bioinformatics 2015, 16(1), 1 (in Chapter 3).
4. Azian Azamimi Abdullah, Md. Altaf-Ul-Amin, Naoki Ono, Nurlisa Yusuf, Ammar Zakaria, Takaaki Nishioka, and Shigehiko Kanaya, 2016, Comparison of Clustering Methods in the Context of Chemical Structure Similarity Based Classification of VOCs. Procedia Chemistry 20 (2016), 40-44 (in Chapter 4).
5. Kang Liu, Azian Azamimi Abdullah, Ming Huang, Takaaki Nishioka, Md. Altaf-Ul-Amin, and Shigehiko Kanaya, 2017, Novel Approach to Classify Plants Based on Metabolite Content Similarity, BioMed Research International, Volume 2017 (2017), Article ID 5296729 (in Chapter 4).

International and domestic conferences

1. Yuki Ohtana, Azian Azamimi Abdullah, Md. Altaf-Ul-Amin, Ming Huang, Naoaki Ono, Tetsuo Sato, Tadao Sugiura, Hisayuki Horai, Yukiko Nakamura, Aki Morita (Hirai), Klaus W. Lange, Nelson K. Kibinge, Tetsuo Katsuragi, Tsuyoshi Shirai, and Shigehiko Kanaya, "Clustering of 3D-Structure Similarity Based Network of Secondary Metabolites Reveals Their Relationships with Biological Activities" GIW ISCB-ASIA 2014, 45, P86, Tokyo, Japan, December 16, 2014 (Poster Presentation) (in Chapter 2).
2. Azian Azamimi Abdullah, Md. Altaf-Ul-Amin, Takaaki Nishioka, Tetsuo Katsuragi, Naoaki Ono, Ammar Zakaria, Ali Yeon Md Shakaff and Shigehiko Kanaya, "Analysis of Volatile Metabolites Emitted by Various Species to Reveal Their Roles in Chemical Ecology and Healthcare", 2015 2nd International Conference on Biomedical Engineering (ICOBE 2015), 30-31st March, Penang, Malaysia (Oral Presentation) (in Chapter 2).
3. Azian Azamimi Abdullah, Md. Altaf-Ul-Amin, Aki Hirai Morita, Tetsuo Sato, Tadao Sugiura, Naoaki Ono, Takaaki Nishioka and Shigehiko Kanaya, "KNApSAcK Metabolite Ecology Database for Investigating the Relationships Between VOCs and Biological Activities", 11th International Conference of the Metabolomics Society (Metabolomics 2015), 29th June – 2nd July 2015, San Francisco, California, USA (Poster Presentation) (in Chapter 2, 3 and 4).
4. Azian Azamimi Abdullah, Shigehiko Kanaya and Md Altaf-Ul-Amin, "Development and Mining of a Volatile Organic Compound (VOC) Database", 11th Asian Conference on Chemical Sensors (ACCS2015), 16-18th November 2015, Penang, Malaysia (Oral Presentation) (in Chapter 2, 3 and 4).
5. Azian Azamimi Abdullah, Md Altaf-Ul-Amin, Naoaki Ono, Nurlisa Yusuf, Ammar Zakaria, Takaaki Nishioka, Shigehiko Kanaya, "Comparison of

Clustering Methods in the Context of Chemical Structure Similarity Based Classification of VOCs", 11th Asian Conference on Chemical Sensors (ACCS2015), 16-18th November 2015, Penang, Malaysia (Oral Presentation) (in Chapter 4).

6. Md. Altaf-Ul-Amin, Kang Liu, Azian Azamimi Abdullah, Aki H. Morita, Makio Shiraishi and Shigehiko Kanaya, "Classification of plants based on chemical structural similarity of metabolite contents obtained from KNApSAcK database", Applied Bioinformatics in Life Sciences, 17-18th March 2016, Leuven, Belgium (Poster Presentation) (in Chapter 4).
7. Azian Azamimi Abdullah, Md. Altaf-Ul-Amin, Ammar Zakaria and Shigehiko Kanaya, "Prediction of Biological Activities of Volatile Metabolites Using Deep Learning", 15th International Conference On BioInformatics (InCOB 2016), 21-23 September 2016, Biopolis, Singapore (Short Oral and Poster Presentation) (in Chapter 4).
8. Azian Azamimi Abdullah, "Comprehensive Understand of Relationship Among Species, Volatile Organic Compound, and Biological Activity", 39th Symposium on Chemoinformatics, 29-30 September 2016, Hamamatsu, Shizouka, Japan (Oral Presentation) (in Chapter 2, 3 and 4).

Awards

1. Molecular Informatics 2014 Best Paper Award

Yuki Ohtana, Azian Azamimi Abdullah, Md Altaf-Ul-Amin, Ming Huang, Naoaki Ono, Tetsuo Sato, Tadao Sugiura, Hisayuki Horai, Yukiko Nakamura, Klaus W Lange, Nelson K Kibinge, Tetsuo Katsuragi, Tsuyoshi Shirai, Shigehiko Kanaya, "Clustering of 3D-Structure Similarity Based Network of Secondary Metabolites Reveals Their Relationships with Biological Activities", Molecular Informatics 2014, 33(11-12), 790-801.

2. INCOB 2016 Best Poster Award (Silver)

Azian Azamimi Abdullah, Md. Altaf-Ul-Amin, Ammar Zakaria and Shigehiko Kanaya, “Prediction of Biological Activities of Volatile Metabolites Using Deep Learning”, 15th International Conference On BioInformatics (InCOB 2016), 21-23 September 2016, Biopolis, Singapore.

Appendices

Appendix A

The description of PubChem fingerprint, as mentioned in Chapter 4.

Bit Position	Bit Substructure
0	>= 4 H
1	>= 8 H
2	>= 16 H
3	>= 32 H
4	>= 1 Li
5	>= 2 Li
6	>= 1 B
7	>= 2 B
8	>= 4 B
9	>= 2 C
10	>= 4 C
11	>= 8 C
12	>= 16 C
13	>= 32 C
14	>= 1 N
15	>= 2 N
16	>= 4 N
17	>= 8 N
18	>= 1 O
19	>= 2 O
20	>= 4 O
21	>= 8 O
22	>= 16 O
23	>= 1 F
24	>= 2 F

25	>= 4 F
26	>= 1 Na
27	>= 2 Na
28	>= 1 Si
29	>= 2 Si
30	>= 1 P
31	>= 2 P
32	>= 4 P
33	>= 1 S
34	>= 2 S
35	>= 4 S
36	>= 8 S
37	>= 1 Cl
38	>= 2 Cl
39	>= 4 Cl
40	>= 8 Cl
41	>= 1 K
42	>= 2 K
43	>= 1 Br
44	>= 2 Br
45	>= 4 Br
46	>= 1 I
47	>= 2 I
48	>= 4 I
49	>= 1 Be
50	>= 1 Mg
51	>= 1 Al
52	>= 1 Ca
53	>= 1 Sc
54	>= 1 Ti
55	>= 1 V

56	>= 1 Cr
57	>= 1 Mn
58	>= 1 Fe
59	>= 1 Co
60	>= 1 Ni
61	>= 1 Cu
62	>= 1 Zn
63	>= 1 Ga
64	>= 1 Ge
65	>= 1 As
66	>= 1 Se
67	>= 1 Kr
68	>= 1 Rb
69	>= 1 Sr
70	>= 1 Y
71	>= 1 Zr
72	>= 1 Nb
73	>= 1 Mo
74	>= 1 Ru
75	>= 1 Rh
76	>= 1 Pd
77	>= 1 Ag
78	>= 1 Cd
79	>= 1 In
80	>= 1 Sn
81	>= 1 Sb
82	>= 1 Te
83	>= 1 Xe
84	>= 1 Cs
85	>= 1 Ba
86	>= 1 Lu

87	>= 1 Hf
88	>= 1 Ta
89	>= 1 W
90	>= 1 Re
91	>= 1 Os
92	>= 1 Ir
93	>= 1 Pt
94	>= 1 Au
95	>= 1 Hg
96	>= 1 Tl
97	>= 1 Pb
98	>= 1 Bi
99	>= 1 La
100	>= 1 Ce
101	>= 1 Pr
102	>= 1 Nd
103	>= 1 Pm
104	>= 1 Sm
105	>= 1 Eu
106	>= 1 Gd
107	>= 1 Tb
108	>= 1 Dy
109	>= 1 Ho
110	>= 1 Er
111	>= 1 Tm
112	>= 1 Yb
113	>= 1 Tc
114	>= 1 U
115	>= 1 any ring size 3
116	>= 1 saturated or aromatic carbon-only ring size 3
117	>= 1 saturated or aromatic nitrogen-containing ring size 3

118	>= 1 saturated or aromatic heteroatom-containing ring size 3
119	>= 1 unsaturated non-aromatic carbon-only ring size 3
120	>= 1 unsaturated non-aromatic nitrogen-containing ring size 3
121	>= 1 unsaturated non-aromatic heteroatom-containing ring size 3
122	>= 2 any ring size 3
123	>= 2 saturated or aromatic carbon-only ring size 3
124	>= 2 saturated or aromatic nitrogen-containing ring size 3
125	>= 2 saturated or aromatic heteroatom-containing ring size 3
126	>= 2 unsaturated non-aromatic carbon-only ring size 3
127	>= 2 unsaturated non-aromatic nitrogen-containing ring size 3
128	>= 2 unsaturated non-aromatic heteroatom-containing ring size 3
129	>= 1 any ring size 4
130	>= 1 saturated or aromatic carbon-only ring size 4
131	>= 1 saturated or aromatic nitrogen-containing ring size 4
132	>= 1 saturated or aromatic heteroatom-containing ring size 4
133	>= 1 unsaturated non-aromatic carbon-only ring size 4
134	>= 1 unsaturated non-aromatic nitrogen-containing ring size 4
135	>= 1 unsaturated non-aromatic heteroatom-containing ring size 4
136	>= 2 any ring size 4
137	>= 2 saturated or aromatic carbon-only ring size 4
138	>= 2 saturated or aromatic nitrogen-containing ring size 4
139	>= 2 saturated or aromatic heteroatom-containing ring size 4
140	>= 2 unsaturated non-aromatic carbon-only ring size 4
141	>= 2 unsaturated non-aromatic nitrogen-containing ring size 4
142	>= 2 unsaturated non-aromatic heteroatom-containing ring size 4
143	>= 1 any ring size 5
144	>= 1 saturated or aromatic carbon-only ring size 5
145	>= 1 saturated or aromatic nitrogen-containing ring size 5
146	>= 1 saturated or aromatic heteroatom-containing ring size 5
147	>= 1 unsaturated non-aromatic carbon-only ring size 5
148	>= 1 unsaturated non-aromatic nitrogen-containing ring size 5

149	>= 1 unsaturated non-aromatic heteroatom-containing ring size 5
150	>= 2 any ring size 5
151	>= 2 saturated or aromatic carbon-only ring size 5
152	>= 2 saturated or aromatic nitrogen-containing ring size 5
153	>= 2 saturated or aromatic heteroatom-containing ring size 5
154	>= 2 unsaturated non-aromatic carbon-only ring size 5
155	>= 2 unsaturated non-aromatic nitrogen-containing ring size 5
156	>= 2 unsaturated non-aromatic heteroatom-containing ring size 5
157	>= 3 any ring size 5
158	>= 3 saturated or aromatic carbon-only ring size 5
159	>= 3 saturated or aromatic nitrogen-containing ring size 5
160	>= 3 saturated or aromatic heteroatom-containing ring size 5
161	>= 3 unsaturated non-aromatic carbon-only ring size 5
162	>= 3 unsaturated non-aromatic nitrogen-containing ring size 5
163	>= 3 unsaturated non-aromatic heteroatom-containing ring size 5
164	>= 4 any ring size 5
165	>= 4 saturated or aromatic carbon-only ring size 5
166	>= 4 saturated or aromatic nitrogen-containing ring size 5
167	>= 4 saturated or aromatic heteroatom-containing ring size 5
168	>= 4 unsaturated non-aromatic carbon-only ring size 5
169	>= 4 unsaturated non-aromatic nitrogen-containing ring size 5
170	>= 4 unsaturated non-aromatic heteroatom-containing ring size 5
171	>= 5 any ring size 5
172	>= 5 saturated or aromatic carbon-only ring size 5
173	>= 5 saturated or aromatic nitrogen-containing ring size 5
174	>= 5 saturated or aromatic heteroatom-containing ring size 5
175	>= 5 unsaturated non-aromatic carbon-only ring size 5
176	>= 5 unsaturated non-aromatic nitrogen-containing ring size 5
177	>= 5 unsaturated non-aromatic heteroatom-containing ring size 5
178	>= 1 any ring size 6
179	>= 1 saturated or aromatic carbon-only ring size 6

180	>= 1 saturated or aromatic nitrogen-containing ring size 6
181	>= 1 saturated or aromatic heteroatom-containing ring size 6
182	>= 1 unsaturated non-aromatic carbon-only ring size 6
183	>= 1 unsaturated non-aromatic nitrogen-containing ring size 6
184	>= 1 unsaturated non-aromatic heteroatom-containing ring size 6
185	>= 2 any ring size 6
186	>= 2 saturated or aromatic carbon-only ring size 6
187	>= 2 saturated or aromatic nitrogen-containing ring size 6
188	>= 2 saturated or aromatic heteroatom-containing ring size 6
189	>= 2 unsaturated non-aromatic carbon-only ring size 6
190	>= 2 unsaturated non-aromatic nitrogen-containing ring size 6
191	>= 2 unsaturated non-aromatic heteroatom-containing ring size 6
192	>= 3 any ring size 6
193	>= 3 saturated or aromatic carbon-only ring size 6
194	>= 3 saturated or aromatic nitrogen-containing ring size 6
195	>= 3 saturated or aromatic heteroatom-containing ring size 6
196	>= 3 unsaturated non-aromatic carbon-only ring size 6
197	>= 3 unsaturated non-aromatic nitrogen-containing ring size 6
198	>= 3 unsaturated non-aromatic heteroatom-containing ring size 6
199	>= 4 any ring size 6
200	>= 4 saturated or aromatic carbon-only ring size 6
201	>= 4 saturated or aromatic nitrogen-containing ring size 6
202	>= 4 saturated or aromatic heteroatom-containing ring size 6
203	>= 4 unsaturated non-aromatic carbon-only ring size 6
204	>= 4 unsaturated non-aromatic nitrogen-containing ring size 6
205	>= 4 unsaturated non-aromatic heteroatom-containing ring size 6
206	>= 5 any ring size 6
207	>= 5 saturated or aromatic carbon-only ring size 6
208	>= 5 saturated or aromatic nitrogen-containing ring size 6
209	>= 5 saturated or aromatic heteroatom-containing ring size 6
210	>= 5 unsaturated non-aromatic carbon-only ring size 6

211	>= 5 unsaturated non-aromatic nitrogen-containing ring size 6
212	>= 5 unsaturated non-aromatic heteroatom-containing ring size 6
213	>= 1 any ring size 7
214	>= 1 saturated or aromatic carbon-only ring size 7
215	>= 1 saturated or aromatic nitrogen-containing ring size 7
216	>= 1 saturated or aromatic heteroatom-containing ring size 7
217	>= 1 unsaturated non-aromatic carbon-only ring size 7
218	>= 1 unsaturated non-aromatic nitrogen-containing ring size 7
219	>= 1 unsaturated non-aromatic heteroatom-containing ring size 7
220	>= 2 any ring size 7
221	>= 2 saturated or aromatic carbon-only ring size 7
222	>= 2 saturated or aromatic nitrogen-containing ring size 7
223	>= 2 saturated or aromatic heteroatom-containing ring size 7
224	>= 2 unsaturated non-aromatic carbon-only ring size 7
225	>= 2 unsaturated non-aromatic nitrogen-containing ring size 7
226	>= 2 unsaturated non-aromatic heteroatom-containing ring size 7
227	>= 1 any ring size 8
228	>= 1 saturated or aromatic carbon-only ring size 8
229	>= 1 saturated or aromatic nitrogen-containing ring size 8
230	>= 1 saturated or aromatic heteroatom-containing ring size 8
231	>= 1 unsaturated non-aromatic carbon-only ring size 8
232	>= 1 unsaturated non-aromatic nitrogen-containing ring size 8
233	>= 1 unsaturated non-aromatic heteroatom-containing ring size 8
234	>= 2 any ring size 8
235	>= 2 saturated or aromatic carbon-only ring size 8
236	>= 2 saturated or aromatic nitrogen-containing ring size 8
237	>= 2 saturated or aromatic heteroatom-containing ring size 8
238	>= 2 unsaturated non-aromatic carbon-only ring size 8
239	>= 2 unsaturated non-aromatic nitrogen-containing ring size 8
240	>= 2 unsaturated non-aromatic heteroatom-containing ring size 8
241	>= 1 any ring size 9

242	>= 1 saturated or aromatic carbon-only ring size 9
243	>= 1 saturated or aromatic nitrogen-containing ring size 9
244	>= 1 saturated or aromatic heteroatom-containing ring size 9
245	>= 1 unsaturated non-aromatic carbon-only ring size 9
246	>= 1 unsaturated non-aromatic nitrogen-containing ring size 9
247	>= 1 unsaturated non-aromatic heteroatom-containing ring size 9
248	>= 1 any ring size 10
249	>= 1 saturated or aromatic carbon-only ring size 10
250	>= 1 saturated or aromatic nitrogen-containing ring size 10
251	>= 1 saturated or aromatic heteroatom-containing ring size 10
252	>= 1 unsaturated non-aromatic carbon-only ring size 10
253	>= 1 unsaturated non-aromatic nitrogen-containing ring size 10
254	>= 1 unsaturated non-aromatic heteroatom-containing ring size 10
255	>= 1 aromatic ring
256	>= 1 hetero-aromatic ring
257	>= 2 aromatic rings
258	>= 2 hetero-aromatic rings
259	>= 3 aromatic rings
260	>= 3 hetero-aromatic rings
261	>= 4 aromatic rings
262	>= 4 hetero-aromatic rings
263	Li-H
264	Li-Li
265	Li-B
266	Li-C
267	Li-O
268	Li-F
269	Li-P
270	Li-S
271	Li-Cl
272	B-H

273	B-B
274	B-C
275	B-N
276	B-O
277	B-F
278	B-Si
279	B-P
280	B-S
281	B-Cl
282	B-Br
283	C-H
284	C-C
285	C-N
286	C-O
287	C-F
288	C-Na
289	C-Mg
290	C-Al
291	C-Si
292	C-P
293	C-S
294	C-Cl
295	C-As
296	C-Se
297	C-Br
298	C-I
299	N-H
300	N-N
301	N-O
302	N-F
303	N-Si

304	N-P
305	N-S
306	N-Cl
307	N-Br
308	O-H
309	O-O
310	O-Mg
311	O-Na
312	O-Al
313	O-Si
314	O-P
315	O-K
316	F-P
317	F-S
318	Al-H
319	Al-Cl
320	Si-H
321	Si-Si
322	Si-Cl
323	P-H
324	P-P
325	As-H
326	As-As
327	C(~Br)(~C)
328	C(~Br)(~C)(~C)
329	C(~Br)(~H)
330	C(~Br)(:C)
331	C(~Br)(:N)
332	C(~C)(~C)
333	C(~C)(~C)(~C)
334	C(~C)(~C)(~C)(~C)

335	C(~C)(~C)(~C)(~H)
336	C(~C)(~C)(~C)(~N)
337	C(~C)(~C)(~C)(~O)
338	C(~C)(~C)(~H)(~N)
339	C(~C)(~C)(~H)(~O)
340	C(~C)(~C)(~N)
341	C(~C)(~C)(~O)
342	C(~C)(~Cl)
343	C(~C)(~Cl)(~H)
344	C(~C)(~H)
345	C(~C)(~H)(~N)
346	C(~C)(~H)(~O)
347	C(~C)(~H)(~O)(~O)
348	C(~C)(~H)(~P)
349	C(~C)(~H)(~S)
350	C(~C)(~I)
351	C(~C)(~N)
352	C(~C)(~O)
353	C(~C)(~S)
354	C(~C)(~Si)
355	C(~C)(:C)
356	C(~C)(:C)(:C)
357	C(~C)(:C)(:N)
358	C(~C)(:N)
359	C(~C)(:N)(:N)
360	C(~Cl)(~Cl)
361	C(~Cl)(~H)
362	C(~Cl)(:C)
363	C(~F)(~F)
364	C(~F)(:C)
365	C(~H)(~N)

366	C(~H)(~O)
367	C(~H)(~O)(~O)
368	C(~H)(~S)
369	C(~H)(~Si)
370	C(~H)(:C)
371	C(~H)(:C)(:C)
372	C(~H)(:C)(:N)
373	C(~H)(:N)
374	C(~H)(~H)(~H)
375	C(~N)(~N)
376	C(~N)(:C)
377	C(~N)(:C)(:C)
378	C(~N)(:C)(:N)
379	C(~N)(:N)
380	C(~O)(~O)
381	C(~O)(:C)
382	C(~O)(:C)(:C)
383	C(~S)(:C)
384	C(:C)(:C)
385	C(:C)(:C)(:C)
386	C(:C)(:C)(:N)
387	C(:C)(:N)
388	C(:C)(:N)(:N)
389	C(:N)(:N)
390	N(~C)(~C)
391	N(~C)(~C)(~C)
392	N(~C)(~C)(~H)
393	N(~C)(~H)
394	N(~C)(~H)(~N)
395	N(~C)(~O)
396	N(~C)(:C)

397	N(~C)(:C)(:C)
398	N(~H)(~N)
399	N(~H)(:C)
400	N(~H)(:C)(:C)
401	N(~O)(~O)
402	N(~O)(:O)
403	N(:C)(:C)
404	N(:C)(:C)(:C)
405	O(~C)(~C)
406	O(~C)(~H)
407	O(~C)(~P)
408	O(~H)(~S)
409	O(:C)(:C)
410	P(~C)(~C)
411	P(~O)(~O)
412	S(~C)(~C)
413	S(~C)(~H)
414	S(~C)(~O)
415	Si(~C)(~C)
416	C=C
417	C#C
418	C=N
419	C#N
420	C=O
421	C=S
422	N=N
423	N=O
424	N=P
425	P=O
426	P=P
427	C(#C)(-C)

428	C(#C)(-H)
429	C(#N)(-C)
430	C(-C)(-C)(=C)
431	C(-C)(-C)(=N)
432	C(-C)(-C)(=O)
433	C(-C)(-Cl)(=O)
434	C(-C)(-H)(=C)
435	C(-C)(-H)(=N)
436	C(-C)(-H)(=O)
437	C(-C)(-N)(=C)
438	C(-C)(-N)(=N)
439	C(-C)(-N)(=O)
440	C(-C)(-O)(=O)
441	C(-C)(=C)
442	C(-C)(=N)
443	C(-C)(=O)
444	C(-Cl)(=O)
445	C(-H)(-N)(=C)
446	C(-H)(=C)
447	C(-H)(=N)
448	C(-H)(=O)
449	C(-N)(=C)
450	C(-N)(=N)
451	C(-N)(=O)
452	C(-O)(=O)
453	N(-C)(=C)
454	N(-C)(=O)
455	N(-O)(=O)
456	P(-O)(=O)
457	S(-C)(=O)
458	S(-O)(=O)

459	S(=O)(=O)
460	C-C-C#C
461	O-C-C=N
462	O-C-C=O
463	N:C-S-[#1]
464	N-C-C=C
465	O=S-C-C
466	N#C-C=C
467	C=N-N-C
468	O=S-C-N
469	S-S-C:C
470	C:C-C=C
471	S:C:C:C
472	C:N:C-C
473	S-C:N:C
474	S:C:C:N
475	S-C=N-C
476	C-O-C=C
477	N-N-C:C
478	S-C=N-[#1]
479	S-C-S-C
480	C:S:C-C
481	O-S-C:C
482	C:N-C:C
483	N-S-C:C
484	N-C:N:C
485	N:C:C:N
486	N-C:N:N
487	N-C=N-C
488	N-C=N-[#1]
489	N-C-S-C

490	C-C-C=C
491	C-N:C-[#1]
492	N-C:O:C
493	O=C-C:C
494	O=C-C:N
495	C-N-C:C
496	N:N-C-[#1]
497	O-C:C:N
498	O-C=C-C
499	N-C:C:N
500	C-S-C:C
501	Cl-C:C-C
502	N-C=C-[#1]
503	Cl-C:C-[#1]
504	N:C:N-C
505	Cl-C:C-O
506	C-C:N:C
507	C-C-S-C
508	S=C-N-C
509	Br-C:C-C
510	[#1]-N-N-[#1]
511	S=C-N-[#1]
512	C-[As]-O-[#1]
513	S:C:C-[#1]
514	O-N-C-C
515	N-N-C-C
516	[#1]-C=C-[#1]
517	N-N-C-N
518	O=C-N-N
519	N=C-N-C
520	C=C-C:C

521	C:N-C-[#1]
522	C-N-N-[#1]
523	N:C:C-C
524	C-C=C-C
525	[As]-C:C-[#1]
526	Cl-C:C-Cl
527	C:C:N-[#1]
528	[#1]-N-C-[#1]
529	Cl-C-C-Cl
530	N:C-C:C
531	S-C:C-C
532	S-C:C-[#1]
533	S-C:C-N
534	S-C:C-O
535	O=C-C-C
536	O=C-C-N
537	O=C-C-O
538	N=C-C-C
539	N=C-C-[#1]
540	C-N-C-[#1]
541	O-C:C-C
542	O-C:C-[#1]
543	O-C:C-N
544	O-C:C-O
545	N-C:C-C
546	N-C:C-[#1]
547	N-C:C-N
548	O-C-C:C
549	N-C-C:C
550	Cl-C-C-C
551	Cl-C-C-O

552	C:C-C:C
553	O=C-C=C
554	Br-C-C-C
555	N=C-C=C
556	C=C-C-C
557	N:C-O-[#1]
558	O=N-C:C
559	O-C-N-[#1]
560	N-C-N-C
561	Cl-C-C=O
562	Br-C-C=O
563	O-C-O-C
564	C=C-C=C
565	C:C-O-C
566	O-C-C-N
567	O-C-C-O
568	N#C-C-C
569	N-C-C-N
570	C:C-C-C
571	[#1]-C-O-[#1]
572	N:C:N:C
573	O-C-C=C
574	O-C-C:C-C
575	O-C-C:C-O
576	N=C-C:C-[#1]
577	C:C-N-C:C
578	C-C:C-C:C
579	O=C-C-C-C
580	O=C-C-C-N
581	O=C-C-C-O
582	C-C-C-C-C

583	Cl-C:C-O-C
584	C:C-C=C-C
585	C-C:C-N-C
586	C-S-C-C-C
587	N-C:C-O-[#1]
588	O=C-C-C=O
589	C-C:C-O-C
590	C-C:C-O-[#1]
591	Cl-C-C-C-C
592	N-C-C-C-C
593	N-C-C-C-N
594	C-O-C-C=C
595	C:C-C-C-C
596	N=C-N-C-C
597	O=C-C-C:C
598	Cl-C:C:C-C
599	[#1]-C-C=C-[#1]
600	N-C:C:C-C
601	N-C:C:C-N
602	O=C-C-N-C
603	C-C:C:C-C
604	C-O-C-C:C
605	O=C-C-O-C
606	O-C:C-C-C
607	N-C-C-C:C
608	C-C-C-C:C
609	Cl-C-C-N-C
610	C-O-C-O-C
611	N-C-C-N-C
612	N-C-O-C-C
613	C-N-C-C-C

614	C-C-O-C-C
615	N-C-C-O-C
616	C:C:N:N:C
617	C-C-C-O-[#1]
618	C:C-C-C:C
619	O-C-C=C-C
620	C:C-O-C-C
621	N-C:C:C:N
622	O=C-O-C:C
623	O=C-C:C-C
624	O=C-C:C-N
625	O=C-C:C-O
626	C-O-C:C-C
627	O=[As]-C:C:C
628	C-N-C-C:C
629	S-C:C:C-N
630	O-C:C-O-C
631	O-C:C-O-[#1]
632	C-C-O-C:C
633	N-C-C:C-C
634	C-C-C:C-C
635	N-N-C-N-[#1]
636	C-N-C-N-C
637	O-C-C-C-C
638	O-C-C-C-N
639	O-C-C-C-O
640	C=C-C-C-C
641	O-C-C-C=C
642	O-C-C-C=O
643	[#1]-C-C-N-[#1]
644	C-C=N-N-C

645	O=C-N-C-C
646	O=C-N-C-[#1]
647	O=C-N-C-N
648	O=N-C:C-N
649	O=N-C:C-O
650	O=C-N-C=O
651	O-C:C:C-C
652	O-C:C:C-N
653	O-C:C:C-O
654	N-C-N-C-C
655	O-C-C-C:C
656	C-C-N-C-C
657	C-N-C:C-C
658	C-C-S-C-C
659	O-C-C-N-C
660	C-C=C-C-C
661	O-C-O-C-C
662	O-C-C-O-C
663	O-C-C-O-[#1]
664	C-C=C-C=C
665	N-C:C-C-C
666	C=C-C-O-C
667	C=C-C-O-[#1]
668	C-C:C-C-C
669	Cl-C:C-C=O
670	Br-C:C:C-C
671	O=C-C=C-C
672	O=C-C=C-[#1]
673	O=C-C=C-N
674	N-C-N-C:C
675	Br-C-C-C:C

676	N#C-C-C-C
677	C-C=C-C:C
678	C-C-C=C-C
679	C-C-C-C-C-C
680	O-C-C-C-C-C
681	O-C-C-C-C-O
682	O-C-C-C-C-N
683	N-C-C-C-C-C
684	O=C-C-C-C-C
685	O=C-C-C-C-N
686	O=C-C-C-C-O
687	O=C-C-C-C=O
688	C-C-C-C-C-C-C
689	O-C-C-C-C-C-C
690	O-C-C-C-C-C-O
691	O-C-C-C-C-C-N
692	O=C-C-C-C-C-C
693	O=C-C-C-C-C-O
694	O=C-C-C-C-C=O
695	O=C-C-C-C-C-N
696	C-C-C-C-C-C-C-C
697	C-C-C-C-C-C(C)-C
698	O-C-C-C-C-C-C-C
699	O-C-C-C-C-C(C)-C
700	O-C-C-C-C-C-O-C
701	O-C-C-C-C-C(O)-C
702	O-C-C-C-C-C-N-C
703	O-C-C-C-C-C(N)-C
704	O=C-C-C-C-C-C-C
705	O=C-C-C-C-C(O)-C
706	O=C-C-C-C-C(=O)-C

707	O=C-C-C-C-C(N)-C
708	C-C(C)-C-C
709	C-C(C)-C-C-C
710	C-C-C(C)-C-C
711	C-C(C)(C)-C-C
712	C-C(C)-C(C)-C
713	Cc1ccc(C)cc1
714	Cc1ccc(O)cc1
715	Cc1ccc(S)cc1
716	Cc1ccc(N)cc1
717	Cc1ccc(Cl)cc1
718	Cc1ccc(Br)cc1
719	Oc1ccc(O)cc1
720	Oc1ccc(S)cc1
721	Oc1ccc(N)cc1
722	Oc1ccc(Cl)cc1
723	Oc1ccc(Br)cc1
724	Sc1ccc(S)cc1
725	Sc1ccc(N)cc1
726	Sc1ccc(Cl)cc1
727	Sc1ccc(Br)cc1
728	Nc1ccc(N)cc1
729	Nc1ccc(Cl)cc1
730	Nc1ccc(Br)cc1
731	Clc1ccc(Cl)cc1
732	Clc1ccc(Br)cc1
733	Brc1ccc(Br)cc1
734	Cc1cc(C)ccc1
735	Cc1cc(O)ccc1
736	Cc1cc(S)ccc1
737	Cc1cc(N)ccc1

738	Cc1cc(Cl)ccc1
739	Cc1cc(Br)ccc1
740	Oc1cc(O)ccc1
741	Oc1cc(S)ccc1
742	Oc1cc(N)ccc1
743	Oc1cc(Cl)ccc1
744	Oc1cc(Br)ccc1
745	Sc1cc(S)ccc1
746	Sc1cc(N)ccc1
747	Sc1cc(Cl)ccc1
748	Sc1cc(Br)ccc1
749	Nc1cc(N)ccc1
750	Nc1cc(Cl)ccc1
751	Nc1cc(Br)ccc1
752	Clc1cc(Cl)ccc1
753	Clc1cc(Br)ccc1
754	Brc1cc(Br)ccc1
755	Cc1c(C)cccc1
756	Cc1c(O)cccc1
757	Cc1c(S)cccc1
758	Cc1c(N)cccc1
759	Cc1c(Cl)cccc1
760	Cc1c(Br)cccc1
761	Oc1c(O)cccc1
762	Oc1c(S)cccc1
763	Oc1c(N)cccc1
764	Oc1c(Cl)cccc1
765	Oc1c(Br)cccc1
766	Sc1c(S)cccc1
767	Sc1c(N)cccc1
768	Sc1c(Cl)cccc1

769	Sc1c(Br)cccc1
770	Nc1c(N)cccc1
771	Nc1c(Cl)cccc1
772	Nc1c(Br)cccc1
773	Clc1c(Cl)cccc1
774	Clc1c(Br)cccc1
775	Brc1c(Br)cccc1
776	CC1CCC(C)CC1
777	CC1CCC(O)CC1
778	CC1CCC(S)CC1
779	CC1CCC(N)CC1
780	CC1CCC(Cl)CC1
781	CC1CCC(Br)CC1
782	OC1CCC(O)CC1
783	OC1CCC(S)CC1
784	OC1CCC(N)CC1
785	OC1CCC(Cl)CC1
786	OC1CCC(Br)CC1
787	SC1CCC(S)CC1
788	SC1CCC(N)CC1
789	SC1CCC(Cl)CC1
790	SC1CCC(Br)CC1
791	NC1CCC(N)CC1
792	NC1CCC(Cl)CC1
793	NC1CCC(Br)CC1
794	ClC1CCC(Cl)CC1
795	ClC1CCC(Br)CC1
796	BrC1CCC(Br)CC1
797	CC1CC(C)CCC1
798	CC1CC(O)CCC1
799	CC1CC(S)CCC1

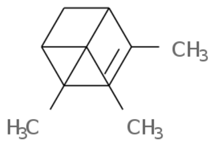
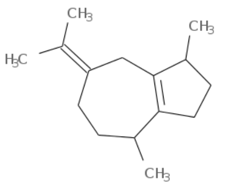
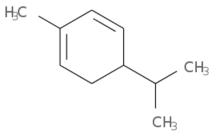
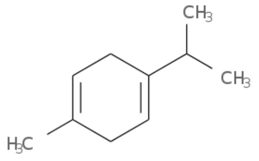
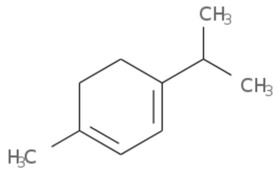
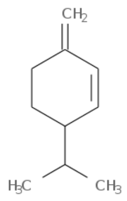
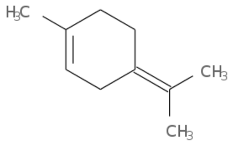
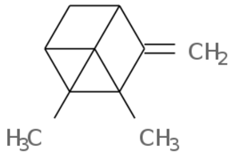
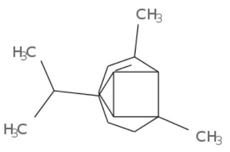
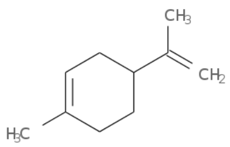
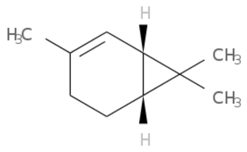
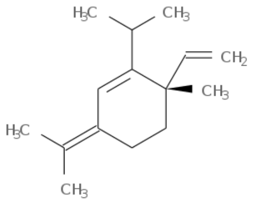
800	CC1CC(N)CCC1
801	CC1CC(Cl)CCC1
802	CC1CC(Br)CCC1
803	OC1CC(O)CCC1
804	OC1CC(S)CCC1
805	OC1CC(N)CCC1
806	OC1CC(Cl)CCC1
807	OC1CC(Br)CCC1
808	SC1CC(S)CCC1
809	SC1CC(N)CCC1
810	SC1CC(Cl)CCC1
811	SC1CC(Br)CCC1
812	NC1CC(N)CCC1
813	NC1CC(Cl)CCC1
814	NC1CC(Br)CCC1
815	ClC1CC(Cl)CCC1
816	ClC1CC(Br)CCC1
817	BrC1CC(Br)CCC1
818	CC1C(C)CCCC1
819	CC1C(O)CCCC1
820	CC1C(S)CCCC1
821	CC1C(N)CCCC1
822	CC1C(Cl)CCCC1
823	CC1C(Br)CCCC1
824	OC1C(O)CCCC1
825	OC1C(S)CCCC1
826	OC1C(N)CCCC1
827	OC1C(Cl)CCCC1
828	OC1C(Br)CCCC1
829	SC1C(S)CCCC1
830	SC1C(N)CCCC1

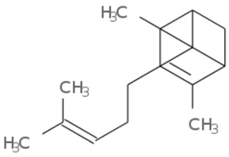
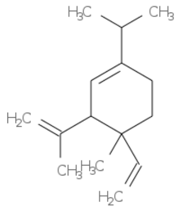
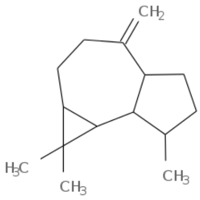
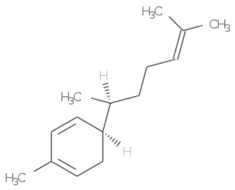
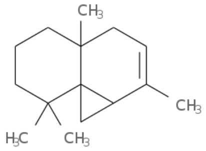
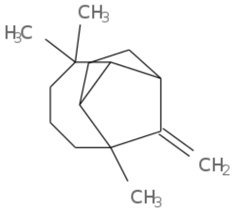
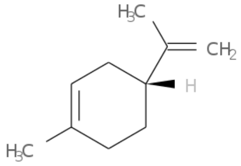
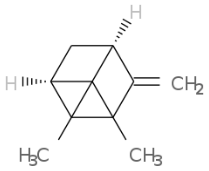
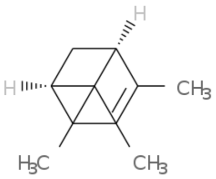
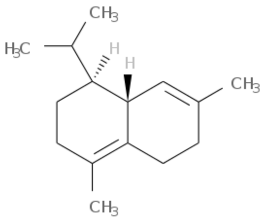
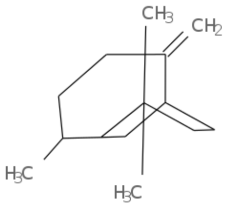
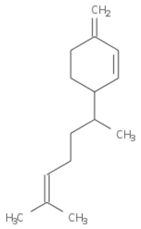
831	SC1C(Cl)CCCC1
832	SC1C(Br)CCCC1
833	NC1C(N)CCCC1
834	NC1C(Cl)CCCC1
835	NC1C(Br)CCCC1
836	ClC1C(Cl)CCCC1
837	ClC1C(Br)CCCC1
838	BrC1C(Br)CCCC1
839	CC1CC(C)CC1
840	CC1CC(O)CC1
841	CC1CC(S)CC1
842	CC1CC(N)CC1
843	CC1CC(Cl)CC1
844	CC1CC(Br)CC1
845	OC1CC(O)CC1
846	OC1CC(S)CC1
847	OC1CC(N)CC1
848	OC1CC(Cl)CC1
849	OC1CC(Br)CC1
850	SC1CC(S)CC1
851	SC1CC(N)CC1
852	SC1CC(Cl)CC1
853	SC1CC(Br)CC1
854	NC1CC(N)CC1
855	NC1CC(Cl)CC1
856	NC1CC(Br)CC1
857	ClC1CC(Cl)CC1
858	ClC1CC(Br)CC1
859	BrC1CC(Br)CC1
860	CC1C(C)CCC1
861	CC1C(O)CCC1

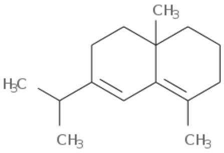
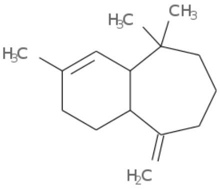
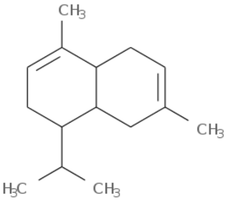
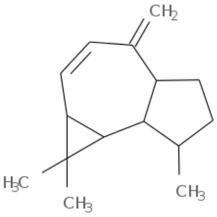
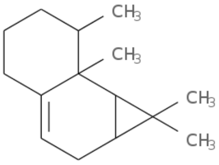
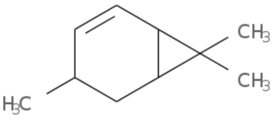
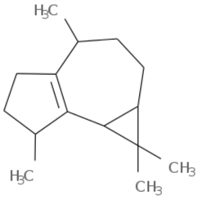
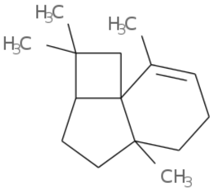
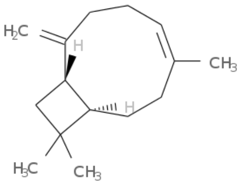
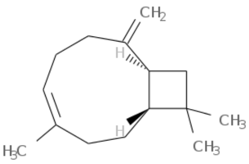
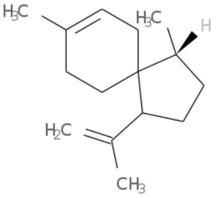
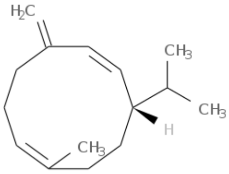
862	CC1C(S)CCC1
863	CC1C(N)CCC1
864	CC1C(Cl)CCC1
865	CC1C(Br)CCC1
866	OC1C(O)CCC1
867	OC1C(S)CCC1
868	OC1C(N)CCC1
869	OC1C(Cl)CCC1
870	OC1C(Br)CCC1
871	SC1C(S)CCC1
872	SC1C(N)CCC1
873	SC1C(Cl)CCC1
874	SC1C(Br)CCC1
875	NC1C(N)CCC1
876	NC1C(Cl)CC1
877	NC1C(Br)CCC1
878	ClC1C(Cl)CCC1
879	ClC1C(Br)CCC1
880	BrC1C(Br)CCC1

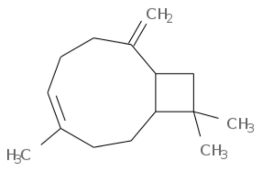
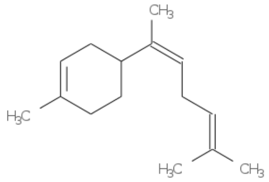
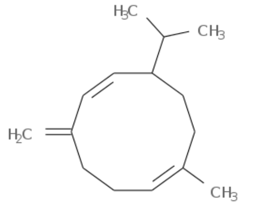
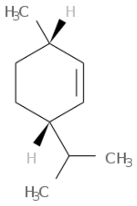
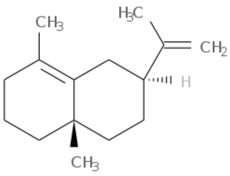
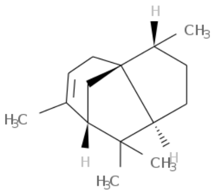
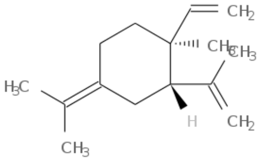
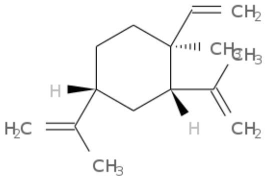
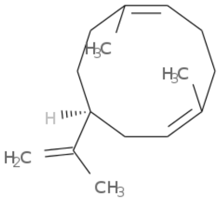
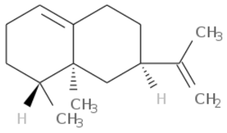
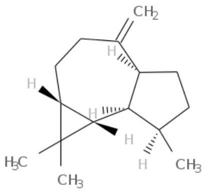
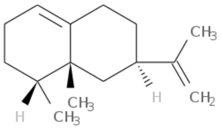
Appendix B

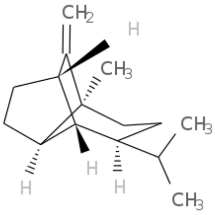
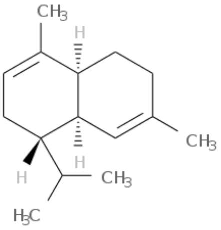
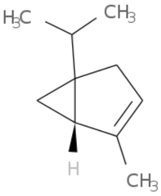
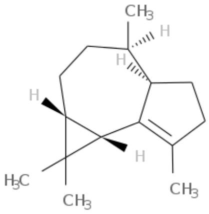
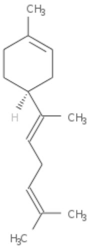
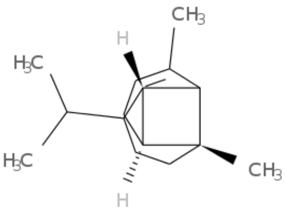
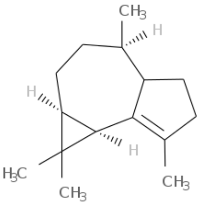
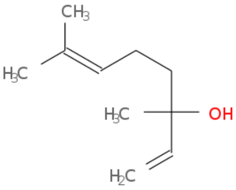
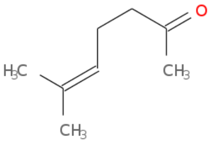
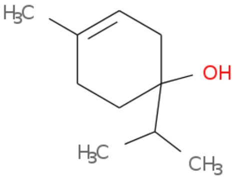
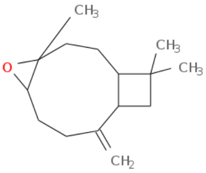
Chemical structures of the VOCs belonging to all clusters (Cluster 1 to Cluster 11), as mentioned in Chapter 4.

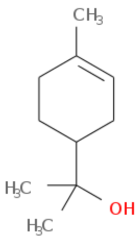
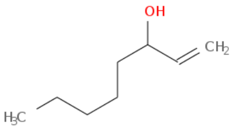
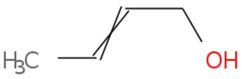
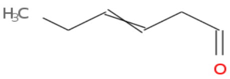
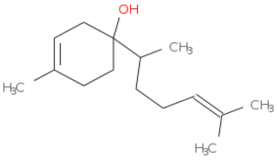
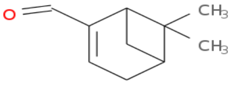
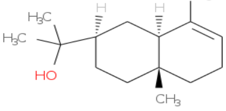
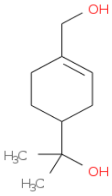
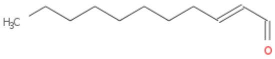
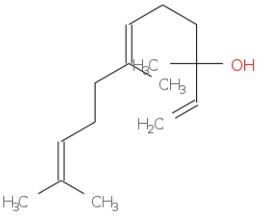
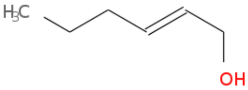
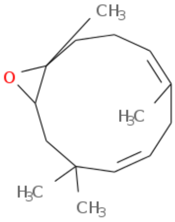
Cluster 1 (55 VOCs)			
 C00000805 Alpha-Pinene	 C00020376 beta-Guaiene	 C00003051 alpha-Phellandrene	 gamma-Terpinen
C00003060  alpha-Terpinene	 C00010872 abeta-Phellandrene	 Terpinolene	 C00000816 beta-Pinene
 C00003118 Copaene	 C00010868 d-Limonene	 (+)-2-Carene	 a-Elemene

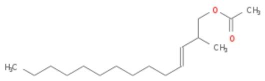
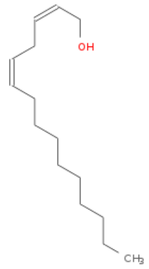
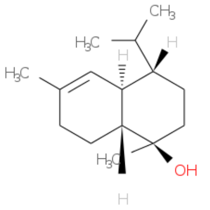
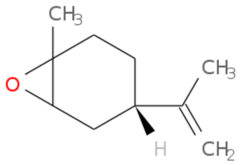
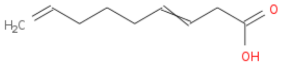
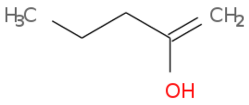
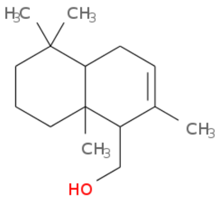
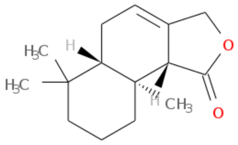
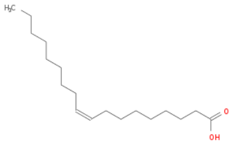
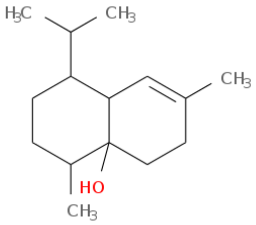
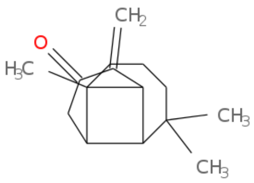
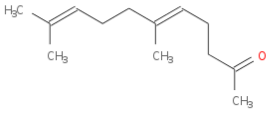
 (E)-alpha-Bergamotene	 C00012011 delta-Elemene	 C00021230 Aromadendrene	 C00003204 alpha-Zingiberene
 C00003194 Thujopsene	 C00003162 Longifolene	 (+)-Limonene	 C00000806 (-)-beta-Pinene
 1-alpha-Pinene	 C00007636 delta-Cadinene	 C00021999 Seychellene	 Sesquiphellandrene

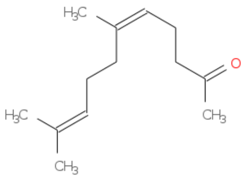
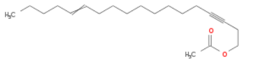
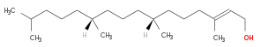
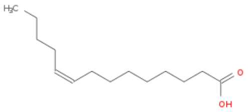
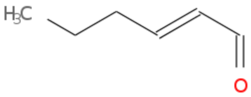
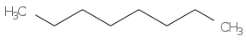
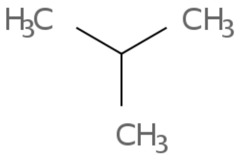
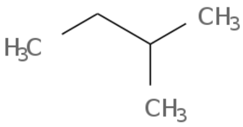
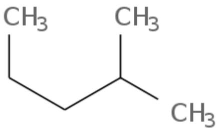
 C00007634 delta-Selinene	 C00021309 alpha-Himachalene	 Cadinene	 Dehydroaromadendrene
 C00017471 Aristolene	 (+)-4-Carene	 C00037332 Isoledene	 alpha-Panasinsen
 C00003110 beta-Caryophyllene	 C00012474 Isocaryophyllene	 C00021580 beta-Acoradiene	 (-)-Germacrene D

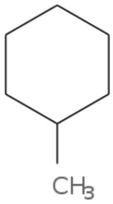
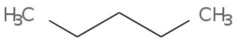
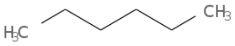
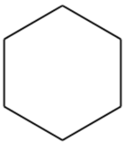
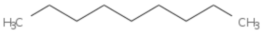
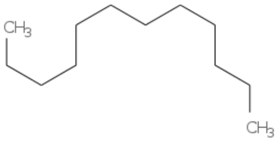
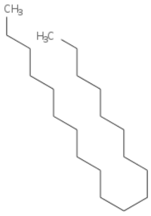
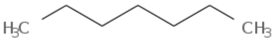
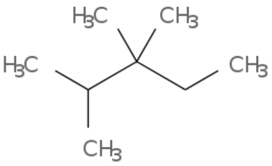
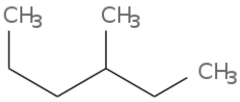
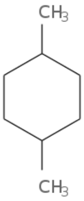
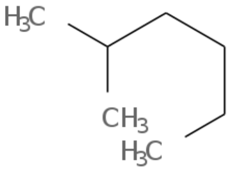
 C00003110 Caryophyllene	 alpha-Bisabolene	 C00011720 Caryophyllene	 2-Menthene
 4,11-Selinadiene	 C00003111 alpha-Cedrene	 C00012012 gamma-Elemene	 C00007453 beta-Elemene
 C00011719 Germacrene A	 C00034741 Valencene	 C00021229 Alloaromadendrene	 C00016975 Eremophilene

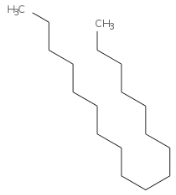
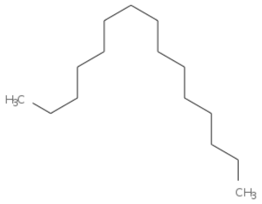
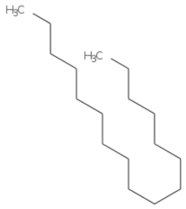
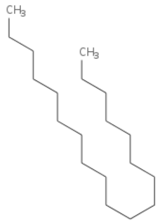
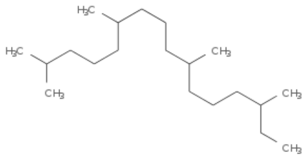
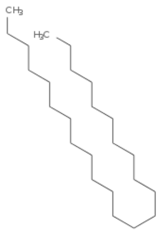
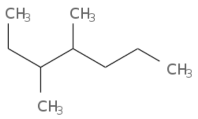
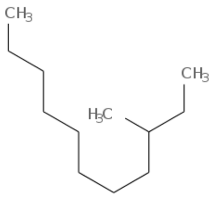
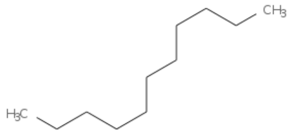
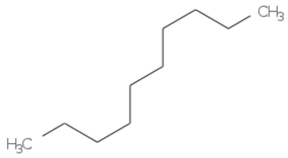
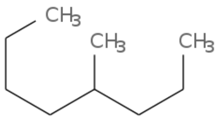
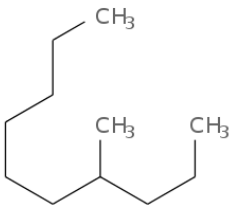
 (+)-Sativene	 C00029671 alpha-Muurolene	 C00000184 alpha-Thujene	 C00021227 alpha-Gurjunene
 C00050259 Trans-alpha-bisabolene	 C00003118 alpha-Copaene	 Gurjunene	
Cluster 2 (33 VOCs)			
 beta-Linalool	 C00034775 6-Methyl-5-hepten-2-one	 C00029544 Terpinen-4-ol	 C00012483 Caryophyllene oxide

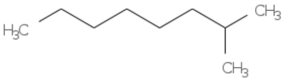
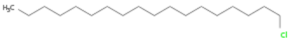
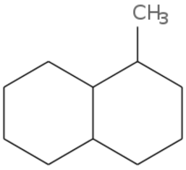
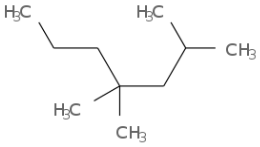
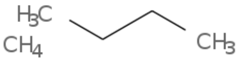
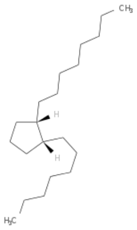
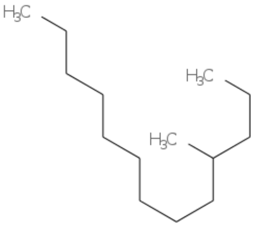
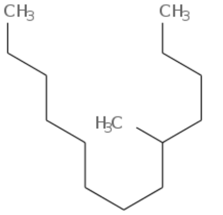
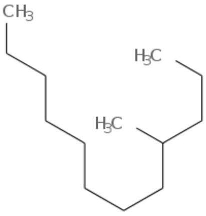
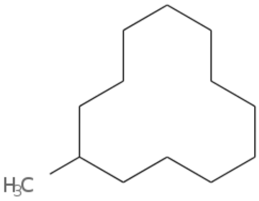
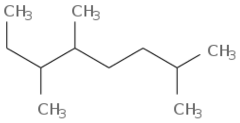
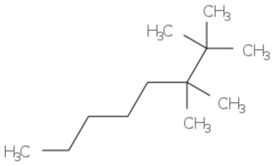
 p-Menth-1-en-8-ol	 C00029423 1-Octen-3-ol	 2-Buten-1-ol	 C00048948 3-Hexenal
 beta-Bisobolol	 C00030803 Myrtenal	 C00000163 alpha-Eudesmol	 p-Menth-1-en-8-ol
 2-Undecenal	 C00003166 Nerolidol	 2-Hexenol	 C00012443 Humulene epoxide

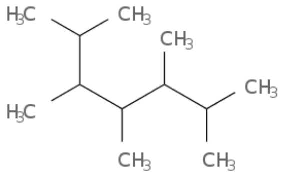
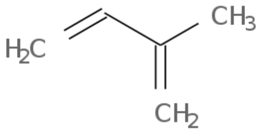
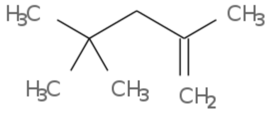
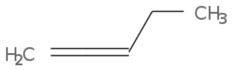
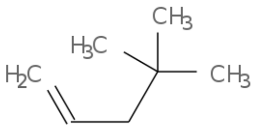
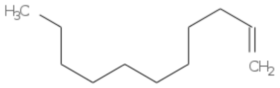
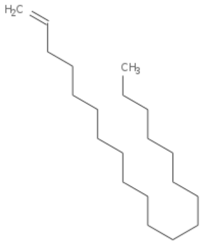
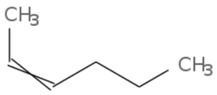
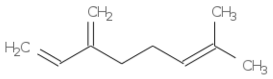
 E-2-Methyl-3-tetradecen-1-ol acetate	 Z,Z-2,5-Pentadecadien-1-ol	 C00020065 alpha-Cadinol	 C00035852 Limonene oxide
 (2E,6E)-2,6-Nonadienoic acid	 1-Penten-2-on	 C00020282 Drimenol	 C00003252 Drimenin
 C00001232 Oleic acid	 C00033734 Cubenol	 Longipinocarvone	 Geranylacetone

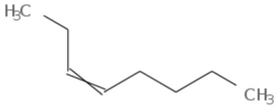
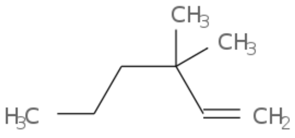
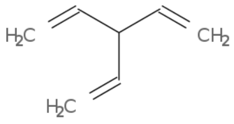
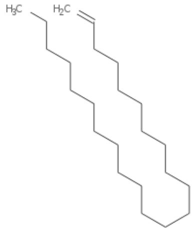
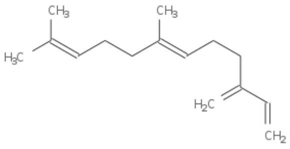
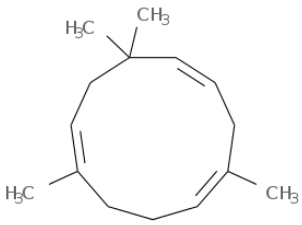
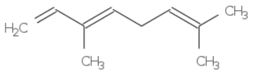
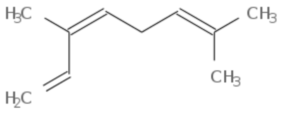
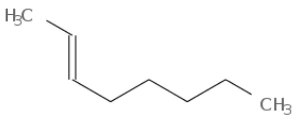
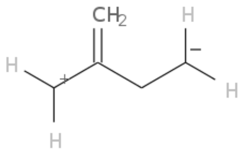
 C00029336 (E)-Geranyl acetone	 Z-3-Octadecen-1-ol acetate	 C00003467 Phytol	 C00001229 Myristoleic acid
 (E)-2-Hexenal			
Cluster 3 (41 VOCs)			
 C00035857 Octane	 Isobutane	 2-Methyl-butane	 2-Methylpentane

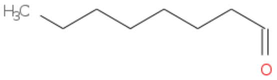
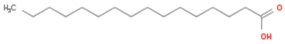
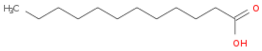
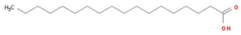
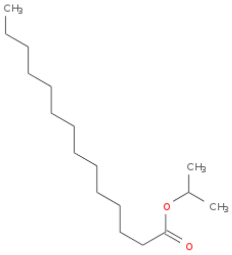
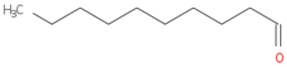
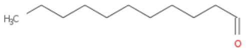
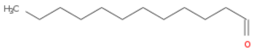
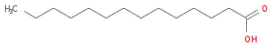
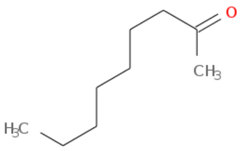
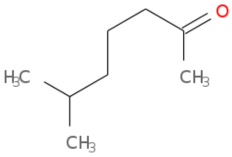
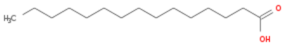
 C00035853 Methylcyclohexane	 Pentane	 Hexane	 C00007453 Cyclohexane
 C00034882 Nonane	 C00001248 Dodecane	 C00030165 Eicosane	 Heptane
 2,3,3-Trimethylpentane	 C00050708 3-Methylhexane	 1,4-Dimethylcyclohexane	 C00035484 2-Methylhexane

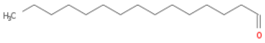
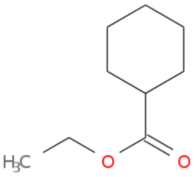
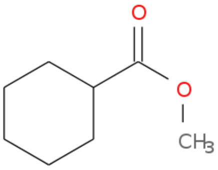
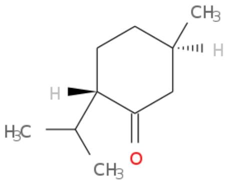
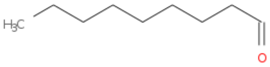
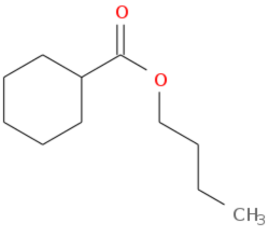
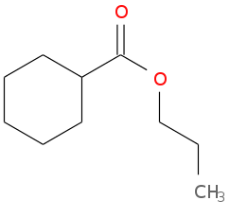
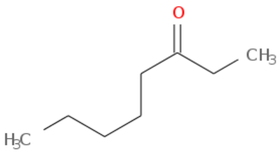
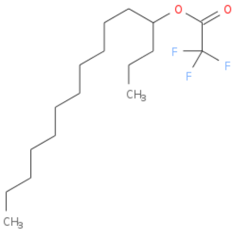
 C00030879 Octadecane	 C00001265 Pentadecane	 C00030472 Heptadecane	 C00030827 Nonadecane
 Phytane	 C00032307 Tetracosane	 2,4-Dimethylheptane	 3-Methylundecane
 C00032443 Undecane	 C00048375 Decane	 4-Methyloctane	 4-Methyl decane

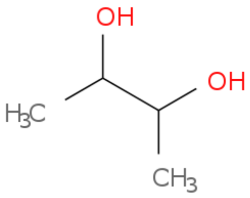
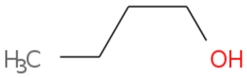
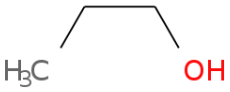
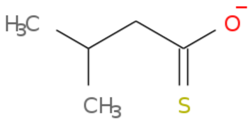
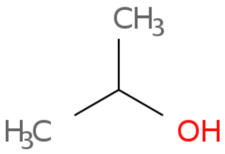
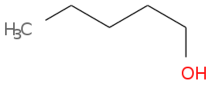
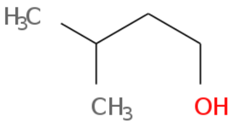
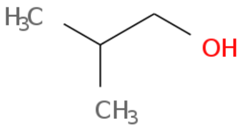
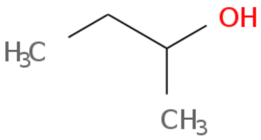
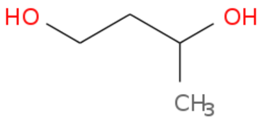
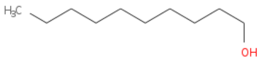
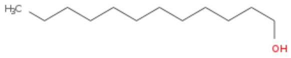
 2-Methyloctane	 1-Chlorooctadecane	 Trans-anti-1-methyl- decahydronaphthalene	 2,2,4-Trimethylheptane
 Alkanes	 8-Isoprostane	 4-Methyltridecane	 5-Methyltridecane
 4-Methyldodecane	 Methylcyclododecane	 2,5,6-Trimethyloctane	 Pentamethylheptane

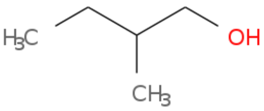
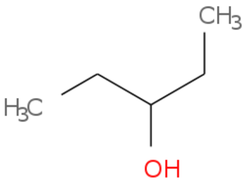
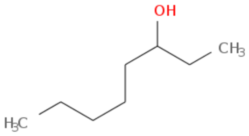
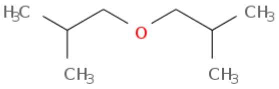
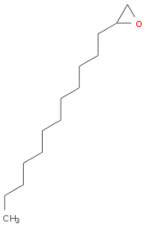
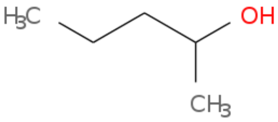
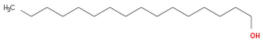
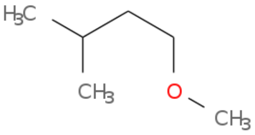
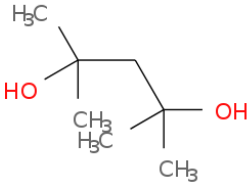
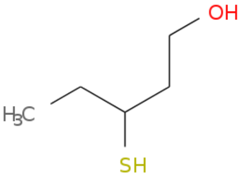
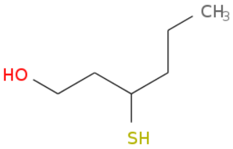
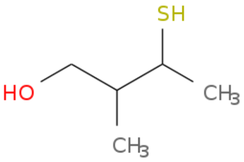
 2,3,4,5,6-Pentamethylheptane			
Cluster 4 (18 VOCs)			
 C00046784 Isoprene	 2,4,4-Trimethyl-1-pentene	 ,2-Pentadiene	 4,4-Dimethyl-1-pentene
 1-Undecene	 1-Eicosene	 2-Hexene	 C00000853 Myrcene

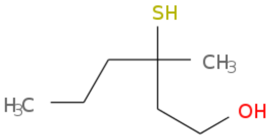
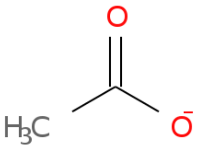
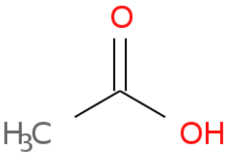
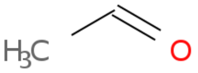
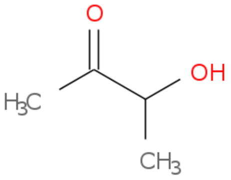
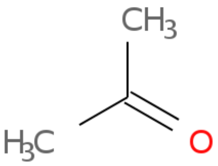
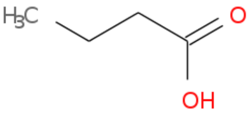
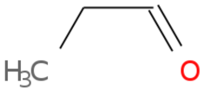
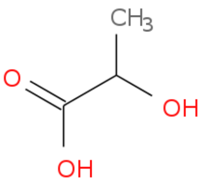
 3-Octene	 3,3-Dimethylhex-1-ene	 1,4-Pentadiene	 Tricosene
 C00003131 beta-Farnesene	 C00003147 alpha-Caryophyllene	 Ocimene	 C00029335 beta-Ocimene
 2-Octene	 2-Methylbut-1-ene		

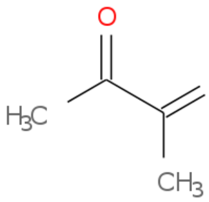
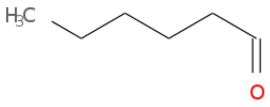
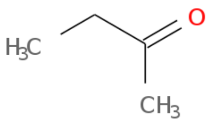
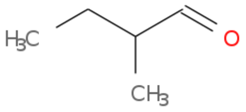
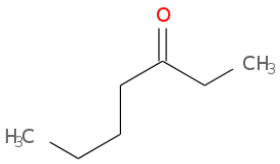
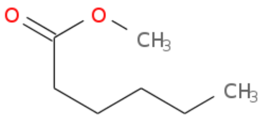
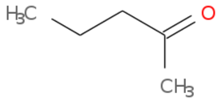
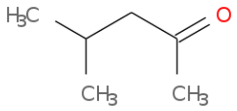
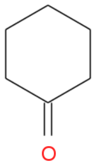
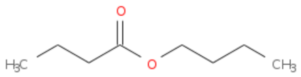
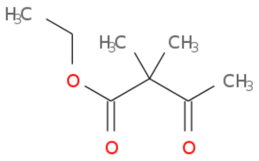
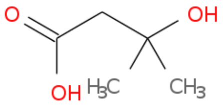
Cluster 5 (21 VOCs)			
 C00030880 Octanal	 Hexadecanoic acid	 Dodecanoic acid	 C00001238 Octadecanoic acid
 Isopropyl myristate	 C00030099 Decanal	 C00032442 Undecanal	 Dodecanal
 C00001228 Tetradecanoic acid	 C00029463 2-Nonanone	 6-Methylheptan-2-one	 C00007423 Pentadecanoic acid

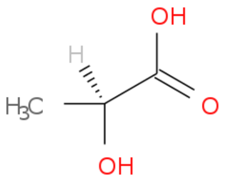
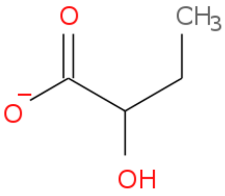
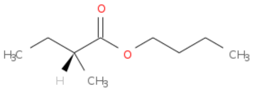
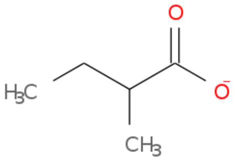
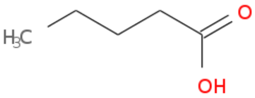
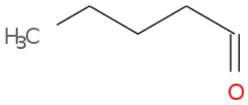
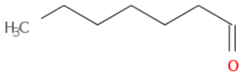
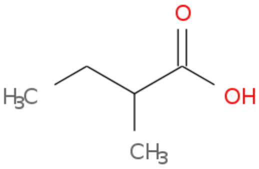
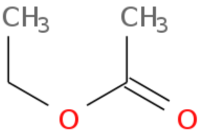
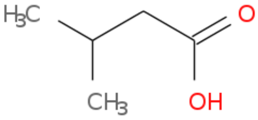
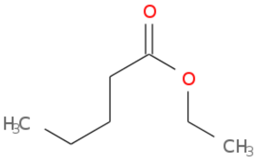
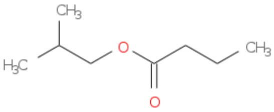
 C00030959 Pentadecanal	 Ethyl cyclohexanecarboxylate	 Methyl cyclohexanecarboxylate	 p-Menthone
 C00030828 Nonanal	 Butyl cyclohexanecarboxylate	 Propyl cyclohexanecarboxylate	 Octan-3-one
 4-Trifluoroacetoxypentadecane			

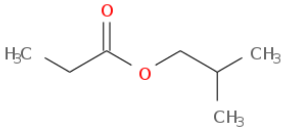
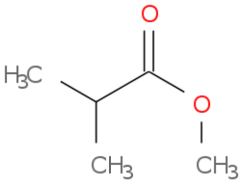
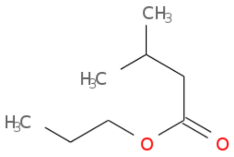
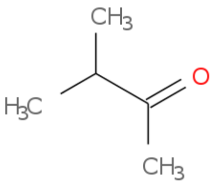
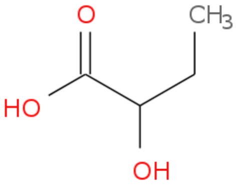
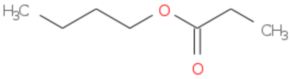
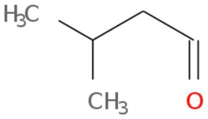
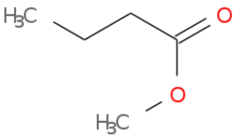
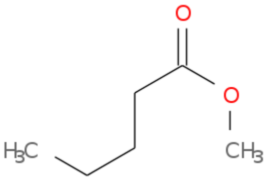
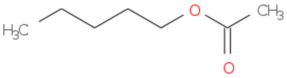
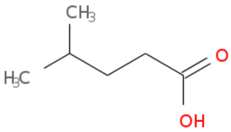
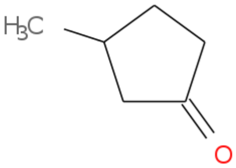
Cluster 6 (25 VOCs)			
 C00050411 2,3-Butanediol	 1-Butanol	 Propanol	 S-Methyl 3-methylbutanethioate
 2-Propanol	 Pentanol	 3-Methyl-1-butanol	 2-Methyl-1-propanol
 2-Butanol	 1, 3-butanediol	 C00030100 Decanol	 C00030152 1-Dodecanol

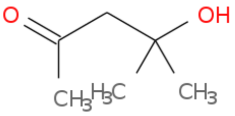
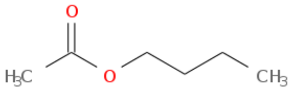
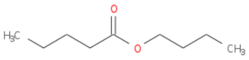
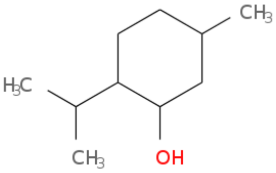
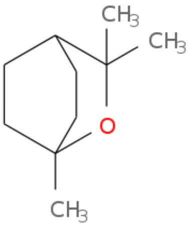
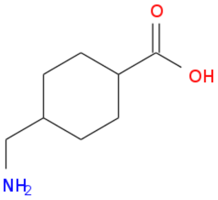
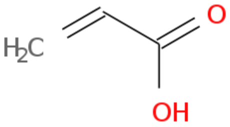
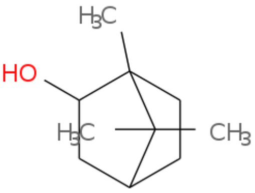
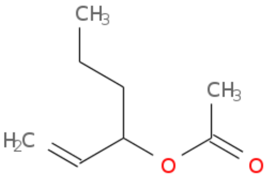
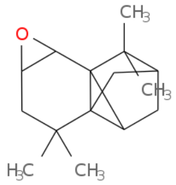
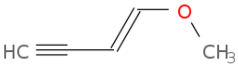
 C00050415 2-Methyl-1-butanol	 3-Pentanol	 C00035495 3-Octanol	 Isobutylether
 2-Dodecyloxirane	 2-Pentanol	 1-Hexadecanol	 Butane-1-methoxy-3-methyl
 2,4-Dimethylpentane	 3-Sulfanyl-pentan-1-ol	 3-Sulfanylhexan-1-ol	 2-Methyl-3-sulfanylbutan-1-ol

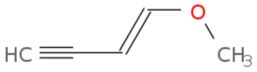
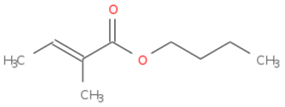
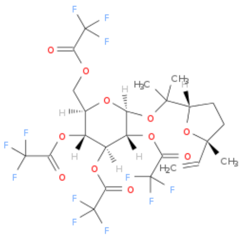
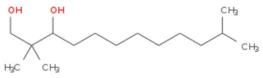
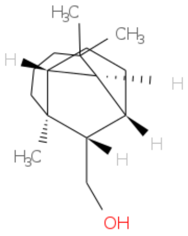
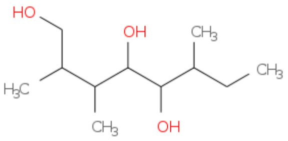
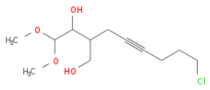
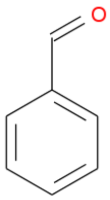
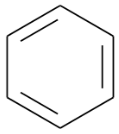
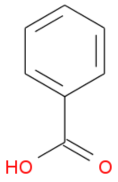
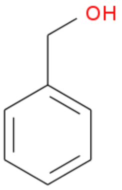
 3-Methyl-3-sulfanylhexan-1-ol			
Cluster 7 (47 VOCs)			
 Acetate	 Acetic acid	 C00007392 Acetaldehyde	 C00050424 Acetoin
 C00048304 Acetone	 Butanoic acid	 C00050492 Propanal	 Propanoic acid

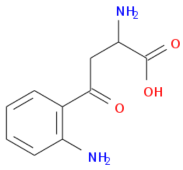
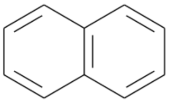
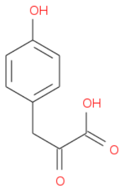
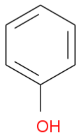
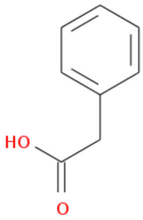
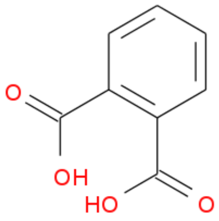
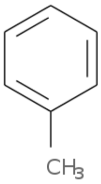
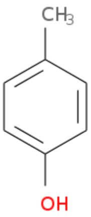
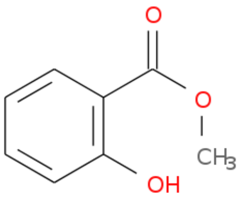
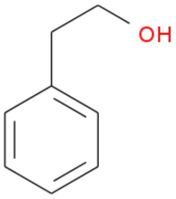
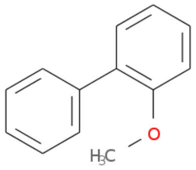
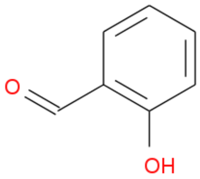
 C00050437 2,3-Butanedione	 C00000357 Hexanal	 2-Butanone	 2-Methyl-butanal
 3-Heptanone	 Methyl hexanoate	 2‑Pentanone	 C00051562 Methyl isobutyl ketone
 Cyclohexanone	 1-Chloro-3-methylbutane	 Ethyl2,2-dimethyl-3-oxobutanoate	 3-Hydroxyisovaleric acid

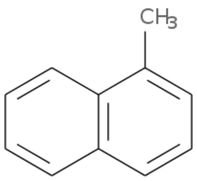
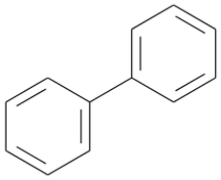
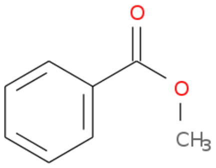
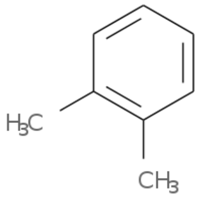
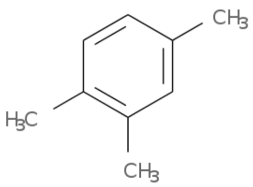
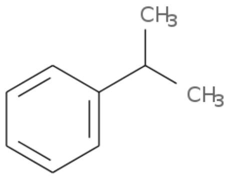
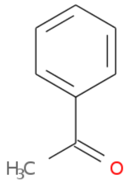
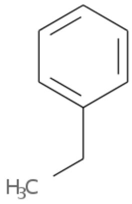
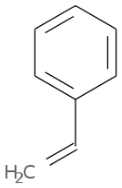
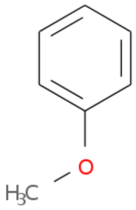
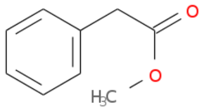
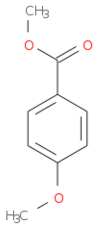
 C00050478 Lactic acid	 2-Hydroxy-butyrate	 Butyl (2R)-2-methylbutanoate	 2-Methylbutanoate
 Pentanoic acid	 Pentanal	 Heptanal	 C00029461 2-Methylbutanoic acid
 C00001308 Ethyl acetate	 C00001189 3-Methylbutanoic acid	 Ethyl valerate	 2-Methylpropyl butanoate

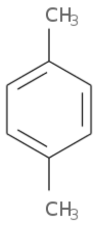
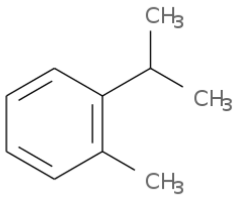
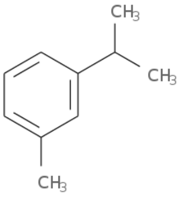
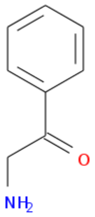
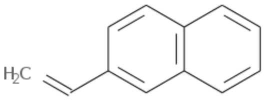
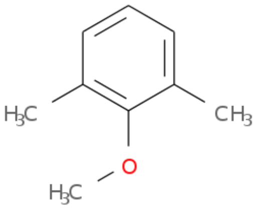
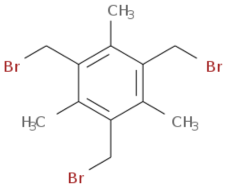
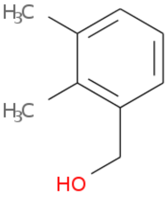
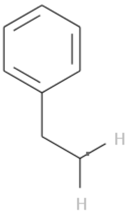
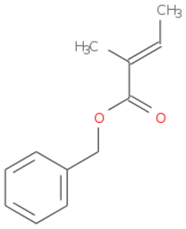
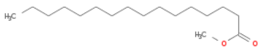
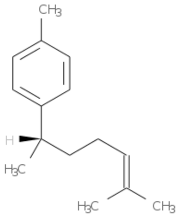
 2-Methylpropyl propanoate	 Methyl 2-methylpropanoate	 Propyl 3-methylbutanoate	 3-Methyl-butan-2-one
 alpha-Hydroxybutyric acid	 Butyl propanoate	 C00048949 3-Methyl-butanal	 Methyl butanoate
 Methyl pentanoate	 Pentyl acetate	 4-Methylpentanoic acid	 3-Methyl-cyclopentanone

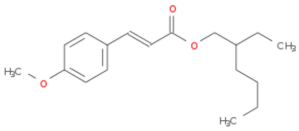
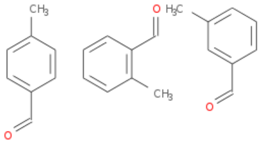
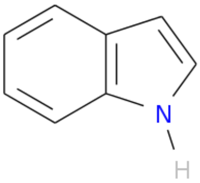
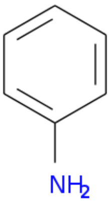
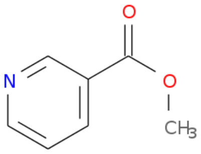
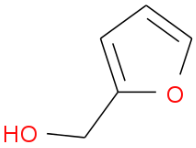
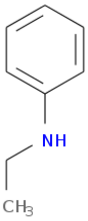
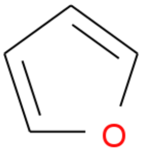
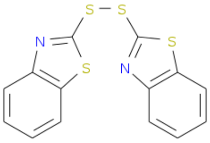
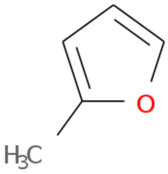
 4-Hydroxy-4-methyl-2-pentanone	 C00048958 Butyl acetate	 Butyl pentanoate	
Cluster 8 (15 VOCs)			
 5-Methyl-2-(1-methylethyl)-cyclohexanol	 C00000136 1,8-Cineole	 Cyclohexanecarboxylic acid	 C00034423 2-Propenoic acid
 C00003028 Borneol	 3-Hexenyl acetate	 Cyclolongifolene oxide	 C00030482 Hexyl tiglate

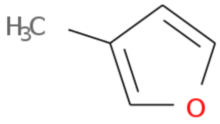
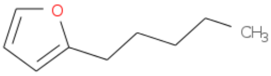
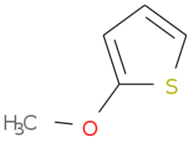
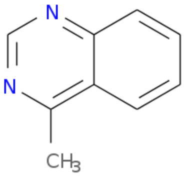
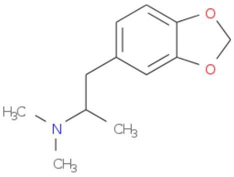
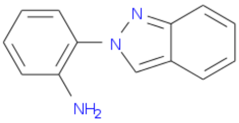
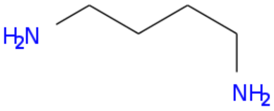
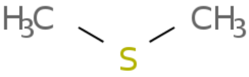
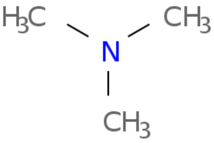
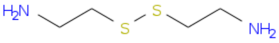
 1-Methoxy-1-buten-3-yne	 Butyl tiglate	 Linalool oxide (furanoid)	 C00011411 2,6,10-Trimethyldodecane
 Isolongifolol	 2,3,6-Trimethyloctane	 1,4-Dimethoxy-2,3-butanediol	
Cluster 9 (42 VOCs)			
 C00034452 Benzaldehyde	 Benzene	 C00000207 Benzoic acid	 C00029811 Benzyl alcohol

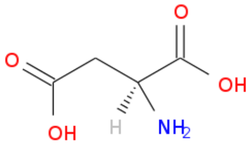
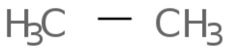
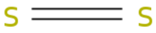
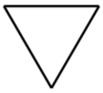
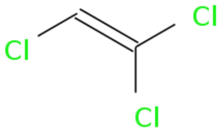
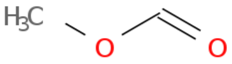
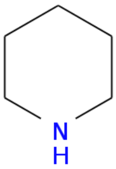
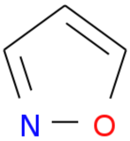
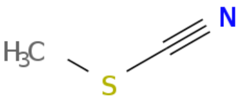
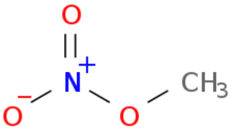
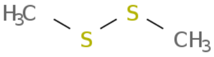
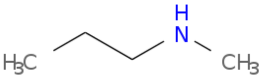
 C00051181 Kynurenine	 C00001259 Naphthalene	 C00007512 p-Hydroxyphenylpyruvic acid	 C00002664 Phenol
 C00000750 Phenylacetic acid	 Phthalic acid	 Toluene	 C00002645 4-Methylphenol
 C00030767 Methyl salicylate	 C00002663 2-Phenylethanol	 o-Phenylanisole	 C00031273 Salicylaldehyde

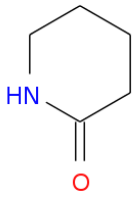
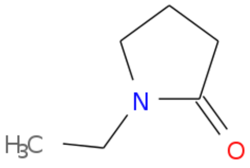
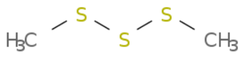
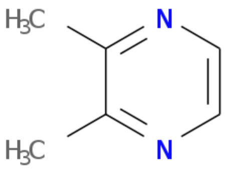
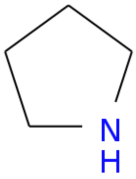
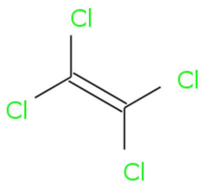
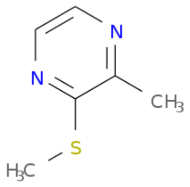
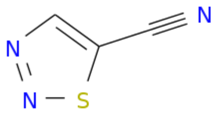
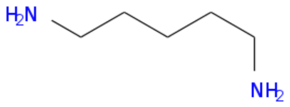
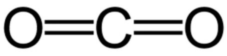
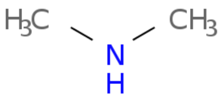
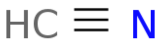
 C00050647 1-Methylnaphthalene	 Biphenyl	 C00034054 Methyl benzoate	 C00035778 1,3-Dimethylbenzene
 Trimethylbenzene	 3-(1-methylethyl)-benzene	 1-Phenyl-ethanone	 Ethylbenzene
 C00037855 Styrene	 Anisole	 C00051576 Methyl phenylacetate	 C00030761 Methyl p-anisate

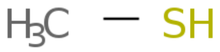
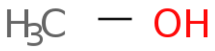
 1,4-Xylene	 1-Methyl-2-(1-methylethyl)-benzene	 beta-Cymene	 2-Aminoacetophenone
 2-Ethenylnaphtalene	 2-Methoxy-1,3-dimethylbenzene	 1,2,4-Trimethylbenzene	 Benzenemethanol
 2-Phenylethyl alcohol	 C00048327 Benzyl tiglate	 C00030755 Methyl hexadecanoate	 C00046607 alpha-Curcumene

 2-Ethylhexyl-4-methoxy-cinnamate	 Tolualdehyde		
Cluster 10 (14 VOCs)			
 C00001418 Indole	 Aniline	 C00044235 Methyl nicotinate	 Furanmethanol
 N-Ethylaniline	 Furan	 Benzothiazole	 C00048276 2-Methylfuran

 3-Methylfuran	 C00034763 2-Pentylfuran	 2-Methoxytiophene	 4-Methyl-quinazo- line
 3,5-Dimethylamphetamine	 2-Indazol-2-ylphenylamine		
Cluster 11 (30 VOCs)			
 C00001428 Putrescine	 Benzene	 C00001433 Trimethylamine	 Cystamine

 Aspartic acid	 Ethane	 Carbon disulfide	 Cyclopropane
 Trichloroethylene	 Methyl formate	 Piperidine	 Isoxazole
 Methyl thiocyanide	 Methyl nitrate	 C00001245 Dimethyl disulfide	 N-Methyl-2-methylpropylamine

 2-Piperidinone	 1-Ethyl-2-pyrrolidinone	 C00001246 Dimethyl trisulfide	 2,3-Dimethyl-pyrazine
 Pyrrolidine	 Tetrachloroethyene	 2-Methyl-pyrazine	 5-Cyano-1,2,3-thiadiazole
 C00001403 Cadaverine	 Carbon dioxide	 C00050440 Dimethylamine	 Hydrogen cyanide

 Methylmercaptan	 C00050480 Methanol		

Appendix C

The performance of all classifiers using all datasets as training and 10-fold cross validation technique, which were mentioned in Chapter 4.

Table 1. *The performance of classifiers (mse) using all datasets as training.*

Model No	Fingerprint + Machine Learning Method	MSE
1	Combine + DNN1 (default)	0.1052596
2	Combine + DNN2	0.107273
3	Combine + DNN3	0.2447039
4	Combine + DNN4	0.5051482
5	Combine + DNN5	0.1619378
6	Combine + RF	0.4212882
7	Combine + GBM	0.3952804
8	Combine + GLM	0.4319557
9	KR + DNN1 (default)	0.1582859
10	KR + DNN2	0.141156
11	KR + DNN3	0.1656269
12	**KR + DNN4**	**0.05420784**
13	KR + DNN5	0.08050456
14	KR + RF	0.4104758
15	KR + GBM	0.1267871
16	KR + GLM	0.3484839
17	PubChem + DNN1 (default)	0.1775599
18	PubChem + DNN2	0.1265089
19	PubChem + DNN3	0.1768494
20	**PubChem + DNN4**	**0.05871162**
21	PubChem + DNN5	0.08168069
22	PubChem + RF	0.4074268
23	PubChem + GBM	0.1214795
24	PubChem + GLM	0.3679084

25	CDK + DNN1 (default)	0.2230534
26	CDK + DNN2	0.220504
27	CDK + DNN3	0.2025835
28	CDK + DNN4	0.1089344
29	CDK + DNN5	0.1698504
30	CDK + RF	0.4555405
31	CDK + GBM	0.1498589
32	CDK + GLM	0.3724793
33	Extended + DNN1	0.2230534
34	Extended + DNN2	0.220504
35	Extended + DNN3	0.2025835
36	Extended + DNN4	0.1089344
37	Extended + DNN5	0.505119
38	Extended + RF	0.4555405
39	Extended + GBM	0.1498589
40	Extended + GLM	0.3724793
41	AP + DNN1 (default)	0.4246078
42	AP + DNN2	0.4150098
43	AP + DNN3	0.4606583
44	AP + DNN4	0.3413434
45	AP + DNN5	0.3742727
46	AP + RF	0.494816
47	AP + GBM	0.3831628
48	AP + GLM	0.4787031
49	Sub + DNN1 (default)	0.3793974
50	Sub + DNN2	0.3491443
51	Sub + DNN3	0.3132125
52	Sub + DNN4	0.2179004
53	Sub + DNN5	0.2655871
54	Sub + RF	0.4480489
55	Sub + GBM	0.4497772
56	Sub + GLM	0.450233

57	Estate + DNN1 (default)	0.3621
58	Estate + DNN2	0.3775214
59	Estate + DNN3	0.4585305
60	Estate + DNN4	0.2531738
61	Estate + DNN5	0.3149532
62	Estate + RF	0.4561216
63	Estate + GBM	0.2974358
64	Estate + GLM	0.4667689
65	MACCS + DNN1 (default)	0.2418533
66	MACCS + DNN2	0.235304
67	MACCS + DNN3	0.1849032
68	**MACCS + DNN4**	**0.07807859**
69	MACCS + DNN5	0.5128929
70	MACCS + RF	0.4293163
71	MACCS + GBM	0.39979038
72	MACCS + GLM	0.3989986

*red color indicates the best result

Table 2. *The performance of classifiers (mse) using 10-fold cross-validation technique.*

Model No	Fingerprint + Machine Learning Method	MSE
1	Combine + DNN1 (default)	0.48404413
2	Combine + DNN2	0.49436516
3	Combine + DNN3	0.5023598
4	Combine + DNN4	0.50643396
5	Combine + DNN5	0.4514478
6	Combine + RF	0.42321494
7	**Combine + GBM**	**0.39837325**
8	Combine + GLM	0.4323252
9	KR + DNN1 (default)	0.48404413
10	KR + DNN2	0.48621503
11	KR + DNN3	0.46792474
12	KR + DNN4	0.5382593
13	KR + DNN5	0.50003135
14	KR + RF	0.4173862
15	KR + GBM	0.4144334
16	KR + GLM	0.43971023
17	PubChem + DNN1 (default)	0.4472546
18	PubChem + DNN2	0.52192795
19	PubChem + DNN3	0.46042094
20	PubChem + DNN4	0.5767418
21	PubChem + DNN5	0.4764082
22	PubChem + RF	0.40837318
23	**PubChem + GBM**	**0.39318013**
24	PubChem + GLM	0.4595151
25	CDK + DNN1 (default)	0.49185374
26	CDK + DNN2	0.54940474
27	CDK + DNN3	0.49814865

28	CDK + DNN4	0.5754862
29	CDK + DNN5	0.509189
30	CDK + RF	0.4635254
31	CDK + GBM	0.43289793
32	CDK + GLM	0.4731141
33	Extended + DNN1	0.4707817
34	Extended + DNN2	0.51699466
35	Extended + DNN3	0.51963675
36	Extended + DNN4	0.64318633
37	Extended + DNN5	0.49146965
38	Extended + RF	0.4361141
39	Extended + GBM	0.41715953
40	Extended + GLM	0.4461066
41	AP + DNN1 (default)	0.5482254
42	AP + DNN2	0.5460825
43	AP + DNN3	0.57297605
44	AP + DNN4	0.5734207
45	AP + DNN5	0.52786994
46	AP + RF	0.4963377
47	AP + GBM	0.49641743
48	AP + GLM	0.510474
49	Sub + DNN1 (default)	0.54112405
50	Sub + DNN2	0.49989194
51	Sub + DNN3	0.5515154
52	Sub + DNN4	0.53961086
53	Sub + DNN5	0.50628924
54	Sub + RF	0.4541103
55	Sub + GBM	0.4492355
56	Sub + GLM	0.48457363
57	Estate + DNN1 (default)	0.52493817
58	Estate + DNN2	0.4919421
59	Estate + DNN3	0.58307207

60	Estate + DNN4	0.5230969
61	Estate + DNN5	0.47620323
62	Estate + RF	0.45589486
63	Estate + GBM	0.45273414
64	Estate + GLM	0.49563873
65	MACCS + DNN1 (default)	0.5004864
66	MACCS + DNN2	0.485334
67	MACCS + DNN3	0.501069
68	MACCS + DNN4	0.5599283
69	MACCS + DNN5	0.510386
70	MACCS + RF	0.43273932
71	**MACCS + GBM**	**0.39979038**
72	MACCS + GLM	0.47322118

*red color indicates the best result

Table 3. *The performance of classifiers (accuracy) using all datasets as training.*

Model No	Fingerprint + Machine Learning Method	Accuracy (%)
1	Combine + DNN1 (default)	87.39002933
2	Combine + DNN2	87.68328446
3	Combine + DNN3	74.78005865
4	**Combine + DNN4**	**91.49560117**
5	Combine + DNN5	83.28445748
6	Combine + RF	57.771261
7	**Combine + GBM**	**94.42815249**
8	Combine + GLM	76.83284457
9	KR + DNN1 (default)	80.64516129
10	KR + DNN2	81.81818182
11	KR + DNN3	81.81818182
12	**KR + DNN4**	**92.08211144**
13	KR + DNN5	91.20234604
14	KR + RF	54.25219941
15	KR + GBM	88.56304985
16	KR + GLM	70.08797654
17	PubChem + DNN1 (default)	80.93841642
18	PubChem + DNN2	81.81818182
19	PubChem + DNN3	79.76539589
20	PubChem + DNN4	91.20234604
21	PubChem + DNN5	90.32258065
22	PubChem + RF	55.42521994
23	PubChem + GBM	88.85630499
24	PubChem + GLM	65.98240469
25	CDK + DNN1 (default)	74.19354839
26	CDK + DNN2	71.84750733
27	CDK + DNN3	76.83284457
28	CDK + DNN4	85.04398827
29	CDK + DNN5	90.32258065

30	CDK + RF	57.771261
31	CDK + GBM	83.87096774
32	CDK + GLM	66.27565982
33	Extended + DNN1	81.52492669
34	Extended + DNN2	74.48680352
35	Extended + DNN3	70.96774194
36	Extended + DNN4	86.51026393
37	Extended + DNN5	83.28445748
38	Extended + RF	52.78592375
39	Extended + GBM	86.2170088
40	Extended + GLM	68.62170088
41	AP + DNN1 (default)	52.19941349
42	AP + DNN2	53.07917889
43	AP + DNN3	50.14662757
44	AP + DNN4	59.53079179
45	AP + DNN5	56.8914956
46	AP + RF	49.5601173
47	AP + GBM	59.53079179
48	AP + GLM	52.78592375
49	Sub + DNN1 (default)	60.11730205
50	Sub + DNN2	61.58357771
51	Sub + DNN3	65.39589443
52	Sub + DNN4	73.90029326
53	Sub + DNN5	68.32844575
54	Sub + RF	51.61290323
55	Sub + GBM	68.32844575
56	Sub + GLM	58.94428152
57	Estate + DNN1 (default)	58.65102639
58	Estate + DNN2	58.94428152
59	Estate + DNN3	46.33431085
60	Estate + DNN4	68.32844575
61	Estate + DNN5	65.1026393

62	Estate + RF	53.07917889
63	Estate + GBM	66.86217009
64	Estate + GLM	55.13196481
65	MACCS + DNN1 (default)	71.84750733
66	MACCS + DNN2	73.31378299
67	MACCS + DNN3	77.12609971
68	MACCS + DNN4	88.56304985
69	MACCS + DNN5	87.39002933
70	MACCS + RF	53.07917889
71	MACCS + GBM	87.09677419
72	MACCS + GLM	60.70381232

*red color indicates the best result

Table 4. *The performance of classifiers (accuracy) using 10-fold cross-validation technique.*

Model No	Fingerprint + Machine Learning Method	Accuracy (%)
1	Combine + DNN1 (default)	48.920146
2	Combine + DNN2	46.910718
3	Combine + DNN3	47.205934
4	Combine + DNN4	44.44432
5	Combine + DNN5	53.690916
6	**Combine + RF**	**57.95232**
7	**Combine + GBM**	**57.67722**
8	**Combine + GLM**	**58.666307**
9	KR + DNN1 (default)	52.468336
10	KR + DNN2	47.317985
11	KR + DNN3	50.685245
12	KR + DNN4	40.425983
13	KR + DNN5	48.402813
14	KR + RF	56.734097
15	KR + GBM	53.760934
16	**KR + GLM**	**58.08204**
17	PubChem + DNN1 (default)	51.247895
18	PubChem + DNN2	43.821302
19	PubChem + DNN3	52.581054
20	PubChem + DNN4	35.308698
21	PubChem + DNN5	49.682412
22	**PubChem + RF**	**57.811296**
23	PubChem + GBM	55.39983
24	PubChem + GLM	56.47231
25	CDK + DNN1 (default)	46.061528
26	CDK + DNN2	40.97237
27	CDK + DNN3	46.9928

28	CDK + DNN4	35.53523
29	CDK + DNN5	44.373882
30	CDK + RF	52.278584
31	CDK + GBM	51.450676
32	CDK + GLM	51.86358
33	Extended + DNN1	46.061528
34	Extended + DNN2	40.97237
35	Extended + DNN3	46.9928
36	Extended + DNN4	35.53523
37	Extended + DNN5	44.373882
38	Extended + RF	52.278584
39	Extended + GBM	51.450676
40	Extended + GLM	51.86358
41	AP + DNN1 (default)	39.69568
42	AP + DNN2	39.742348
43	AP + DNN3	40.6057
44	AP + DNN4	41.487348
45	AP + DNN5	42.50953
46	AP + RF	50.130326
47	AP + GBM	49.249876
48	AP + GLM	51.675236
49	Sub + DNN1 (default)	42.447615
50	Sub + DNN2	44.32884
51	Sub + DNN3	40.49191
52	Sub + DNN4	39.712846
53	Sub + DNN5	44.513887
54	Sub + RF	51.382804
55	Sub + GBM	50.69634
56	Sub + GLM	55.37688
57	Estate + DNN1 (default)	43.221298
58	Estate + DNN2	47.891808
59	Estate + DNN3	37.531152

60	Estate + DNN4	42.794597
61	Estate + DNN5	48.037446
62	Estate + RF	52.78824
63	Estate + GBM	51.56761
64	Estate + GLM	51.391643
65	MACCS + DNN1 (default)	46.25236
66	MACCS + DNN2	45.327207
67	MACCS + DNN3	45.650625
68	MACCS + DNN4	41.63565
69	MACCS + DNN5	45.148638
70	MACCS + RF	52.279365
71	MACCS + GBM	56.29321
72	MACCS + GLM	55.346507

*red color indicates the best result

Appendix D

The source codes below were used to generate heatmap plot in Chapter 4.

R package　　　：ChemmineR

```
source("http://bioconductor.org/biocLite.R") # Sources the biocLite.R installation
script.
biocLite("ChemmineR") # Installs the package.
library("ChemmineR") # Loads the package
#setwd(".")
mywd <-"."
setwd(mywd)

library(gplots)
library(gclus) # for order.hclust
library(RColorBrewer)
options(expressions=10000)
##===============================================================================
=======
par.margin <- function(margin="smart"){
  if(margin=="smart"){
    par(cex=0.8, mgp=c(2,1,0), xaxs="i",yaxs="i", mar=c(3,3,2,1)+0.1)
  }

  if(margin=="narrow"){
    par(cex=0.8, mgp=c(2,0.5,0), xaxs="i",yaxs="i", mar=c(3,3,2,1)+0.1)
  }

  if(margin=="none"){
    par(cex=0.8, mgp=c(0,0,0), xaxs="i",yaxs="i", mar=c(0,0,0,0))
  }
  if(margin=="sqrt"){
    par(cex=0.8, mgp=c(0,0,0), xaxs="i",yaxs="i", mar=c(3,3,3,3)+0.1)
  }
}
##==============================================================================
=======
optMatrix <-
function(ofname=FALSE,mat1=mat,mysep=",",bcnames=TRUE,brnames=TRUE,brreverse=FAL
SE){
  fout <-  file(ofname,"w")
  if(bcnames==TRUE){
    if(brnames==TRUE){
      cnames <- "LABEL"
    }else{
      cnames <- c()
    }
    cnames <- c( cnames,dimnames(mat1)[[2]] )
    writeLines(paste(cnames), fout , sep=mysep)
    writeLines("",fout)
  }
  if(brreverse==FALSE){
    for(i in 1:dim(mat1)[1]){
      if(brnames==TRUE){
        rnames <- dimnames(mat1)[[1]]
        writeLines(paste(rnames[i]),fout,sep=mysep)
      }
      z <- mat1[i,]
      writeLines(paste(z), fout, sep=mysep)
      writeLines("",fout)
    }
  }else{
    for(i in rev(1:dim(mat1)[1])){
      if(brnames==TRUE){
```

```
      rnames <- dimnames(mat1)[[1]]
      writeLines(paste(rnames[i]),fout,sep=mysep)
     }
     z <- mat1[i,]
     writeLines(paste(z), fout, sep=mysep)
     writeLines("",fout)
   }
  }
  close(fout)
}

##==========================================================================
=======
optHeatmap.2 <- function(ofname="test.csv",mat,hv=hv,mysep=mysep, ...){
  res <- hv$carpet

optMatrix(ofname=ofname,mat1=t(res),mysep=mysep,bcnames=TRUE,brnames=TRUE,brreve
rse=TRUE)

  return (res)
}

##==========================================================================
=======
colorList <- function(r0=0,g0=1,b0=0,r1=0,g1=0,b1=0,num){
        lst <- c()
        red  =r0+(r1-r0)*(0:num)/num
        green=g0+(g1-g0)*(0:num)/num
        blue =b0+(b1-b0)*(0:num)/num

        lst <- rgb(red=red,green=green, blue=blue)
        return (lst)
}
```

```
##==============================================================================
=======
myColor <- function(mat,colmat,num=1000){
        if(FALSE){
                num <- 1000
                mat <- matrix((-19:20)/20,nrow=4)
                colmat <- matrix(0,ncol=8,nrow=4)
                colmat[1,] <- c( 0,  9, 9, 9,  0,  0, 1, 0)
                colmat[2,] <- c(-1.2,  0, 0, 1,  0,  1, 1, 1)
                colmat[3,] <- c( 0,  1, 1, 1,  1.2,  1, 0, 0)
                colmat[4,] <- c( 0,  0, 0, 0,  0,  9, 9, 9)
        }

        ncolor <- dim(colmat)[1]
        xmin <- min(mat,na.rm=TRUE)
        xmax <- max(mat,na.rm=TRUE)
        mattmp <-c()

        cat(xmin)
        cat(xmax)

        if(xmin<=colmat[2,1] && colmat[ncolor-1,5]<=xmax){
                mattmp <- matrix(0,ncol=8+1,nrow=ncolor)
                a0=xmin
                a1=colmat[2,1]
                seg=floor(num*(a1-a0)/(xmax-xmin))
                mattmp[1,] <- c( a0, colmat[1,6:8], a1, colmat[1,6:8], seg)
                for(i in seq(2,ncolor-1)){
                        a0=colmat[i,1]
                        a1=colmat[i,5]
                        seg=floor(num*(a1-a0)/(xmax-xmin))
                        mattmp[i,]=c(colmat[i,],seg)
                }
                a0=colmat[ncolor-1,5]
                a1=xmax
                seg=floor(num*(a1-a0)/(xmax-xmin))
               mattmp[ncolor,] <- c( a0, colmat[ncolor,2:4], a1, colmat[ncolor,2:4],
seg)
                #print("col:case1")
        }else if(colmat[2,1]<=xmin && colmat[ncolor-1,5]<=xmax){
                mattmp <- matrix(0,ncol=8+1,nrow=ncolor-1)
                for(i in seq(2,ncolor-1)){
                        a0=colmat[i,1]
                        a1=colmat[i,5]
                        seg=floor(num*(a1-a0)/(xmax-colmat[2,1]))
                        mattmp[i-1,]=c(colmat[i,],seg)
                }
                a0=colmat[ncolor-1,5]
                a1=xmax
                seg=floor(num*(a1-a0)/(xmax-colmat[2,1]))
                mattmp[ncolor-1,] <- c( a0, colmat[ncolor,2:4], a1,
colmat[ncolor,2:4], seg)
                #print("col:case2")
        }else if(xmin<=colmat[2,1] && xmax<=colmat[ncolor-1,5]){
                mattmp <- matrix(0,ncol=8+1,nrow=ncolor-1)
                a0=xmin
                a1=colmat[2,1]
                seg=floor(num*(a1-a0)/(colmat[ncolor-1,5]-xmin))
                mattmp[1,] <- c( a0, colmat[1,6:8], a1, colmat[1,6:8], seg)
                for(i in seq(2,ncolor-1)){
                        a0=colmat[i,1]
                        a1=colmat[i,5]
                        seg=floor(num*(a1-a0)/(colmat[ncolor-1,5]-xmin))
                        mattmp[i,]=c(colmat[i,],seg)
                }
                #print("col:case3")
        }else if(colmat[2,1]<=xmin && xmax<=colmat[ncolor-1,5]){
                mattmp <- matrix(0,ncol=8+1,nrow=ncolor-2)
                for(i in seq(2,ncolor-1)){
                        a0=colmat[i,1]
                        a1=colmat[i,5]
                        seg=floor(num*(a1-a0)/(colmat[ncolor-1,5]-colmat[2,1]))
                        mattmp[i-1,]=c(colmat[i,],seg)
                }
                #print("col:case4")
        }else{
                #print("col:case5")
        }
        #browser()
        numlist<-c()
        #numlist <- c(numlist,0)
        for(i in seq(1,dim(mattmp)[1],1)){
                if((mattmp[i,9])>0){
                        numlist<-c(numlist,colorList(
                                                        r0=mattmp[i,2],
                                                        g0=mattmp[i,3],
                                                        b0=mattmp[i,4],
                                                        r1=mattmp[i,6],
                                                        g1=mattmp[i,7],
                                                        b1=mattmp[i,8],
                                                        num=mattmp[i,9]
                                                        )
                                        )
                }
        }
        #numlist
        return (numlist)
}
```

```
#===============================================================================
======
plotFigure <-
function(plotfunction,height=35/2.5,width=35/2.5,directory=".",filename="plot1",
dev="x11",openfile=TRUE,closefile=TRUE,mymargin="sqrt",...){
        if(openfile==TRUE){
                if(dev=="pdf"){
                        filename=paste(filename,".pdf",sep="")
                        cat(filename,"\n")

#                       if(.Platform$OS.type =="windows"){
#
        pdf(file=file.path(directory,filename),bg="white",height=height,width=wi
dth)
#                               par(family = "Japan1GothicBBB")
#                       }else if(capabilities("aqua")){
#
        quartz(file=file.path(directory,filename),type="pdf",
height=height,width=width) # ?P???F?C???`
#                               par(family="HiraKakuProN-W3")
#                       }


        pdf(file=file.path(directory,filename),bg="white",height=height,width=wi
dth)
                        par(family = "Japan1GothicBBB")
            }
                if(dev %in% c("jpeg","jpg")){
                        filename=paste(filename,".jpg",sep="")
                        cat(filename,"\n")
                        jpeg(file=file.path(directory,filename))
                }
                if(dev=="png"){
                        filename=paste(filename,".png",sep="")
                        cat(filename,"\n")
                        png(file=file.path(directory,filename))
                }
        }

        par.margin(margin=mymargin)
        plotfunction(...)

        if(closefile==TRUE){
                if(dev!="x11"){
                        dev.off()
                }
        }
}

readMat <- function(directory=".",ifname="test.txt"){
##----------------------------------------------------------------------------
-------
## Read names at first to aboid converting the white space into the dot automatically

  data <- read.delim(ifname,
header=F,sep="\t",row.names=1,as.is=c(TRUE,TRUE),strip.white=FALSE)
  mat <- as.matrix(data)
  colname <- mat[1,]
  mat <- mat[-c(1),]
  rowname <- rownames(mat)

  ## Read contents
  data <- read.delim(ifname, header=T,sep="\t",row.names=1)
  mat <- as.matrix(data)
  colnames(mat) <- colname
```

```
 rownames(mat) <- rowname

 cat(colnames(mat))
 cat(rownames(mat))

 return (mat)
}

selectMat <-
function(thresh1=FALSE,thresh2=FALSE,thresh4=FALSE,boptcsv=TRUE,ifname=NULL){
 if(thresh1!=FALSE){
  matlimit <- c()

  vecrange <- apply(mat, 1, function(x) (max(x) - min(x)) )
  idxVecUpper <- c()
  idxVecUpper <- (1:length(vecrange))[vecrange > thresh1]

  if(length(idxVecUpper)==0){
   mat <- mat
  }else{
   matlimit <- mat[idxVecUpper,]
   mat <- matlimit
  }
 }

 if(thresh2!=FALSE){
  vecrange2 <- apply(mat, 2, function(x) (max(x) - min(x)) )
  idxVecUpper <- c()
  idxVecUpper <- (1:length(vecrange2))[vecrange2 > thresh2]

  if(length(idxVecUpper)==0){
   mat <- mat
  }else{
      matlimit <- mat[,idxVecUpper]
      mat <- c()
      mat <- matlimit
    }
  }

  if(thresh4!=FALSE){
    vecrange4 <- apply(mat, 2, sum )
    idxVecUpper <- c()
    idxVecUpper <- (1:length(vecrange4))[vecrange4 > thresh4]
    if(length(idxVecUpper)==0){
      mat <- mat
    }else{
      matlimit <- mat[,idxVecUpper]
      mat <- c()
      mat <- matlimit
    }
  }

  ##------------------ output the matrix data of the target
  if(boptcsv==TRUE){
    if(ifname==NULL){
      ofname<-"debug_mat.txt"
    }else{
      ofname <- ifname
      ofname <- sub(".dat","",ofname)
      ofname <-
sprintf("%s%_thresh04i_%04i.tsv",ofname,floor(thresh1),floor(thresh2))
    }
    optMatrix(ofname=ofname2,mat1=mat,bcnames=TRUE,brnames=TRUE)
  }

  return (mat)
```

```
}


myTitle <- function(ifname="test.txt",thresh1=FALSE,thresh2=FALSE,thresh4=FALSE){
  ##------------------ set the title for the plot
  if(thresh1==FALSE){
    if(thresh2==FALSE){
      mytitle <- sprintf("%s\n",ifname)
    }else{
      mytitle <- sprintf("%s: thresh2=%i\n",ifname,thresh2)
    }
  }else{
    if(thresh2==FALSE){
      mytitle <- sprintf("%s: thresh1=%i\n",ifname,thresh1)
    }else{
      mytitle <- sprintf("%s: thresh1=%i, thresh2=%i\n",ifname,thresh1,thresh2)
    }
  }

  return (mytitle)
}

##==============================================================================
=======
heatmapPlot <- function (mat=mat,mycol=heat.colors(n=12),mytitle=NULL,
                      bboxplot=FALSE,bhist=FALSE,bopttsv=FALSE,boptheatmap=FALSE,
                      boptgroups=FALSE,
                      bsortRow=TRUE,bsortCol=TRUE,bnum=TRUE,
    bdendRow=FALSE,bdendCol=FALSE,bAttachGroup=TRUE,
                      distmethod="manhattan",hclustmethod="ward.D2",
    ncolhr=5,ncolhc=5,bshift=TRUE,
                      i0=10,i1=100,i2=20,...
){
##------------------ attach the # of the degree to the names
```

```
  sum1 <- apply(mat,1,sum,na.rm=TRUE)
  sum2 <- apply(mat,2,sum,na.rm=TRUE)
  if(bnum==TRUE){
   dimnames(mat)[[1]]=paste(sum1, dimnames(mat)[[1]])
   dimnames(mat)[[2]]=paste(sum2, dimnames(mat)[[2]])
  }

##------------------ sort within the dendrogram by the degree
  if(bsortRow==TRUE){
   mat <- mat[order(sum1),         ]
  }
  if(bsortCol==TRUE){
   mat <- mat[        ,order(sum2)]
  }

 ##------------------ distant
 d_row<-dist( mat ,method=distmethod)
 d_col<-dist(t(mat),method=distmethod)
  print(head(d_col))



  ##------------------ store original names of mat
  rownamesorg <- rownames(mat)
  colnamesorg <- colnames(mat)

  mycolhc="white"
  mycolhr="white"
  dend="none"
 ##------------------ hclust
  if(bdendRow==TRUE){
    hc_row<-hclust(d_row^2, method=hclustmethod)
    ##------------------ group row
    if(bAttachGroup==TRUE){
     l <- seq(i0,i1,by=i2)
     groups_row <- cutree(hc_row,k=l)
     for(i in 1:length(l)){
      clr <- cutree(hc_row,k=l[i])
      hc_row$labels <- paste(hc_row$labels,clr,sep="=")
     }
    }
    ##------------------ set new names with clustering results
    rownames(mat) <- hc_row$labels
    ##------------------ sort within the dendrogram by the degree
    if(bsortRow==TRUE){
     dd_row <- as.dendrogram(hc_row)
     dd_row.reorder <- reorder(dd_row,order(sum1))

     iii <- order.dendrogram(dd_row.reorder)
     jjj <- seq(1,length(iii),2)
     kkk <- iii[jjj]
     if(bshift==TRUE){
      rownames(mat)[kkk] <-
 paste(paste(rep(("-"),18),collapse=""),rownames(mat)[kkk])
     }

     rowv=dd_row.reorder
     dend="row"
    }else{
     rowv=NULL
    }

    ## ----------------- set colors for the heatmap color bar
    color1<-colorRampPalette(brewer.pal(12,"Set3"))(ncolhr)
    mycolhr <- color1
    mycolhr <- mycolhr[as.vector(cutree(hc_row,k=ncolhr))]
```

```
 }else{
   rowv=1:(dim(mat)[2])
   mycolhr=rep("white",dim(mat)[1])
 }


 if(bdendCol==TRUE){
   print(d_col[is.na(d_col)])
   hc_col<-hclust(d_col^2, method=hclustmethod)
   ##------------------ group col
   if(bAttachGroup==TRUE){
     l <- seq(i0,i1,by=i2)
     groups_col <- cutree(hc_col,k=l)
     for(i in 1:length(l)){
       clc <- cutree(hc_col,k=l[i])
       hc_col$labels <- paste(hc_col$labels,clc,sep="=")
     }
     ##------------------ set new names with clustering results
     colnames(mat) <- hc_col$labels
   }
   ##------------------ sort within the dendrogram by the degree
   if(bsortCol==TRUE){
     dd_col <- as.dendrogram(hc_col)
     dd_col.reorder <- reorder(dd_col,order(sum2))

     iii <- order.dendrogram(dd_col.reorder)
     jjj <- seq(1,length(iii),2)
     kkk <- iii[jjj]
     if(bshift==TRUE){
       colnames(mat)[kkk] <-
paste(colnames(mat)[kkk],paste(rep(("-"),20),collapse=""))
     }
```

```
      colv=dd_col.reorder
      dend="column"
    }else{
      colv<-1:(dim(mat)[1])
    }

    ## ---------------- set colors for the heatmap color bar
    color2<-colorRampPalette(brewer.pal(12,"Set3"))(ncolhc)
    mycolhc <- color2
    mycolhc <- mycolhc[as.vector(cutree(hc_col,k=ncolhc))]
  }else{
    colv=NULL
    mycolhc=rep("white",length(d_col))
  }

  if(bdendCol==TRUE && bdendRow==TRUE){
    dend="both"
  }

#   cat(head(mat[,1:5]))
#   print("device=")
#   dev.cur()
#   print("desu")
  #print("mycol=")
  #print(summary(mycolhc))
  #print(summary(mycolhr))
  #print("desu")

 ## ---------------- draw a heatmap
 hv <- heatmap.2(mat, col=mycol,scale='none',
             trace='none',keysize=1.0,
             density.info="none",key=TRUE,
             dendrogram=dend,
             Colv=colv,Rowv=rowv,
             ColSideColors=mycolhc, RowSideColors=mycolhr,
                         margin=c(12,12),cexCol=0.22,cexRow=0.22,
#                                add.expr = c(abline(h=seq(0.5,1000, 1),
v=seq(0.5,1000, 1),lty=1,lwd=0.1,col='gray'),
#                                          abline(h=seq(0.5,1000, 5),
v=seq(0.5,1000, 5),lty=1,lwd=0.2,col='black' ),
#                                          abline(h=seq(0.5,1000,10),
v=seq(0.5,1000,10),lty=1,lwd=0.5,col='black') ),
             ...
 )
#             add.expr = abline(h=seq(0.5,1000, 1), v=seq(0.5,1000,
1),lty=1,lwd=0.05,col='gray'),
#                                ...
#             )
                                 ## mat    : matrix for clustering
                                 ## margin : the margins (see par(mar= *)) for column
and row names
                                 ## cexCol, cexRow: used as cex.axis in for the row
or column axis labeling.
                                 ## add.expr=abline() : horizontal and virtical
lines

 ## ----------------- output the data of heatmap
  if(bopttsv==TRUE){
   optHeatmapToTsv(hv=hv,ifname=ifname,thresh1=thresh1,thresh2=thresh2)
  }

 ## ----------------- output the group data for heatmap
 if(boptgroups==TRUE){
   # row
   ofname <- ifname
   ofname <- sub(".dat","",ofname)
   ofname <- sprintf("%s_reac_groups.tsv",ofname)

   clr <- c()
   for(i in 1:length(l)){
     tmp <- cutree(hc_row,k=l[i])
     clr <- rbind(clr,tmp)
   }
   rownames(clr) <- l
   colnames(clr) <- rownamesorg

   res <- optMatrix(ofname=ofname,mat1=t(clr),mysep="\t",
                bcnames=TRUE,brnames=TRUE,brreverse=TRUE)
   # col
   ofname <- ifname
   ofname <- sub(".dat","",ofname)
   ofname <- sprintf("%s_comp_groups.tsv",ofname)

   clc <- c()
   for(i in 1:length(l)){
     tmp <- cutree(hc_col,k=l[i])
     clc <- rbind(clc,tmp)
   }
   rownames(clc) <- l
   colnames(clc) <- colnamesorg

   res <- optMatrix(ofname=ofname,mat1=t(clc),mysep="\t",
                bcnames=TRUE,brnames=TRUE,brreverse=TRUE)
 }
}
##============================================================================
=======
optHeatmapToTsv <-
function(hv=hv,ifname=ifname,thresh1=thresh1,thresh2=thresh2,boptheatmap=TRUE){
   if(boptheatmap==TRUE){
     ofname <- ifname
     ofname <- sub(".dat","",ofname)
     if(thresh1==FALSE){
       if(thresh2==FALSE){
         ofname <- sprintf("%s_heatmap.tsv",ofname)
       }else{
         ofname <- sprintf("%s_heatmap_thresh1_%04i.tsv",ofname,floor(thresh1))
       }
     }else{
       if(thresh2==FALSE){
         ofname <- sprintf("%s_heatmap_thresh2_%04i.tsv",ofname,floor(thresh2))
       }else{
         ofname <- 
sprintf("%s_heatmap_thresh1_%04i_2_%04i.tsv",ofname,floor(thresh1),floor(thresh2
))
       }
     }
     optHeatmap.2(ofname=ofname,mat=mat,hv=hv,mysep="\t",brreverse=TRUE)
   }
}
#=============================================================================
======
ofname <- "curry2"
thresh1=FALSE
thresh2=FALSE
thresh4=FALSE
bhist=TRUE

##---------------------- setting the color matrix for heatmap
colmat <- matrix(0,ncol=8,nrow=3)
colmat[1,] <- c(  0,   9, 9, 9,   0,  1, 1, 1)
colmat[2,] <- c(  0,   1, 1, 1,   1,  1, 0, 0)
```

```
colmat[3,] <- c(  0,  1, 0, 0,  0, 9, 9, 9)
##---------------------- plot heatmap

ifname="curry2"

vocset <- read.SDFset("voc_set.sdf")
view(vocset)
length(vocset)
#apvoc <- sdf2ap(vocset)
#sapply(cid(apvoc), function(x) cmp.similarity(apvoc[1], apvoc[x])) ## Run
cmp.similarity in loop as custom similarity search function

fpvoc <- fp2bit(vocset) # Convert base 64 encoded fingerprints to binary matrix
fpma <- as.matrix(fpvoc) # Converting a fingerprint database to a matrix
write.csv(fpma, "binarymatrix.csv", row.names=TRUE)
#fpSim(fpvoc[1], fpvoc[2]) # Pairwise compound structure comparisons

fpSim(fpvoc[1], fpvoc, method="Tanimoto") #Similarity searching and returning
Tanimoto, Dice, Cosine, Tversky similarity coefficients:
#fpSim(fpvoc[1], fpvoc, method="Tversky", cutoff=0.4, top=4, alpha=0.5, beta=1) #
Under method one can choose from several predefined similarity measures including
Tanimoto (default), Euclidean, Tversky or Dice
#fpSim(fpvoc[1], fpvoc, method="Tversky", alpha=0.7, beta=0.7)
#cosine <- function(a, b, c, d) c/sqrt(a*b) #Example for using a custom similarity
function:
#fpSim(fpvoc[1], fpvoc, method=cosine)
simMAT <- sapply(cid(fpvoc), function(x) fpSim(fpvoc[x], fpvoc, method="Tanimoto",
sorted=FALSE)) # Compute similarity matrix
write.csv(simMAT, "tanimotomatrix.csv", row.names=TRUE)
mat <- 1-simMAT

mycol <- myColor(mat,colmat)
mycol<-rev(heat.colors(12))
```

```
if(thresh1==FALSE && thresh2==FALSE && thresh4==FALSE){
}else{
  mat <-
selectMat(thresh1=FALSE,thresh2=FALSE,thresh4=FALSE,boptcsv=TRUE,ifname=ifname)
}

##------------------ output the histgram of the mat
if(bhist==TRUE) hist(mat,nclass=100)



ifname="heat_map1"
ofname <- "heat_map1"
plotFigure(heatmapPlot,height=21/2.5,width=21/2.5,

mat=mat,filename=ofname,directory=mywd,dev="pdf",openfile=TRUE,closefile=TRUE,
         xlab="",ylab="",main="",mycol=mycol,
         bsortRow=TRUE,bsortCol=TRUE,
         bdendRow=TRUE,bdendCol=TRUE,bAttachGroup=TRUE,
bopttsv=TRUE,mymargin="sqrt",
         bboxplot=FALSE,bhist=FALSE,boptheatmap=TRUE,i0=11,i1=11,i2=5,bnum=TRUE,
         ncolhr=50,ncolhc=50,bshift=TRUE,
         distmethod="euclidean")

ifname="heat_map2"
ofname <- "heat_map2"
plotFigure(heatmapPlot,height=34/2.5,width=34/2.5,

mat=mat,filename=ofname,directory=mywd,dev="pdf",openfile=TRUE,closefile=TRUE,
         xlab="",ylab="",main="",mycol=mycol,
         bsortRow=TRUE,bsortCol=TRUE,
         bdendRow=TRUE,bdendCol=TRUE,bAttachGroup=FALSE,
bopttsv=TRUE,mymargin="sqrt",
         bboxplot=FALSE,bhist=FALSE,boptheatmap=TRUE,i0=11,i1=11,i2=5,bnum=FALSE,
         ncolhr=50,ncolhc=50,bshift=FALSE,
         distmethod="euclidean")


# heatmapPlot(xlab="",ylab="",main="",mycol=mycol,
#          bsortRow=TRUE,bsortCol=TRUE,
#          bdendRow=TRUE,bdendCol=TRUE,
#
bboxplot=FALSE,bhist=FALSE,boptheatmap=TRUE,k_cutree=20,i0=5,i1=15,i2=5,bnum=TRU
E,
#         distmethod="euclidean")

#------------------------

#------------------------
```