Master's Thesis

Clustering Analysis of Soil Microbial Community at Global Scale

Tetsushi Tanaka

March 6, 2019

Graduate School of Information Science
Nara Institute of Science and Technology

本論文は奈良先端科学技術大学院大学情報科学研究科に
修士（工学）授与の要件として提出した修士論文である．

田中 徹士

審査委員：

金谷 重彦 教授	（主指導教員）
笠原 正治 教授	（副指導教員）
MD.ALTAF-UL-AMIN 准教授	（副指導教員）
小野 直亮 准教授	（副指導教員）
黄 銘 助教	（副指導教員）

グローバルスケールでの土壌微生物コミュニティのクラスタリング解析*

田中 徹士

内容梗概

土壌中には巨大な数の細菌が存在していて、それらの存在や機能が土壌特性に影響を与えている。細菌はマイクロバイオームとよばれる複雑なコミュニティを形成している。土壌マイクロバイオームの生態系は植物や動物に比べて多くのことがわかっていない。土壌マイクロバイオームが世界全体でどのように異なるか、そして細菌組成が環境とどのように関連しているのかといった疑問がある。近年、次世代シーケンサーの発展によってマイクロバイオームデータは蓄積されつつあり、情報科学的なアプローチによって大規模で網羅的なマイクロバイオーム解析が必要とされている。本研究ではマイクロバイオームのデータベース、Earth Microbiome Project（EMP）を使用して、幅広い地域の様々な環境のデータを比較解析した。細菌の遺伝的距離に基づいた距離、UniFrac 距離を計算し、クラスタリング解析を行った。クラスタリング結果を細菌の機能や特徴の視点から生態学的に解釈した。水田とワイン園のクラスタで有意に存在する細菌は、先行研究で挙げられていた細菌と共通していたことが確認された。加えて、モンゴルの草原や森林、バイオフィルターなどに特徴的な細菌群を新たに明らかにした。さらに、クラスタと気候区分の関係について調べた。本研究は細菌を基準にした土壌管理の知見につながることが期待される。

キーワード

土壌微生物, マイクロバイオーム, メタゲノム, UniFrac 距離, クラスタリング解析

*奈良先端科学技術大学院大学 情報科学研究科 情報科学専攻 修士論文
March 6, 2019.

Clustering Analysis of Soil Microbial Community at Global Scale *

Tetsushi Tanaka

Abstract

There is a huge number of bacteria in the soil and their existence and function affect the soil properties. Bacteria form a complex community called microbiome. Compared to the ecosystem of plants and animals, we still know little about soil microbial ecosystem. How the soil microbiomes are different throughout the world and how they relate to the region and the environment is a major interest. Recently, the development of next-generation sequencer has been enabled to accumulate metagenome profile data, and a large scale and comprehensive microbiome analysis are required by an information scientific approach. In this study, we compared and analyzed the data from various environments on a global scale using the microbiome database, the Earth Microbiome Project (EMP). We calculated the distance based on genetic distance, named UniFrac distance and did clustering analysis. Clustering results were ecologically interpreted from the view of the function and the characteristics of bacteria. We revealed the characteristic of groups of bacteria related to paddy, vineyard, grasslands in Mongolian, forests, and biofilter. Furthermore, we investigated the relationship between clusters and climate zones. This research is expected to lead to knowledge of soil management based on soil microbiome.

Keywords:

Soil Microbiology, Microbiome, Metagenome, UniFrac Distance, Clustering Analysis

[*] Master's Thesis, Graduate School of Information Science,
Nara Institute of Science and Technology, March 6, 2019

Contents

List of Figures

1 Introduction

1.1 Soil Microbiology

In soil, a huge number of species of microorganism exist. The term microorganism or microbes refers to a living thing that is too small to be seen with the eye and include bacteria, fungi, archaea, and protists. Their existences and the functions play an important roles in maintaining soil fertility through recycling nutrients and influencing their availability to plants, improving soil structure, affecting plants and soil environment. As an example, some bacteria and fungus perform to increase the bioavailability of nutrients such as nitrogen and phosphate [1]. The diazotrophic bacteria and fungus perform to transform nitrogen from the atmospheric gas to usable combined nitrogen compounds, which is essential for plants growth [2]. Moreover, soil microbial biomass and microbial metabolites contribute to the agglomeration of soil [3].

1.2 Metagenome Analysis and Operational Taxonomic Unit

A huge number of microbes create complex communities having specific and exclusive relationships. This complex community composed of microbes is called microbiome. Until recently, the analysis of microorganism was performed by culture-dependent methodologies. This approach was limited to culturable microbial species and it was not suitable for examining the microbial composition in a sample. The development of next generation DNA sequencing opened up microbiome analysis and made it possible to determine the taxonomic composition of many samples without isolation culture, and research the enormous biodiversity and complex ecology of microbial ecosystems. This DNA sequencing approach to the study of the microbiome is called metagenomics. The brief process of metagenomics analysis is shown in Figure 1.1.

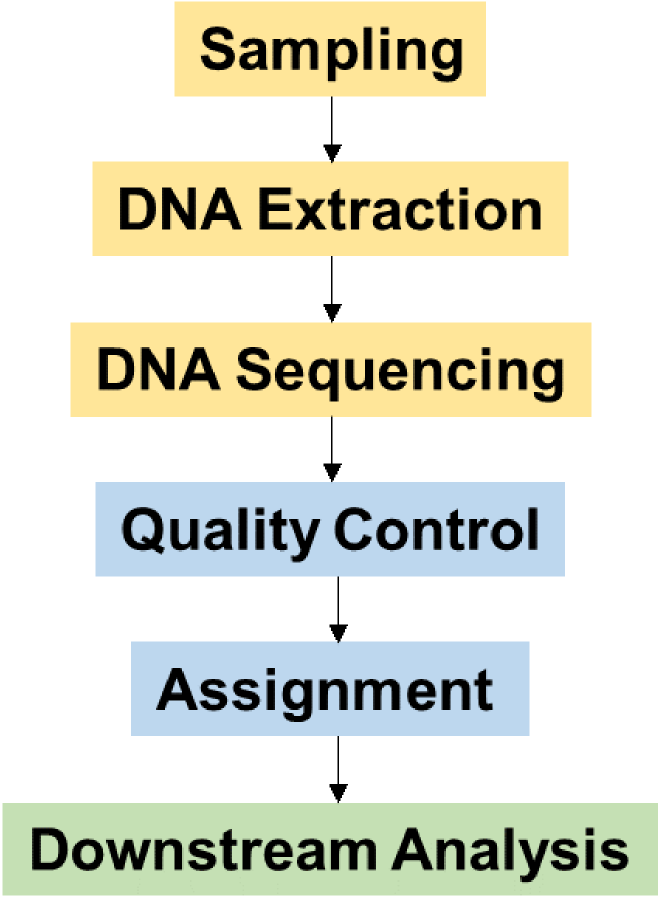


Figure 1.1: Yellow colored represent biological processing. Blue one represents the bioinformatic processing. Downstream Analysis includes clustering analysis and statistical test.

Amplicon sequencing and OTU-based analysis are one of the main approaches to microbiome research. Amplicon sequencing reads marker gene, which is highly conserved across taxa and, cost less than whole-genome sequencing. The target gene is commonly 16S rRNA gene for bacteria and, internal transcribed spacer (ITS) region and 18S rRNA genes for fungi. After DNA extraction, the target gene is amplified by PCR, which generates copies of the target sequence. Not all PCR products will be used to analysis because of sequencing error and chimeras (result from a combination of two or more sequence templates and synthesized when prematurely terminated fragments reanneal to other template DNA during PCR), therefore filtering of the sequences depending on certain criteria is needed. One of the simple approaches is to discard sequences based on their length. Sequences having remarkably longer than the average length tend to be a chimeric sequence. Tools such as PyroNoise [6], Denoise [5], and AmpliconNoise [6] are applied to control sequencing. As Chimera detection bioinformatics tools, UCHIME [7] and Perseus [6] are available. After quality control, reads are clustered by similarity with or without referencing external reference sequences collection, such as SILVA [8], RDP [9], Greengenes [10] and NCBI [11]. The approach to cluster sequences without reference database is called de novo OTU picking. In De novo clustering process, reads are clustered against one another. On the other hand, closed-reference OTU picking and open-reference OTU picking are to cluster with referencing reference sequences collection; the former exclude any reads which do not hit a sequence in a database, the latter conduct de novo OTU picking on such sequences. There are several algorithms to divide a set of sequences into clusters. The UCLUST algorithm divides a set of sequences into clusters under the condition that a cluster is defined by the centroid and every sequence in the cluster must have a similarity above a given identity threshold with the centroid (Figure 1.2 - Figure 1.3). In closed-reference OTU picking and open-reference OTU picking, a reference sequence is used instead of a representative sequence (Figure 1.4). After the OTU assignment, an OTU table is obtained. An OTU table is a matrix that gives the number of reads per OTU per sample.

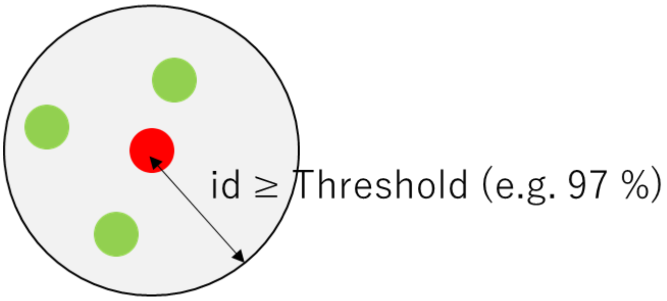


Figure 1.2: Every sequence (green and red circle) in the cluster must have similarity above a given identity threshold with the centroid (red circle)

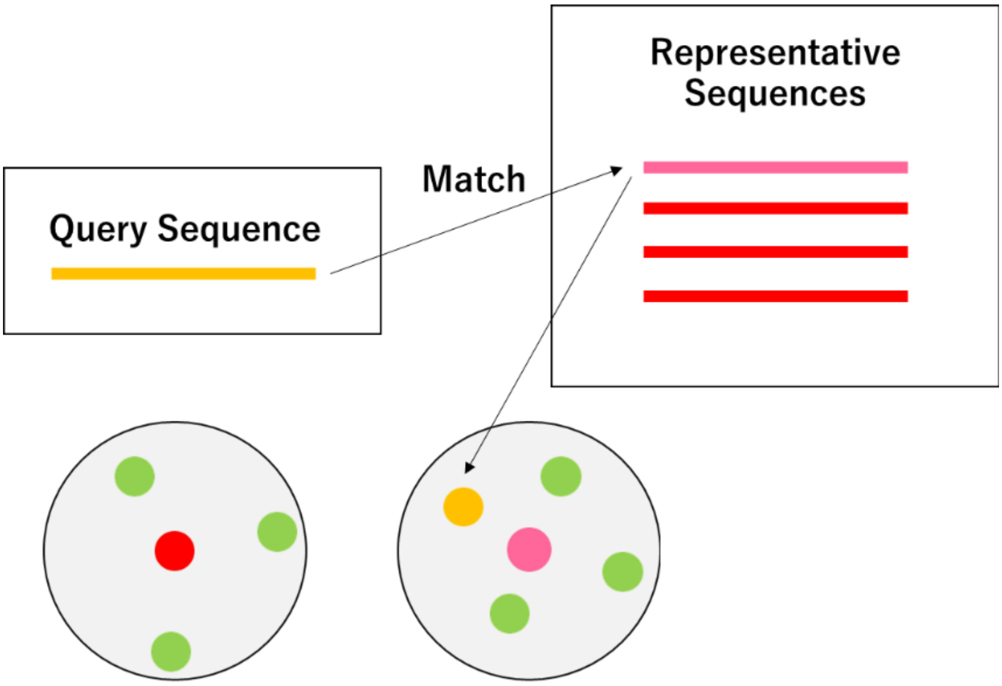


Figure 1.3: If a match is found to representative sequences, the query is assigned to that cluster, otherwise, the query becomes the seed of a new cluster (de novo and open-reference OTU picking) or is removed (closed-reference OTU picking)

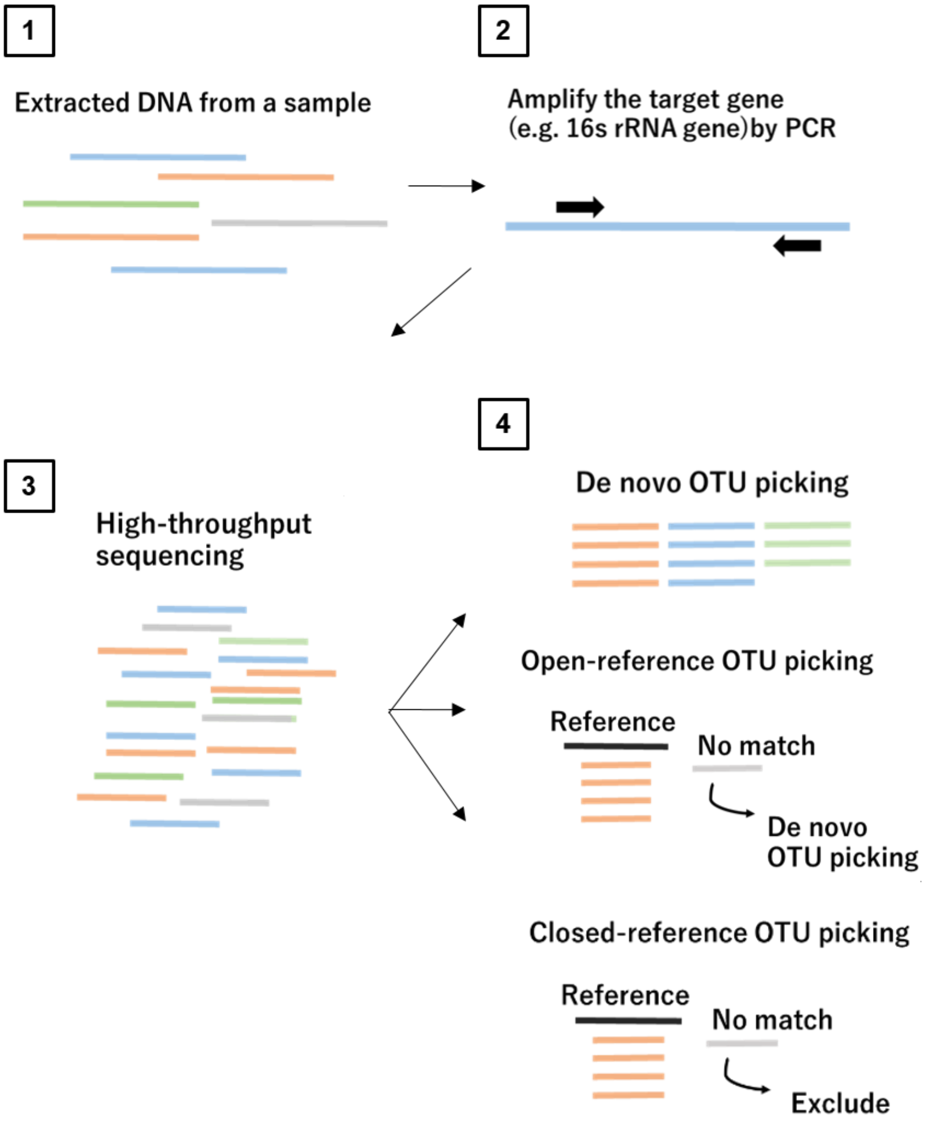


Figure 1.4: The process from DNA extraction to OTU picking

In order to understand the structure of microbial communities and their interactions within the context of ecological or environmental metadata, bioinformatics and statistical approach are necessary due to the big data. In recent years, microbiomes have been studied in various fields, for example, in the medical field, studies on the relationship between intestinal bacteria and diseases such as obesity have been reported [12].

1.3 Soil Microbial Ecology

Although the biological diversity across the globe of vegetation and animal is clarified by long-term studies by ecologists, most of the diversity and regionality of soil microbiome remain undescribed. With metagenome analysis, we are starting to investigate how soil microbiome is different across the globe and how microbial composition is related to ecological attributes. Previous studies have shown that generally only a few bacteria are in common among samples [13–15], but some taxa which are abundant in individual soil may also be abundant in soil, even when those soil are from distant places. In many research, soil microbiome has been compared within a small area or limited environment. Association between soil pH and Carbon / Nitrogen ratio and microbiome in vineyards were studied [16]. In alpine, differences in microbiome structure to altitude, season, vegetation were investigated [17]. A few comparative analyses have been conducted on global scale, but only diversity index or top abundant taxa were used to compare the differences [14, 18].

1.4 Outline and Purpose of this Work

There may be characteristics of microbiome composition that explains the ecology attributes and we aimed to clarify those characteristics. As previously stated, the comparative analysis of microbiome has performed on a small spatial scale in a limited environment or region, or on a global scale only using limited indicators and taxa. In this work, we compared soil microbiomes collected from various regions and environments on a global scale, using the taxon composition and genetic distance. We used The Earth Microbiome Project (EMP) as a dataset and analyzed 4998 samples, including 23 countries and regions and 48674 bacterial OTUs. We clarified the relationship between bacterial composition and environment and considered from the viewpoint of the bacterial ecosystem. Finding the characteristics of microbial community composition will lead to soil management based on soil microbiome. Today, materials that use biological functions including microorganisms instead of chemical fertilizers have attracted attention from the environmental viewpoint, so soil management using bacteria have become important.

1.5 The Earth Microbiome Project

The Earth Microbiome Project (EMP, http://www.earthmicrobiome.org/) is an open database of microbiome research, having over 200,000 samples collected from numerous type, including human, animal, soil, plant, marine, and so on. EMP developed the standard protocol (http://www.earthmicrobiome.org/protocols-and-standards/), including DNA extraction, Illumina amplicon protocol, bioinformatics processing, and enables to compare many samples across the world. As a default, the bioinformatics processing is on QIIME [19], which is an open-source bioinformatics software. Quality control of reads and OTU picking are handled by QIIME. QIIME command is described in the Appendix.

2 Materials & Methods

2.1 Dataset

The OTU table was downloaded from EMP database (https://qiita.ucsd.edu/emp/). Data were stored as Biological Observation Matrix (BIOM) format, which is designed to be a general-use format for representing biological sample by observation contingency tables and a recognized standard for EMP. In this work, BIOM format was converted into a data frame using R package biomformat[1], and then the analysis was performed. The codes are described in the Appendix. The dataset contained 4998 soil microbiome sample collected from around the world including 23 countries and regions (Figure 2.1, Table 2.1). All data had the information about “Land-use”, what the land was used for. The land-use labels were classified into 8 categories, including cropland, forest, urban, polar, grassland, tundra, shrubland, and others. The number of samples of each the categories is as follow Table 2.2. The sum of OTUs abundances in a sample was transformed to 1 for normalization.

[1] https://bioconductor.org/packages/release/bioc/htm.html

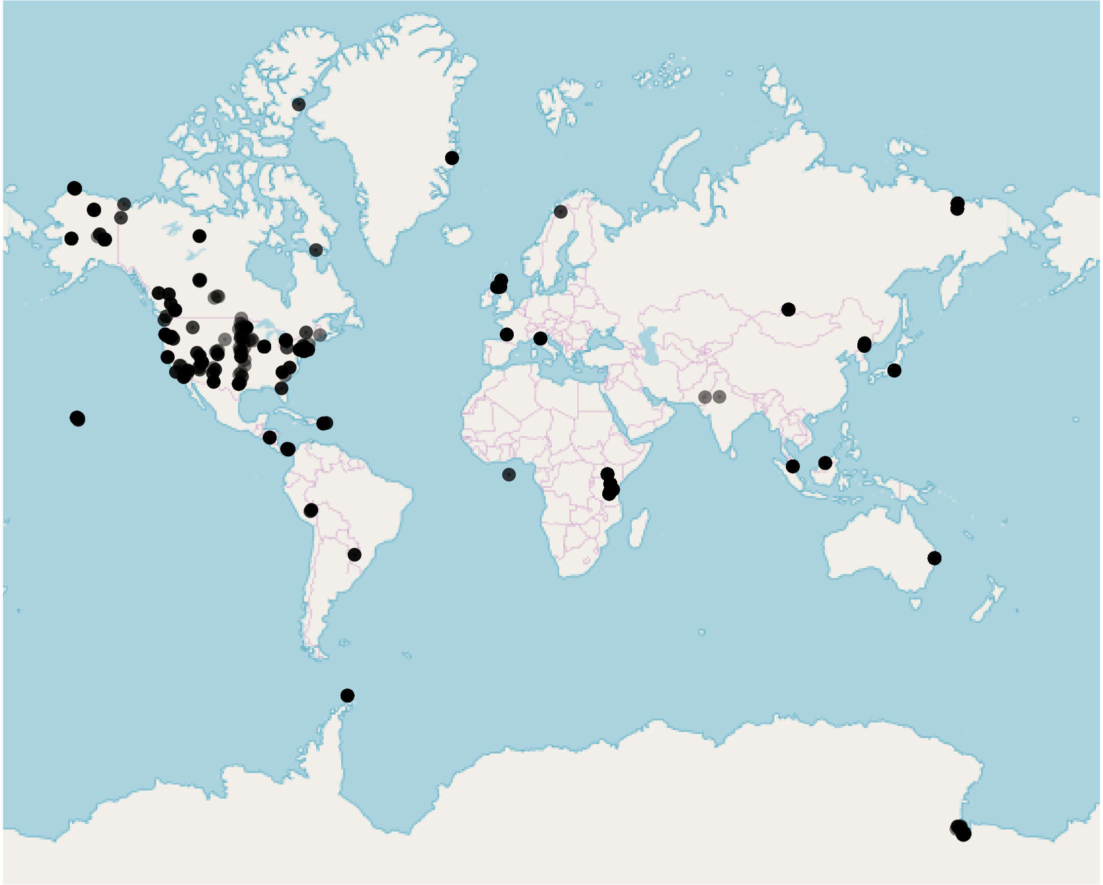


Figure 2.1: Sampling area

Table 2.1: Location and land-use categories of samples

Countries and regions	Number of samples	Samples Categories
Antarctic	173	Polar, Tundra
Argentina	4	Grassland
Australia	292	Cropland
British Virgin Islands	31	Forest
Canada	55	Tundra, Grassland, Shrubland, Urban, Others
China	22	Shrubland
Denmark (Geenland)	20	Tundra
France	15	Cropland
India	2	Others(Desert)
Italy	48	Cropland
Japan	629	Cropland
Kenya	77	Forest, Cropland, Others(Rangeland)
Malaysia	34	Forest
Mongolia	229	Grassland
Nicaragua	61	Cropland
Panama	43	Forest
Peru	6	Forest
Russia	76	Tundra
Singapore	25	Forest
Sweden	2	Tundra
Tanzania	128	shrubland
United Kingdom	929	Cropland Urban
United States	2097	All

Table 2.2: Land-use categories

Land-use categories	Number of samples
Cropland	1487
Forest	405
Urban	1774
Polar	161
Grassland	314
Tundra	414
Shrubland	278
Others	165

2.2 Variable Reduction

We performed variable selection using random forest (RF). RF is a group learning algorithm developed in 2001 by L. Breiman, randomly constructing multiple decision trees and combining those results to predict. RF can be used for classification, regression and feature selection. R package randomForest[2] provides two different importance measure, "MeanDecreaseAccuracy" (MDA) and " MeanDecreaseGini " (MDG), which can be used to rank variables for variable selection. MDA ranks the importance of a variable by measuring the change in prediction accuracy when the values of the variable are randomly permuted compared to the original observation. MDG quantifies the importance by the sum of all decreases in Gini impurity due to a given variable, normalized by the number of trees. We used MeanDecreaseAccuracy criteria in the study. We set the land-use categories of samples as classes labels and trained RF classifier using all OTUs abundances. We evaluated out-of-bag error rate with from 4 to 32768 highest important OTUs.

2.3 Unifrac Distance and Clustering

After variable selection, we calculated the distance among samples using selected OTUs abundance for clustering. UniFrac distance metric has been used for comparing microbial communities and calculate a distance between pairs of samples based on taxa amount or existence. Consider two microbiome samples A and B. Suppose there is a rooted tree with n branches. b_i is the length of branch i and p_i^A and p_i^B are the taxa proportions descending from the branch i for sample A and B, respectively. The UniFrac metric measures the phylogenetic distance between taxa in a phylogenetic tree as a percentage of the branch length of a tree derived from one sample or another sample. Unweighted UniFrac distance [21] is defined as

$$d_U = \sum_{i=1}^{n} \frac{b_i \left| I\left(p_i^A > 0\right) - I\left(p_i^B > 0\right) \right|}{b_i} \tag{2.1}$$

[2] https://cran.r-project.org/web/packages/randomForest/index.html

Function *I(.)* is the indicator function, which takes 1 if the condition is satisfied, and 0 otherwise. The distance d_U ignores the taxa abundance. On the other hand, weighted UniFrac distance [21] takes abundance of taxa into account and is defined as

$$d_W = \sum_{i=1}^{n} \frac{b_i \left| p_i^A - p_i^B \right|}{b_i \left(p_i^A + p_i^B \right)} \tag{2.2}$$

The distance d_W uses the absolute proportion difference $|p_i^A - p_i^B|$, instead of presence/absence of data. As a consequence, the value of d_W is dominated by branches with large proportions and is less sensitive to the abundance changes on the branches with small proportions. The generalized UniFrac distance [22] is a generalized version of UniFrac distance, defined as

$$d^{(\alpha)} = \sum_{i=1}^{n} \frac{b_i \left(p_i^A + p_i^B \right)^{\alpha} \left| \frac{p_i^A - p_i^B}{p_i^A + p_i^B} \right|}{b_i \left(p_i^A + p_i^B \right)^{\alpha}} \tag{2.3}$$

To reduce the effect of the weight on branches with large proportions, the distance use the relative difference $\left| \frac{p_i^A - p_i^B}{p_i^A + p_i^B} \right|$ and has a parameter α to controll the weight on abundant lineages so the distance is not dominated by highly abundant lineages. It is reported that the generalized UniFrac Distance is generally more robust. We chose generalized UniFrac Distance metric and parameter α=0.5 in this work and used R package GUniFrac[3].

2.4 Statistical Test

After clustering, we summed up the OTUs abundances at the genus level for each cluster. To identify the bacterias characterizing the clusters, we performed analysis of variance (ANOVA), which provides a statistical test whether the OTU population means of several groups are equal. In this work, we chose non-parametric ANOVA (kruskal wallis test) because the taxa abundance data do not follow a Gaussian distribution. For the genus that had a significant difference, all samples were sorted in descending order of the abundance, and the percentage of clusters to which the top 10 % samples belonged was examined in Figure 2.2 A. For those

[3] https://cran.rproject.org/web/packages/GUniFrac/index.html

genera with high proportions (top 10) for each cluster, their biological functions and characteristics were investigated Figure 2.2 B.

2.5 Shannon Diversity Index

In order to evaluate the bacterial diversity in a sample, the shannon index was calculated. This index is often used in ecosystem analysis. The Shannon index S can be calculated as

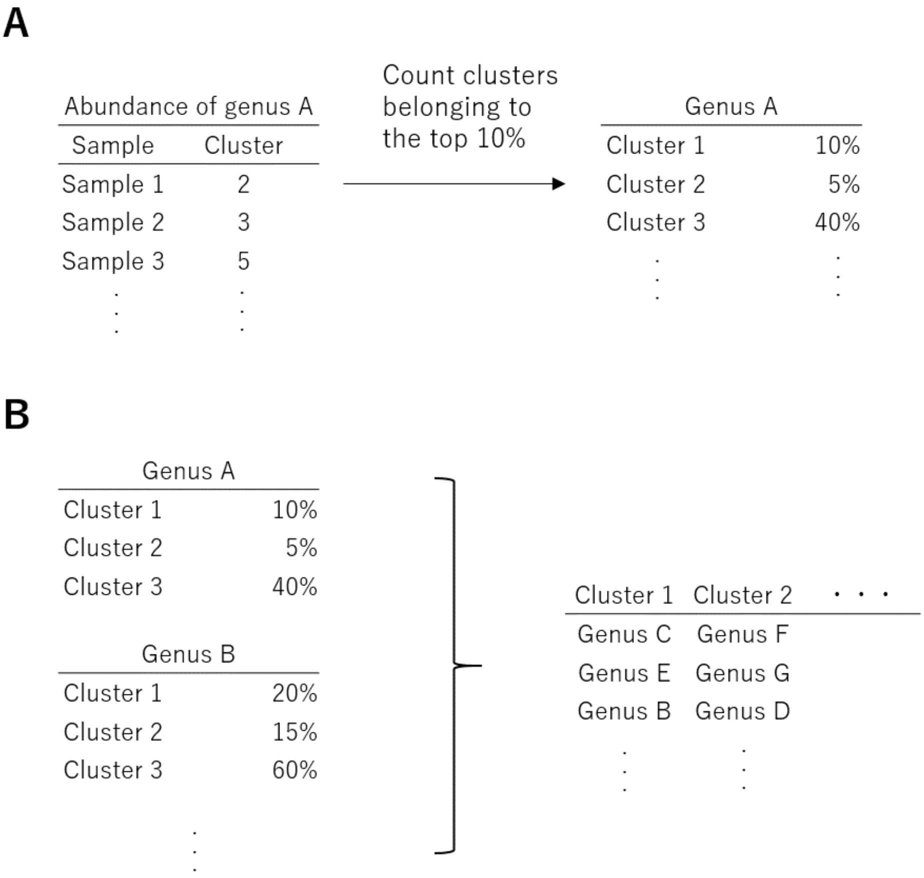
$$S = -\sum_{i=1}^{N} p_i \ln p_i \tag{2.4}$$

where p_i is the proportion of genes relative to the total amount of genera and N is the number of genera.

2.6 Climate Zones

Wladimir koppen has proposed the five vegetation-based climate zones which is one of the most widely used climate classification system, which has been updated by Kottek et al. [23]. The zones we used in this work are as follows: (I) tropical, (Ⅱ) arid, (Ⅲ) temperate, (Ⅳ) continental, and (V) polar. We investigated whether clusters were related to climate zones. In order to count multiple samples at one point as one sample,Two sampling points where the difference of latitude/longitude was less than 1 degree were summarized as the same one point. We aggregated the proportion of climate zones in each cluster using R packages kgc[4].

[4] https://cran.r-project.org/web/packages/kgc/index.html



Abundance of genus A

Sample	Cluster
Sample 1	2
Sample 2	3
Sample 3	5
⋮	⋮

Genus A

Cluster 1	10%
Cluster 2	5%
Cluster 3	40%
⋮	⋮

Genus A

Cluster 1	10%
Cluster 2	5%
Cluster 3	40%

Genus B

Cluster 1	20%
Cluster 2	15%
Cluster 3	60%

⋮

Cluster 1	Cluster 2	· · ·
Genus C	Genus F	
Genus E	Genus G	
Genus B	Genus D	
⋮	⋮	

Figure 2.2: In each genus, we sorted in descending order of abundance, and calculated the cluster to which the top 10 % sample belongs (A). We conducted the above for all genus and picked up 10 genus that had a high ratio in each cluster (B). Their bacteriological properties were investigated.

3 Result

3.1 Variable Reduction

The top 2048 important OTUs, which is based on MeanDecreaseAccuracy criteria in RF had the lowest OOB error rate as shown in Figure 3.1. Therefore, we used those 2048 OTUs for the downstream analysis. Selected 2048 OTUs had composed of 1459 genus.

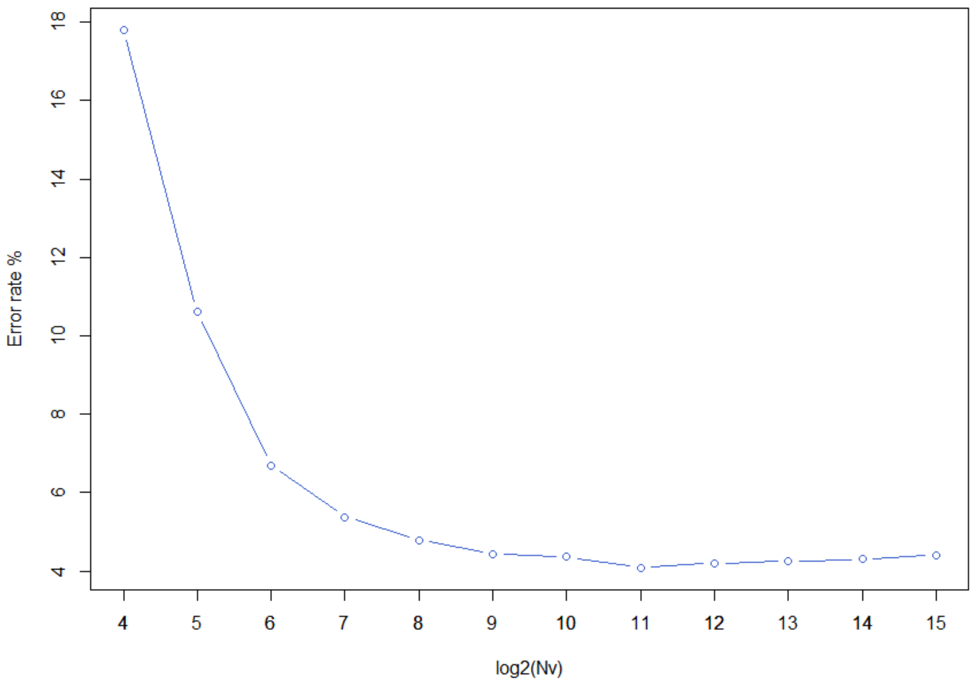


Figure 3.1: Nv is the number of variables. The error rate was the smallest when Nv was 2048.

3.2 Clustering

We calculated generalized UniFrac Distance among samples with selected OTUs and applied a hierarchical clustering method to systematize the difference of soil samples regarding the microbiome composition. We tentatively classified 4998 samples into 11 clusters as shown in the dendrogram Figure 3.2.

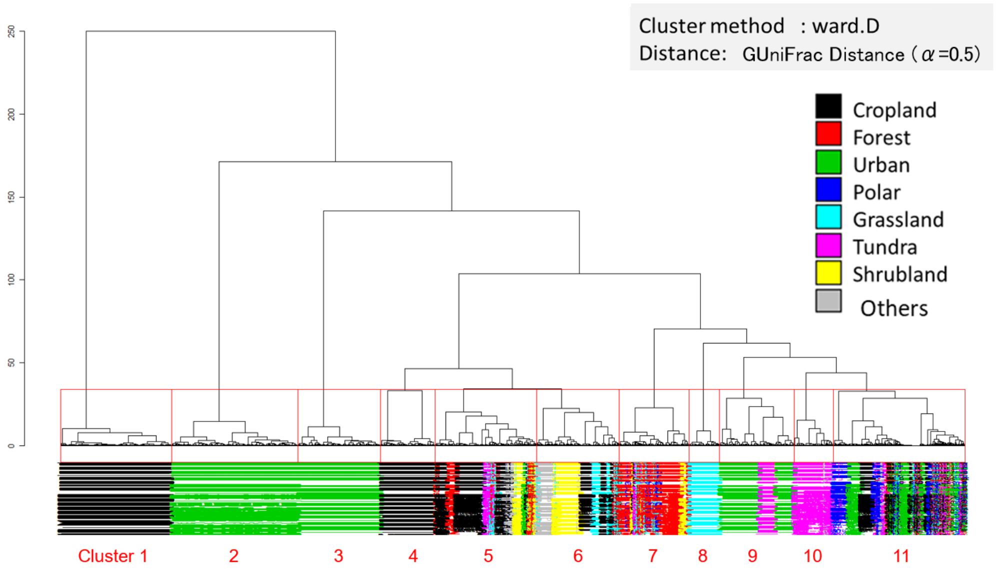


Figure 3.2: UniFrac distance based on ward's hierarchical clustering of soil microbiome. Land use categories represented by colors.

3.3 Clusters and Samples

All of cluster 1 samples were composed of soil samples of rice field in Japan. Cluster 2 samples were taken from the area in Fermilab Nature Area in Illinois, USA (http://www.fermilabnaturalareas.org). All samples in cluster 3 were sampled from biofilter (samples of sand from slow sand filter water purification system). Samples in cluster 4 were composed of soils of vineyard in the United States and France. Cluster 5 included Cropland, Tundra, Forest and Shrubland categories, and took from Nicaraguan coffee plantation and Tropical moist broadleaf forest in Kenya, and so on. Cluster 6 consisted of Cropland, Forest, Grassland, Shrubland, and others. Samples were collected from the dam, dried land soil in India, corn farm in Italy and Shrubland in Tanzania, and so on. In cluster 7, 74 % of the samples were from Forest categories. A wide range of type of forest samples was included, such as conifer forests in the USA, broadleaf forests in Canada, and tropical forests in Panama and Puerto Rico. Other categories in cluster 7 included Cropland, Grassland, Polar, Shrubland, Tundra. All of cluster 8 samples were composed of shrubland samples in Mongolian. Cluster 9 consisted of Alaskan tundra soil and biofilter. In cluster 10, samples taken from Alaska tundra and Mexican desert were included. Cluster 11 contained samples of all eight categories Table 3.1.

Table 3.1: Clusters and contained samples

Cluster	Description of contained samples
1 (614 samples)	Rice field in Japan. Rhizosphere soil
2 (698)	Urban soil in the USA (park)
3 (455)	Biofilter (samples of sand from slow sand filter water purification system)
4 (304)	Vineyard in USA and France Rhizosphere soil
5 (560)	Tropical moist broadleaf forest in Panama and Puerto Rico
	Coffee plantation, farm and rangeland soil in Nicaragua Tropical moist broadleaf forest in Kenya Barley cropland in Australia Californian Grassland soil Tundra in Greenland, Alaska Temperate grasslands, Savannas, and Shrubland in British Columbia in Canada Tanana Valley Forrest in the USA Tropical shrubland in Hawaii Garden and park soil in Manhattan, Brooklyn, and Staten Island (NY) Grassland and agricultural soil in the UK Montane shrubland in China
6 (457)	Shrubland in Tanzania Montane grassland in Mongolia Dam in Utah in the USA Agricultural field in Texas Maize field in Italy Wheat, soy filed in Ontario in Canada Grassland, Forest, dry soil in Minnesota, Nebraska and so on in the USA Tropical shrubland in Hawaii Dry soil in India

Cluster	Description of contained samples
7 (386)	Coniferous forest in Oregon Forest soil in Malaysia Lambir National Park Agricultural soil in the UK Tundra in Alaska Tanana Valley Forest in Alaska Forest soil in USA, Puerto Rico and Peru Polar desert Temperate broadleaf and mixed forest in Canada Tundra in Alaska Tropical shrubland in Hawaii Forest in Malaysia Forest in Panama Tropical moist broadleaf forest in Panama and Puerto Rico Temperate grasslands, savannas, and shrubland biome in Canada
8 (168)	Montane grassland in Mongolia
9 (413)	Biofilter (samples of sand from slow sand filter water purification system) Tundra in Alaska
10 (215)	Tundra in Alaska Desert soil in México
11 (728)	Polar desert Tundra Barley cropland in Australia Urban soil in the USA Paddy soil in Japan Oil contaminated soil in Polar Biofilter (samples of sand from slow sand filter water purification system) Garden and park soil in Manhattan, Brooklyn, and Staten Island (NY)

Cluster	Description of contained samples
	Dam Tropical shrubland in Hawaii Surface soil collected near from Brazilian Antarctic Station Comandante Ferraz And so on

Table 3.2: Clusters and land-use categories

	Cropland	Forest	Urban	Polar	Grassland	Tundra	Shrubland	Others
Cluster 1	614							
Cluster 2			698					
Cluster 3			455					
Cluster 4	304							
Cluster 5	284	86	23		18	47	68	34
Cluster 6	115	8			84		153	97
Cluster 7	4	286		10	18	13	55	
Cluster 8					168			
Cluster 9			325			88		
Cluster 10					1	207		7
Cluster 11	166	25	278	151	25	59	2	22

3.4 Clusters and Bacteria

The genus name and Proportion, which was the top 10 in the statistical test, were as shown in the table. Clusters 1, 2, 3, 4 had genera with a high proportion exceeding 80 %. Cluster 1 had the top 32 genera had more than 90 % ratio. The functions and characteristics of the bacteria significantly contained in each cluster are as follows Table 3.3 - 3.13.

Table 3.3: Top 10 genera in cluster 1

Genus	Ratio	Characteristics	Ref
Methanocella	0.99	Methanogenic archaea isolated from paddy in Japan	[24]
Methanolinea	0.99	Produce methane under strictly anaerobic condition	[25]
Methylosarcina	0.99	Anaerobic methanogens that produce methane	[26]
DCE29	0.99		
Methanomassiliicoccus	0.99	Methanogen	[27]
Anaeromyxobacter	0.99	All strains share is the ability to reduce soluble and amorphous ferric iron as well as other oxidized metal species	[28]
Methanospirillum	0.99	Include methane-producing archaeon isolated from puddly soil	[29]
Candidatus Methanoregula	0.98	Methanogen	[30]
SHD.14	0.98		
Methylomonas	0.98	Live in water where methane exists. It has methane monooxygenase and energy can be obtained by oxidizing methane and methanol	[31]
Desulfomonile	0.98	Strict anaerobic and sulfate-reducing bacterium	[32]
Sporomusa	0.98	Appear to involve in the first step for methanogenic degradation in paddy field	[33]
Desulfovirga	0.98	Sulfate-reducing bacterium	[34]
Anaerolinea	0.96		
SJA.88	0.96		

Genus	Ratio	Characteristics	Ref
Anabaena	0.96	Cyanobacteria and can be used for fixing nitrogen in paddy fields	[35]
Desulfococcus	0.96	Sulfate-reducing bacteria and live in water	[36]
Methanosaeta	0.96	Methanogenic archaea and use Acetate	[37]
BSV43	0.96		
Blvii28	0.95		
Formivibrio	0.95		
Magnetospirillum	0.94		
Chlamydomonas	0.94	Photosynthetic organisms	
Methylococcus	0.94	Exists in water in which methane is present, and oxidizes methane and methanol to obtain energy	[38]
G07	0.94		
Candidatus Methylomirabilis	0.93	Methanotrophs that metabolize methane as their only source of carbon and energy	[39]
Microvirgula	0.93		
Treponema	0.93		
GOUTA19	0.93		
Syntrophobacter	0.92	Decompose propionic acid. Since growth is inhibited in the presence of hydrogen, both hydrogen-consuming bacteria such as methane bacteria and sulfate-reducing bacteria must be present	[40]
WCHB1.84	0.91		
Methylocaldum	0.90	Methane-oxidizing bacteria	[41]

Table 3.4: Top 10 genera in cluster 2

Genus	Ratio	Characteristics	Ref
DA101	0.82		
Beijerinckia	0.80	Non-symbiotic, aerobic and nitrogen-fixing bacterium that inhabits the soil and the leaf area. Glucose, fructose, and sucrose are used as carbon sources.	[42]
Xenophilus	0.71		
Herbidospora	0.70		
Georgfuchsia	0.69	Strictly anaerobic betaproteobacterium	[43]
Asteroleplasma	0.60	Anaerobic bacteria	[44]
Rhodopila	0.59	Anoxygenic phototrophic bacteria and growth preferably under anaerobic conditions in the light but can grow aerobically in the dark	[45]
Blastomonas	0.57	Photoheterotrophic, strictly aerobic bacteria	[46]
Actinocorallia	0.56	Aerobium	[47]
Pilimelia	0.55	Aerobium	[47]

Table 3.5: Top 10 genera in cluster 3

Genus	Ratio	Characteristics	Ref
Nitrosopumilus	0.86	Oxidize ammonia to nitrite	[48]
Synechococcus	0.66	Autotrophic organism and the main source of primary production in poor nutrition	[49]
Hyphomicrobium	0.66	Performs denitrification with methanol and formic acid as a carbon source	[50]
Nitrospira	0.65	Nitrite-oxidizing bacteria	[51]
Polynucleobacter	0.64		
Pseudanabaena	0.62	The dominant species in the reservoir	[52]
Phaselicystis	0.62		
Candidatus Rhodoluna	0.62		
Pedomicrobium	0.60	Ubiquitous bacterium dominant in biofilms of man-made aquatic environments such as water distribution systems and bioreactors	[53]
Chthoniobacter	0.58		

Table 3.6: Top 10 genera in cluster 4

Genus	Ratio	Characteristics	Ref
Steroidobacter	0.44	Abundant microbiota of grapevine root reported in previous research	[16]
Glycomyces	0.43	A relatively minor actinomycete isolated from plant roots in farm soils.	[54]
Niastella	0.41		
Planctomyces	0.37	Abundant microbiota of grapevine root reported in previous research	[16]
Variovorax	0.36	Include plant-growth-promoting rhizobacteria species	[55]
Skermanella	0.35	The dominant bacteria in grapevine soil	[56]
Lacibacter	0.34		
Sarcandra	0.33		
Cellvibrio	0.32		
Aeromicrobium	0.32		

Table 3.7: Top 10 genera in cluster 5

Genus	Ratio	Characteristics	Ref
Solirubrobacter	0.62	Include species isolated in a farm in America and ginseng soil in Korea	[57]
Kribbia	0.56		
Actinoallomurus	0.48	Produce antibacterial or antifungal compounds	[58]
Streptacidiphilus	0.44		
Lapillicoccus	0.44		
Planococcus	0.43		
Knoellia	0.43	Aerobic or microaerophilic bacteria	[59]
Labrys	0.43		
Mesorhizobium	0.41	Root nodule bacteria	[60]
Amycolatopsis	0.41		

Table 3.8: Top 10 genera in cluster 6

Genus	Ratio	Characteristics	Ref
Rubrobacter	0.69	Among actinomycetes, there are many bacterial species that are aerobic and resistant to radiation	[61]
Sciscionella	0.61	Aerobic, marine actinomycete	[62]
Chloroflexus	0.60	Photosynthetic bacteria and live in various kinds of environments such as hot springs, lakes, river water, sediments, and oceans and high salinity environment	[63]
Roseiflexus	0.60	Live photoheterotrophically in anaerobic conditions or chemoheterotrophically under the dark aerobic conditions	[64]
Actinopolyspora	0.60	Isolated from the saline and arid surroundings of an oil field in the Sultanate of Oman	[65]
Devriesea	0.60		
Planomonospora	0.60	Some species are capable of tolerating high salinity	[66]
Candidatus Chloracidobacterium	0.60		
Succinivibrio	0.59		
Calditerrivibrio	0.59		

Table 3.9: Top 10 genera in cluster 7

Genus	Ratio	Characteristics	Ref
Candidatus Xiphinematobacter	0.57		
Pedosphaera	0.56		
Acidicapsa	0.55	Include species that isolated from acidic soil of a deciduous forest	[67]
Acidophila	0.55		
Granulicella	0.54		
Nevskia	0.52		
Xanthobacter	0.44		
Nitrobacter	0.43	Nitrite-oxidizing bacteria and reported to be robust to lower pH than other nitrite-oxidizing bacteria	[68]
Burkholderia	0.43	Include species that have potential for agricultural or environmental purposes, such as biological control	[69]
Candidatus Koribacter	0.42		
Acidocella	0.41	Acidophilic bacteria	
Mucilaginibacter	0.40		
Candidatus Solibacter	0.40		
Acidisoma	0.39	Acidophilic bacteria	

Table 3.10: Top 10 genera in cluster 8

Genus	Ratio	Characteristics	Ref
Serratia	0.33	Live in anaerobic environment and include pathogens	
Paenibacillus	0.33	Produce antimicrobial substances to suppress pathogens	[70]
Bacillus	0.32	Contain so many species universally present in water, the soil and so on. Many species adapt to various extreme environments such as high pH, low temperature, high salt concentration and high pressure	[71]
Erwinia	0.32	Facultative anaerobic bacteria and include many phytopathogen	[72]
Rahnella	0.29		
Lysinibacillus	0.28		
Gluconacetobacter	0.26		
JG37.AG.70	0.24		
Yersinia	0.23	Include a facultative intracellular pathogen of mammals	[73]
Raoultella	0.21		

Table 3.11: Top genera in cluster 9

Genus	Ratio	Characteristics	Ref
Delftia	0.56	Have the ability of Extracellular electron transfer	[74]
Armatimonas	0.54		
Flavobacterium	0.52		
Rhodobacter	0.52	Grow photosynthetically in heavy metal contaminated environments	[75]
Ramlibacter	0.49		
Acidovorax	0.49	Phytopathogen	[76]
Haliscomenobacter	0.49		
Novosphingobium	0.47		
Arthrobacter	0.46		
Pelomonas	0.46		

Table 3.12: Top genera in cluster 10

Genus	Ratio	Characteristics	Ref
Demequina	0.40		
Paludibacter	0.38	Strictly anaerobic and chemoorganotrophic bacteria	[77]
Acetobacterium	0.36	Make Hydrogen oxidized and carbon dioxide reduced to acetic acid.	[78]
Sterolibacterium	0.34		
Propionispira	0.33		
Methanobacterium	0.33	Anaerobic methanogenic bacteria	[79]
Ethanoligenens	0.32		
Cellulomonas	0.30		
Crenothrix	0.29	Belong to the iron bacteria and consume methane	[80]
Actinotalea	0.28		

Table 3.13: Top 10 genera in cluster 11

Genus	Ratio	Characteristics	Ref
Iamia	0.40		
HB2.32.21	0.39		
B.42	0.38		
HTCC	0.36		
Marinobacter	0.36	Found in sea water and A number of species can degrade hydrocarbons	[81]
Segetibacter	0.36		
Rubricoccus	0.36		
Erythrobacter	0.35		
HTCC2207	0.35		
Ardenscatena	0.34		

Table 3.14: Shannon diversity in the clusters

Cluster	S
1	6.00
2	5.05
3	5.03
4	5.10
5	5.09
6	4.87
7	4.43
8	2.74
9	3.87
10	4.14
11	2.76

3.5 Climate Zone

We calculated the percentage of clusters in each climate division, as shown in Table 3.15. Samples of clusters 5, 6, 7, and 11 were collected from a wide area and therefore included a wide range of climate zones. Especially arid in cluster 6 and polar in cluster 11 were included with high rate.

Table 3.15: Cluster proportion of each climate zone

Cluster	Tropical	Arid	Temperate	Continental	Polar
1	0%	0 %	1 %	0 %	0 %
2	0%	0 %	0 %	20 %	0 %
3	0%	0 %	2 %	0 %	0 %
4	0%	0 %	43 %	5 %	0 %
5	42%	2 %	11 %	18 %	3 %
6	39%	83 %	23 %	16 %	0 %
7	16%	0 %	11 %	15 %	9 %
8	0%	0 %	0 %	7 %	0 %
9	0%	0 %	0 %	5 %	0 %
10	0%	2 %	0 %	3 %	6 %
11	3%	13 %	10 %	11 %	82 %
	100 %	100 %	100 %	100 %	100 %

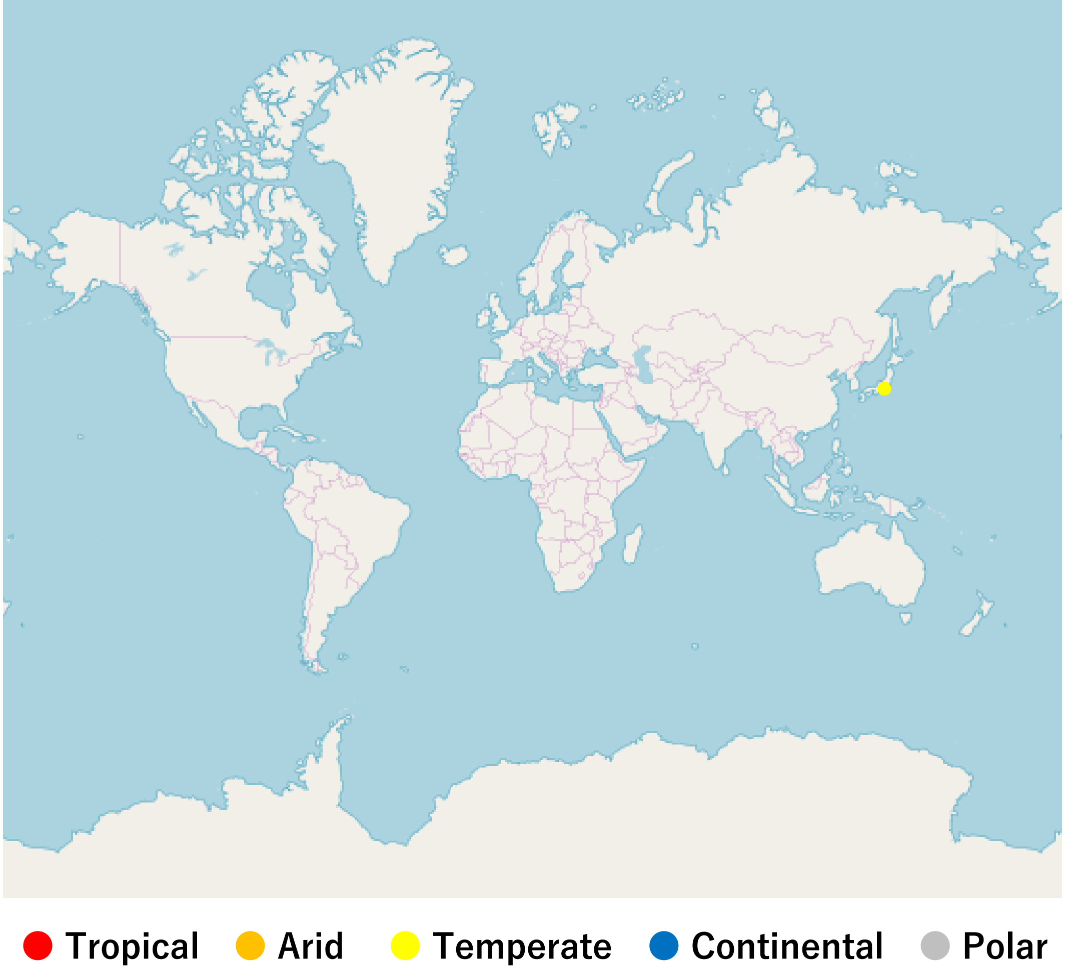


Figure 3.3: Climate zones of cluster 1

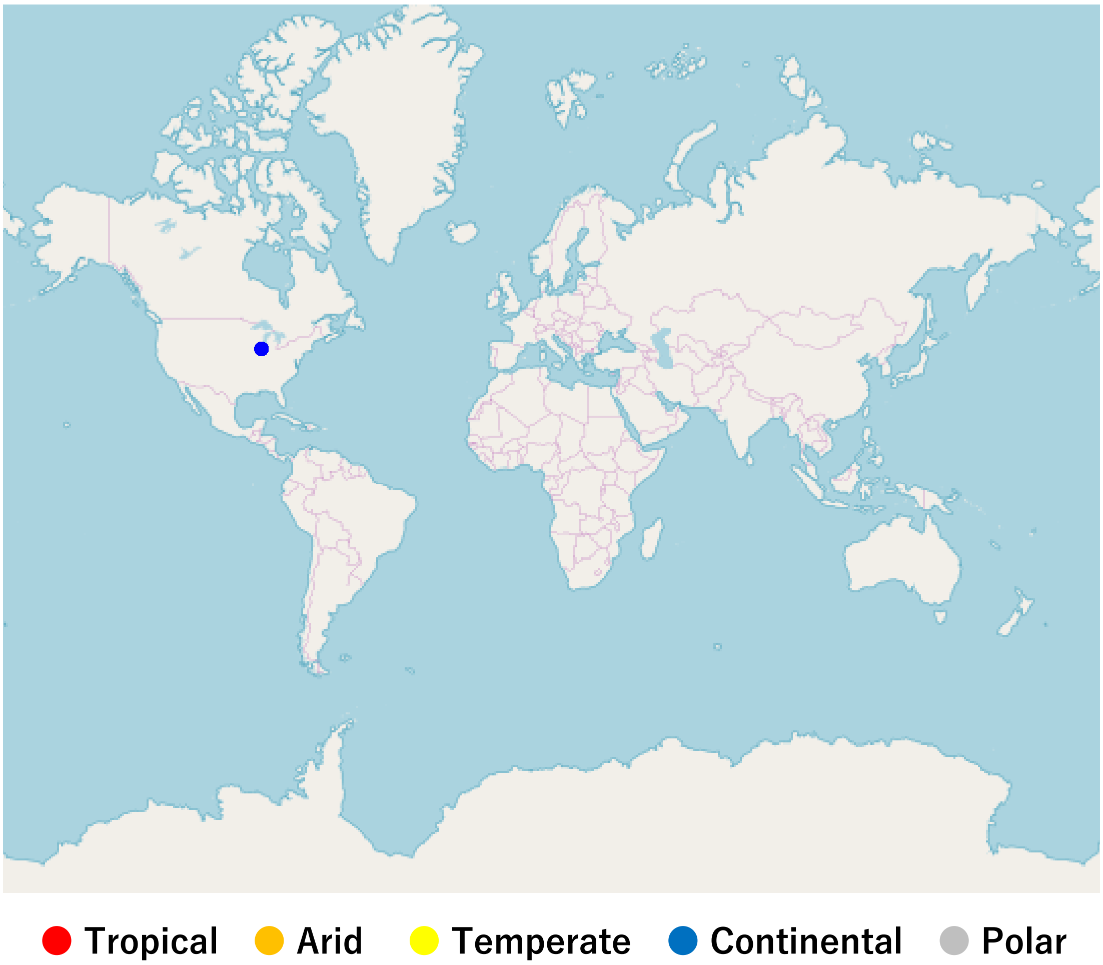


Figure 3.4: Climate zones of cluster 2

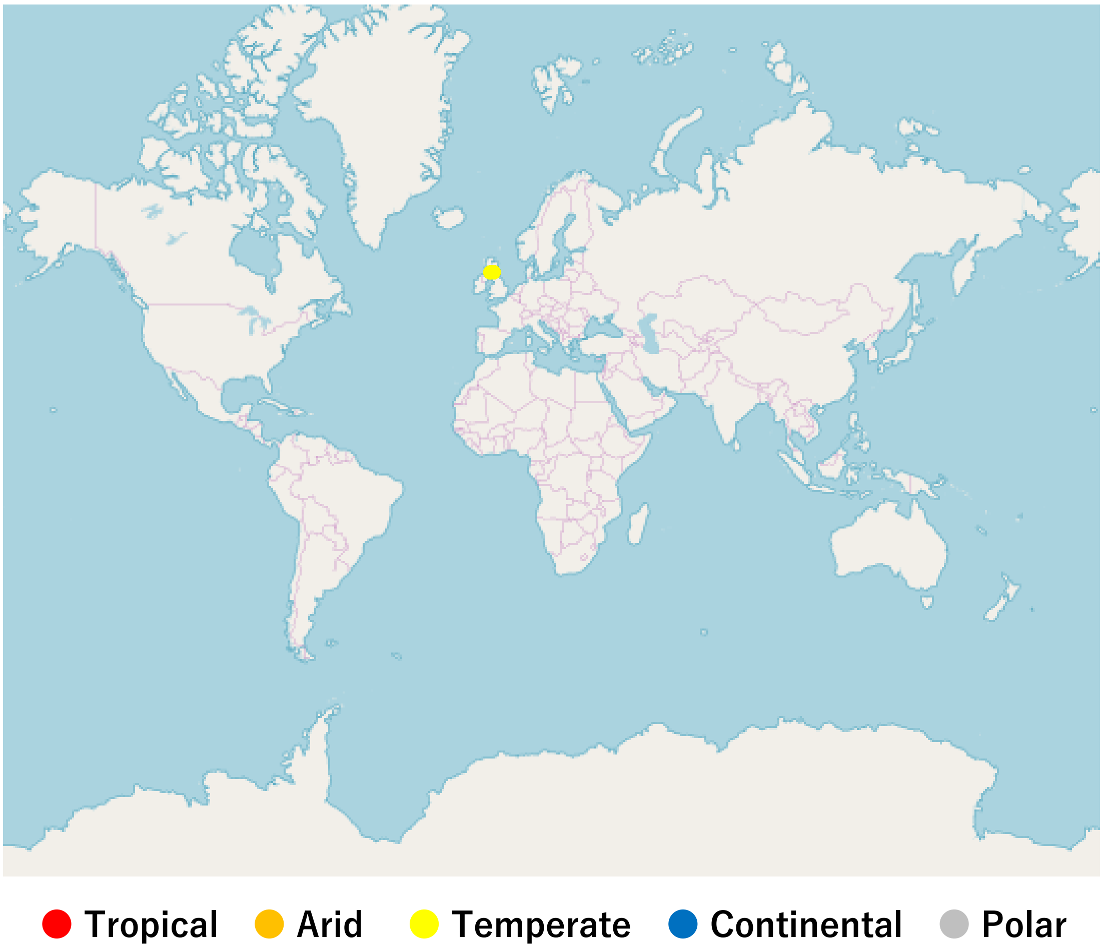


Figure 3.5: Climate zones of cluster 3

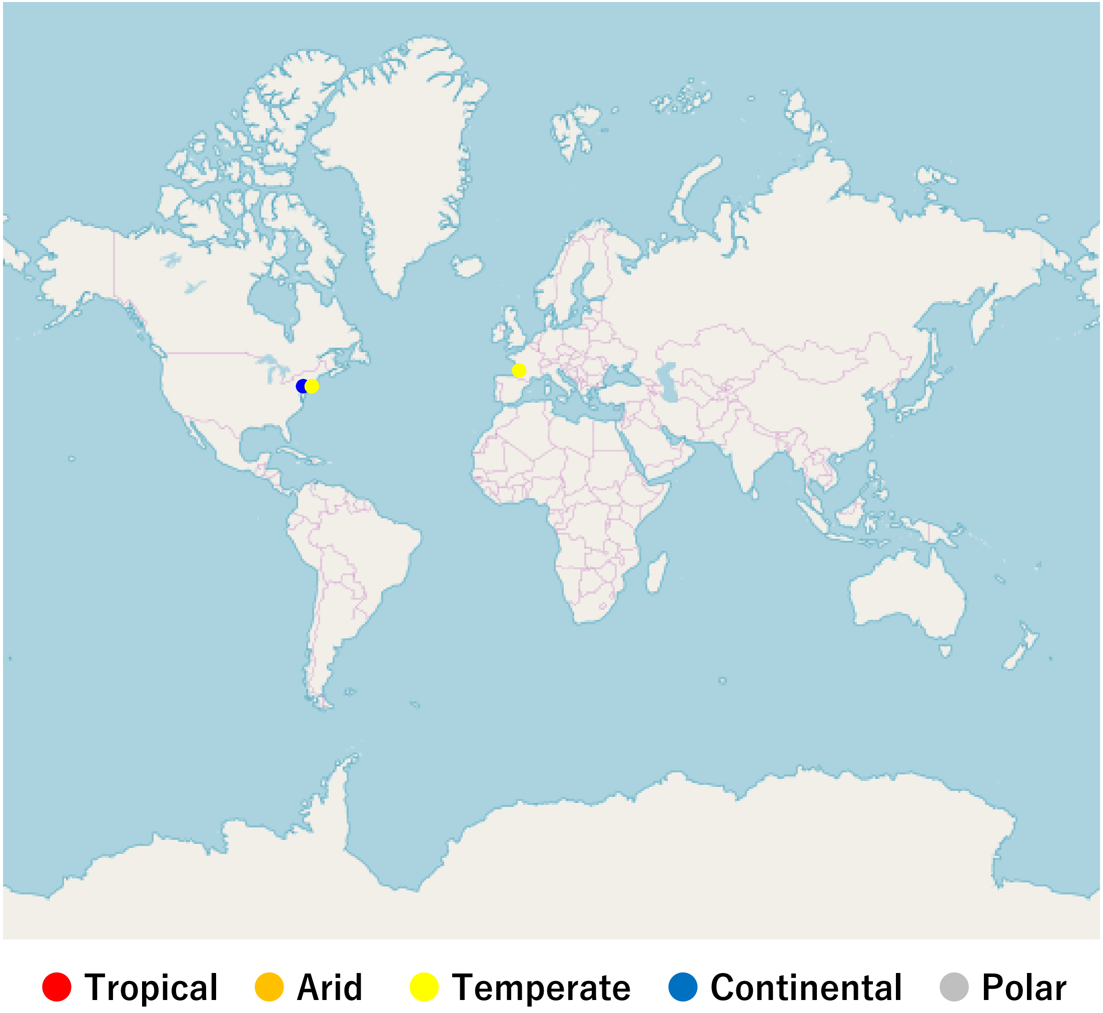


Figure 3.6: Climate zones of cluster 4

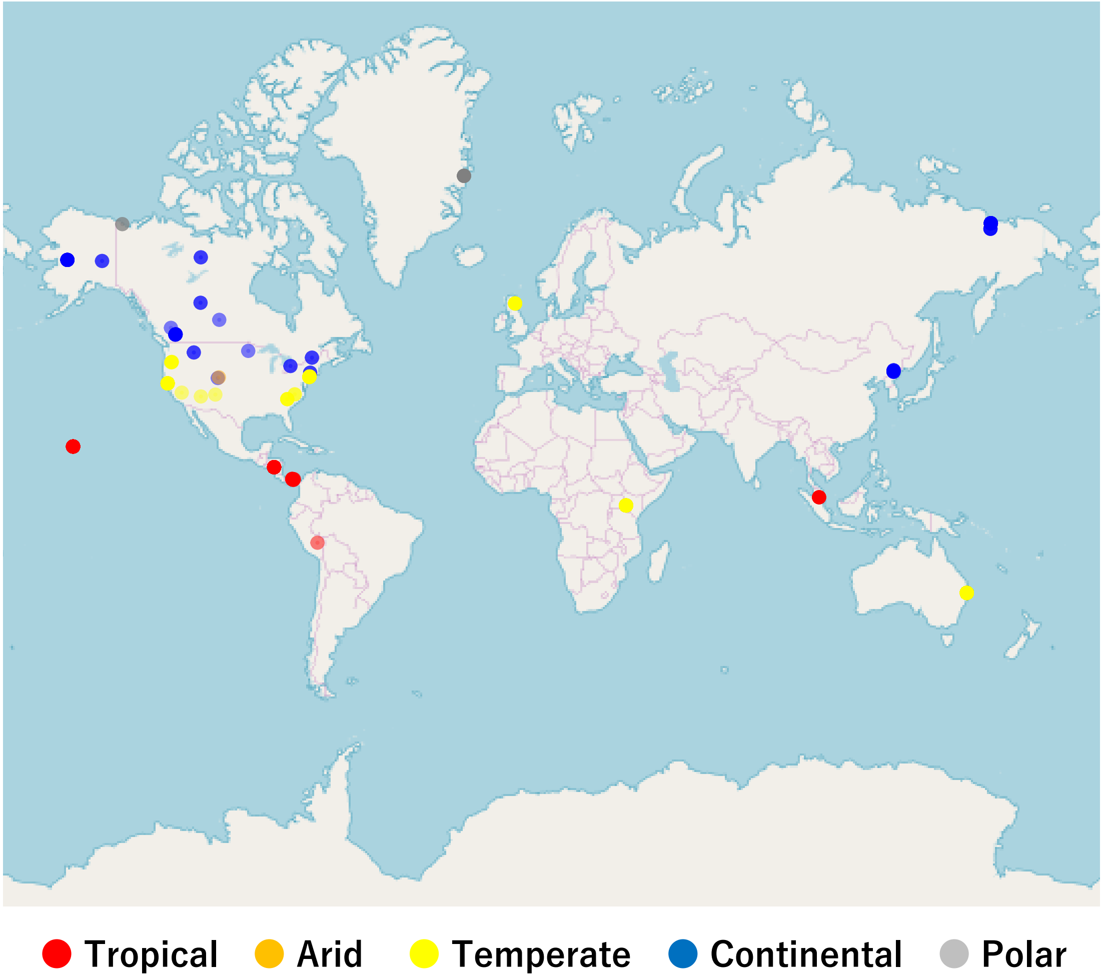


Figure 3.7: Climate zones of cluster 5

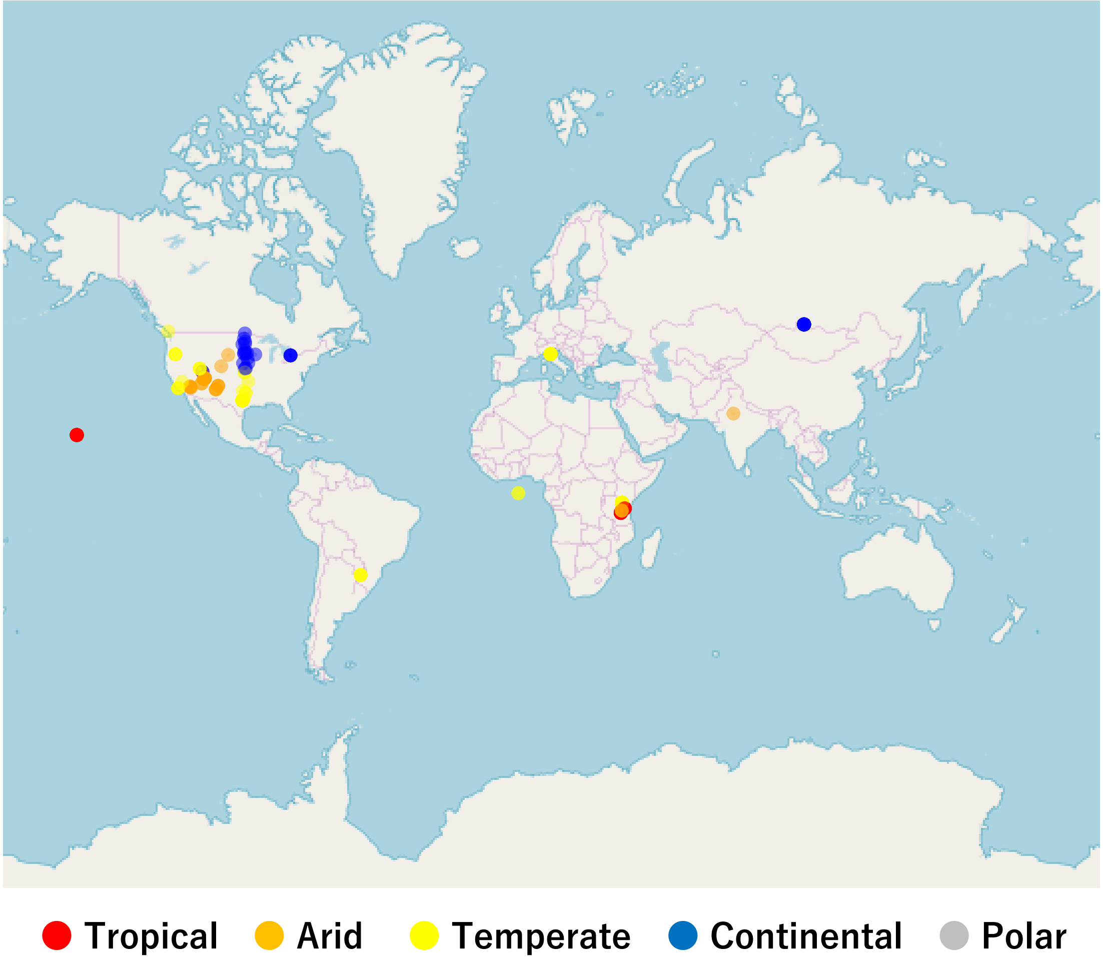


Figure 3.8: Climate zones of cluster 6

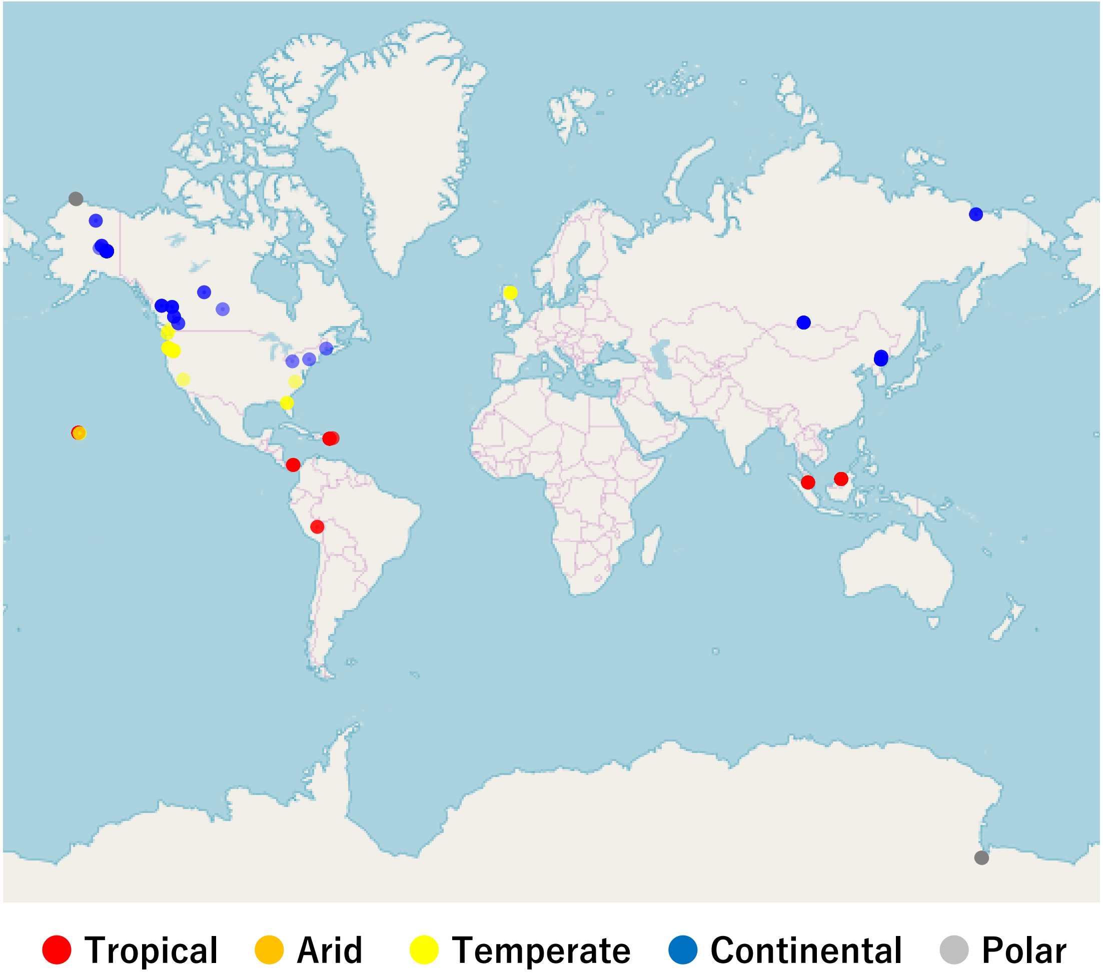


Figure 3.9: Climate zones of cluster 7

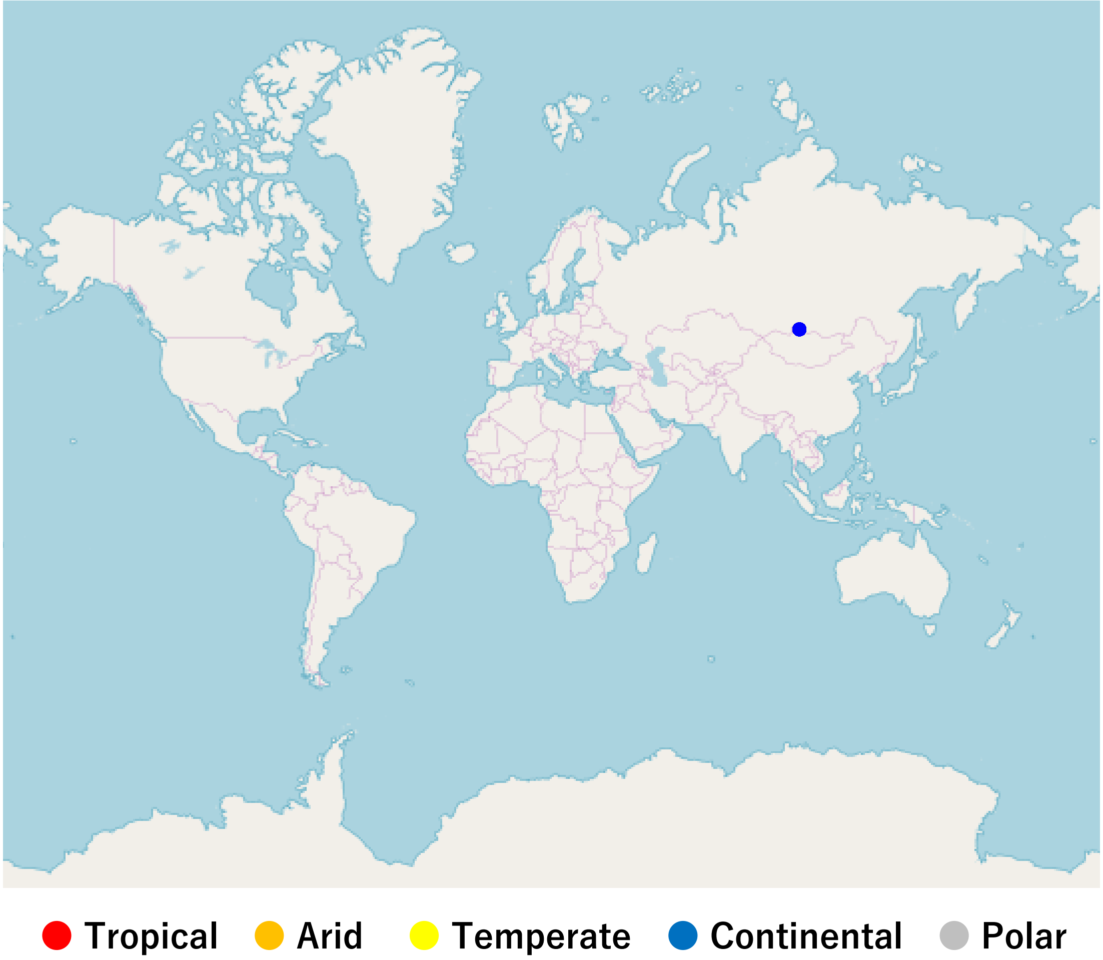


Figure 3.10: Climate zones of cluster 8

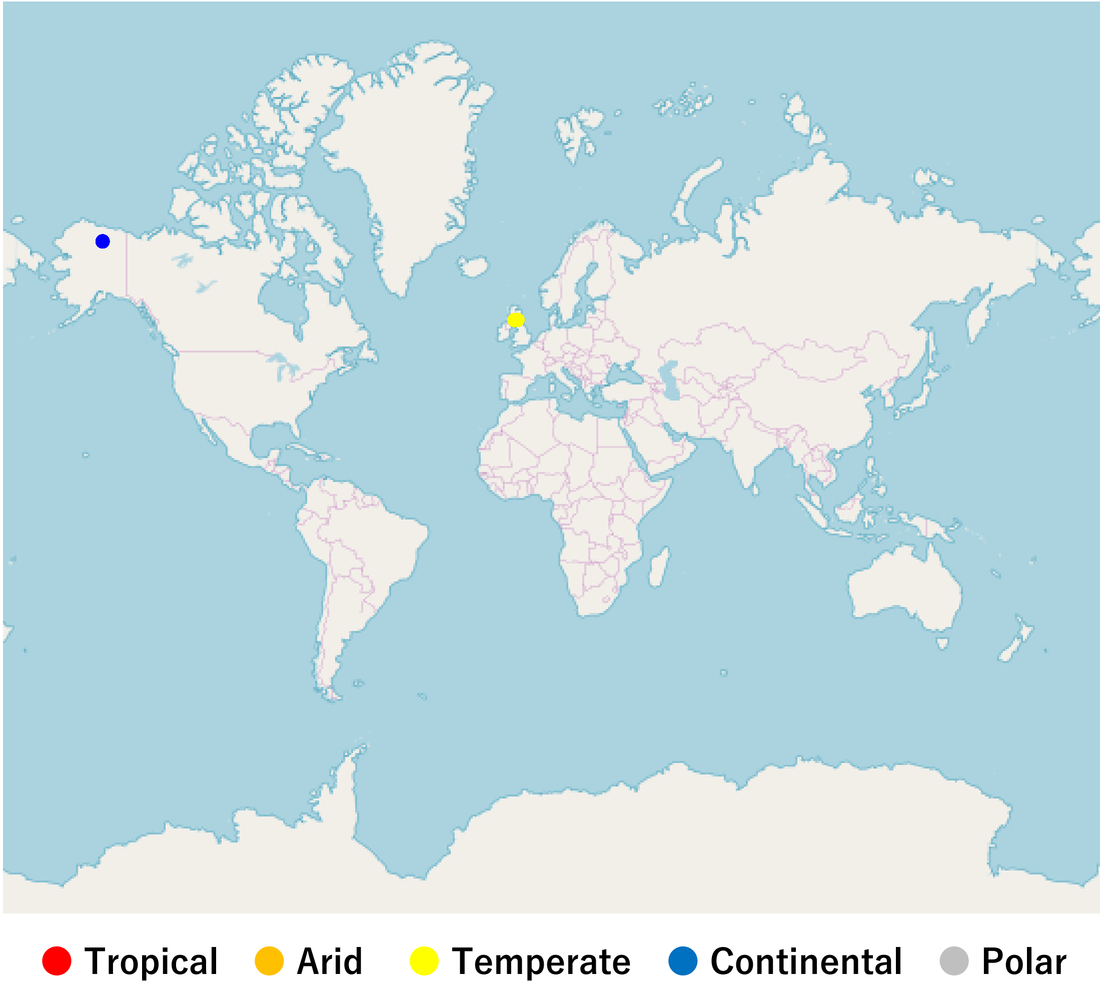


Figure 3.11: Climate zones of cluster 9

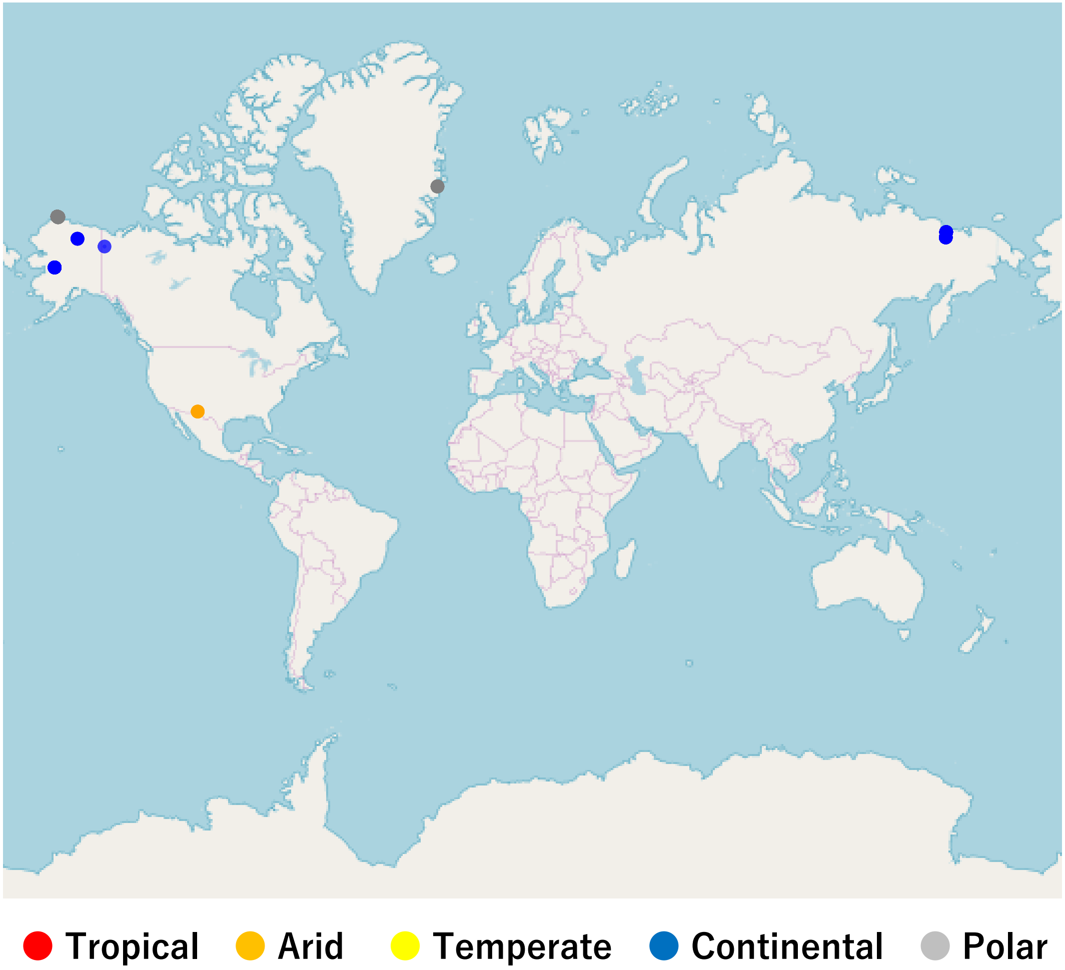


Figure 3.12: Climate zones of cluster 10

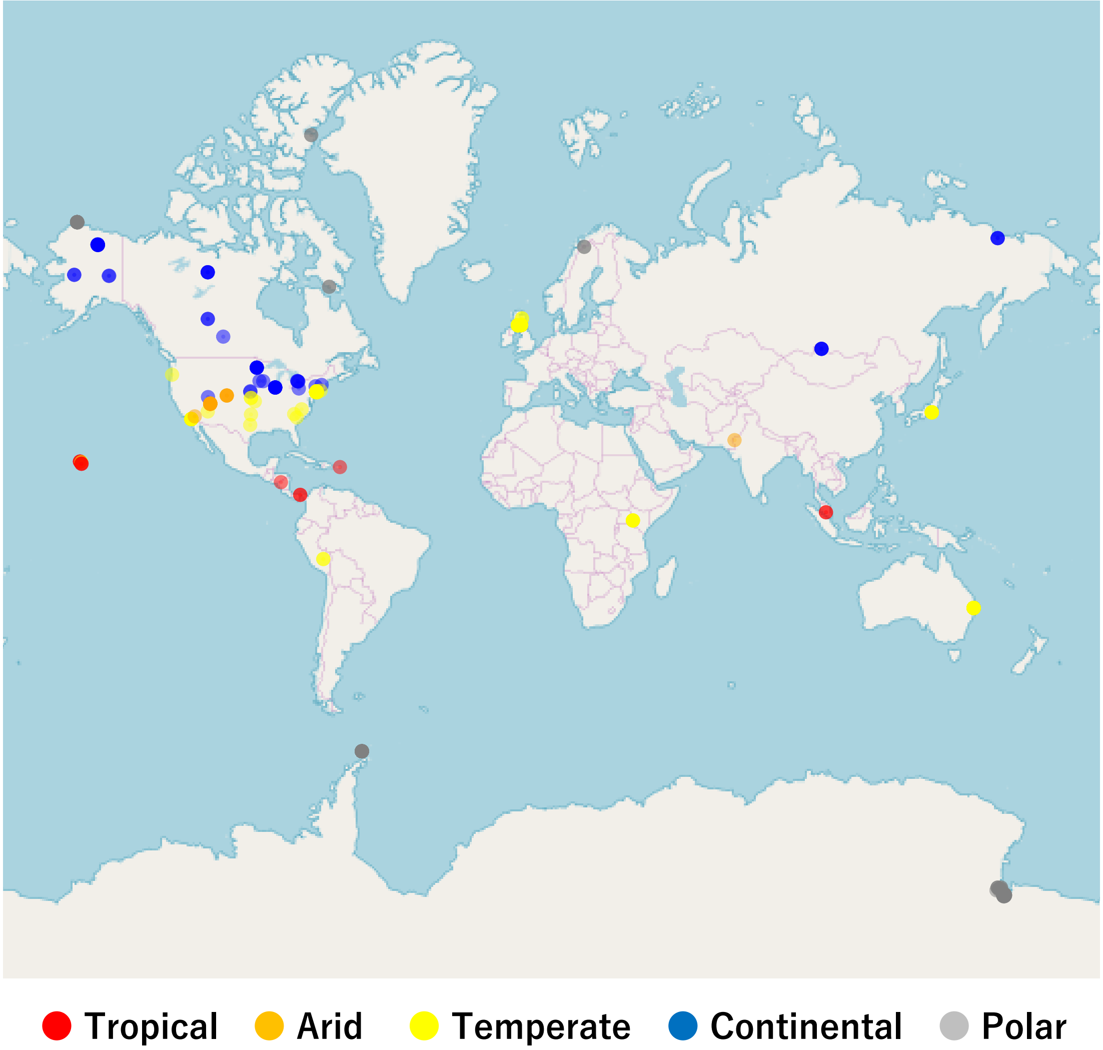


Figure 3.13: Climate zones of cluster 11

4 Discussion

4.1 Ecological Interpretation of Clustering

We did Ecological interpretation by characterizing each cluster from the function and characteristics of bacteria. For clusters 1, 2, 3, 4, 6, 7, and 8, the ecological interpretation was made. On the other hand, clusters 5, 9, 10, and 11 can not be ecologically interpreted. We expect those interpretation will be brought by the discovery of new features and ecological feature of bacteria in the future.

4.1.1 Cluster 1

Methanocella, Methanolinea, Methylosarcina, and Methanomassiliicoccus belong to Methanogen and produce methane in an anaerobic environment. Methylomonas, Methylococcus and Candidatus Methylomirabilis belongs to methanotrophs that grow on methane as their sole source of carbon and energy. Desulfomonile, Desulfococcus is a sulfate-reducing bacterium, which uses sulfate as the final electron acceptor and reduces it to hydrogen sulfide. Anabaena and Chlamydomonas is an autotrophic organism that photosynthesis with cyanobacteria. In paddy fields, methane production reaction occurs in anaerobic environment using fertilizer and organic matter produced by cyanobacteria Methanogen is involved in the process. Generated methane is released as a gas, but a part of it is decomposed by methane oxidizing bacteria. The electrons obtained by the anaerobic methane oxidation are used for the reduction of iron and sulfate, involving iron reducing bacteria, sulfate and nitrate reducing bacteria. Based on the above, the significant genus in Cluster 1, abundant genus was a bacterial flora peculiar to paddy field environment [82].

4.1.2 Cluster 2

Beijerinckia performs nitrogen fixation to glucose or fructose as a carbon source in an aerobic environment. Blastomonas is an aerobic photo heterotrophic bacterium. In addition, Asteroleplasma, Rhodopila, Actinocorallia belong to aerobic bacteria. Characteristics of soil bacteria in cluster 2 are aerobic bacteria. The aerobic condition is generally a good feature of soil from the viewpoint of plant growth. Cells of plant roots absorb oxygen and breathe to discharge carbon dioxide. If oxygen is not present or at a low level in the soil, the plant would become oxygen deficient and adversely affects the growth. When anaerobic bacteria propagate in the soil, sulfate ions are reduced to form hydrogen sulfide or iron sulfide causing root rot. Therefore, it is important to maintain the soil in an aerobic environment, and measures such as improvement of breathability and drainage and fertilization of soil improvement are necessary. The sample in Cluster 2 is a natural park in the state of Illinois, USA, and the environment where wild birds such as wild birds can live by the activities of many volunteers is maintained. It is thought that the soil is managed as well, and it is considered that the aerobic bacterial flora was characteristic, reflecting it.

4.1.3 Cluster 3

Synechococcus is an autotrophic organism that can live even under poor nutrition. Hyphomicrobium and Nitrosopumilus utilize organic substances (methanol, ammonia, formic acid, etc.) with a simple chemical structure. In addition, A (the dominant species in the reservoir) and B (reported ubiquitous bacterium dominant in biofilms of man-made aquatic environments) were significantly abundant. Samples in cluster 3 were a sample taken from the water purification system and were expected to be the non-natural environment and free of organic matter. In order to adapt to such an environment, it seems to be building a network in which autotrophs and bacteria using simple compounds coexist.

4.1.4 Cluster 4

It was reported that Steroidobacter and Planctomyces existed abundantly in a comparative analysis limited to wine and it was confirmed in this work that the comparison scale analysis on the global scale including other kinds of soil also showed the same result. Skermanella has been analyzed in the winery regions in China and reported to be abundant in the soil. This suggests the existence of a group of bacteria inherent in the wine area regardless of location. Planctomyces is a bacterium isolated by farm and Variovorax include a Plant growth promoting rhizobacteria spicies. These seem to be bacteria characteristic of farm crops, not only wine.

4.1.5 Cluster 6

Sciscionella, Chloroflexus, Actinopolyspora, Planomonospora have salt tolerance such as living in sea water. Samples in Cluster 6 included American dams, corn farms in Italy, wheat in Canada, dry soil in India, shrublands in Tanzania, and so on. Maize belongs to C4 plants that can maintain high photosynthetic activity even under strong light, high temperature, and dryness, and prefers relatively dry soil. Wheat and soybeans are vulnerable to over-humidity. Salt accumulation occurred in dry soil, which is considered to be the reason that salt tolerant bacteria were significantly abundant.

4.1.6 Cluster 7

Acidicapsa, Acidopila, Acidocella, Acidisoma isolated and live in an acidic environment. Nitrobacter belongs to nitrite-oxidizing bacteria and reported to be robust to lower pH than other nitrite-oxidizing bacteria. 74 % of the samples in cluster 7 were samples of the forest category. For some reasons, forest soil seems to be acidic. One is the pH adjustment. It is performed regularly for the growth of crops in agricultural land, but not in forest soil, so the forest soil is generally much more acidic than agricultural land. In addition, forest soil derived from volcanic ash is acidic. Furthermore, today deforestation is proceeding, affecting depletion of organic matter and soil buffering effect and causing soil acidification. The soil samples in cluster 7 are acidic and it seems to be reflected in the microbial layer.

4.1.7 Cluster 8

Serratia, Erwinia, Yersinia are anaerobic and pathogenic bacteria. Paenibacillus produce antimicrobial substances to suppress pathogens. Samples in cluster 8 had the smallest shannon divesity among clusters (Table 3.14). Under anaerobic conditions with low bacterial diversity, pathogenic bacteria tend to propagate. It is considered that cluster 8 soil have an environment where pathogenic bacteria tend to reproduce because of no soil management.

4.2 Climate Zones and Clusters

Samples which belonged to arid climate zone were included in clusters 5, 6, 10, and 11, and 83 % of those samples belonged to cluster 6. It seems that the soil in the arid area is in a dry state, and it is related to the large abundance of salt-tolerant bacteria. Cluster 11 has features with a lower diversity index than other clusters, and 82 % of the polar climate zone was included. It seems that less abundance of bacteria is involved.

5 Conclusion

There are many types and numbers of bacteria in the soil, forming a community called microbiome. Compared to the ecosystem of plants and animals, we still know little about soil microbial ecosystem. How the soil microbiomes are different throughout the world and how they relate to the region and the environment is a major interest. The comparative analysis of microbiome had performed on a small spatial scale in a limited environment or region, or on a global scale using only limited indicators and taxa. In this work, we compared soil microbiomes collected from various regions and environments on a global scale. We used EMP database did clustering using the distance based on bacterial systematic distance. We clarified the bacteria characteristic to clusters by a statistical test and examined the result of clustering from the functions and ecological features. The bacteria that were significant in the paddy field and the cluster of the vineyard were common to the bacteria mentioned in the previous study. In addition, we have revealed a group of bacteria characteristic of Mongolian grasslands, forests, bio filters and others. Furthermore, we investigated the relationship between clusters and climate zones. This research is expected to deepen the understanding of the ecology of the soil bacterial flora and lead to knowledge for soil management based on bacteria.

Acknowledgements

計算システムズ生物学研究室では充実した研究生活を送ることができました。
ブラジル、サンパウロ大学への訪問、カナダでの学会発表など、貴重なチャンスをくださり、自分のペースで研究をのびのびできる環境を与えてくださった、金谷教授に深く感謝申し上げます。バイオ系出身で、研究で右も左もわからない中、小野准教授の助けなく研究を進めることができなかったと思います。ありがとうございました。ゼミなどで的確なアドバイスをくださったアミン准教授、黄助教に感謝申し上げます。一緒にマイクロバイオーム研究を行い、またカナダの学会発表やブラジル訪問で大変お世話になりました京都府立大学のアンドレ フレイリクルス先生に感謝申し上げます。筋トレやＡＩなどを私に勧め指南してくれた同学年の浅野幸之助には大変感謝しております。研究室のメンバー一同のおかげでよい研究室生活を送ることができました。感謝いたします。
最後に、大学院まで多大な支援をいただいた家族に深く感謝し、ここに謝辞とさせていただきます。

References

[1] Jovel, Juan, et al. "Characterization of the gut microbiome using 16S or shotgun metagenomics." Frontiers in microbiology 7 (2016): 459.

[2] Dixon, Ray, and Daniel Kahn. "Genetic regulation of biological nitrogen fixation." Nature Reviews Microbiology 2.8 (2004): 621.

[3] Burns, Richard G., and Julie A. Davies. "The microbiology of soil structure." Biological Agriculture & Horticulture 3.2-3 (1986): 95-113.

[4] Quince, Christopher, et al. "Accurate determination of microbial diversity from 454 pyrosequencing data." Nature methods 6.9 (2009): 639.

[5] Reeder, Jens, and Rob Knight. "Rapidly denoising pyrosequencing amplicon reads by exploiting rank-abundance distributions." Nature methods 7.9 (2010): 668.

[6] Quince, Christopher, et al. "Removing noise from pyrosequenced amplicons." BMC bioinformatics 12.1 (2011): 38.

[7] Edgar, Robert C., et al. "UCHIME improves sensitivity and speed of chimera detection." Bioinformatics 27.16 (2011): 2194-2200.

[8] Yilmaz, Pelin, et al. "The SILVA and "all-species living tree project (LTP)" taxonomic frameworks." Nucleic acids research 42.D1 (2013): D643-D648.

[9] Wang, Qiong, et al. "Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy." Applied and environmental microbiology 73.16 (2007): 5261-5267.

[10] McDonald, Daniel, et al. “An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea.” The ISME journal 6.3 (2012): 610.

[11] Federhen, Scott. “The NCBI taxonomy database.” Nucleic acids research 40.D1 (2011): D136-D143.

[12] Kung, His-Chung, et al. “Stratification of Human Gut Microiome and Building a SVM-Based Classifier.” 2018 IEEE 18th International Conference on Bioinformatics and Bioengineering (BIBE). IEEE, 2018.

[13] Ramirez, Kelly S., et al. “Biogeographic patterns in below-ground diversity in New York City’s Central Park are similar to those observed globally.” Proc. R. Soc. B 281.1795 (2014): 20141988.

[14] Fierer, Noah, et al. “Cross-biome metagenomic analyses of soil microbial communities and their functional attributes.” Proceedings of the National Academy of Sciences 109.52 (2012): 21390-21395.

[15] Maestre, Fernando T., et al. “Increasing aridity reduces soil microbial diversity and abundance in global drylands.” Proceedings of the National Academy of Sciences 112.51 (2015): 15684-15689.

[16] Zarraonaindia, Iratxe, et al. “The soil microbiome influences grapevine-associated microbiota.” MBio 6.2 (2015): e02527-14.

[17] Howard, Mia M., Terrence H. Bell, and Jenny Kao-Kniffin. “Soil microbiome transfer method affects microbiome composition, including dominant microorganisms, in a novel environment.” FEMS microbiology letters 364.11 (2017).

[18] Delgado-Baquerizo, Manuel, et al. “A global atlas of the dominant bacteria found in soil.” Science 359.6373 (2018): 320-325.

[19] Caporaso, J. Gregory, et al. “QIIME allows analysis of high-throughput community sequencing data.” Nature methods 7.5 (2010): 335.

[20] Lozupone, Catherine, and Rob Knight. “UniFrac: a new phylogenetic method for comparing microbial communities.” Applied and environmental microbiology 71.12 (2005): 8228-8235.

[21] Lozupone, Catherine A., et al. “Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities.” Applied and environmental microbiology 73.5 (2007): 1576-1585.

[22] Chen, Jun, et al. “Associating microbiome composition with environmental covariates using generalized UniFrac distances.” Bioinformatics 28.16 (2012): 2106-2113.

[23] Kottek, Markus, et al. “World map of the Köppen-Geiger climate classification updated.” Meteorologische Zeitschrift 15.3 (2006): 259-263.

[24] Sakai, Sanae, et al. “Methanocella paludicola gen. nov., sp. nov., a methane-producing archaeon, the first isolate of the lineage ‘Rice Cluster I’, and proposal of the new archaeal order Methanocellales ord. nov.” International Journal of Systematic and Evolutionary Microbiology 58.4 (2008): 929-936.

[25] Sakai, Sanae, et al. “Methanolinea mesophila sp. nov., a hydrogenotrophic methanogen isolated from rice field soil, and proposal of the archaeal family Methanoregulaceae fam. nov. within the order Methanomicrobiales.” International journal of systematic and evolutionary microbiology 62.6 (2012): 1389-1395.

[26] Wei, Meng, et al. “Methane oxidation and response of Methylobacter/Methylosarcina methanotrophs in flooded rice soil amended with urea.” Applied soil ecology 101 (2016): 174-184.

[27] Dridi, Bédis, et al. “Methanomassiliicoccus luminyensis gen. nov., sp. nov., a methanogenic archaeon isolated from human faeces.” International journal of systematic and evolutionary microbiology 62.8 (2012): 1902-1907.

[28] Sanford, Robert A., James R. Cole, and James M. Tiedje. “Characterization and description of Anaeromyxobacter dehalogenans gen. nov., sp. nov., an aryl-halorespiring facultative anaerobic myxobacterium.” Applied and environmental microbiology 68.2 (2002): 893-900.

[29] Tonouchi, Akio. “Isolation and characterization of a motile hydrogenotrophic methanogen from rice paddy field soil in Japan.” FEMS microbiology Letters 208.2 (2002): 239-243.

[30] Bräuer, Suzanna L., et al. “Methanoregula boonei gen. nov., sp. nov., an acidiphilic methanogen isolated from an acidic peat bog.” International journal of systematic and evolutionary microbiology 61.1 (2011): 45-52.

[31] Kalyuzhnaya, Marina G., et al. “Methylomonas scandinavica sp. nov., a new methanotrophic psychrotrophic bacterium isolated from deep igneous rock ground water of Sweden.” Systematic and applied microbiology 22.4 (1999): 565-572.

[32] Sun, Baolin, James R. Cole, and James M. Tiedje. “Desulfomonile limimaris sp. nov., an anaerobic dehalogenating bacterium from marine sediments.” International journal of systematic and evolutionary microbiology 51.2 (2001): 365-371.

[33] Kato, Souichiro, et al. “Methanogenic degradation of lignin-derived monoaromatic compounds by microbial enrichments from rice paddy field soil.” Scientific reports 5 (2015): 14295.

[34] Tanaka, Kazuhiro, et al. “Desulfovirga adipica gen. nov., sp. nov., an adipate-degrading, gram-negative, sulfate-reducing bacterium.” International journal of systematic and evolutionary microbiology 50.2 (2000): 639-644.

[35] Oh, Hee Mock, J. Maeng, and G-Yull Rhee. “Nitrogen and carbon fixation byAnabaena sp. isolated from a rice paddy and grown under P and light limitations.” Journal of applied phycology 3.4 (1991): 335-343.

[36] Janssen, Peter H., and B. Schnik. “Catabolic and anabolic enzyme activities and energetics of acetone metabolism of the sulfate-reducing bacterium Desulfococcus biacutus.” Journal of bacteriology 177.2 (1995): 277-282.

[37] Patel, Girishchandra B., and G. Dennis Sprott. “Methanosaeta concilii gen. nov., sp. nov.(“Methanothrix concilii”) and Methanosaeta thermoacetophila nom. rev., comb. nov.” International Journal of Systematic and Evolutionary Microbiology 40.1 (1990): 79-82.

[38] Patel, Ramesh N., and Derek S. Hoare. “Physiological studies of methane and methanol-oxidizing bacteria: oxidation of C-1 compounds by Methylococcus capsulatus.” Journal of bacteriology 107.1 (1971): 187-192.

[39] Wu, Ming L., et al. “Ultrastructure of the denitrifying methanotroph “ Candidatus Methylomirabilis oxyfera, ” a novel polygon-shaped bacterium.” Journal of bacteriology 194.2 (2012): 284-291.

[40] Wallrabenstein, Christina, Elisabeth Hauschild, and Bernhard Schink. “Syntrophobacter pfennigii sp. nov., new syntrophically propionate-oxidizing anaerobe growing in pure culture with propionate and sulfate.” Archives of Microbiology 164.5 (1995): 346-352.

[41] Takeuchi, Mio, et al. “Methylocaldum marinum sp. nov., a thermotolerant, methane-oxidizing bacterium isolated from marine sediments, and emended description of the genus Methylocaldum.” International Journal of systematic and evolutionary microbiology 64.9 (2014): 3240-3246.

[42] Dedysh, Svetlana N., and Peter F. Dunfield. “B eijerinckiaceae.” Bergey’s Manual of Systematics of Archaea and Bacteria (2015): 1-4.

[43] Weelink, Sander AB, et al. “A strictly anaerobic betaproteobacterium Georgfuchsia toluolica gen. nov., sp. nov. degrades aromatic compounds with Fe (III), Mn (IV) or nitrate as an electron acceptor.” FEMS microbiology ecology 70.3 (2009): 575-585.

[44] Robinson, I. M., and E. A. Freundt. “Proposal for an amended classification of anaerobic mollicutes.” International Journal of Systematic and Evolutionary Microbiology 37.1 (1987): 78-81.

[45] Imhoff, J. F., H. G. Trüper, and N. Pfennig. “Rearrangement of the species and genera of the phototrophic “ purple nonsulfur bacteria ” .” International Journal of Systematic and Evolutionary Microbiology 34.3 (1984): 340-343.

[46] Hiraishi, Akira, Hiroshi Kuraishi, and Kazuyoshi Kawahara. “Emendation of the description of Blastomonas natatoria (Sly 1985) Sly and Cahill 1997 as an aerobic photosynthetic bacterium and reclassification of Erythromonas

ursincola Yurkov et al. 1997 as Blastomonas ursincola comb. nov." International journal of systematic and evolutionary microbiology 50.3 (2000): 1113-1118.

[47] Ruan, J. "Bergey's Manual of Systematic Bacteriology Volume 5 and the study of Actinomycetes systematic in China." Wei sheng wu xue bao= Acta microbiologica Sinica 53.6 (2013): 521-530.

[48] Könneke, Martin, et al. "Isolation of an autotrophic ammonia-oxidizing marine archaeon." Nature 437.7058 (2005): 543.

[49] Christaki, U., et al. "Dynamic characteristics of Prochlorococcus and Synechococcus consumption by bacterivorous nanoflagellates." Microbial Ecology 43.3 (2002): 341-352.

[50] 加藤暢夫, and 緒方浩一. "C1 化合物資化性菌とその代謝." 化学と生物 14.3 (1976): 138-146.

[51] Ehrich, Silke, et al. "A new obligately chemolithoautotrophic, nitrite-oxidizing bacterium, Nitrospira moscoviensis sp. nov. and its phylogenetic relationship." Archives of Microbiology 164.1 (1995): 16-23.

[52] Gao Jingsi, Zhu Jia, Wang Maowei, Dong Wenyi. (2018). Dominance and Growth Factors of Pseudanabaena sp. in Drinking Water Source Reservoirs, Southern China. Sustainability. 10. 3936. 10.3390/su10113936.

[53] Sly, L. I., Vullapa Arunpairojana, and M. C. Hodgkinson. "Pedomicrobium manganicum from drinking-water distribution systems with manganese-related "dirty water" problems." Systematic and Applied Microbiology 11.1 (1988): 75-84.

[54] Qin, Sheng, et al. "Glycomyces endophyticus sp. nov., an endophytic actinomycete isolated from the root of Carex baccans Nees." International journal of systematic and evolutionary microbiology 58.11 (2008): 2525-2528.

[55] Sun, Shi-Lei, et al. "The plant growth-promoting rhizobacterium Variovorax boronicumulans CGMCC 4969 regulates the level of indole-3-acetic acid syn-

thesized from indole-3-acetonitrile." Applied and environmental microbiology (2018): AEM-00298.

[56] Wei, Yu-jie, et al. "High-throughput sequencing of microbial community diversity in soil, grapes, leaves, grape juice and wine of grapevine from China." PloS one 13.3 (2018): e0193097.

[57] Kim, Myung Kyum, et al. "Solirubrobacter soli sp. nov., isolated from soil of a ginseng field." International journal of systematic and evolutionary microbiology 57.7 (2007): 1453-1455.

[58] Iorio, Marianna, et al. "Antibacterial paramagnetic quinones from Actinoallomurus." Journal of natural products 80.4 (2017): 819-827.

[59] Groth, Ingrid, et al. "Knoellia sinensis gen. nov., sp. nov. and Knoellia subterranea sp. nov., two novel actinobacteria isolated from a cave." International journal of systematic and evolutionary microbiology 52.1 (2002): 77-84.

[60] Lu, Yang Li, et al. "Mesorhizobium shangrilense sp. nov., isolated from root nodules of Caragana species." International journal of systematic and evolutionary microbiology 59.12 (2009): 3012-3018.

[61] Chen, Mao-Yen, et al. "Rubrobacter taiwanensis sp. nov., a novel thermophilic, radiation-resistant species isolated from hot springs." International journal of systematic and evolutionary microbiology 54.5 (2004): 1849-1855.

[62] Tian, Xin-Peng, et al. "Sciscionella marina gen. nov., sp. nov., a marine actinomycete isolated from a sediment in the northern South China Sea." International journal of systematic and evolutionary microbiology 59.2 (2009): 222-228.

[63] Dworkin, Martin. The Prokaryotes: Vol. 6: Proteobacteria: Gamma Subclass. Springer Science Business Media, 2006.

[64] Hanada, Satoshi, et al. "Roseiflexus castenholzii gen. nov., sp. nov., a thermophilic, filamentous, photosynthetic bacterium that lacks chlorosomes." In-

ternational journal of systematic and evolutionary microbiology 52.1 (2002): 187-193.

[65] Al-Mueini, Ratiba, et al. “Hydrocarbon degradation at high salinity by a novel extremely halophilic actinomycete.” Environmental Chemistry 4.1 (2007): 5-7.

[66] Wang, Zichao, et al. “Effects of salinity on performance and microbial community structure of an anoxic-aerobic sequencing batch reactor.” Environmental technology 36.16 (2015): 2043-2051.

[67] Matsuo, Heizo, et al. “Acidicapsa acidisoli sp. nov., from the acidic soil of a deciduous forest.” International journal of systematic and evolutionary microbiology 67.4 (2017): 862-867.

[68] Han, Shun, et al. “Nitrospira are more sensitive than Nitrobacter to land management in acid, fertilized soils of a rapeseed-rice rotation field trial.” Science of The Total Environment 599 (2017): 135-144.

[69] El-Banna, N., and G. Winkelmann. “Pyrrolnitrin from Burkholderia cepacia: antibiotic activity against fungi and novel activities against streptomycetes.” Journal of Applied Microbiology 85.1 (1998): 69-78.

[70] Grady, Elliot Nicholas, et al. “Current knowledge and perspectives of Paenibacillus: a review.” Microbial cell factories 15.1 (2016): 203.

[71] Wuytack, Elke Y., and Chris W. Michiels. “A study on the effects of high pressure and heat on Bacillus subtilis spores at low pH.” International journal of food microbiology 64.3 (2001): 333-341.

[72] Toth, Ian K., et al. “Soft rot erwiniae: from genes to genomes.” Molecular plant pathology 4.1 (2003): 17-30.

[73] Young, Vincent B., Stanley Falkow, and Gary K. Schoolnik. “The invasin protein of Yersinia enterocolitica: internalization of invasin-bearing bacteria by eukaryotic cells is associated with reorganization of the cytoskeleton.” The Journal of cell biology 116.1 (1992): 197-207.

[74] Jangir, Yamini, et al. “Isolation and characterization of electrochemically active subsurface Delftia and Azonexus species.” Frontiers in microbiology 7 (2016): 756.

[75] Giotta, Livia, et al. “Heavy metal ion influence on the photosynthetic growth of Rhodobacter sphaeroides.” Chemosphere 62.9 (2006): 1490-1499.

[76] Gardan, L., et al. “Acidovorax anthurii sp. nov., a new phytopathogenic bacterium which causes bacterial leaf-spot of anthurium.” International journal of systematic and evolutionary microbiology 50.1 (2000): 235-246.

[77] Ueki, Atsuko, et al. “Paludibacter propionicigenes gen. nov., sp. nov., a novel strictly anaerobic, Gram-negative, propionate-producing bacterium isolated from plant residue in irrigated rice-field soil in Japan.” International journal of systematic and evolutionary microbiology 56.1 (2006): 39-44.

[78] Balch, William E., et al. “Acetobacterium, a new genus of hydrogen-oxidizing, carbon dioxide-reducing, anaerobic bacteria.” International Journal of Systematic and Evolutionary Microbiology 27.4 (1977): 355-361.

[79] Hobson, P. N., and B. G. Shaw. “The bacterial population of piggery-waste anaerobic digesters.” Water Research 8.8 (1974): 507-516.

[80] Oswald, Kirsten, et al. “Crenothrix are major methane consumers in stratified lakes.” The ISME journal 11.9 (2017): 2124.

[81] Brito, Elcia Margareth S., et al. “Characterization of hydrocarbonoclastic bacterial communities from mangrove sediments in Guanabara Bay, Brazil.” Research in microbiology 157.8 (2006): 752-762.

[82] 武田潔. “水田土壌中のメタン生成・酸化細菌の特徴と生態.” Microbes and environments 13.1 (1998): 39-44.

Appendix

biom フォーマットをデータフレームに変換するソースコードを lstlisting 5.1に示す。

Listing 5.1: Convert biom format to data frame

```
library(devtools)
library(biomformat)

file <- read_biom(biom_file)
taxonomy <- observation_metadata(file)
otu_table <- biom_data(file)
otu_table <- as.data.frame(otu_table)
#Normalization dato to 0-1 range
otu_table <- apply(table, 2, function(d){d/sum(d)})
```

UniFrac 距離を計算するソースコードを lstlisting 5.2に示す。

Listing 5.2: UniFrac distance

```
library(GUniFrac)
library(phyloseq)
library(ape)

#download ref from "https://github.com/biocore/qiime-default-
    reference/blob/master/qiime_default_reference/gg_13_8_otus/
    taxonomy/97_otu_taxonomy.txt.gz"
#OR use taxonomy defined in Listing 5.1
ref <- ref[colnames(otu_table),]
otu_table <- as.matrix(otu_table)

TAX <- tax_table(ref)
OTU <- otu_table(t(otu_table), taxa_are_rows = TRUE)
physeq <- phyloseq(OTU, TAX)
tree <- rtree(ntaxa(physeq), rooted=TRUE, tip.label=taxa_names(
    physeq))
OTU_t <- t(OTU)
OTU_t <- OTU_t[,tree$tip.label]
gu <- GUniFrac(OTU_t, tree, alpha = c(0, 0.5, 1))
```