

Mémoire

Auteur : Misztak, Agnieszka

Promoteur(s) : Baurain, Denis

Faculté : Faculté des Sciences

Diplôme : Master en bioinformatique et modélisation, à finalité approfondie

Année académique : 2020-2021

URI/URL : <http://hdl.handle.net/2268.2/12550>

Avertissement à l'attention des usagers :

Tous les documents placés en accès ouvert sur le site le site MatheO sont protégés par le droit d'auteur. Conformément aux principes énoncés par la "Budapest Open Access Initiative"(BOAI, 2002), l'utilisateur du site peut lire, télécharger, copier, transmettre, imprimer, chercher ou faire un lien vers le texte intégral de ces documents, les disséquer pour les indexer, s'en servir de données pour un logiciel, ou s'en servir à toute autre fin légale (ou prévue par la réglementation relative au droit d'auteur). Toute utilisation du document à des fins commerciales est strictement interdite.

Par ailleurs, l'utilisateur s'engage à respecter les droits moraux de l'auteur, principalement le droit à l'intégrité de l'oeuvre et le droit de paternité et ce dans toute utilisation que l'utilisateur entreprend. Ainsi, à titre d'exemple, lorsqu'il reproduira un document par extrait ou dans son intégralité, l'utilisateur citera de manière complète les sources telles que mentionnées ci-dessus. Toute utilisation non explicitement autorisée ci-avant (telle que par exemple, la modification du document ou son résumé) nécessite l'autorisation préalable et expresse des auteurs ou de leurs ayants droit.



Toward automated classification of bacterial metabarcoding samples by machine learning

Agnieszka Misztak

August 26th ,2021

Promoter: Denis Baurain

Contents

Summary	6
List of abbreviations	7
Introduction	8
Microbiome	8
Brief overview of the machine learning algorithms	10
Support Vector Machines (SVM)	10
Naive Bayes (NB)	12
Random Forest (RF)	14
K Nearest Neighbours (k-NN)	15
Linear Discriminant Analysis (LDA)	16
Association Rule Learning	17
Objective: Towards sample source identification	18
Materials and Methods	19
Data and the preprocessing	19
Feature annotation	20
Association Rule Learning	20
Classification preprocessing	21
Diversity metrics	21
Category arranging	21
Classification	22
Results	22
Database selection and taxonomic features	22
Classification - PCoA and LDA weighted analysis	25
Classification - KPCA and LDA unweighted analysis	29
Association rules	33
Discussion	36
Conclusions	42
Note	43
Supplementary materials	43
Supplementary Figures	43
Supplementary Tables	44
References	44

List of Figures

1	Global scope of sample provenance within EMP on the ontology level 2: samples came from 7 continents and 43 countries. After (Thompson et al. 2017).	9
2	The SVM linear and radial classifier result on two simulated datasets. A. Data points following a linear distribution, which can be divided using a linear approach; B. Three out of an infinite number of possible hyperplanes dividing the data points from two categories with different margins; C. Optimal division of the two categories using a linear approach; D. Nonlinear distribution of the data from two classes; E. Linear attempt to divide non linearly distributed points; F. SVM using the kernel trick optimally divides two datasets in a nonlinear problem.	11
3	The Gaussian RBF Kernel function visualization. X_1 , X_2 and X_3 represent random vectors or data points, y represents a landmark.	12
4	The visualization of the Naive Bayes classifier algorithm steps. A. Imaginary training data; B. New datum introduced into the dataset; C. Radius considered during the calculations of the probabilities.	13
5	The decision tree mode of action for classification problems. A. The relationship between Petal Width and Petal Length of three iris species from the famous Fisher's or Anderson's iris data containing 150 data points (50 points/species); B. Decision tree for classification of iris specimens into three species trained using the data from panel A. Each node shows predicted class, predicted probability of the class and percentage of the observations in the node. Due to misclassification of a few virginica irises as versicolor species, the data become unbalanced between the nodes.	15
6	The mode of action of the k-NN algorithm. A. The example of visualization of the training set; B. The new data point that is being classified; C. The k-nearest neighbours to the new data point selected for classification process by the algorithm; D. New data point with assigned category based on the plurality vote.	16
7	Example of transactions given to Apriori algorithm for strong rules association mining. After https://www.kdnuggets.com/	17
8	The method of calculating the support of two items from the list of transactions for further steps in the Apriori algorithm.	18
9	The method of calculating the confidence of two items from the list of transactions (fig. 7).	18
10	The method of calculating the lift of two items from the list of transactions.	18
11	The visualization of the analysis preprocessing and classification workflow as described in the Materials and Methods section. The presence/absence and abundance matrices were generated separately from the annotated feature tables for taxonomic levels from Genus to Phylum and processed the same way in steps described in E (apart from the distance calculations) and F. After establishing the proper categories on the smaller datasets the steps E and F were performed again for the global set consisting of all analysed samples.	23
12	Distribution of the reads among samples downloaded from NCBI SRA database after quality trimming. The majority of the samples retained most of their reads, the mean number of reads per sample was 80,889.	24

13	The comparison of the taxonomic annotations and number of features annotated with the use of Silva database in four datasets. The panels from A-E represent results at different taxonomic ranks. S=Silva, GG=Greengenes and E=Ezbio. Five panels (A-E) correspond to taxonomic ranks from Genus (A) to Phylum (E). The values in the table provide the percentage of the ASVs without annotation at the given taxonomic level for all databases. There were significant differences in the assigned taxonomy between the databases. Generally, as the taxonomic level increases, the number of uniquely assigned ASVs by all databases decreases and more similarities are observed. All the databases agreed on the assignment in 50% of the ASVs at the very general Phylum level. The lowest percentage of unannotated ASVs was observed with the use of the Silva database in all taxonomic ranks. The barplots represent richness - a number of features annotated with the Silva database. annotated	26
14	The richness and dominance metrics calculated for the four datasets. A. A richness (number of features in each category) is calculated based on the presence/absence matrix at the Genus taxonomic rank. B. The dominance index 'DBP,' or Berger-Parker index calculated at the Genus taxonomic level, is a special case of relative dominance with rank 1, meaning, it describes only the most abundant taxon and its proportion in relation to all detected microorganisms ($0.5 = 50\%$ of all detected groups). Among analysed categories distributed between the four datasets, the DBP index varies. The categories such as gills, spider nests or shoots were strongly dominated by the most common group, while others like acorn barnacles, bird oral samples, periphyton or rhizobiome of non-flowering trees were highly diverse. There was a visible negative correlation between the number of features in the categories (A) and the dominance index DBP (B).	27
15	The result of the PCoA analysis of the global set of 1567 samples based on Bray-Curtis distance calculated from the abundance matrix (weighted data) at Genus taxonomic rank obtained from analysis with QIIME2. There is a visible overlap between the samples belonging to many of the 45 categories, which prevents highly accurate classification of the data points.	28
16	The visualization of the four classifiers' predictions of the categories areas for the training set created from the weighted dataset at the Genus taxonomic rank. The points represent the taxonomic data from bacterial sequencing samples transformed with the LDA method, their colours match the colour areas that mark the algorithm's prediction zones for 45 categories. A-radial SVM, B-Naive Bayes, C-RF, D-k-NN. Thanks to the distances obtained with the linear discriminants the categories are well separated, which facilitates the classification. The best results were obtained with the RF classifier (C) - 97.6 % of correct predictions (kappa).	30
17	The visualization of the four classifiers' predictions of the categories areas for the test set created from the weighted dataset at the Genus taxonomic rank. The points represent the transformed with the LDA method taxonomic data from bacterial sequencing samples, their colours match the colour areas that mark the algorithm's prediction zones for 45 categories. A-radial SVM, B-Naive Bayes, C-RF, D-k-NN. The best results were obtained with the RF classifier (C) - 97.6 % of correct predictions.	31
18	The result of the kernel PCA analysis of the global set of 1567 samples based on the Jaccard distance calculated from the presence/absence matrix at Genus taxonomic rank obtained from analysis with QIIME2. Similarly to the plot based on PCoA for the weighted data (fig. 15), there is a visible overlap between the 45 categories distributions, which prevents highly accurate classification of the data point.	32

19	The visualization of the four classifiers' predictions of the categories areas for the training set created from the unweighted dataset at the Order taxonomic rank. The points represent the taxonomic data from bacterial sequencing samples transformed with the LDA method, their colours match the colour areas that mark the algorithm's prediction zones for 45 categories. A-radial SVM, B-Naive Bayes, C-RF, D-k-NN. Thanks to the distances obtained with the linear discriminants the categories are well separated, which facilitates the classification. The NB classifier (B) suffered from a low density of data on the entire presented surface and failed to produce consistent classification areas. The best results were obtained with the RF classifier (C) - 97.6 % of correct predictions.	33
20	The visualization of the four classifiers' predictions of the categories areas for the test set created from the unweighted dataset at the Order taxonomic rank. The points represent the transformed with the LDA method taxonomic data from bacterial sequencing samples, their colours match the colour areas that mark the algorithm's prediction zones for 45 categories. A-radial SVM, B-Naive Bayes, C-RF, D-k-NN. The best results were obtained with the RF classifier (C) - 97.6 % of correct predictions.	34
21	The shared and unique rules at each investigated taxonomic rank among four datasets. The highest number of unique rules is observed for environment-related data at the Family rank. At the Genus rank, there are no shared associations between the datasets observed.	37
22	The significant associations detected in two datasets. The visualized rules were detected in the environment- (A) and plant-related (B) datasets on the Order rank. The node with the highest degree corresponds to the Order Sphingobacteriales, the Order with the highest number of detected associations among all datasets on all taxonomic ranks.	38
23	The significant associations detected in two datasets. The visualized rules were detected in the animal (A) and animal-gut-related (B) datasets on the Order rank. The richness of the host-associated datasets was smaller and less significant associations were detected.	39

List of Tables

1	The distribution of the samples and the categories between the datasets.	19
2	The distribution of the samples between the categories.	22
3	The mean number of the features at each taxonomic rank in the four investigated datasets, rounded up to the nearest tenth place.	25
4	The accuracy and kappa metrics of the classification, performed with the use of four classifiers, based on the principal components obtained from weighted data throughout taxonomic ranks.	29
5	The accuracy and kappa of the classification, performed with the use of four classifiers, based on the linear combinations obtained from weighted data throughout taxonomic ranks.	29
6	The kappa and accuracy of the classification process using four ML classifiers based on the results from KPCA analysis of unweighted data on five taxonomic ranks. . . .	32
7	The kappa and accuracy of the classification process using four ML classifiers based on the results from LDA analysis of unweighted data on five taxonomic ranks. . . .	32
8	Number of statistically significant rules detected for each dataset at five taxonomic ranks, number of unique (appearing at least once) antecedents and number of unique consequents at each taxonomic rank. The highest number of significant rules were detected in the environment-related set at Order rank, whereas the lowest number was observed in the environment-related set at the Genus rank.	36

Summary

The studies of the bacterial communities are increasingly popular. Thanks to the continuous decrease in price of NGS services, curiosity is the limit. It is reflected in the diversity of the metabarcoding data available. Recently a collaborative Earth Microbiome Project had begun a creation of Earth's multiscale microbial diversity catalogue unifying the effort of almost 100 independent studies for standardization of the protocol for bacterial communities analyses. However, in the public databases there is a substantial amount of the metabarcoding data that were generated throughout the years with the use of different sequencing primers targeting different hypervariable regions.

The information about bacterial communities compositions accumulated in those metabarcoding samples could serve e.g. for identification of the origin of the sample. This work aims at establishing a base process for combining the analysis of the metabarcoding data obtained using various protocols. In the process of selection, out of over a million sequencing runs, 1567 individually processed paired-end reads samples were merged into 45 fine-scaled categories falling into four general datasets: animal-, animal-gut-, environment-, and plant-related. Next, they were processed using popular QIIME2 software without OTU clustering. Three general databases containing 16S rRNA taxonomic information, and their efficacy at five taxonomic ranks, have been tested in order to optimize the taxonomic identification of amplicon sequence variants. The above-mentioned datasets were tested for classification accuracy using two different dimensionality reduction techniques, Principal Component Analysis and Linear Discriminant Analysis applied on the similarity/dissimilarity matrices obtained separately from an abundance and presence/absence matrices. The aptitude of machine learning in establishing the taxonomic-based classification of the sample sources has been tested with four different algorithms, radial SVM, Naive Bayes, Random Forest and k-Nearest Neighbours. The LDA transformed similarity matrix created at Order rank provided the best and most confident classification with corrected accuracy of 97.6%. Additionally, to examine whether there exist taxonomic relationships among the microorganisms detected in the aforementioned studies, the association rule learning algorithms 'Apriori' has been utilized. Number of co-occurrences of microorganisms on different taxonomic ranks was detected and several different taxa forming highly connected nodes were observed. Those taxa can be regarded as putative keystone taxa and considered for further investigation in different niches.

List of abbreviations

ASV - amplicon sequence variant
BLAST - Basic Local Alignment Search Tool
DBP - Berger–Parker index
DMP - McNaughton's dominance
Eclat - Equivalence Class Clustering and bottom-up Lattice Traversal
EMP - Earth Microbiome Project
k-NN - k-nearest neighbours
KPCA - kernel principal component analysis
LDA - linear discriminant analysis
LEfSe - Linear discriminant analysis Effect Size
ML - machine learning
NCBI - The National Center for Biotechnology Information
NB - Naive Bayes
OTU - operational taxonomic unit
PCA - principal component analysis
PCoA - principal coordinate analysis
RF - Random Forest
rRNA - ribosomal RNA
SRA - sequence read archive
SVM - Support Vector Machine

Introduction

Microbiome

‘Life on earth is such a good story you cannot afford to miss the beginning... Beneath our superficial differences we are all of us walking communities of bacteria. The world shimmers, a pointillist landscape made of tiny living beings.’

Lynn Margulis

Even though we have not realized it for millennia, for better or worse microorganisms have always accompanied us. Thanks to great scientists like Antonie van Leeuwenhoek, Louis Pasteur or Robert Koch we could finally see them and begin to understand their omnipresence. The field of microbiology did arise and progressed throughout the years. New derivative fields to microbiology started emerging, among them microbial ecology, pioneered by Sergei Winogradsky and Martinus Willem Beijerinck (Konopka 2009). Microbial ecology began to study the microbes in the environment and their interactions with each other. The more we understood about the microbial communities living all around us the more we began to grasp their importance. Right now we know that microorganisms are creating their own ecosystems and influence the niche in which they find themselves. And so, the microorganisms give the taste to the blue cheese (Zuckermandl and Pauling 1965), slowly consume the marvel of human technology (Mann, Wells, and Blasco 1996), change our behaviour (K. V.-A. Johnson and Foster 2018) and much more. In order to efficiently discuss any idea, the name and definition are required. The term ‘microbiome’ is becoming increasingly popular as a word defining the community of the microorganisms (micro + biome) colonizing a certain niche. However, certain discrepancies are observed as this term is also used to describe the collective genomes or genetic materials (microbe + ome) of the microorganisms. The dispute over the original meaning and the correct usage of the word continues. Many publications (Ursell et al. 2012; Liu 2016; Zhu, Wang, and Li 2010) ascribe the term to the Nobel laureate Joshua Lederberg (Lederberg and McCray 2001). However, as the older and older potential original sources are being revealed it became clear that the term has been around for at least 130 years (France 1886). Recent works suggested a standardization of the definitions - ‘microbiota’ as a term that comprises all living members forming the microbiome, which also describes the ‘Theatre of activity’, such as microbial structural elements (nucleic acids, lipids, proteins...), microbial metabolites and external structural elements (Berg et al. 2020). Until a consensus is coined and potentially a new term describing a community of microorganisms arise for the purpose of this work I will continue to use ‘microbiome.’

A quick look into available literature reveals the popularity of microbiome studies. Just since the beginning of the year 2021 over 9000 publications on this topic have been released, almost half of them focusing on the human microbiome. In 2012 it was even proposed to treat the microbiome of the human gut as a new organ (Baquero and Nombela 2012). The NCBI SRA database holds almost 500,000 records of human microbiome samples from different sites of the body [July 2020]. Moreover, NCBI SRA holds over 600,000 records collected from other sources like soil, water, rhizobiome or animal gut. The studies within which those samples were collected usually focus on describing: the microbial communities observed in a particular niche (e.g. topsoil or glacier microbiomes Anesio et al. (2017)), the differences between the microbiomes of the same category (e.g. differences in patients with hypertension, or between roots of different plants (Yan et al. 2017; Qu et al. 2020) or their fluctuations within one source under conditions of interest (e.g. human gut microbiome in healthy and sick individuals or wheat microbiome under four management strategies (Shreiner, Kao, and Young 2015; Gdanetz and Trail 2017)).

In 2010 the Earth Microbiome Project (EMP) was founded to construct a microbial map of the Earth. In order to understand the patterns in microbial ecology EMP was launched as a collaborative effort pulling together over 500 investigators aiming at developing standardized methods for collection, curation and analysis of the collected data. In 2017 EMP released a

publication describing a meta-analysis of almost 28,000 samples collected in a frame of 97 independent studies (Thompson et al. 2017) all around the world (fig. 1).

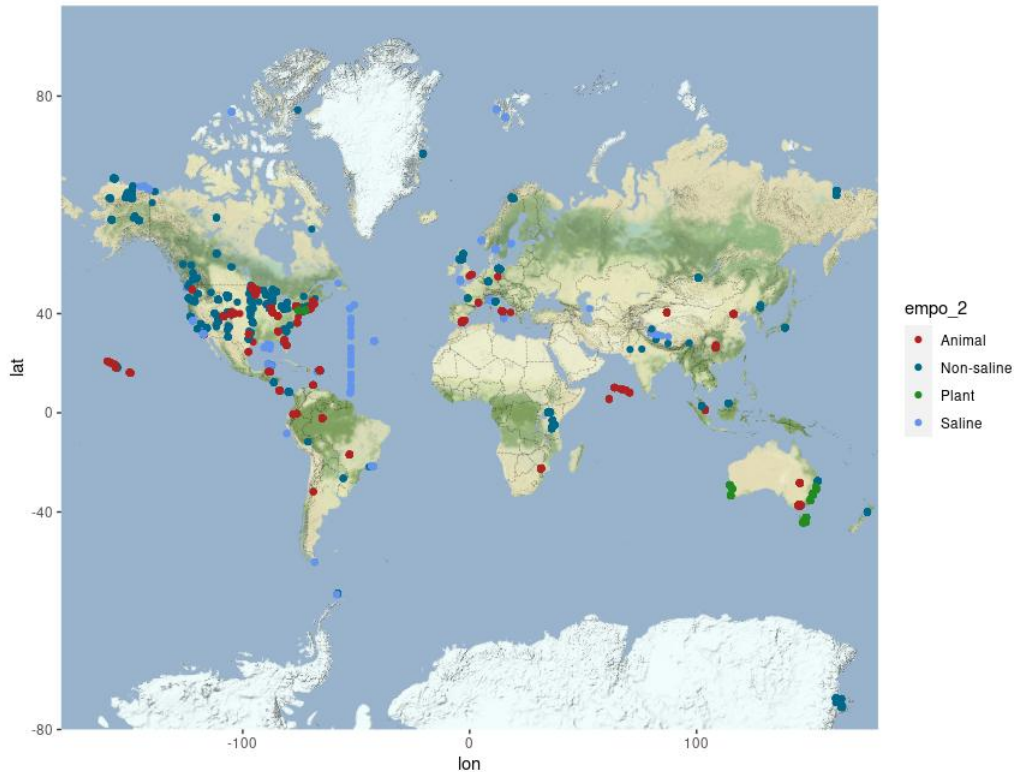


Figure 1: Global scope of sample provenance within EMP on the ontology level 2: samples came from 7 continents and 43 countries. After (Thompson et al. 2017).

Thanks to the crowdfunding approach of this project and the unified procedure of sample collection, handling and sequencing, the researchers generated more than two billion sequences of 16S rRNA region V4 (primer pair 515f–806r (Caporaso et al. 2012)). The scale of EMP allowed for observing interesting trends on a macroscale of microbial ecology, and once again confirmed that generally free-living microbial communities are richer and more diverse than host-associated ones (Ley et al. 2008). Moreover, within the frame of this study, the researchers explored the possibility of identifying the provenance of the sample based on its composition using a Machine Learning algorithm - Random Forest, explained in more detail in the next section. They achieved an accuracy of 91% when only the major classes (animal-associated, plant-associated, saline free-living, or non-saline free-living) were investigated and 84% when the affiliation to a total of 17 fine-scaled classes was considered. Last but not least, the EMP shared the carefully collected and curated catalogue, featuring metadata gathered in compliance with MIMARKS (Yilmaz et al. 2011), EBI (Madeira 2019), and Qiita (Gonzalez et al. 2018). The EMP Ontology for microbial environments were mapped to the ENVO (Buttigieg et al. 2016), UBERON (Mungall et al. 2012), plant ontology (PO) (Cooper et al. 2013), fungal anatomy ontology (FAO), and ontology of microbial phenotypes (OMP) (Chibucos et al. 2014), which will for sure facilitate future meta-analyses.

Brief overview of the machine learning algorithms

Machine learning (ML) is the study of computer algorithms that improve automatically through experience and by the use of data (Mitchell and Learning 1997). The algorithms can be divided into three groups: supervised, unsupervised and reinforced. The first group is defined by the use of labelled datasets. Meaningful labels allow the algorithms to learn cause and effect relationships in the dataset. Their success can be measured e.g. in the form of the accuracy of the predictions (fraction of the correctly foreseen answers). They are used for problems such as classification, so assigning test data into correct categories e.g. recognizing the genetic variants contributing to the breast cancer risk (Behravan et al. 2018), or regression problems that help to understand the relationship between dependent and independent variables such as predicting good credit customers (Hargreaves 2019).

Unsupervised learning algorithms deal with data that is not labelled, therefore it is up to the algorithm to discover the patterns. They can be categorized into two groups aimed at problems of data clustering or association, respectively. The examples of the popular algorithms performing unsupervised ML are hierarchical clustering (Ward’s linkage, average linkage etc.), association mining (‘Apriori’ and ‘Eclat’) or dimensionality reduction techniques (such as PCA, MDS or LDA)(Wang 2001).

The reinforced learning algorithms attempt to determine the set of operations that the algorithm should undertake in order to optimize the reward in a given environment (Hu et al. 2020). The reinforced learning also deals with unlabeled data, therefore it has to divide the algorithm’s attention between exploration and exploitation of the already known space. With each action, the algorithm interacts with the environment, changes the state it has found itself in or remains unchanged and finally receives feedback containing information about the result of the last action. The feedback allows the algorithm to determine whether the action was rewarded or penalized. The algorithm learns about the environment and determines the best set of actions (Nian, Liu, and Huang 2020).

Support Vector Machines (SVM)

SVM was first introduced in 1992 (Boser, Guyon, and Vapnik 1992). The algorithm was created based on the Vapnik-Chervonenkis theory introduced by Vladimir Vapnik and Alexey Chervonenkis (Vapnik and Chervonenkis 1974). It is a robust non-probabilistic binary linear classifier that attempts to create an optimal separation between data from two different categories (fig. 2 A) by creating an N-dimensional hyperplane (hyperplane being a plane that is one dimension lower than the plane that is currently considered). In the case of the multiclass classification problem, SVM breaks down the task into multiple smaller two-class classifications and solves it pair by pair. The selection is being made based on the available data points, from which the algorithm selects the so-called support vectors - data points localized at the edge of the space occupied by the data belonging to one category, neighbouring the other category. SVM investigates a number of possible separations (fig. 2 B) and chooses the one that has the largest possible margin, meaning, the points are as far away as possible from the boundary and we will trust that it will also work well for new observations in the future (fig. 2 C). If the data points are not separable by a simple linear approach but rather follow different distributions (fig. 2 D) the use of the linear classifier will most likely lead to bad separation and misclassification (fig. 2 E). The algorithm can utilize the ‘kernel trick’ to obtain better separation in the nonlinear SVM variation by mapping the data into a higher dimension (fig. 2 F). There are several kernel functions, which were implemented for SVM, however, the Gaussian RBF Kernel is one of the most popular. How does it work? Radial basis function Kernel is defined mathematically as:

$$k(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right)$$

In layman terms, y is a landmark, vector or a data point, in the centre of the space occupied by the data belonging to the category following a radial distribution. The rest of the points are being seen in relation to this point, to be more precise, how far away from the landmark they are after the transformation using the kernel function and therefore mapping to the higher dimension (e.g. fig. 3). Mapping with the kernel function will cause the data points to create a hill with the landmark, the tallest value, localized at the peak. Each point in the data will undergo such transformation. The bigger the distance between the given data point x and the landmark y , the higher the numerator in the kernel function will be (division by $2\sigma^2$ will decrease the number somewhat, but the large number will remain large). Following the equation, the exponent to the power of a large negative number is close to zero. In the opposite case, the small distance between the data point x and the landmark y will result in a small value in the numerator of the function, the division by denominator will once again affect the value just a bit. Exponent to the power of a small negative number has a value close to 1. Therefore the data points will be mapped close to the peak. Using the values calculated for each data point in the previous step, one can create a radial margin that allows for good separation of the data points belonging to two categories. The parameter σ influences the size of the created radial space encompassing the data points from one category. Therefore it has to be optimized not to include the data points belonging to the other categories.

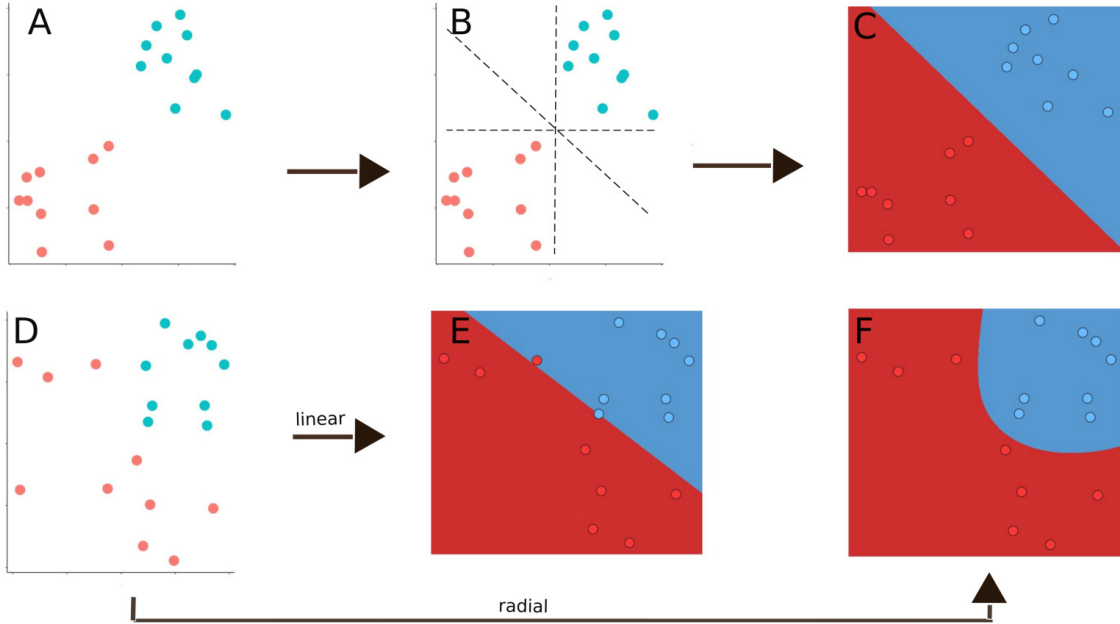


Figure 2: The SVM linear and radial classifier result on two simulated datasets. A. Data points following a linear distribution, which can be divided using a linear approach; B. Three out of an infinite number of possible hyperplanes dividing the data points from two categories with different margins; C. Optimal division of the two categories using a linear approach; D. Nonlinear distribution of the data from two classes; E. Linear attempt to divide non linearly distributed points; F. SVM using the kernel trick optimally divides two datasets in a nonlinear problem.

The Kernel SVM is a powerful, memory-efficient algorithm providing high performance on nonlinear problems, it is also not biased by outliers and not very sensitive to overfitting. Additionally, the linear version of the SVM allows for establishing the importance of the individual features. The same can not be said about the radial SVM version, as mapping to the higher dimension makes the

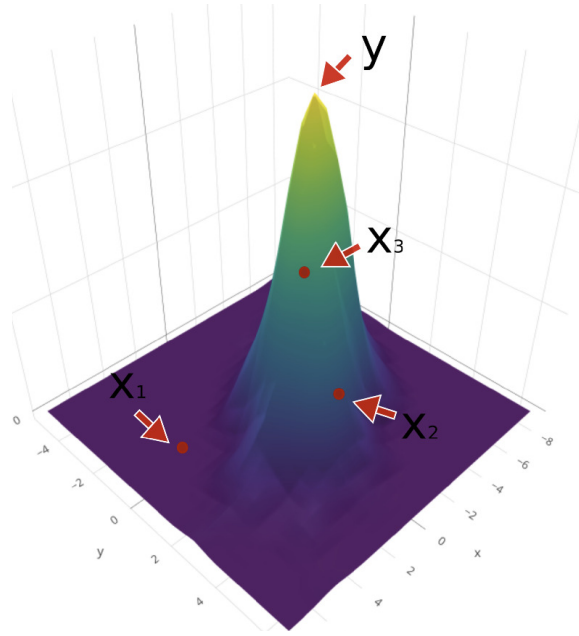


Figure 3: The Gaussian RBF Kernel function visualization. X_1 , X_2 and X_3 represent random vectors or data points, y represents a landmark.

data unrelated to the input space. Moreover, SVM classifiers tend to underperform when applied to a dataset with a number of features exceeding the number of training data samples.

Naive Bayes (NB)

The Naive Bayes classifier is a probabilistic classifier based on applying the Bayes theorem (named after Reverend Thomas Bayes) with feature independence assumption. The Bayes theorem is mathematically defined as:

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

As a reminder, Bayes theorem is defined by Oxford Dictionary as:

‘...a theorem describing how the conditional probability of each of a set of possible causes for a given observed outcome can be computed from knowledge of the probability of each cause and the conditional probability of the outcome of each cause...’
(Dictionary 1989).

To explain how the NB classifier is making decisions about category assignment for the new data, we must use an example. Imagine a study, in which the researchers were investigating differences between spotted dogs and cows. They have gathered data from 22 cows and 31 dogs. They have measured the number of spots counted on the fur of each animal and the maximal speed the individuals were observed to move with. Fig. 4 A.

The researchers proceeded with training the NB classifier to be able to establish, based on the collected data, which animal they were dealing with. Next, they tested the method with a new datum (Fig. 4 B). Two so-called posterior probabilities had to be calculated:

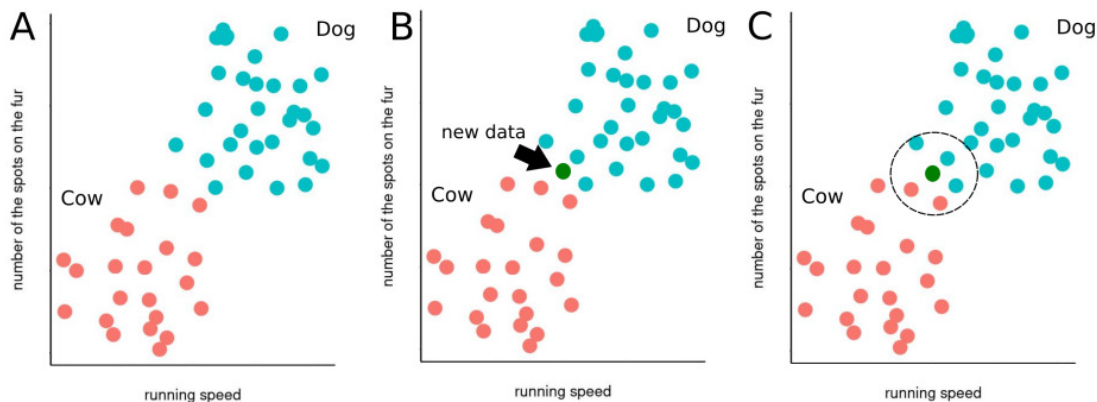


Figure 4: The visualization of the Naive Bayes classifier algorithm steps. A. Imaginary training data; B. New datum introduced into the dataset; C. Radius considered during the calculations of the probabilities.

So, what was the probability that the new point represented a dog. . .

$$P(Dog|B) = \frac{P(B|Dog) * P(Dog)}{P(B)}$$

. . . or what was the probability that the new point represented a cow:

$$P(Cow|B) = \frac{P(B|Cow) * P(Cow)}{P(B)}$$

In order to solve this problem we calculate the probabilities on the right side of the equation. $P(Dog)$ is called a prior probability and is easiest to calculate:

$$P(Dog) = \frac{NumberOfObservationInCategory}{TotalNumberOfObservations} = \frac{30}{22 + 31} = 0.57$$

$P(B)$ is called a marginal likelihood and it is the unconditional probability of observing the data over all possible models. To calculate it, the radius (for 2D data or sphere for 3D data and so on) around the new data point, has first to be determined by the algorithm or manually (Fig. 4 C). The radius will influence the outcome of the classification, therefore it has to be selected with caution. Here, it shows training animals that are similar in terms of the investigated features (here number of spots and running speed) to the data for the new animal. In this example within the selected radius, there are two very fast and very spotted cows and three lazy dogs with relatively small amounts of dots on the fur compared to other dogs. The marginal likelihood is calculated as follows:

$$P(B) = \frac{NumberOfSimilarObservationswithintheRadius}{TotalNumberOfObservations} = \frac{5}{53} = 0.094$$

$P(B|Dog)$ is called just the ‘likelihood.’ To calculate it, we must consider the points within the radius (Fig. 4 C). However, now we consider the number of dogs within the radius in relation to all observations of dogs:

$$P(B|Dog) = \frac{NumberOfDogsWithintheRadius}{TotalNumberOfObservationsForDogs} = \frac{3}{31} = 0.097$$

Finally we can calculate the posterior probability:

$$P(Dog|B) = \frac{0.097 * 0.57}{0.094} = 0.59$$

Similar calculations have to be performed for the cows:

$$P(Cow) = \frac{22}{22 + 31} = 0.42$$

$$P(B) = \frac{5}{53} = 0.094$$

$$P(B|Cow) = \frac{2}{22} = 0.091$$

$$P(Cow|B) = \frac{0.091 * 0.42}{0.094} = 0.41$$

There is a 59% probability that the new datum represents a dog and 41% that it belongs to a cow. The posterior probability, regardless of the number of the considered categories, sums up to 1 (100%). Here, the classifier would therefore decide to assign it as a dog.

In summary, NB is, like SVM, an efficient algorithm not biased by outliers and performing well for nonlinear problems providing a probabilistic approach. This type of classifier does not offer the method for feature importance evaluation. However, as the name states, it is naive, meaning it assumes independence and identical statistical relevance of the features. Apart from that, the choice of the radius has to be decided carefully as it will strongly influence the performed calculations and therefore, the classification itself.

Random Forest (RF)

To understand the idea behind the RF classifier it is crucial to understand the classification performed with the use of a decision tree. The decision tree is a visual representation of the decisions taken by an algorithm, which divides the dataset into consecutively smaller subsets following the values of the given attributes. The decisions are later used as guidelines for classification. The process is presented on Fig. 5. Fig. 5 A is a visualization of the relationship between Petal Width and Petal Length of three iris species from the famous Fisher's or Anderson's iris data (Fisher 1936; Anderson 1935). Panel B shows the decision tree for the classification of iris to one of three species based on the above-mentioned measurements and the appropriate probabilities. Without making any splits, we can classify the new data into any of the species with 33% confidence (since the dataset contains 50 observations for each species, 150 observations in total, therefore $(\frac{50}{150} * 100 = 33)$). If the new data indicates that the iris specimen has a petal length smaller than 2.5 cm then following the decision tree guidelines we can say with 100% confidence that it is an iris from the species *setosa*. If the specimen has a petal length bigger than 2.5 cm and petal width smaller than 1.8 cm then at 91% it is an iris from species *versicolor* and at 9% *virginica*. However, if the petal width is bigger than 1.8 cm then at 98% we are dealing with an iris from species *virginica* and only at 2% from species *versicolor*. The RF algorithm developed by Leo Breiman and Adele Cutler (Breiman 2001; Liaw and Wiener 2013) is a so-called ensemble learning algorithm. It means that it combines multiple smaller algorithms in order to increase its predictive power. RF utilizes a number of decision trees (multiple 'trees' creating a 'forest'). Each of the trees is trained on a subset

of the training set (the training data are reused) given to the algorithm, therefore multiple decision trees are built, each using different data. The category for the new data point is predicted separately by each of the trained decision trees and RF finally decides which category to assign by counting the trees' votes and picking up the one with the most votes.

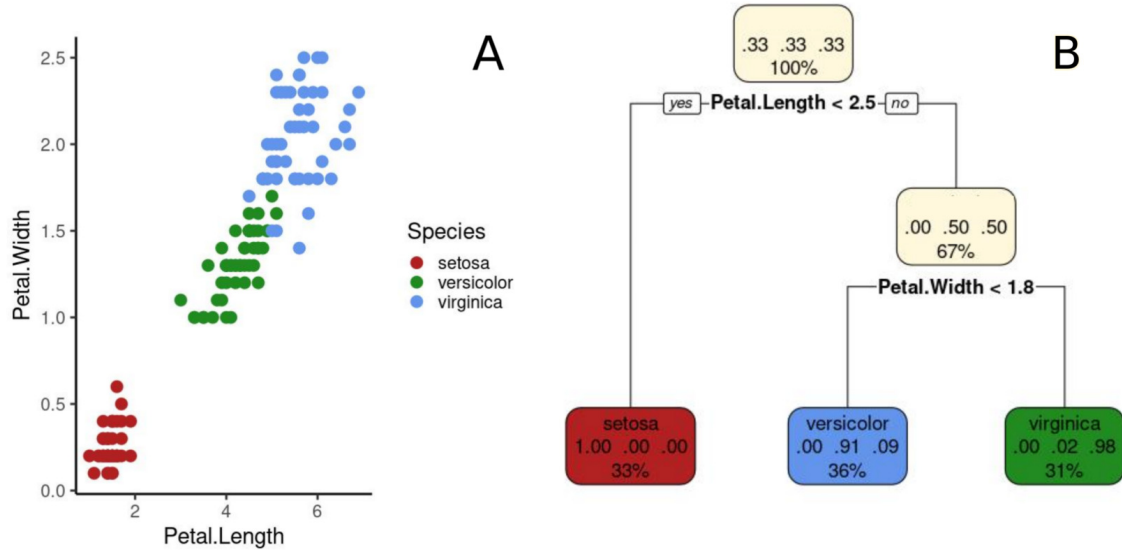


Figure 5: The decision tree mode of action for classification problems. A. The relationship between Petal Width and Petal Length of three iris species from the famous Fisher's or Anderson's iris data containing 150 data points (50 points/species); B. Decision tree for classification of iris specimens into three species trained using the data from panel A. Each node shows predicted class, predicted probability of the class and percentage of the observations in the node. Due to misclassification of a few virginica irises as versicolor species, the data become unbalanced between the nodes.

The RF classifier is an accurate and flexible ensemble algorithm, performing very well on a number of linear and nonlinear problems. However it can be overfitted easily, is computationally demanding and is unable to determine the significance of the features.

K Nearest Neighbours (k-NN)

The algorithm was developed by Evelyn Fix and Joseph Hodges (Fix and Hodges 1951). It has a relatively straightforward mode of action. In order to classify new data points to a category the classifier first identifies the k nearest neighbours (k being a parameter that has to be optimized e.g. five - five neighbours) and the category is assigned by a plurality vote (one vote/neighbour) of those neighbours - more precisely, the category that is the most common among the neighbours is being assigned to the new data point (fig. 6).

K-NN is a very simple but efficient algorithm. However, it does suffer from the 'curse of dimensionality'. It is impossible to retain the density of the dataset at a level high enough to keep up with the exponential growth of the data space, once more and more dimensions are considered. Since k-NN requires the points to be close to each other in each considered dimension, it becomes more and more efficient. As a result, points eventually stop being close to each other and the classification fails. K-NN is also sensitive to noisy data, outliers and scale of the variables. It is therefore important to perform data cleaning and feature scaling beforehand.

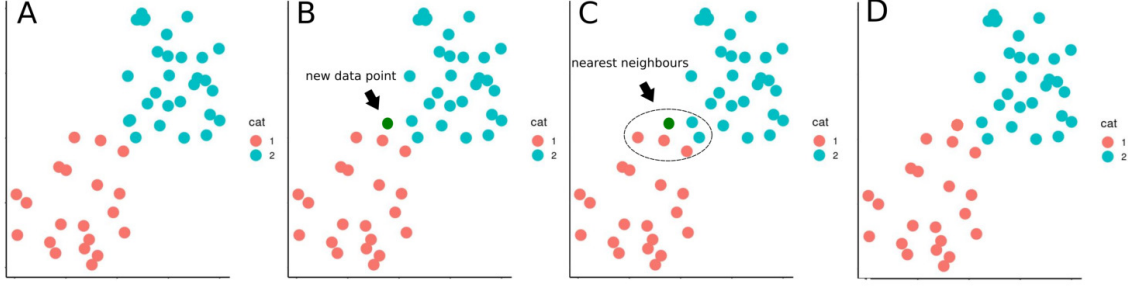


Figure 6: The mode of action of the k-NN algorithm. A. The example of visualization of the training set; B. The new data point that is being classified; C. The k-nearest neighbours to the new data point selected for classification process by the algorithm; D. New data point with assigned category based on the plurality vote.

Linear Discriminant Analysis (LDA)

LDA is a dimensionality reduction technique closely related to PCA. While PCA is looking for an axis that maximizes the variance in the projected data, LDA focuses on maximizing the separability of the given categories. The algorithm searches for a new axis to project data, which will at the same time maximize the distance between the means (μ) of the categories normalized by the intra-category variance (s) (defined as the sum of square differences between the projected samples and their category mean). This linear discriminant is defined as the linear function for solving the eigenvalue problem, which, for two categories, can be mathematically expressed as:

$$\frac{(\mu_1 - \mu_2)^2}{s_1^2 + s_2^2}$$

Ideally, the distance between the means of two categories should be large, while maintaining a small intra-category variance. The result for the two classes is a projection vector with the best eigenvalue. In the case of more categories, the function has to be generalized. The algorithm starts with calculating the d-dimensional mean vectors for each category, followed up by computing the intra- and inter-categories (with respect to the mean of all categories) variance matrices. Next the eigenvalues are calculated using the generalized equation, e.g. for three categories:

$$\frac{d_1^2 + d_2^2 + d_3^2}{s_1^2 + s_2^2 + s_3^2}$$

d = distance of the category mean to the mean of all categories

The eigenvectors corresponding to the highest eigenvalue are then used to transform the data. Because the categories of the training data have to be known, LDA is referred to as a supervised algorithm, as opposed to PCA, which is considered to be an unsupervised method. LDA can be used as a pre-processing step in the process of pattern classification, or as a classifier itself - its variation, quadratic discriminant analysis (QDA), can perform classification decisions in a nonlinear way, therefore providing more flexibility.

Association Rule Learning

Association rule learning is a rule-based ML method developed based on Rakesh Agrawal, Tomasz Imielinski and Arun Swami concept of strong rules (Piatetsky-Shapiro 1991). Algorithms ‘Apriori’ and ‘Eclat’ are both used for discovering interesting relations between variables in datasets (also called transactions). They both mine frequent items in sets in search of strong rules (with parameters above-set threshold). Eclat creates combinations that occur together in the given data, without investigating the pairs of items that were never observed together. In contrast, Apriori generates all the possible combinations of the so-called ‘frequent’ items (and assumes that the datasets containing frequent items are also frequent) and verifies their parameters afterwards. Generally, the Apriori algorithm is slower and designed for larger datasets, while Eclat works better on smaller datasets. There are three common parameters measured in the algorithms:

- Support: measures how popular an item is among all transactions
- Confidence: measures how likely it is to observe two items together
- Lift: also measures how likely is it to observe two items together, yet corrected by taking under consideration the popularity of the items

For example, let the transactions (datasets) be a list of combinations of items ordered by eight different schools: the honey bees, a globe, the plants, some distilled water, the test tubes and a microscope (fig. 7).

Set 1				
Set 2				
Set 3				
Set 4				
Set 5				
Set 6				
Set 7				
Set 8				

Figure 7: Example of transactions given to Apriori algorithm for strong rules association mining. After <https://www.kdnuggets.com/>.

In order to calculate the possible relationship between e.g. the globe (so-called ‘antecedent’) and the test tubes (in association mining called a ‘consequent’), we have to calculate the three above-mentioned parameters for them. The first informative parameter, the support, would be calculated as presented in fig. 8. The number of distinct appearances (no duplicates within datasets) of the item has to be divided by the total number of investigated transactions.

Next, the confidence is calculated by the proportion of transactions containing test tubes, in which also the globe appears to the number of all transactions with test tubes fig. 9.

$$\begin{aligned}\text{Support } \{\text{📊}\} &= \frac{4}{8} \\ \text{Support } \{\text{🌍}\} &= \frac{6}{8}\end{aligned}$$

Figure 8: The method of calculating the support of two items from the list of transactions for further steps in the Apriori algorithm.

$$\text{Confidence } \{\text{📊} \Rightarrow \text{🌍}\} = \frac{\text{Sets with } \{\text{📊}, \text{🌍}\}}{\text{Sets with } \{\text{📊}\}} = \frac{3}{4}$$

Figure 9: The method of calculating the confidence of two items from the list of transactions (fig. 7).

Finally, the lift is calculated by dividing the values from the previous steps, confidence by support fig. 10. The final result of lift, in this case, is 1.5. There is a certain association between the test tubes and the globes since the lift exceeded one, but the enrichment is not very substantial. It is worth mentioning that the rules can be bidirectional (A->B and B->A), but it is not always the case. Here the lift for the matching relationship between globe and test tubes would be 0.66. Therefore this rule (test tubes and globe) is unidirectional.

$$\text{Lift } \{\text{📊} \Rightarrow \text{🌍}\} = \frac{\text{Confidence } \{\text{📊}, \text{🌍}\}}{\text{Support } \{\text{🌍}\}} = \frac{0.75}{0.5}$$

Figure 10: The method of calculating the lift of two items from the list of transactions.

Objective: Towards sample source identification

Undoubtedly, despite all the progress in the field of microbiology, the world of microorganisms remains an enigma. We are just at the beginning of the road to a full understanding of the scope of microorganisms around us. The microbiome sequencing studies, whose popularity continuously increase, allowed us to investigate the microbial content of the samples in considerable depth. The new knowledge made us realize how little we knew about the bacteria living in different environments and how imperfect our cultivation methods are. Thanks to the collective effort of many scientists supported by multiple academic institutions, the first Earth multiscale microbial diversity catalogue had come to life. The insight from such a promising database can result in various applications e.g. identify a sample's origin based on its microbial profile in forensic medicine. The tools that could aid such an effort would be ML algorithms, which have become more and more ubiquitous in our lives. However, to combine the metabarcoding sample information generated by various investigators and following various protocols is not straightforward. To obtain a more detailed category resolution and therefore more in-depth perception thanks to assorted 16S rRNA sequencing data, simple usage of ML classification tools might not be enough. In this work, I used available NCBI Sequence Read Archive data from metabarcoding experiments originating from multiple studies around the world to investigate the aptitude of different ML algorithms in establishing the taxonomic-based classification of the sample source. Additionally, I examined whether there exist taxonomic relationships among the microorganisms detected in the aforementioned studies.

Materials and Methods

Data and the preprocessing

The metadata of the publicly available microbiome data was downloaded from NCBI Sequence Read Archive (SRA - <https://www.ncbi.nlm.nih.gov/sra>). Based on the metadata file the samples were selected based on the following criteria:

- 16S rRNA sequencing data (from any primer pairs)
- Paired-end reads
- Illumina MiSeq: 1406 files
- Illumina HiSeq: 123 files
- Illumina MiniSeq: 17 files
- Illumina NovaSeq: 6 files
- Illumina NextSeq: 9 files
- 454 GS: 3 files
- No information about platform: 3 files
- Non-human samples
- Natural, non-tempered with or contaminated sources (e.g. non-contaminated or non-treated water from the reservoir)
- Available in at least three replicates (biological or technical, which were processed as an individual samples)

The number of samples coming from the sources that were overrepresented, such as popular model organisms (e.g. mice, chicken), were limited to half of initially selected. For each of the considered sequencing files, the corresponding NCBI Biosamples were consulted to verify the sampling source and eliminate inaccuracies, e.g. listing crab shell swab as an aquatic sample or specifying the area of the animal body that was swabbed. Additionally, a number of searches was conducted to introduce more information for the process of establishing the categories. Using biosamples and bioproject information, together with wide literature searches, otherwise 163 individual categories, were merged into 45 categories. It was accomplished by following assumptions such as the plant-eating animals gut microbiome is similar or monocots/dicots rhizobiome will differ in microbial structure etc. The global dataset consisted of 1567 samples. It was then divided into four datasets: environment-, plant-, animal- and animal-gut-related samples. Next, due to large differences in the sequencing depths across studies (i.e. 1000 fold difference between the smallest and largest file), all the files were subsampled at the maximal level of 200,000 paired reads (later merged into 100,000 joined-reads). The smallest file has 10,510 paired reads. The reads were quality trimmed with the use of Trimmomatic-0.32 software using a sliding window of 10 and average quality threshold of 18 (Bolger, Lohse, and Usadel 2014). This part corresponds to fig. 11 A.

Table 1: The distribution of the samples and the categories between the datasets.

Dataset	Number of samples	Number of categories
Animal	385	18
Animal-gut	550	5
Environment	379	12
Plant	253	10

Feature annotation

The samples were imported into QIIME2 software (Bolyen et al. 2019) for processing. Since different microbiome studies might target different variable regions of 16S rRNA (often without providing the information about the target region in the metadata of SRA files) and the samples were sequenced on different machines, the error rates were different. Therefore the samples were processed on a sample-by-sample basis using the DADA2 package (Callahan et al. (2016)) implemented in QIIME2. DADA2 first counts the number of identical copies of amplicon sequence variants (ASVs) and then groups similar variants together (dereplication) following the error model calculated at the beginning of the process. Additionally, DADA2 removes the chimeric sequences from the samples. After feature tables containing ASVs as features were created, they were merged into four big feature tables, one for each dataset (corresponding to the datasets described in Classification preprocessing). The taxonomic assignment was performed using consensus BLAST of at least 10 top hits with 51% agreement based on the 16S rRNA sequences available in three reference databases used separately: SILVA (Quast et al. 2012), GreenGenes (DeSantis et al. 2006) and Ezbio (Yoon et al. 2017). The consensus BLAST generates multi-level results. Meaning, if the comparison of the sequence of interest to the database yields fewer sequences than the threshold, compatible at specific taxonomic rank (e.g. Genus), then the taxonomy can not be assigned. Therefore, consensus BLAST will attempt to match a higher taxonomic rank based on the best hits that might have been incompatible on the lower taxonomic level. For example, in the case of the search with parameters: 3 top hits with 100% agreement, with the following similar matches returned. . . :

- d_Bacteria; p_Firmicutes; c_Bacilli; o_Bacillales; f_Bacillaceae; g_Bacillus,
- d_Bacteria; p_Firmicutes; c_Bacilli; o_Bacillales; f_Bacillaceae; g_Piscibacillus,
- d_Bacteria; p_Firmicutes; c_Bacilli; o_Lactobacillales; f_Lactobacillaceae; g_Weisella;

. . . consensus BLAST would look for full compatibility between the results (100% agreement), starting at Genus rank and progressing higher with each step. Here, since all of the three hits are required to be in agreement, the assigned taxonomy would be d_Bacteria; p_Firmicutes; c_Bacilli. Due to the nature of the consensus BLAST all generated feature tables contained multi-level annotations, since the lowest level annotations were not always possible to obtain. The procedure will repeat until an assignment is found on any of the taxonomic ranks (Domain-Species) or the sequence will be marked as “Unassigned.” Further analysis was performed on the feature tables containing ASV abundances annotated with the SILVA database, which provided the best ratio of annotated vs unannotated ASVs. The taxonomic assignment of ASVs from different regions of 16S rRNA allowed for (previously impossible) feature table collapse (summing up the number of occurrences describing the same taxa inside the sample) separately at taxonomic levels from phylum to genus. The statistical significance of the differences between datasets was tested by pairwise t-test with Bonferroni correction with $\alpha=0.05$. For further analysis, only the collapsed ASVs with an abundance above 10 were considered. This part corresponds to fig. 11 B. For further analysis the filtered feature tables were exported at five taxonomic levels (Genus - Phylum), the abundance of each taxon from a higher rank was a sum of its children taxa. Based on the example above, if the taxonomy d_Bacteria;p_Firmicutes;c_Bacilli was assigned to 10 ASVs and d_Bacteria;p_Firmicutes;c_Clostridia to 20 ASVs, when considering the higher taxonomic rank d_Bacteria;p_Firmicutes would have an abundance of 30.

Association Rule Learning

The association rules were mined using Apriori R implementation (Hahsler et al. 2011) and visualized using arulesViz package (Hahsler 2017) at five taxonomic levels: Genus, Family, Order, Class and Phylum. The lists of transactions (all groups of microorganisms successfully annotated at the analysed taxonomic rank) were prepared from the feature table obtained from QIIME2 software. The rules were mined separately for the four datasets: plant-, environment-, animal and

animal-gut-related with the minimal support of 0.03 and minimal confidence of 0.2, as well as for the global dataset with minimal support of 0.02 and minimal confidence of 0.2 in order to unveil global rules. Afterwards, the rules were tested for significance using Fisher’s Exact Test with $\alpha=0.01$ and Bonferroni correction. The intersections, so the same rules observed in different datasets were visualized using UpSetR (Conway, Lex, and Gehlenborg 2017). This part corresponds to fig. 11 C.

Classification preprocessing

The feature tables generated in QIIME2 at five taxonomic levels (Genus - Phylum) were processed as abundance data and turned into a presence/absence matrix (0 when the ASV abundance was below 10, and 1 if above). For the ordination methods (operations on a community data matrix, e.g. dimensionality reduction techniques) feature tables are not suitable in terms of data structure, since those techniques require a data following linear trend. Since neither the abundance data (in this case), nor the presence/absence data were linear, the raw data (matrix with samples in rows and ASVs abundances in each sample in the columns) were transformed into distances with Jaccard distance for unweighted (presence/absence) and Bray-Curtis distance for weighted (abundance) analysis (resulting in sample x sample similarity/dissimilarity matrices) using vegan package (Oksanen et al. 2015) for further analysis. The data was divided into a training set and test set in a 4:1 ratio. In order to simplify the complexity and reduce the computation power required, instead of using the raw feature table data, the distance matrices were used to perform the kernel PCA (KPCA) for presence/absence data with kernlab package (Karatzoglou et al. 2004), PCoA (due to the fact that Bray-Curtis distance represents dissimilarity) with stats core R package (Team et al. 2013) and LDA analysis with MASS package (Venables and Ripley 2013). The resulting values of PC1 and PC2 for KPCA, PCo1 and PCo2 for PCoA and LD1 and LD2 for LDA analysis were used as data points for ML algorithms (as they contained the most information and allowed for clear visualization). This part corresponds to fig. 11 E.

Full information about assigned categories can be found in STable1.

Diversity metrics

The diversity metrics: DBP, DMN, absolute, simpson and richness were calculated in R programming language (Team et al. 2013) using the Phyloseq package (McMurdie and Holmes 2013). The significance of the diversity metrics between categories was calculated using a pairwise t-test with the $\alpha=0.05$ and Bonferroni correction. Additionally, the Linear discriminant analysis Effect Size (LEfSe) was performed to determine the features most likely to explain differences between categories (organisms present in different categories) using LEfSe software with factorial Kruskal-Wallis test at the significance level of 0.05 (Segata et al. 2011). This part corresponds to fig. 11 D.

Category arranging

The categories used for the analyses were established based on the information given in the corresponding biosample files (STable1). In order to increase the accuracy of the classification, while retaining biological meaning, highly similar sample sources (in terms of variance and means, rendering them inseparable by PCA, PCoA and LDA) were categorized together, e.g. :

- Aquatic animal skin samples = amphibian skin + axolotl skin + fish skin
- Oral-herbivore = alpaca gut + herbivorous bird gut + sheep gut
- Carnivore-gut = tiger gut + coyote gut + alligator gut
- Rhizobiome_monocots = banana rhizobiome + rice rhizobiome + maize rhizobiome
- Marine&beach = marine and beach sand samples
- Restina&marine_sediment = restinga soil samples and marine sediment samples

The classification of the data into established categories was performed for four datasets (animal-, animal-gut-, environment- and plant-related) separately in order to optimize the categories content (merging and separating categories when applicable). The optimal categories were then used in a global set for the classification process.

Table 2: The distribution of the samples between the categories.

Dataset	Animal	Animal-gut	Environment	Plant
Categories and their abundance	acorn barnacle (n=3), aquatic animal oral (n=6), aquatic animal skin (n = 102), bat skin (n=3), bird oral (n=3), bird skin (n=6), carnivore oral (n=6), cat eyes (n=6), coral (n=75), fox skin (n=3), gills (n=6), hemolymph n(=3), herbivore oral (n=6), insect whole body (n=103), nematode (n=12), reproductive tract (n=3), sheepskin (n=3), sponge (n=39)	carnivore gut (n=94), herbivore gut (n=249), omnivore gut (n=172), insectivore gut (n=23), geovore (n=12)	periphyton (n=3), marine&beach (n=84), restinga&marine sediment (n=25), freshwa- ter&sediment (n=120), salt marsh (n=3), soil (n=111), rock (n=7), honey (n=7), air (n=3), halite&salt flat (n=6), spider nest (n=7), saltern (n=3)	flower (n=12), fruit (n=3), phyllosphere (n=41), pollen (n=9), rhizobiome monocots (n=59), rhizobiome dicots (n=101), rhizobiom tree (n=4), seed (n=18), shoot (n=3), stalk&stem (n=3)

Classification

The datasets were classified using four ML algorithms: radial SVM, NB, RF and k-NN, trained and optimized in the caret package (Kuhn 2008) using repeated cross-validation approach. The Accuracy (number of correct predictions in relation to all predictions), Kappa (accuracy metric taking into account the random chance of assigning the point to a given category), specificity and sensitivity for all categories were calculated in the caret package. This part corresponds to fig. 11 F.

Results

Database selection and taxonomic features

The analysed 1567 samples after subsampling and quality trimming consisted of 291,199,806 joined reads in total. The smallest file consisted of 102 and the largest of 100,000 reads, the average was 80,889 joined reads/sample (fig. 12). Microbial counts in different environments are dependent on a number of factors like temperature, availability of nutrients as so on. Therefore, due to the fact that the global set consisted of samples from niches that might naturally consist of a lower number of microorganisms (e.g. salt marsh), even the samples with low reads were kept.

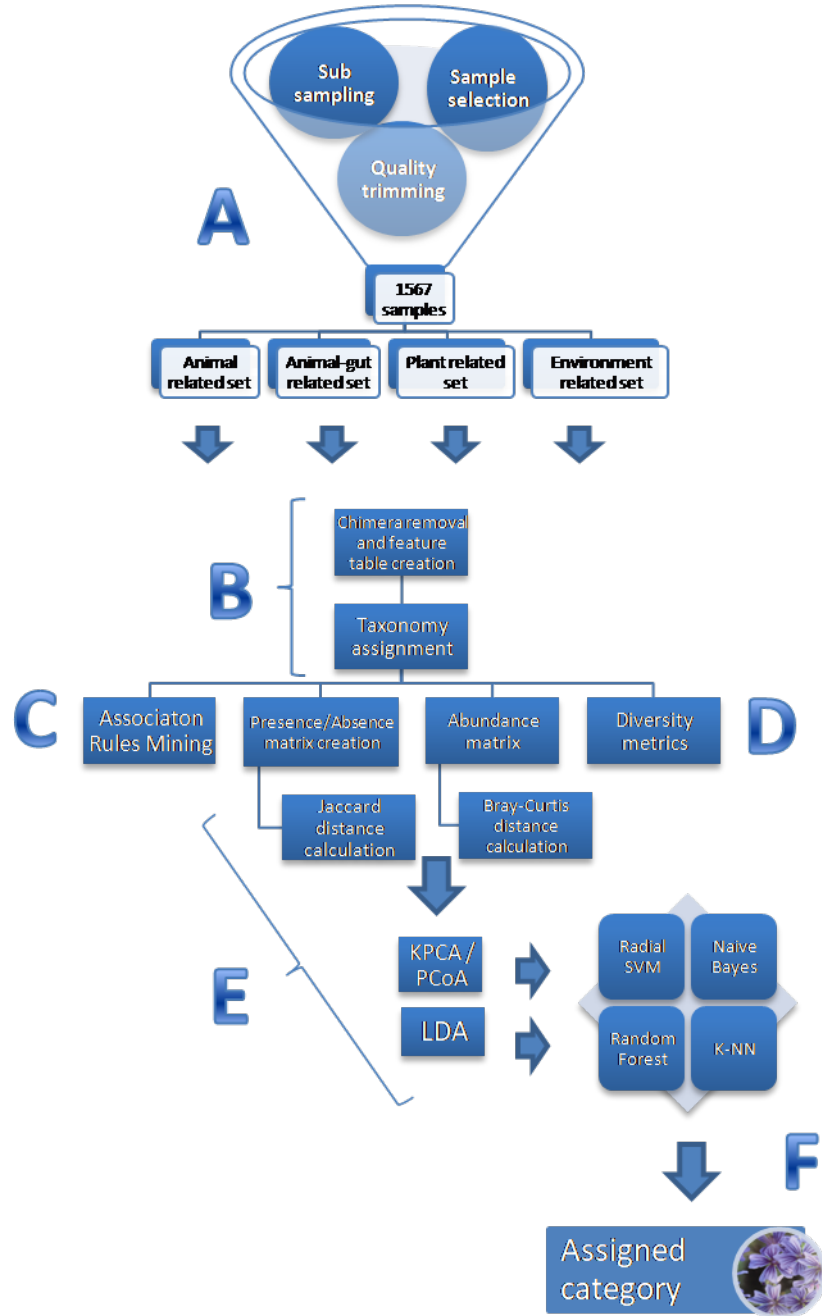


Figure 11: The visualization of the analysis preprocessing and classification workflow as described in the Materials and Methods section. The presence/absence and abundance matrices were generated separately from the annotated feature tables for taxonomic levels from Genus to Phylum and processed the same way in steps described in E (apart from the distance calculations) and F. After establishing the proper categories on the smaller datasets the steps E and F were performed again for the global set consisting of all analysed samples.

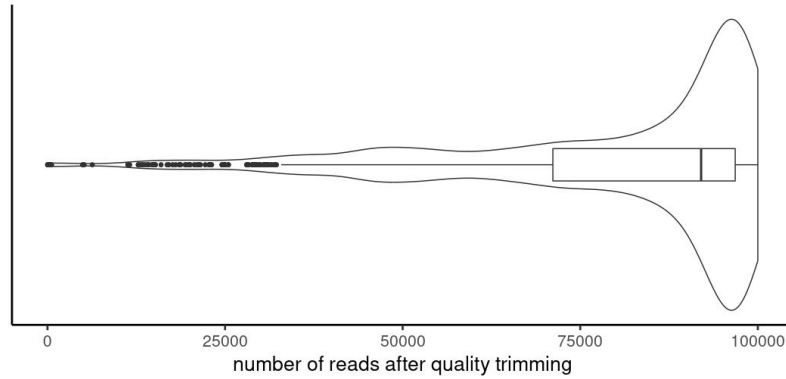


Figure 12: Distribution of the reads among samples downloaded from NCBI SRA database after quality trimming. The majority of the samples retained most of their reads, the mean number of reads per sample was 80,889.

After dereplication in the QIIME2 software, the reads were annotated using the consensus BLAST approach against three publicly available databases: Silva, Greengenes and Ezbio. The results are summed up in fig. 13. In terms of unannotated amplicon sequence variants (ASVs), the taxonomic assignment with the use of the Ezbio database yielded the worst results at all analysed taxonomic ranks, missing from almost 40 to 12.38% of the reads (Genus - Phylum). The lowest percentage of unannotated ASVs was observed with the use of the Silva database - here, the percentage of the unannotated reads at five taxonomic ranks varied from 15.43 to 6.55% of the total (fig. 13).

Generally, the higher the analysed rank, the more similarities were observed between the databases and the fewer unique assignments were detected. However, it is important to take into account that the increase in the taxonomic rank is creating the similarities only formally, as the higher taxonomic rank will accommodate multiple diverse groups from the lower ranks. Nevertheless, Silva generated the most diverse taxonomic results with a maximum of 38% of the unique taxonomic assignments at the Genus level fig. 13. There was little to almost no consistency (fig. 13) between uniquely Silva-Greengenes or Silva-Ezbio assignment results. However, Greengenes-Ezbio databases agreed on a taxonomic label in 4 to 22% of the obtained groups from Genus to Phylum rank. Due to the lowest percentage of unannotated reads and the highest observed diversity of the assigned taxonomy, further analyses were performed based on the feature table annotated by the Silva database.

The four analysed datasets were significantly different from each other in terms of the number of annotated ASVs, with the exception of animal-associated sets (animal and animal-gut samples), between which there were no differences ($p\text{-value} > 0.01$). The environment-related, therefore non-host related set, was the richest (fig. 13 ; tbl. 3) at all taxonomic ranks, followed by plant-related and animal-related sets. The environment- and plant-related sets were significantly richer ($p\text{-value} < 0.01$) than both animal-related sets (fig. 13). The species dominance metrics calculated using different indices were consistent and varied between the established 45 categories (fig. 14). Between most of the categories the differences were not statistically significant (STable2). At the Genus rank, only five categories had a DBP index above 0.75: shoot, flower, gill, spider nest and stalk&stem, while in a majority (78% of the categories) the most common genus accounted for up to 50% of the detected ASVs. Full information about the dominance metrics can be found in STable3. Additionally, at the taxonomic rank of Phylum, 15 categories did not share any taxon, while at the rank of the Genus this number did rise to 24 (data not shown). There was no taxonomic group at any taxonomic rank that would be present in all the samples from an established category and unique to it, even on the Genus rank.

Table 3: The mean number of the features at each taxonomic rank in the four investigated datasets, rounded up to the nearest tenth place.

Dataset	Phylum	Class	Order	Family	Genus
animal	9.1	13.7	28.3	38.4	51.8
animal-gut	8.4	12.3	24.0	33.9	52.1
environment	17.3	32.3	60.4	78.7	101.1
plant	11.3	21.3	41.1	55.4	75.6

Classification - PCoA and LDA weighted analysis

Dimensionality reduction methods like PCoA, PCA or LDA are often utilized as very useful preprocessing steps before the actual classification. They allow for simplification, enable visualization in 2D, at the same time reducing the redundancy and minimizing the information loss. They also reduce the required computational power, opening a way for big data analyses. For the process of sample classification with the use of ML algorithms, the selected data were processed in two ways: as abundance data, taking under consideration the amounts of the detected ASVs in each sample at every analysed rank, and as presence/absence data, therefore considering only the presence of the given taxonomic group at the respective ranks. The weighted analysis started with the abundance matrix, which was used to perform PCoA and LDA analysis. The training set, comprising 80% of the data obtained from the most informative principal coordinates (PCo1, PCo2 in PCoA) and linear combinations (LD1 and LD2) were used separately to train four classifiers. The accuracy and kappa (accuracy corrected by random chance) for the classification process based on weighted (abundance) data are presented in tbl. 4 and tbl. 5. Generally, the dissimilarities based on the microbial composition of the samples did not allow for enough resolution to differentiate between the categories. Analyses performed at all taxonomic ranks showed considerable overlap between the categories in the PCoA analysis as seen e.g. in fig. 15 and SFig1-4. The classification based on PCoA varied in kappa from 14 to 23.8% of proper predictions throughout the taxonomic ranks (tbl. 4). In the case of the classification based on the LDA analysis, which aims to maximize the variance and differences between the means of the categories, the kappa varied from 71.7 for NB classifier at Class rank to 97.6% for RF at Genus rank (tbl. 5 ; fig. 16 C and fig:best-classification-weighted-test C). The optimized parameters for the global set are listed in STable4. However, categories were unbalanced (tbl. 2) - there was a considerable difference between the number of samples between them, from 3 (e.g. cat eyes) to 249 (herbivore gut). The LEfSe analysis performed at the taxonomic ranks from Phylum to Family indicated very strong influence (LDA score $\log_{10} > 4$) of 26 (Phylum) to 147 (Family) investigated taxa on categories differentiation (STable5-8). Most of the categories had a very high (100%) sensitivity (rate of true positives) and specificity (rates of true negatives) of the classification. The most problematic categories in the global set classification were: seeds, insectivore-guts, shoots, insects whole body and aquatic animals skin, for which sensitivity dropped even to 0 at certain taxonomic ranks. Additionally, k-NN and NB classifiers were unable to properly distinguish (at most or all taxonomic ranks) categories with a low number of samples (SFig5), such as hemolymph, sheepskin, reproductive tract, acorn barnacles, fruits, pollen, salt marsh and saltern. The specificity dropped in the case of categories such as aquatic animals skin, herbivore or omnivore gut at the higher taxonomic ranks (i.e. Phylum and Class). Even though the above-mentioned metrics varied between the four classifiers, overall, the classifications based on the analyses performed at the lower taxonomic ranks (i.e. Genus, Family and Order) were more accurate. The classifier with the highest mean sensitivity (calculated over values at all taxonomic ranks) on the abundance data was Random Forest (82.3%), followed by NB (68.2%), Radial SVM (64.2%) and finally k-NN (62.2%). All classifiers throughout all taxonomic ranks had the same mean specificity of 99.7%.

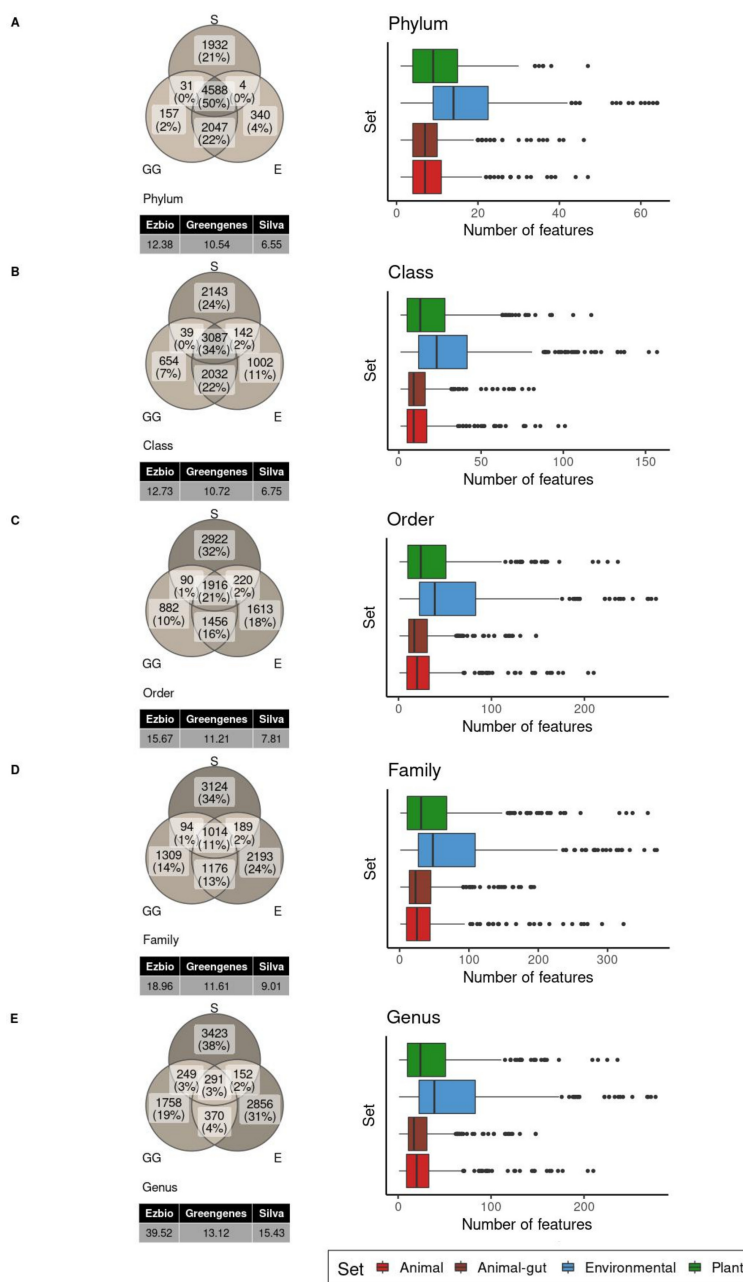


Figure 13: The comparison of the taxonomic annotations and number of features annotated with the use of Silva database in four datasets. The panels from A-E represent results at different taxonomic ranks. S=Silva, GG=Greengenes and E=Ezbio. Five panels (A-E) correspond to taxonomic ranks from Genus (A) to Phylum (E). The values in the table provide the percentage of the ASVs without annotation at the given taxonomic level for all databases. There were significant differences in the assigned taxonomy between the databases. Generally, as the taxonomic level increases, the number of uniquely assigned ASVs by all databases decreases and more similarities are observed. All the databases agreed on the assignment in 50% of the ASVs at the very general Phylum level. The lowest percentage of unannotated ASVs was observed with the use of the Silva database in all taxonomic ranks. The barplots represent richness - a number of features annotated with the Silva database. annotated

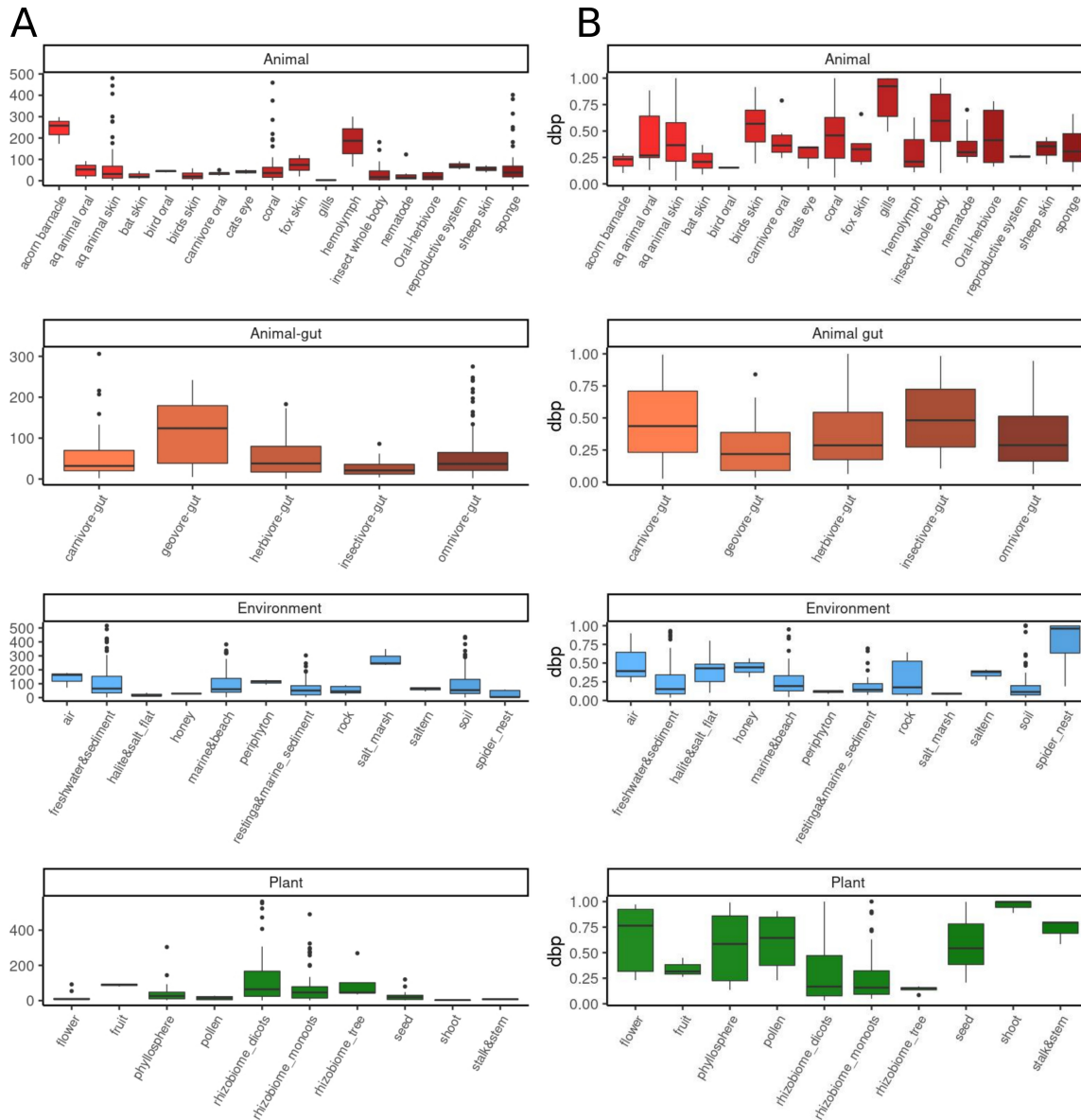


Figure 14: The richness and dominance metrics calculated for the four datasets. A. A richness (number of features in each category) is calculated based on the presence/absence matrix at the Genus taxonomic rank. B. The dominance index ‘DBP,’ or Berger–Parker index calculated at the Genus taxonomic level, is a special case of relative dominance with rank 1, meaning, it describes only the most abundant taxon and its proportion in relation to all detected microorganisms ($0.5 = 50\%$ of all detected groups). Among analysed categories distributed between the four datasets, the DBP index varies. The categories such as gills, spider nests or shoots were strongly dominated by the most common group, while others like acorn barnacles, bird oral samples, periphyton or rhizobiome of non-flowering trees were highly diverse. There was a visible negative correlation between the number of features in the categories (A) and the dominance index DBP (B).

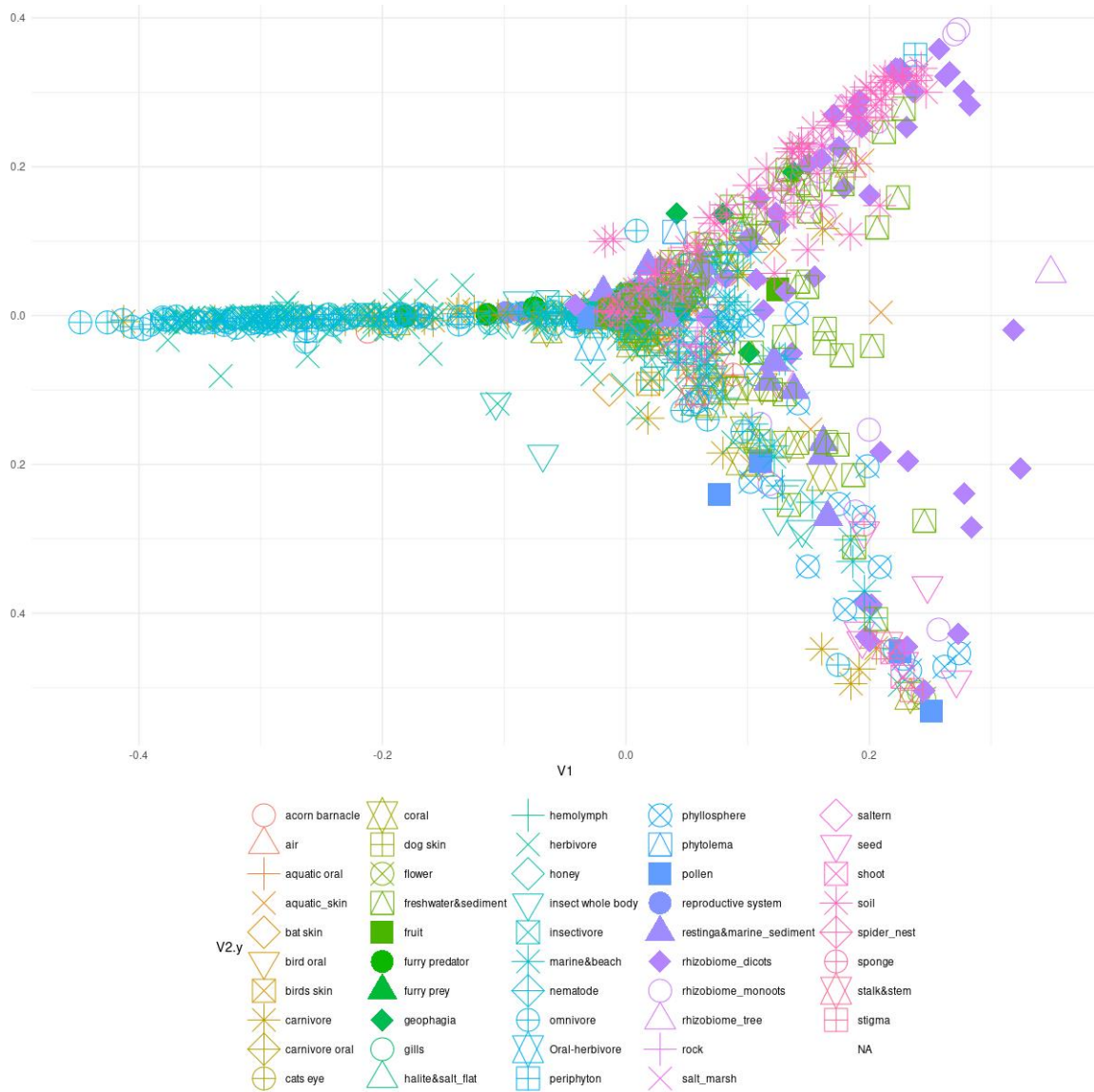


Figure 15: The result of the PCoA analysis of the global set of 1567 samples based on Bray-Curtis distance calculated from **the abundance matrix** (weighted data) at Genus taxonomic rank obtained from analysis with QIIME2. There is a visible overlap between the samples belonging to many of the 45 categories, which prevents highly accurate classification of the data points.

Table 4: The accuracy and kappa metrics of the classification, performed with the use of four classifiers, based on the principal components obtained from weighted data throughout taxonomic ranks.

Rank	radial SVM	Naive Bayes	Random Forest	K-NN
Genus	19.8/28.1	20.8/21.1	14/23.7	17/23.3
Family	19.8/25.2	22.5/22.1	15.8/22.1	16.1/23.1
Order	19.3/25.6	23.8/24.9	17.1/21.5	18.3/23.3
Class	15.5/23.3	21.3/20.2	18.1/20.2	15.8/20.5
Phylum	15.1/19.6	20.6/19.2	17.1/15.8	14.1/18.6

Table 5: The accuracy and kappa of the classification, performed with the use of four classifiers, based on the linear combinations obtained from weighted data throughout taxonomic ranks.

Rank	radial SVM	Naive Bayes	Random Forest	K-NN
Genus	96.9/97.2	94.5/94.9	97.6/97.8	94.2/95.6
Family	95.6/95.6	90.1/90.9	94.2/94.6	91.8/92.1
Order	92.5/93.1	88.1/89	90.8/91.5	88.4/89.9
Class	74.4/76.3	71.7/73.2	77.5/79.2	75.1/77.3
Phylum	84/85.2	79.8/81.4	84/85.2	81.6/93.9

Classification - KPCA and LDA unweighted analysis

The unweighted analysis started with the presence/absence matrix obtained from QIIME2 study of the 1567 samples, which was used to calculate the Jaccard distance, followed by KPCA and LDA analysis and classification with four different classifiers (similarly as before with the abundance data). The accuracy and kappa for the classification process based on unweighted data are presented in tbl. 6 and tbl. 7. Once again, the purely principal components-based classification was not enough to obtain high accuracy of the classification on any taxonomic rank investigated, due to insufficient variance between the categories. The best results in the case of the unweighted analysis were generated by k-NN on the Genus taxonomic rank - 21.8%, and the worst by NB at Phylum taxonomic rank - 8.8% (tbl. 6). The NB classifier failed to produce consistent classification areas when applied to the global data after LDA transformation (consisting of 45 categories) (fig. 20 B), instead it has created a very narrow fit for each category. However, it performed better on separate datasets (e.g. environment-related) (SFig6). The analysis based on LDA results yielded higher kappa and sensitivity. The highest kappa parameter - 97.6% of the correct predictions, resulted from RF classification at the Order level (tbl:unLDA-kappa ; fig. 19 C and fig. 20 C), the lowest kappa - 56.2% of the correct predictions, was obtained with the use of NB at the Phylum level. The optimized parameters are listed in STable4. In the case of unweighted data, the shoot samples were misclassified (as insects whole-body) by all used classifiers, with the exception of NB on genus rank. Additionally, there were multiple misclassifications at higher taxonomic ranks (i.e. Genus and Family) in the categories like carnivore-, herbivore-, omnivore-gut, rhizobiome-monocots and dicots, insects whole body or phyllosphere (SFig7). The specificity dropped mostly in the case of omnivore and herbivore gut data. The sensitivity and specificity of the classifiers in the case of the unweighted data over five taxonomic ranks was similar, as in the case of the weighted data. The classifier with the highest mean sensitivity throughout all taxonomic ranks on the presence/absence data was again RF (81.8%), followed by NB (66.9%), Radial SVM (63.9%) and finally again k-NN (62%). All classifiers throughout all taxonomic ranks had the same mean specificity of 99.6%.

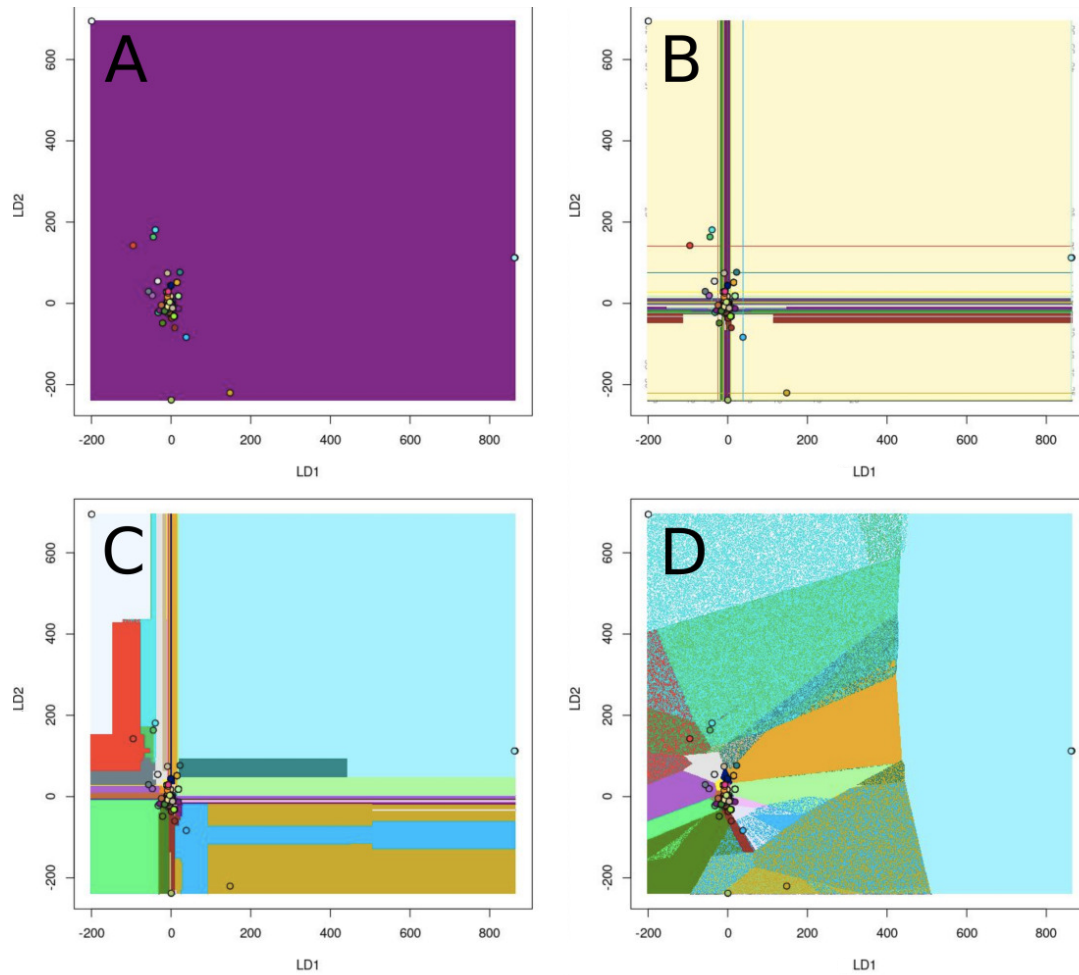


Figure 16: The visualization of the four classifiers' predictions of the categories areas for the **training** set created from the **weighted** dataset at the Genus taxonomic rank. The points represent the taxonomic data from bacterial sequencing samples transformed with the LDA method, their colours match the colour areas that mark the algorithm's prediction zones for 45 categories. A-radial SVM, B-Naive Bayes, C-RF, D-k-NN. Thanks to the distances obtained with the linear discriminants the categories are well separated, which facilitates the classification. The best results were obtained with the RF classifier (C) - 97.6 % of correct predictions (kappa).

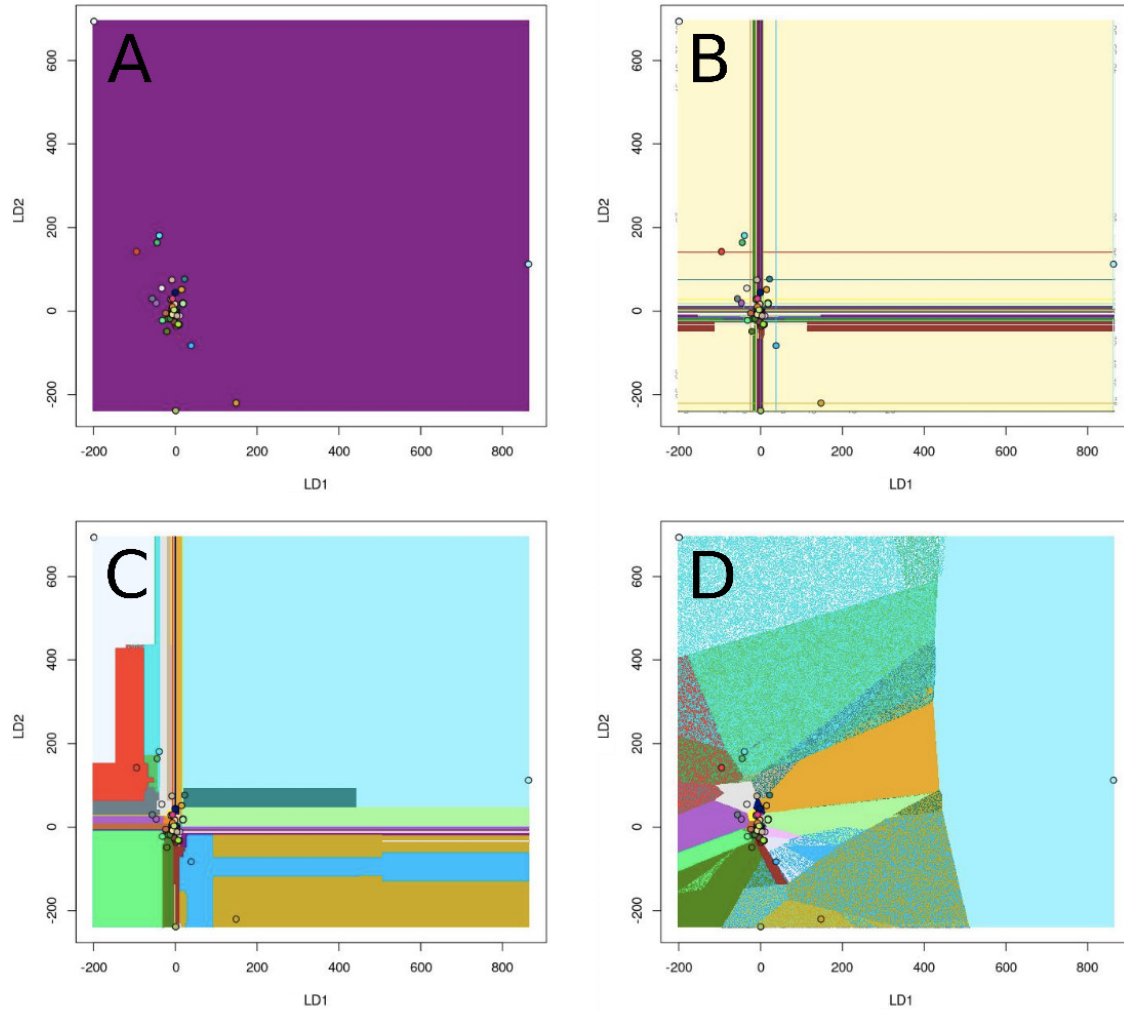


Figure 17: The visualization of the four classifiers' predictions of the categories areas for the **test** set created from the **weighted** dataset at the Genus taxonomic rank. The points represent the transformed with the LDA method taxonomic data from bacterial sequencing samples, their colours match the colour areas that mark the algorithm's prediction zones for 45 categories. A-radial SVM, B-Naive Bayes, C-RF, D-k-NN. The best results were obtained with the RF classifier (C) - 97.6 % of correct predictions.

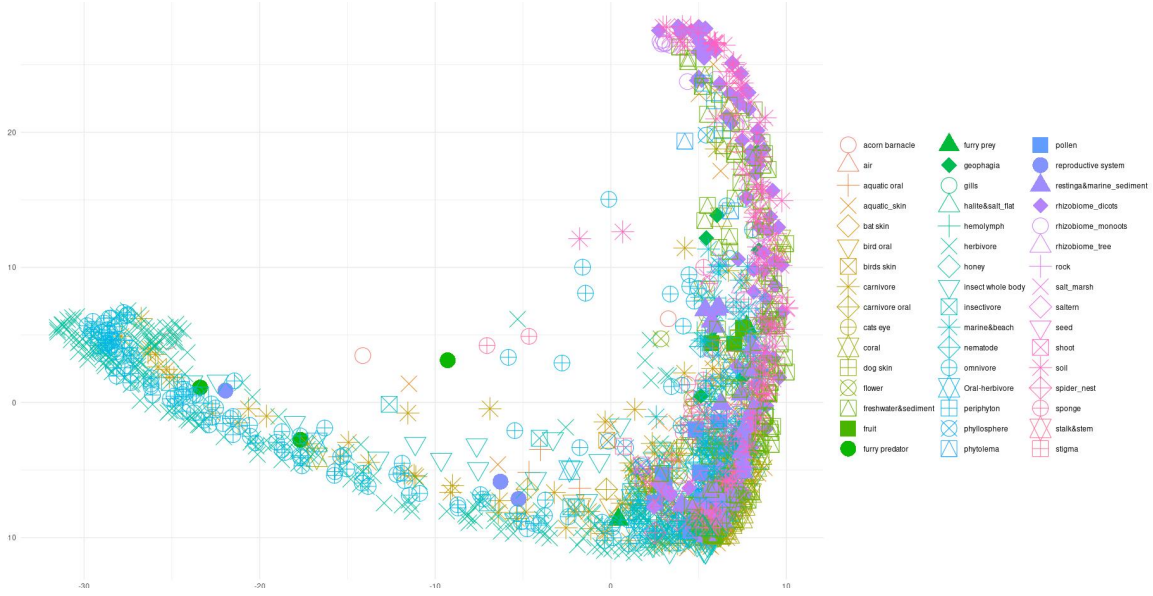


Figure 18: The result of the kernel PCA analysis of the global set of 1567 samples based on the Jaccard distance calculated from the **presence/absence matrix** at Genus taxonomic rank obtained from analysis with QIIME2. Similarly to the plot based on PCoA for the weighted data (fig. 15), there is a visible overlap between the 45 categories distributions, which prevents highly accurate classification of the data point.

Table 6: The kappa and accuracy of the classification process using four ML classifiers based on the results from KPCA analysis of unweighted data on five taxonomic ranks.

Rank	radial SVM	Naive Bayes	Random Forest	K-NN
Genus	18.2/ 28.6	18.1/ 25.3	20.7/27.7	21.8/25.5
Family	15.7/ 24.8	15.4/23.7	16.9/24.5	16.4/23.9
Order	15.7/23.3	14.9/20.2	15.6/24.5	16.1/20.4
Class	11.7/23.3	11.5/20	14.7/23	13.2/20.4
Phylum	9.3/20.8	8.8/19.5	15.8/23.6	14/19.8

Table 7: The kappa and accuracy of the classification process using four ML classifiers based on the results from LDA analysis of unweighted data on five taxonomic ranks.

Rank	radial SVM	Naive Bayes	Random Forest	K-NN
Genus	88.4/89.2	90.4/91.1	90/90.8	88.7/89.9
Family	94.6/95	94.5/94.5	97.3/97.5	94.6/95.6
Order	94.2/94.6	95.9/96.2	97.6/97.8	93.5/93.5
Class	69.8/72.2	69.8/71.9	75.4/77.6	71.7/74.8
Phylum	59.1/62.8	56.2/59.6	92.9/65.6	60.7/62.1

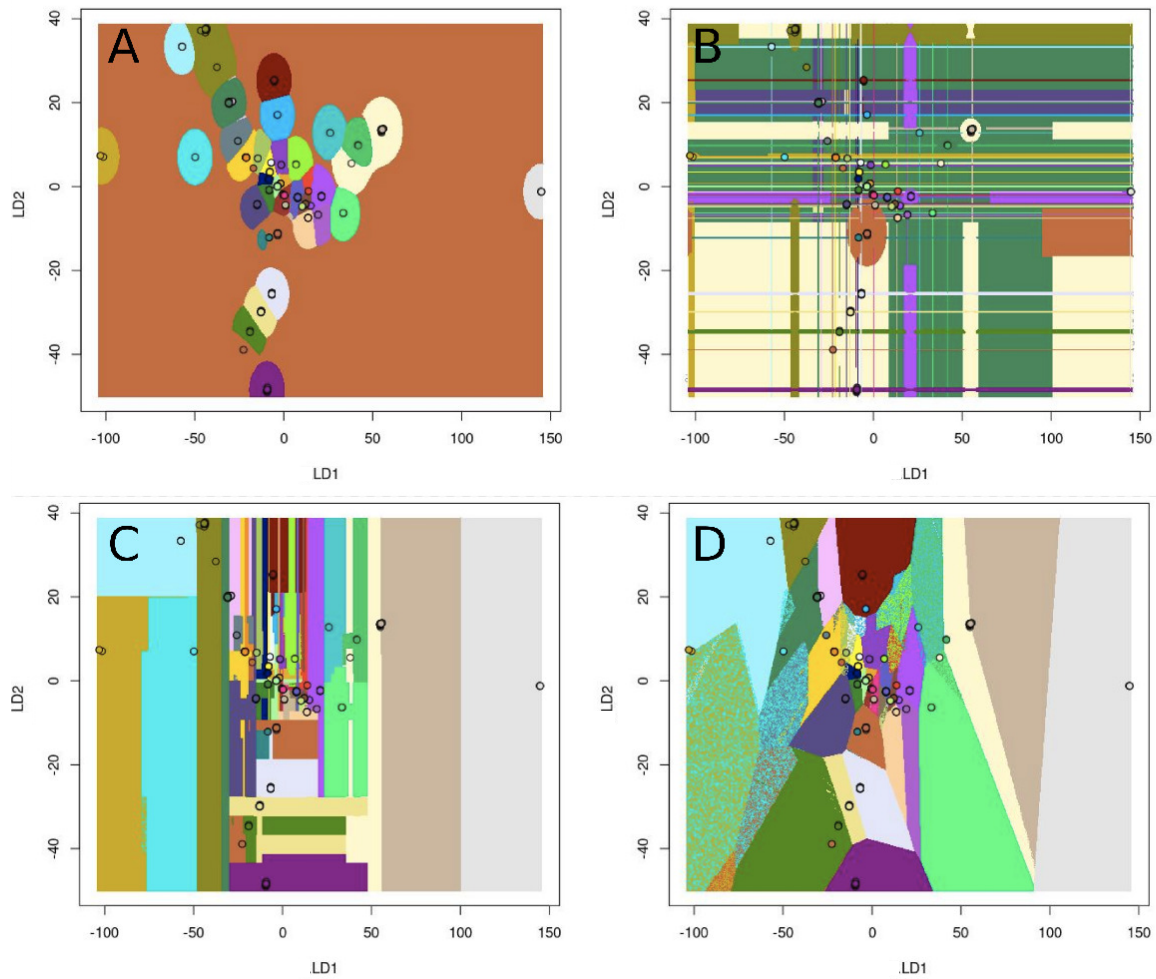


Figure 19: The visualization of the four classifiers' predictions of the categories areas for the **training set** created from the **unweighted** dataset at the Order taxonomic rank. The points represent the taxonomic data from bacterial sequencing samples transformed with the LDA method, their colours match the colour areas that mark the algorithm's prediction zones for 45 categories. A-radial SVM, B-Naive Bayes, C-RF, D-k-NN. Thanks to the distances obtained with the linear discriminants the categories are well separated, which facilitates the classification. The NB classifier (B) suffered from a low density of data on the entire presented surface and failed to produce consistent classification areas. The best results were obtained with the RF classifier (C) - 97.6 % of correct predictions.

Association rules

All four datasets separately and the global dataset were analysed with the Apriori algorithm to investigate whether significantly enriched cooccurrences of microorganisms could be detected. The highest number of significant rules were detected in the environment-related set at Order rank (233), the lowest number was observed also in the environment-related set, at Genus rank (2) (tbl. 8). Most of the 1326 detected rules at all taxonomic ranks were bilateral (924) with identical lift values. However, there were also 400 only one-sided rules. Out of 117 rules uncovered in the global set (four datasets merged), 78 rules also appeared when datasets were considered separately. However, 39

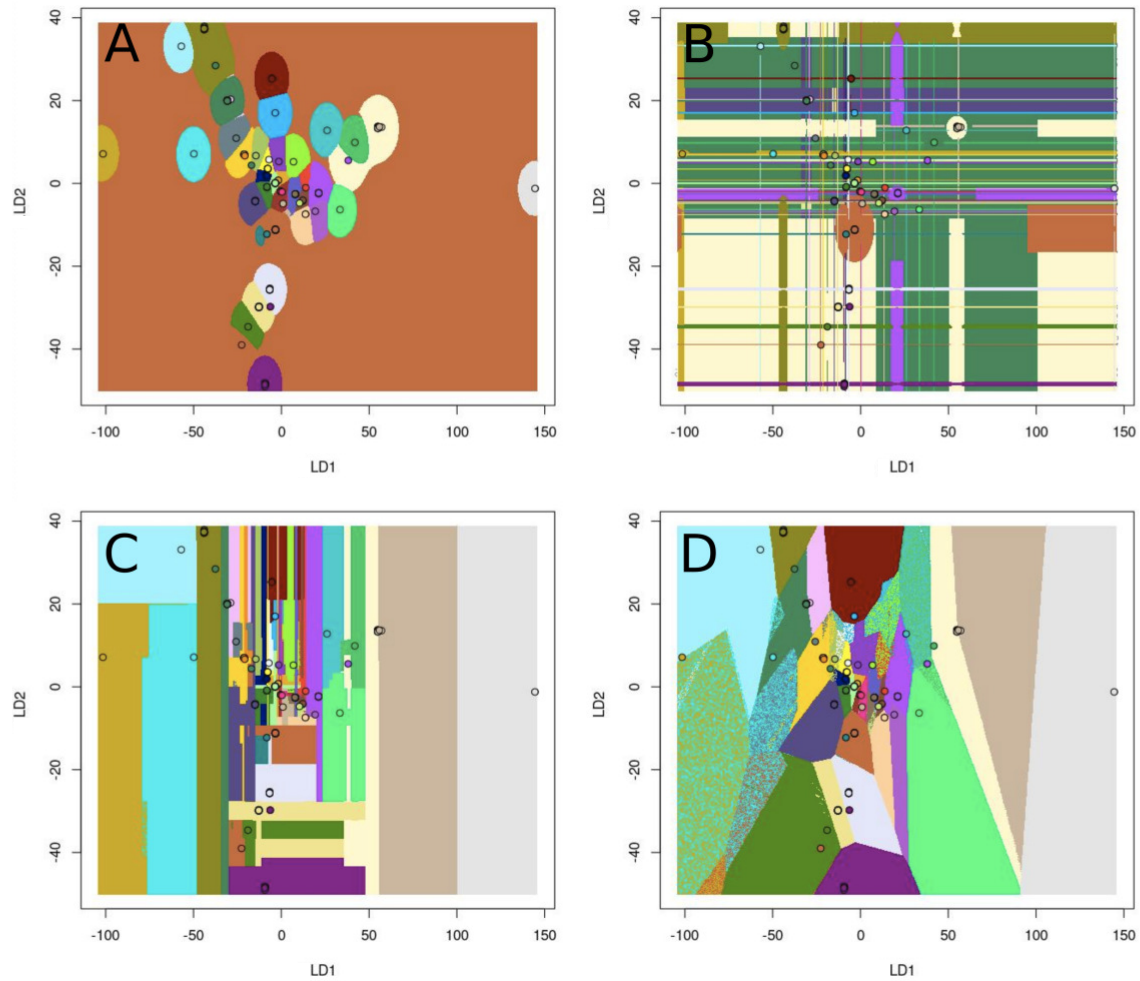


Figure 20: The visualization of the four classifiers' predictions of the categories areas for the **test set** created from the **unweighted** dataset at the Order taxonomic rank. The points represent the transformed with the LDA method taxonomic data from bacterial sequencing samples, their colours match the colour areas that mark the algorithm's prediction zones for 45 categories. A-radial SVM, B-Naive Bayes, C-RF, D-k-NN. The best results were obtained with the RF classifier (C) - 97.6 % of correct predictions.

rules could only be observed in the global set. At all taxonomic ranks, there were shared association rules between datasets (fig. 21), e.g. the shared rules at the Family rank were as follows:

- Total x Environment
p_Proteobacteria;c_Alphaproteobacteria;o_Sphingomonadales;**f_Sphingomonadaceae**
=>p_Cyanobacteria;c_Cyanobacteriia;o_Chloroplast;**f_Chloroplast**
- Animal x Plant
p_Proteobacteria;c_Alphaproteobacteria;o_Sphingomonadales;**f_Sphingomonadaceae**
<=>p_Proteobacteria;c_Alphaproteobacteria;o_Rhizobiales;**f_Rhizobiaceae**
- Animal x Environment
p_Proteobacteria;c_Alphaproteobacteria;o_Rhizobiales;**f_Rhizobiaceae**
=>p_Proteobacteria;c_Alphaproteobacteria;o_Sphingomonadales;**f_Sphingomonadaceae**
p_Proteobacteria;c_Alphaproteobacteria;o_Sphingomonadales;**f_Sphingomonadaceae**
=>p_Bacteroidota;c_Bacteroidia;o_Flavobacteriales;**f_Flavobacteriaceae**
- Plant x Environment
p_Proteobacteria;c_Gammaproteobacteria;o_Burkholderiales;**f_Burkholderiaceae**
<=>p_Bacteroidota;c_Bacteroidia;o_Chitinophagales;**f_Chitinophagaceae**
p_Bacteroidota;c_Bacteroidia;o_Chitinophagales;**f_Chitinophagaceae**
<=>p_Proteobacteria;c_Gammaproteobacteria;o_Burkholderiales;**f_Comamonadaceae**
p_Proteobacteria;c_Gammaproteobacteria;o_Burkholderiales;**f_Comamonadaceae**
<=>p_Proteobacteria;c_Alphaproteobacteria;o_Sphingomonadales;**f_Sphingomonadaceae**
p_Proteobacteria;c_Gammaproteobacteria;o_Burkholderiales;**f_Comamonadaceae**
<=>p_Bacteroidota;c_Bacteroidia;o_Sphingobacteriales;**f_Sphingobacteriaceae**
p_Proteobacteria;c_Gammaproteobacteria;o_Burkholderiales;**f_Comamonadaceae**
=>p_Cyanobacteria;c_Cyanobacteriia;o_Chloroplast;**f_Chloroplast**
p_Proteobacteria;c_Alphaproteobacteria;o_Rhizobiales;**f_Rhizobiaceae**
=>p_Proteobacteria;c_Alphaproteobacteria;o_Sphingomonadales;**f_Sphingomonadaceae**
p_Proteobacteria;c_Alphaproteobacteria;o_Rhizobiales;**f_Xanthobacteraceae**
=>p_Proteobacteria;c_Alphaproteobacteria;o_Sphingomonadales;**f_Sphingomonadaceae**

Note: ‘p’ = Phylum, ‘c’ = Class, ‘o’ = Order, ‘f’=Family. Notice that some rules are one-sided (=>), whereas others are bilateral (<=>).

All the considered rules had the lift > 1. The environment-related set contained the most statistically significant rules at the taxonomic ranks from Phylum to Order. The majority of the unveiled microbial associations were interconnected with each other (fig. 22, fig. 23), forming a network. The most broadly connected bacterium order was the Sphingobacteriales, which occurred significantly more often with 19 different orders in environment-related set: Vicinamibacterales, Bacteroidales, Chitinophagales, Cytophagales, Flavobacteriales, Sphingobacteriales, Oligoflexales, c_Anaerolineae;o_SBR1031, Gemmatimonadales, Tepidisphaerales, Caulobacterales, Rhizobiales, Rhodospirillales, Rickettsiales, Sphingomonadales, Burkholderiales, Cellvibrionales, Gammaproteobacteria_Incertae_Sedis, Steroidobacterales, Opitutales, Pedosphaerales and Verrucomicrobiales (fig. 22 A). The lift values for the above-mentioned rules were between 1.07 to 2.63. Additionally, in the global set at the Phylum rank, there were three rules with the double antecedent (here, significantly increased co-occurrence of two taxa with one other taxon), indicating stronger ties between three bacteria Phyla: Actinobacteriota, Firmicutes and Proteobacteria. The full list of mined association rules and their parameters is available in STable9-32.

Table 8: Number of statistically significant rules detected for each dataset at five taxonomic ranks, number of unique (appearing at least once) antecedents and number of unique consequents at each taxonomic rank. The highest number of significant rules were detected in the environment-related set at Order rank, whereas the lowest number was observed in the environment-related set at the Genus rank.

Dataset	Phylum	Class	Order	Family	Genus
animal	19\10\6	4\3\3	35\14\10	5\3\3	4\4\4
animal-gut	22\8\7	8\5\5	51\18\10	36\15\13	4\3\3
environment	130\25\16	109\28\21	233\51\36	85\32\23	2\2\2
plant	86\13\13	12\5\5	199\34\29	147\33\33	18\10\7
global	61\19\7	23\9\5	31\14\7	1\1\1	0\0\0

Discussion

The rapid growth of microbiome related studies resulted in a wide availability of raw sequencing data from multiple sources, e.g. animal-derived, food-associated or environmental samples . The component of the prokaryotic ribosome 30S subunit, namely 16S rRNA is a popular, universal and conserved marker used in taxonomic identification studies. It consists of nine hypervariable regions (V1-V9), which exhibit different levels of conservation (Bukin et al. (2019)). The market of the metabarcoding is right now largely dominated by Illumina sequencing, which usually generates reads up to 300bp (“Maximum read length for Illumina sequencing platforms Illumina” (n.d.)). Due to the size constraint, metabarcoding of the samples is limited to one or two hypervariable regions. Multiple studies have shown that the choice of the this region for sequencing can substantially influence the estimates of the taxonomic diversity (Yu et al. (2008), Klindworth et al. (2013), Yang, Wang, and Qian (2016), Bukin et al. (2019)), e.g. Bakin et al. have proven that sequencing with the use of primers targeting V2-V3 regions is more precise than sequencing targeting regions V3-V4, i.e. the former can be used for species and genus identification, unlike the latter. Another study (Mizrahi-Man, Davenport, and Gilad (2013)) studied the coverage and confidence thresholds of taxonomic identification using seven distinct study designs on five taxonomic ranks, from Genus to Phylum. They recommended the use of primers targeting either hypervariable regions V3 or V4 for interrogating bacterial communities. For the sake of that study the researchers have defined a statistical measure, the coverage, as the number of times the rank is predicted with confidence greater or equal to the threshold (with 5% of false positive rate), divided by the total number of sequences with a prediction at any level. The obtained results showed that the aforementioned coverage for the rank of Order and higher remained above 90% for paired-end reads and dropped below 70% at Genus rank. Similar conclusions were presented by Bokulich et al., 2018, who showed that most methods perform well from Phylum to Family rank identification, but performance decreases at Genus and Species ranks. It is important to remember that identification at higher taxonomic rank is by essence less precise, as the higher ranks encompass multiple different lower ones. Still, it is also more certain, especially when taking under consideration many various primer pairs targeting different hypervariable regions of 16S rRNA. This is why care has to be taken when choosing the taxonomic rank at which the results are reported, so as to present the maximally precise, but also maximally certain taxonomic compositions.

Among the metabarcoding samples used in the current study, only 120 out of 1567 come with metadata about the sequencing primers. Among those 120 that did, 60 samples targeted V4 region, 27 samples V3-V4 regions, 21 samples V3 region, 3 samples V2 region, 3 samples V3-V5 samples regions, 3 samples V5-V6 regions, and 3 samples V5-V7 regions. Due to the uncertainty of the

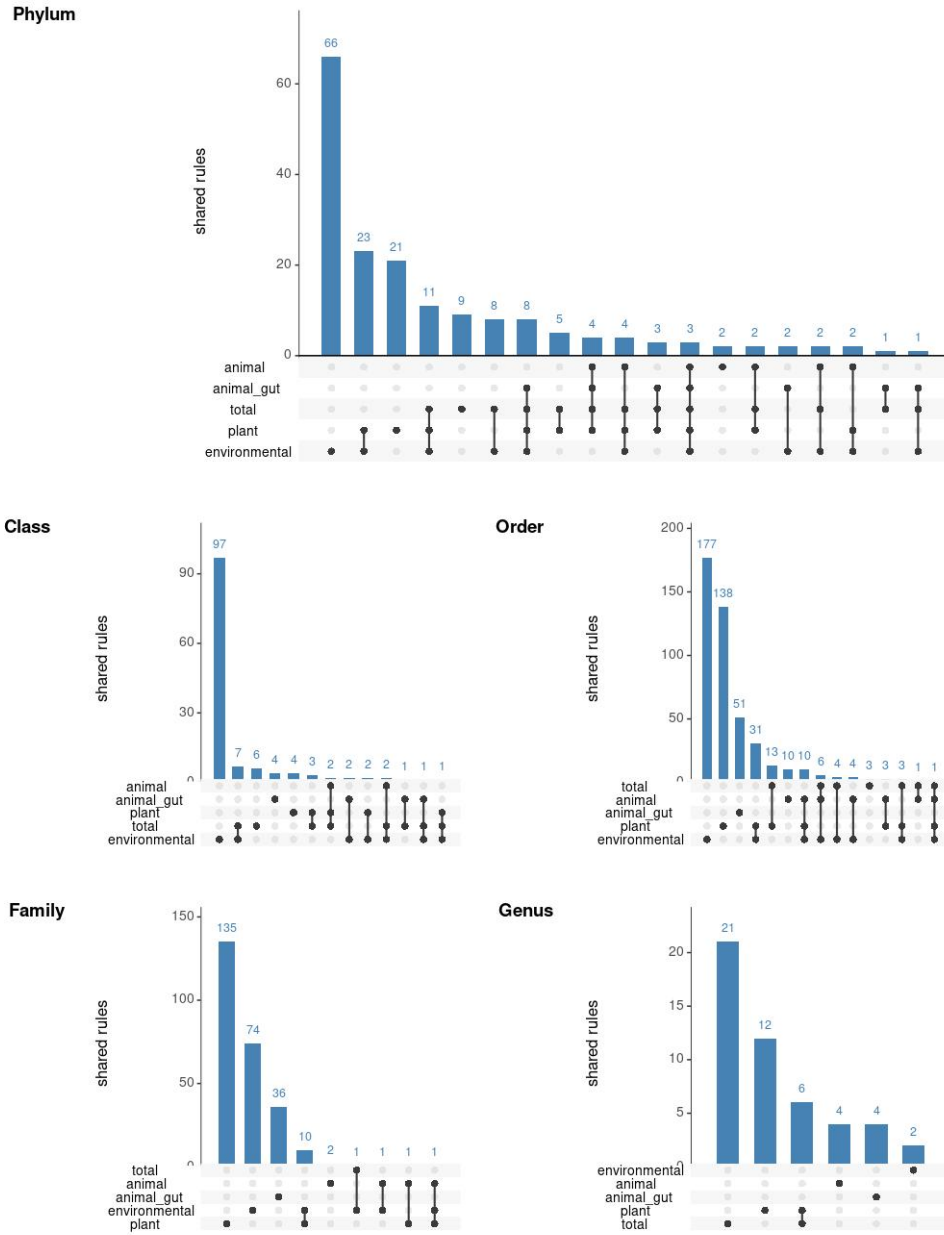
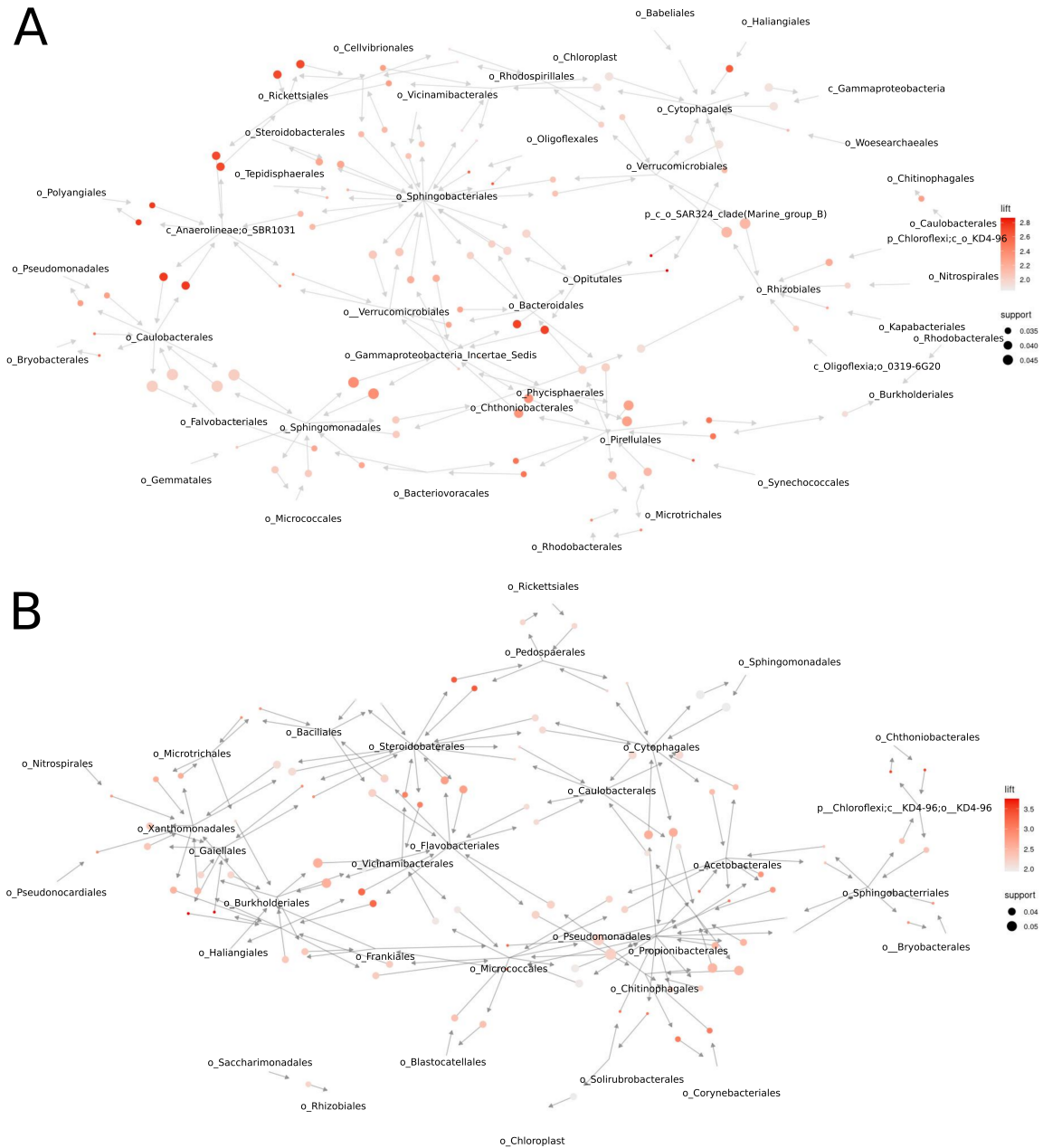


Figure 21: The shared and unique rules at each investigated taxonomic rank among four datasets. The highest number of unique rules is observed for environment-related data at the Family rank. At the Genus rank, there are no shared associations between the datasets observed.



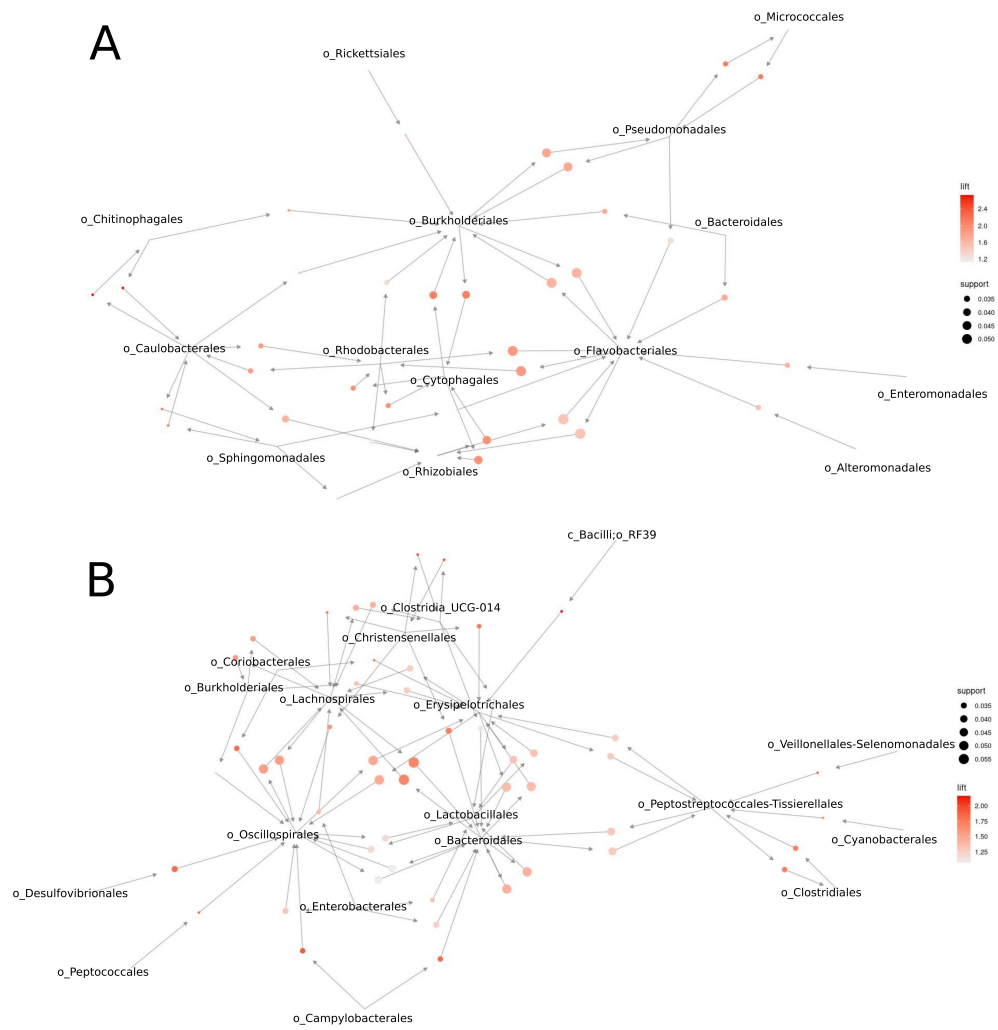


Figure 23: The significant associations detected in two datasets. The visualized rules were detected in the animal (A) and animal-gut-related (B) datasets on the Order rank. The richness of the host-associated datasets was smaller and less significant associations were detected.

hypervariable regions sequenced in most samples, the raw sequences were not grouped into OTUs and processed together, but rather processed separately with the use of the consensus BLAST with the optimized parameters as suggested in Bokulich et al. (2018), against three different databases. Multiple studies have evaluated different databases in terms of their usability for ASV identification. The tested databases were non-specialized like Greengenes, Silva, BLAST16S, UNITE, EzBio (Bokulich et al. 2018; Edgar 2018; Park and Won 2018) or specialized like DAIRYdb or Bee Gut Microbiota-Database (BGM-Db) (Meola et al. 2019; Xue et al. 2019). However, neither of the tested databases proved to be universal. The outcome of the identifications depended strongly on the datasets considered and the method used (BLAST, NaiveBayes, UCLUST, etc.). Therefore in this work, due to the multifariousness of the analysed data, three non-specialized databases (Silva, Greengenes and Ezbio) were used as a reference for taxonomic identification of ASVs with consensus BLAST. The best results (highest diversity and the percentage of annotated ASVs) were obtained with the Silva database at all tested taxonomic ranks, and therefore further processing was performed on feature tables annotated with the use of this database. Since the analysed dataset consisted of the ASVs originating from different hypervariable regions of 16S rRNA, simple clustering into OTUs would lead to many orphaned ASVs creating their own OTU, even though taxonomically, they would belong to the same organism as members of other OTUs. Therefore, instead of OTU creation, after BLAST annotation of the sequences, the abundances of ASVs belonging to the same group of organisms (be that Genera, Families, etc. at specific taxonomic rank) were added to each other to create the final collapsed feature table. The final feature tables were a starting point for further analysis. Overall, at all examined taxonomic ranks, the animal- and animal-gut-related data were the least rich out of the considered datasets. The mean richness was higher for plant-related data, which consisted of free-living (e.g. rock, marine sediment) and host-related (rhizobiomes) data. The highest mean number of species were identified in the only free-living environment-related dataset. Those results are congruent with the results obtained by Earth Microbiome Project (Thompson et al. 2017), during which the same dependency between the habitat and the detected species richness have been observed. The abundance-based dominance indices were unanimous and showed that the categories with smaller organism richness were also in general more strongly dominated by the most abundant species. This dependency has already been hypothesized to be a result of competition for resources or the consequence of the regional species pool effect (Akatoov and Perevozov 2011). This hypothesis proposes that the species richness of the habitat is dependent on its size and age, which influences the speciation opportunities. Hence, the larger and older the particular habitat is, the more associated species can be observed.

The microbiome composition study results in the generation of high-dimension data. While performing classifications with the use of ML algorithms on the ‘raw’ feature table is doable, this sort of data is usually redundant and at least part of it is non-informative towards the designated scientific goal. What is more, algorithms like k-NN suffer from the ‘curse of dimensionality’ and high-dimension data impair their performance. There are various approaches to preprocessing data that reduce the data space, e.g. approach implemented in QIIME2 software (Bolyen et al. 2019), which performs UniFrac analysis of the data before further processing. UniFrac is a distance metric commonly used in ecological studies, measuring relatedness of the organisms, while including the phylogenetic information. Oudah and Henschel (2018) have created a feature selection algorithm that would allow users to select the most relevant features while discarding the lower taxonomic rank if they were redundant in relation to their parent rank. The same authors have also presented a PCoA analysis of over 30,000 environmental samples as published by Henschel, Anwar, and Manohar (2015) and the failure of this technique in fine-scale differentiation of the samples originating from the various sources. The PCoA analysis for weighted data and PCA analysis of the unweighted data performed in this study, were also based on the relatedness metrics. However, due to the unique character of the meta-analysis based on different 16S rRNA regions attempted here, it was impossible to employ UniFrac, but rather proceed as described in Materials and Methods metrics. Nonetheless,

the results based solely on PCA/PCoA showed very low prediction accuracy (14.1.-23.8% for abundance and 8.8-21.8% for presence/absence based data). There was no real possibility to separate the categories efficiently regardless of the applied ML algorithm, due to considerable overlap between the categories. This result is still better than an average from zero classifier (assigning all new data points to the most abundant category), still, it remains unsatisfactory. The second method used in this study, LDA, performed much better. As it focuses not only on reducing dimensionality by maximizing the variance, but also maximizing the means of the investigated category, LDA is a very useful tool for data preprocessing. It has been used with success to compress data in studies targeting Neural Network-based heart failures predictions (Burse et al. 2019) or radial kernel-based breast cancer detection (Omondiagbe, Veeramani, and Sidhu 2019). Here, LDA transformation of the data allowed the ML algorithms to predict the category with a much better kappa metric at each investigated taxonomic rank, for both weighted and unweighted data. The best results were obtained at Genus rank for weighted data and at Order rank for unweighted data with the use of RF algorithm, which frequently produces superior results when compared to other supervised algorithms such as SVM or NB classifiers (Uddin et al. 2019). Although, for some particular applications like image processing, SVM classifier matches Random Forest in the accuracy (Kranjčević et al. 2019).

Due to the blend of the data originating from the various, mostly unknown primers, targeting different hypervariable regions, choosing to perform analyses on the Genus rank seems misguided. As already proven hypervariable regions perform differently and many are incapable of providing a good (comparable to full 16S rRNA sequence) resolution up to the Genus rank (J. S. Johnson et al. 2019) and might therefore impair the future scaled-up analyses. Another matter that has to be taken into consideration to create a good foundation for automatic bacterial sample metabarcoding classification, is equalization of the category abundances, or at least, enriching the ones with the least samples ($n=3$). What is more, in this study, many interesting categories had to be excluded due to an insufficient number of samples, even though the need for replicates has been highlighted as a necessity (Andr  n et al. 2008; Prosser 2010). As the microbiome studies remain increasingly popular, it is my hope that the availability of metabarcoded samples originating from less popular and still unstudied sources would increase and thanks to the method described here they could be incorporated into the microbiome analysis and used for category classification.

The association rules learning is an unsupervised machine learning technique that checks for the dependencies between transactions (here the transactions correspond to the species richness in the analysed samples). The results of this work have unveiled a number of co-occurrences between microorganisms at all analysed taxonomic ranks in samples originating from the different sources. Generally, more rules were detected in the environment- and plant-related samples, which can be partially explained by higher species richness in the host-free and rhizobiome data as shown on the data from EMP (Thompson et al. 2017). In all datasets and at all taxonomic ranks a number of highly connected taxa has been identified. The most broadly connected taxa were Order Sphingobacteriales co-occurring significantly more often with 19 other Orders. Interestingly, Sphingobacteriales is a keystone taxon - defined by Paine in 1966 as important and non-redundant components of a community influencing its structure and integrity (Paine 1969). The keystone taxa can be inferred by performing a co-occurrence analysis (correlation-based) or with association rule mining (Centler et al. 2020). Studies up to now have established a list of keystone taxa in 11 ecosystems, like grasslands, soil, plant, aquatic etc., through computational inferences and empirical evidence (Banerjee, Schlaeppli, and Heijden 2018). Recently, researchers have established a number of connections between mycobacterial Phyla (Ma et al. 2020) using the data accumulated within the Earth Microbiome Project. Ten other taxa appeared in the analysis forming a highly connected node (connecting more than one other taxon at Order rank) and were identified before as a keystone taxa:

- Burkholderiales
- Chloroflexi

- Flavobacteriales
- Oceanospirillales
- Pseudomonadales
- Rhizobiales
- Solirubrobacterales
- Sphingobacteriales
- Verrucomicrobiales
- Xanthomonadales

What is more, all aforementioned keystone taxa apart from two (Chloroflexi, Solirubrobacterales) appeared also in LEfSe analysis, proving that they were important for explaining the differences between established categories. Other taxa that formed nodes with 10 or more connections were Sphingomonadales, Propionibacteriales, Caulobacterales, Cytophagales, Acetobacterales, Chthoniobacterales, Rickettsiales, Polyangiales, Pedosphaerales, Micrococcales, Steroidobacterales and Pirellulales. Apart from the latter five, the highly connected taxa also appeared relevant in LEfSe analysis. Many of those species are heterotrophs, forming complex consortia, taking part in biodegradation, biotransformation and carbon cycle processes (Sun et al. 2019; Morohoshi et al. 2018; Hug et al. 2013). They appear in a wide range of environments, from sludge, through soil, bacterioneuston (community of bacteria present in surface microlayers) and bacterioplankton to oil spillage (e.g. The Persian Gulf) (Taylor et al. 2014; Kim et al. 2014; Somee et al. 2021). To establish whether the identified taxa plays a role as keystone taxa (in other environments than the ones already studied) the detailed analysis of correlation is required, possibly with supplementary metabarcoding sequencing analyses. The keystone taxa are crucial for the proper functioning and structure of the environment they inhabit, therefore the empirical evidence of changes under different conditions is necessary to prove their status.

Conclusions

Public databases are a rich source of metabarcoding data performed in separate experiments using different sets of primers and targeting different hypervariable regions of 16S rRNA. They can be used in a meta-analysis without clustering into OTUs based on sequence similarities, but rather using the assigned taxonomy, to sum the reads belonging to the same organism in the final feature table. Optimization of the used for this task database is therefore crucial. Another matter is the choice of the taxonomic rank, in which the results are communicated. While researchers generally aim at providing the most precise and accurate data at the lowest possible rank (optimally Species or Genus), it is not always possible. The study design and, most importantly the choice of the primers, is crucial for the high resolution of identification of ASVs. It has been reported that primers targeting two, instead of one, hypervariable regions deliver better results. Still, the most precise and certain results are obtained when the full length of the 16S rRNA is sequenced, which for now remains difficult to do for a metabarcoding experiment.

The meta-analysis performed in this work aimed at establishing a method for fine-scale categorization of sample sources. The PCA and PCoA-based analysis of the distances proved that there is not enough variance in the investigated data to distinguish between the categories. However, the application of the LDA analysis proved successful in the given task. Similar corrected accuracy were obtained for weighted data at Genus rank and for unweighted data at Order rank - 97.6% of correct predictions. Despite this, in my opinion, the data reporting at the Order rank with unweighted is the most appropriate, precisely due to the meta-analysis specifics mentioned above.

Association rule learning analysis of the four sets and a global set have detected a number of co-occurrences of microorganisms on different taxonomic ranks. Several different taxa were observed

to form highly connected nodes. Those taxa can be regarded as putative keystone taxa and considered for further investigation in different niches.

Note

All the figures presented in this work were created in R programming language (Team et al. 2013) and Inkscape (Bah 2007). Figure 1 was created after (Thompson et al. 2017) using location data provided by the authors in the supplementary materials. Figures 7-10 were inspired by <https://www.kdnuggets.com/>. The remaining 18 visualizations are original.

Supplementary materials

Supplementary Figures

- SFig1. The result of the PCoA analysis of the global set of 1567 samples based on Bray-Curtis distance calculated from the abundance matrix (weighted data) at Family taxonomic rank obtained from analysis with QIIME2. There is a visible overlap between the samples belonging to many of the 45 categories, which prevents highly accurate classification of the data point.
- SFig2. The result of the PCoA analysis of the global set of 1567 samples based on Bray-Curtis distance calculated from the abundance matrix (weighted data) at Order taxonomic rank obtained from analysis with QIIME2. There is a visible overlap between the samples belonging to many of the 45 categories, which prevents highly accurate classification of the data point.
- SFig3. The result of the PCoA analysis of the global set of 1567 samples based on Bray-Curtis distance calculated from the abundance matrix (weighted data) at Class taxonomic rank obtained from analysis with QIIME2. There is a visible overlap between the samples belonging to many of the 45 categories, which prevents highly accurate classification of the data point.
- SFig4. The result of the PCoA analysis of the global set of 1567 samples based on Bray-Curtis distance calculated from the abundance matrix (weighted data) at Phylum taxonomic rank obtained from analysis with QIIME2. There is a visible overlap between the samples belonging to many of the 45 categories, which prevents highly accurate classification of the data point.
- SFig5. The detailed sensitivity of the classification of four classifiers throughout taxonomic ranks for all 45 categories for analysis based on the unweighted data. The k-NN classifier showed the lowest sensitivity, RF the highest. Categories with a low number of samples were classified less accurately, especially by k-NN and NB classifiers. None of the used classifiers (with the exception of NB on genus rank) classified the shoot samples properly.
- SFig6. NaiveBayes classification of training set for Environmental-related and Plant-related sets at Order rank.
- SFig7. The detailed specificity of the classification performed by four classifiers throughout taxonomic ranks for all 45 categories for analysis based on the unweighted data. The lowest specificity was obtained for aquatic animal skin, insect whole body and herbivore-, carnivore- and omnivore gut categories for all classifiers throughout the taxonomic ranks (Phylum and Class).
- SFig8. The detailed sensitivity of the classification of four classifiers throughout taxonomic ranks for all 45 categories for analysis based on the weighted data. The k-NN classifier showed the lowest sensitivity, RF the highest. Categories with a low number of samples were classified less accurately, especially by k-NN and Naive Bayes classifiers.

- SFig9. The detailed specificity of the classification performed by four classifiers throughout taxonomic ranks for all 45 categories for analysis based on the weighted data. The lowest specificity was obtained for aquatic animals skin, insects whole body and herbivore gut categories for all classifiers at the highest taxonomic ranks (Phylum and Class).

Supplementary Tables

- STable1. The metadata of 1567 samples downloaded from NCBI SRA databases analysed in this study.
- STable2. Dominance metrics differences significance between 45 categories calculated with pairwise t.test with bonferroni correction at Genus rank. The metrics in order: DBP, DMN, absolute, simpson.
- STable3. Detailed dominance metrics of all samples at all analysed ranks.
- STable4. Optimized parameters for all classifiers at all analysed ranks.
- STable5 - 8. Results of LEfSe analysis at Phylum - Family rank.
- STable9-32. Results of Association Rules Learning Analysis of all four sets and global set at all taxonomic ranks.

References

- Akatov, VV, and AG Perevozov. 2011. "Relationship Between Dominance and Richness of Local Species: An Analysis of the Underlying Reasons with Arboreal and Avian Communities of the West Caucasus as an Example." *Zhurnal Obshchei Biologii* 72 (2): 111–26.
- Anderson, Edgar. 1935. "The Irises of the Gaspé Peninsula." *Bull. Am. Iris Soc.* 59: 2–5.
- Andr  n, O, H Kirchmann, T K  tterer, J Magid, EA Paul, and DC Coleman. 2008. "Visions of a More Precise Soil Biology." *European Journal of Soil Science* 59 (2): 380–90.
- Anesio, Alexandre M, Stefanie Lutz, Nathan AM Christmas, and Liane G Benning. 2017. "The Microbiome of Glaciers and Ice Sheets." *Npj Biofilms and Microbiomes* 3 (1): 1–11.
- Bah, Tavmjong. 2007. *Inkscape: Guide to a Vector Drawing Program*. prentice hall press.
- Bahram, Mohammad, Falk Hildebrand, Sofia K Forslund, Jennifer L Anderson, Nadejda A Soudzilovskaia, Peter M Bodegom, Johan Bengtsson-Palme, et al. 2018. "Structure and Function of the Global Topsoil Microbiome." *Nature* 560 (7717): 233–37.
- Banerjee, Samiran, Klaus Schlaeppli, and Marcel GA van der Heijden. 2018. "Keystone Taxa as Drivers of Microbiome Structure and Functioning." *Nature Reviews Microbiology* 16 (9): 567–76.
- Baquero, Fernando, and Cesar Nombela. 2012. "The Microbiome as a Human Organ." Wiley Online Library.
- Behravan, Hamid, Jaana M Hartikainen, Maria Tengstrom, Katri Pylkas, Robert Winqvist, Veli-Matti Kosma, and Arto Mannermaa. 2018. "Machine Learning Identifies Interacting Genetic Variants Contributing to Breast Cancer Risk: A Case Study in Finnish Cases and Controls." *Scientific Reports* 8 (1): 1–13.
- Berg, G, D Rybakova, D Fischer, T Cernava, MCC Verg  s, T Charles, X Chen, et al. 2020. "Microbiome Definition Re-Visited: Old Concepts and New Challenges. Microbiome 8: 103."
- Bokulich, Nicholas A, Benjamin D Kaehler, Jai Ram Rideout, Matthew Dillon, Evan Bolyen, Rob Knight, Gavin A Huttley, and J Gregory Caporaso. 2018. "Optimizing Taxonomic Classification of Marker-Gene Amplicon Sequences with QIIME 2's Q2-Feature-Classifer Plugin." *Microbiome* 6 (1): 1–17.
- Bolger, Anthony M, Marc Lohse, and Bjoern Usadel. 2014. "Trimmomatic: A Flexible Trimmer for Illumina Sequence Data." *Bioinformatics* 30 (15): 2114–20.

- Bolyen, Evan, Jai Ram Rideout, Matthew R Dillon, Nicholas A Bokulich, Christian C Abnet, Gabriel A Al-Ghalith, Harriet Alexander, et al. 2019. "Reproducible, Interactive, Scalable and Extensible Microbiome Data Science Using QIIME 2." *Nature Biotechnology* 37 (8): 852–57.
- Boser, Bernhard E, Isabelle M Guyon, and Vladimir N Vapnik. 1992. "A Training Algorithm for Optimal Margin Classifiers." In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, 144–52.
- Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32.
- Bukin, Yu S, Yu P Galachyants, IV Morozov, SV Bukin, AS Zakharenko, and TI Zemskaya. 2019. "The Effect of 16s rRNA Region Choice on Bacterial Community Metabarcoding Results." *Scientific Data* 6 (1): 1–14.
- Burse, Kavita, Vishnu Pratap Singh Kirar, Abhishek Burse, and Rashmi Burse. 2019. "Various Preprocessing Methods for Neural Network Based Heart Disease Prediction." In *Smart Innovations in Communication and Computational Sciences*, 55–65. Springer.
- Buttigieg, Pier Luigi, Evangelos Pafilis, Suzanna E Lewis, Mark P Schildhauer, Ramona L Walls, and Christopher J Mungall. 2016. "The Environment Ontology in 2016: Bridging Domains with Increased Scope, Semantic Density, and Interoperation." *Journal of Biomedical Semantics* 7 (1): 1–12.
- Callahan, Benjamin J, Paul J McMurdie, Michael J Rosen, Andrew W Han, Amy Jo A Johnson, and Susan P Holmes. 2016. "Dada2: High-Resolution Sample Inference from Illumina Amplicon Data." *Nature Methods* 13 (7): 581–83.
- Caporaso, J Gregory, Christian L Lauber, William A Walters, Donna Berg-Lyons, James Huntley, Noah Fierer, Sarah M Owens, et al. 2012. "Ultra-High-Throughput Microbial Community Analysis on the Illumina HiSeq and MiSeq Platforms." *The ISME Journal* 6 (8): 1621–24.
- Centler, Florian, Sarah Gunnigmann, Ingo Fetzer, and Annelie Wendeberg. 2020. "Keystone Species and Modularity in Microbial Hydrocarbon Degradation Uncovered by Network Analysis and Association Rule Mining." *Microorganisms* 8 (2): 190.
- Chibucos, Marcus C, Adrienne E Zweifel, Jonathan C Herrera, William Meza, Shabnam Eslamfam, Peter Uetz, Deborah A Siegle, James C Hu, and Michelle G Giglio. 2014. "An Ontology for Microbial Phenotypes." *BMC Microbiology* 14 (1): 1–8.
- Conway, Jake R, Alexander Lex, and Nils Gehlenborg. 2017. "UpSetR: An r Package for the Visualization of Intersecting Sets and Their Properties." *Bioinformatics*.
- Cooper, Laurel, Ramona L Walls, Justin Elser, Maria A Gandolfo, Dennis W Stevenson, Barry Smith, Justin Preece, et al. 2013. "The Plant Ontology as a Tool for Comparative Plant Anatomy and Genomic Analyses." *Plant and Cell Physiology* 54 (2): e1–1.
- DeSantis, Todd Z, Philip Hugenholtz, Neils Larsen, Mark Rojas, Eoin L Brodie, Keith Keller, Thomas Huber, Daniel Dalevi, Ping Hu, and Gary L Andersen. 2006. "Greengenes, a Chimera-Checked 16s rRNA Gene Database and Workbench Compatible with ARB." *Applied and Environmental Microbiology* 72 (7): 5069–72.
- Dictionary, Oxford English. 1989. "Oxford English Dictionary." *Simpson, Ja & Weiner, Esc.*
- Edgar, Robert C. 2018. "Accuracy of Taxonomy Prediction for 16s rRNA and Fungal ITS Sequences." *PeerJ* 6: e4652.
- Fisher, Ronald A. 1936. "The Use of Multiple Measurements in Taxonomic Problems." *Annals of Eugenics* 7 (2): 179–88.
- Fix, Evelyn, and Joseph Lawson Hodges. 1951. "Nonparametric Discrimination: Consistency Properties." *Randolph Field, Texas, Project*, 21–49.
- France. 1886. *Journal Officiel de La Republique Francaise*. Journaux officiels. <https://books.google.be/books?id=8aJAAAAYAAJ>.
- Gdanetz, Kristi, and Frances Trail. 2017. "The Wheat Microbiome Under Four Management Strategies, and Potential for Endophytes in Disease Protection." *Phytobiomes* 1 (3): 158–68.

- Gonzalez, Antonio, Jose A Navas-Molina, Tomasz Kosciolk, Daniel McDonald, Yoshiki Vazquez-Baeza, Gail Ackermann, Jeff DeReus, et al. 2018. "Qiita: Rapid, Web-Enabled Microbiome Meta-Analysis." *Nature Methods* 15 (10): 796–98.
- Hahsler, Michael. 2017. "arulesViz: Interactive Visualization of Association Rules with r." *R J.* 9 (2): 163.
- Hahsler, Michael, Sudheer Chelluboina, Kurt Hornik, and Christian Buchta. 2011. "The Arules r-Package Ecosystem: Analyzing Interesting Patterns from Large Transaction Data Sets." *The Journal of Machine Learning Research* 12: 2021–25.
- Hargreaves, Carol Anne. 2019. "Machine Learning Application to Identify Good Credit Customers." *International Journal of Advanced Engineering and Technology*.
- Henschel, Andreas, Muhammad Zohaib Anwar, and Vimitha Manohar. 2015. "Comprehensive Meta-Analysis of Ontology Annotated 16s rRNA Profiles Identifies Beta Diversity Clusters of Environmental Bacterial Communities." *PLoS Computational Biology* 11 (10): e1004468.
- Hu, Junyan, Hanlin Niu, Joaquin Carrasco, Barry Lennox, and Farshad Arvin. 2020. "Voronoi-Based Multi-Robot Autonomous Exploration in Unknown Environments via Deep Reinforcement Learning." *IEEE Transactions on Vehicular Technology* 69 (12): 14413–23.
- Hug, Laura A, Cindy J Castelle, Kelly C Wrighton, Brian C Thomas, Itai Sharon, Kyle R Frischkorn, Kenneth H Williams, Susannah G Tringe, and Jillian F Banfield. 2013. "Community Genomic Analyses Constrain the Distribution of Metabolic Traits Across the Chloroflexi Phylum and Indicate Roles in Sediment Carbon Cycling." *Microbiome* 1 (1): 1–17.
- Johnson, Jethro S, Daniel J Spakowicz, Bo-Young Hong, Lauren M Petersen, Patrick Demkowicz, Lei Chen, Shana R Leopold, et al. 2019. "Evaluation of 16s rRNA Gene Sequencing for Species and Strain-Level Microbiome Analysis." *Nature Communications* 10 (1): 1–11.
- Johnson, Katerina V-A, and Kevin R Foster. 2018. "Why Does the Microbiome Affect Behaviour?" *Nature Reviews Microbiology* 16 (10): 647–55.
- Karatzoglou, Alexandros, Alexandros Smola, Kurt Hornik, and Achim Zeileis. 2004. "Kernlab-an S4 Package for Kernel Methods in r." *Journal of Statistical Software* 11 (1): 1–20.
- Kim, Ye-Eun, Hyeokjun Yoon, Young-Hyun You, Hyun Kim, Yeonggyo Seo, Miae Kim, Ju-Ri Woo, et al. 2014. "Diversity and Characteristics of Rhizosphere Microorganisms Isolated from the Soil Around the Roots of Three Plants Native to the Dokdo Islands." *Journal of Life Science* 24 (4): 461–66.
- Klindworth, Anna, Elmar Pruesse, Timmy Schweer, Jörg Peplies, Christian Quast, Matthias Horn, and Frank Oliver Glöckner. 2013. "Evaluation of General 16s Ribosomal RNA Gene PCR Primers for Classical and Next-Generation Sequencing-Based Diversity Studies." *Nucleic Acids Research* 41 (1): e1–1.
- Konopka, Allan. 2009. "Ecology, Microbial." Pacific Northwest National Lab.(PNNL), Richland, WA (United States).
- Kranjčević, Nikola, Damir Medak, Robert Župan, and Milan Rezo. 2019. "Machine Learning Methods for Classification of the Green Infrastructure in City Areas." *ISPRS International Journal of Geo-Information* 8 (10): 463.
- Kuhn, Max. 2008. "Building Predictive Models in r Using the Caret Package." *Journal of Statistical Software* 28 (1): 1–26.
- Lederberg, Joshua, and Alexa T McCray. 2001. "Ome SweetOmics—a Genealogical Treasury of Words." *The Scientist* 15 (7): 8–8.
- Ley, Ruth E, Catherine A Lozupone, Micah Hamady, Rob Knight, and Jeffrey I Gordon. 2008. "Worlds Within Worlds: Evolution of the Vertebrate Gut Microbiota." *Nature Reviews Microbiology* 6 (10): 776–88.
- Liaw, Andy, and M Wiener. 2013. "Documentation for r Package randomForest." *PDF*. Retrieved 15: 191.
- Liu, Xiaoni. 2016. "Focus: Microbiome: Microbiome." *The Yale Journal of Biology and Medicine* 89 (3): 275.

- Ma, Bin, Yiling Wang, Shudi Ye, Shan Liu, Erinne Stirling, Jack A Gilbert, Karoline Faust, et al. 2020. "Earth Microbial Co-Occurrence Network Reveals Interconnection Pattern Across Microbiomes." *Microbiome* 8: 1–12.
- Madeira, F. 2019. "Mi Park." *Lee J., Buso N., Gur T., Madhusoodanan N., Basutkar P., Tivey ARN, Potter SC, Finn RD, Et Al. The EMBL-EBI Search and Sequence Analysis Tools APIs in.*
- Mann, H, W Wells, and S Blasco. 1996. "Rusticles from the RMS Titanic."
- "Maximum read length for Illumina sequencing platforms Illumina." n.d.
<https://emea.support.illumina.com/bulletins/2020/04/maximum-read-length-for-illumina-sequencing-platforms.html>.
- McMurdie, Paul J, and Susan Holmes. 2013. "Phyloseq: An r Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data." *PloS One* 8 (4): e61217.
- Meola, Marco, Etienne Rifa, Noam Shani, Celine Delbès, Heene Berthoud, and Christophe Chassard. 2019. "DAIRYdb: A Manually Curated Reference Database for Improved Taxonomy Annotation of 16s rRNA Gene Sequences from Dairy Products." *BMC Genomics* 20 (1): 1–16.
- Mitchell, Tom M, and Machine Learning. 1997. "The McGraw-Hill Companies." *Inc., New York.*
- Mizrahi-Man, Orna, Emily R Davenport, and Yoav Gilad. 2013. "Taxonomic Classification of Bacterial 16s rRNA Genes Using Short Sequencing Reads: Evaluation of Effective Study Designs." *PloS One* 8 (1): e53608.
- Morohoshi, Tomohiro, Taishiro Oi, Haruna Aiso, Tomohiro Suzuki, Tetsuo Okura, and Shunsuke Sato. 2018. "Biofilm Formation and Degradation of Commercially Available Biodegradable Plastic Films by Bacterial Consortia in Freshwater Environments." *Microbes and Environments*, ME18033.
- Mungall, Christopher J, Carlo Torniai, Georgios V Gkoutos, Suzanna E Lewis, and Melissa A Haendel. 2012. "Uberon, an Integrative Multi-Species Anatomy Ontology." *Genome Biology* 13 (1): 1–20.
- Nian, Rui, Jinfeng Liu, and Biao Huang. 2020. "A Review on Reinforcement Learning: Introduction and Applications in Industrial Process Control." *Computers & Chemical Engineering* 139: 106886.
- Oksanen, Jari, F Guillaume Blanchet, Roeland Kindt, Pierre Legendre, Peter R Minchin, RB O'hara, Gavin L Simpson, Peter Solymos, M Henry H Stevens, and Helene Wagner. 2015. "Vegan: Community Ecology Package. R Package Version 2.2-1."
- Omondigbe, David A, Shanmugam Veeramani, and Amandeep S Sidhu. 2019. "Machine Learning Classification Techniques for Breast Cancer Diagnosis." In *IOP Conference Series: Materials Science and Engineering*, 495:012033. 1. IOP Publishing.
- Oudah, Mai, and Andreas Henschel. 2018. "Taxonomy-Aware Feature Engineering for Microbiome Classification." *BMC Bioinformatics* 19 (1): 1–13.
- Paine, Robert T. 1969. "A Note on Trophic Complexity and Community Stability." *The American Naturalist* 103 (929): 91–93.
- Park, Sang-Cheol, and Sungho Won. 2018. "Evaluation of 16s rRNA Databases for Taxonomic Assignments Using a Mock Community." *Genomics & Informatics* 16 (4).
- Piatetsky-Shapiro, Gregory. 1991. "Discovery, Analysis, and Presentation of Strong Rules." *Knowledge Discovery in Databases*, 229–38.
- Prosser, James I. 2010. "Replicate or Lie." *Environmental Microbiology* 12 (7): 1806–10.
- Qu, Qian, Zhenyan Zhang, WJGM Peijnenburg, Wanyue Liu, Tao Lu, Baolan Hu, Jianmeng Chen, Jun Chen, Zhifen Lin, and Haifeng Qian. 2020. "Rhizosphere Microbiome Assembly and Its Impact on Plant Growth." *Journal of Agricultural and Food Chemistry* 68 (18): 5024–38.
- Quast, Christian, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. 2012. "The SILVA Ribosomal RNA Gene Database Project: Improved Data Processing and Web-Based Tools." *Nucleic Acids Research* 41 (D1): D590–96.
- Segata, Nicola, Jacques Izard, Levi Waldron, Dirk Gevers, Larisa Miropolsky, Wendy S Garrett, and Curtis Huttenhower. 2011. "Metagenomic Biomarker Discovery and Explanation." *Genome Biology* 12 (6): 1–18.

- Shreiner, Andrew B, John Y Kao, and Vincent B Young. 2015. "The Gut Microbiome in Health and in Disease." *Current Opinion in Gastroenterology* 31 (1): 69.
- Somee, Maryam Rezaei, Seyed Mohammad Mehdi Dastgheib, Mahmoud Shavandi, Leila Ghanbari Maman, Kaveh Kavousi, Mohammad Ali Amoozegar, and Maliheh Mehrshad. 2021. "Distinct Microbial Community Along the Chronic Oil Pollution Continuum of the Persian Gulf Converge with Oil Spill Accidents." *Scientific Reports* 11 (1): 1–15.
- Sun, Haohao, Takashi Narihiro, Xueyan Ma, Xu-Xiang Zhang, Hongqiang Ren, and Lin Ye. 2019. "Diverse Aromatic-Degrading Bacteria Present in a Highly Enriched Autotrophic Nitrifying Sludge." *Science of the Total Environment* 666: 245–51.
- Taylor, Joe D, Samuel D Cottingham, Jack Billinge, and Michael Cunliffe. 2014. "Seasonal Microbial Community Dynamics Correlate with Phytoplankton-Derived Polysaccharides in Surface Coastal Waters." *The ISME Journal* 8 (1): 245–48.
- Team, R Core et al. 2013. "R: A Language and Environment for Statistical Computing."
- Thompson, Luke R, Jon G Sanders, Daniel McDonald, Amnon Amir, Joshua Ladau, Kenneth J Locey, Robert J Prill, et al. 2017. "A Communal Catalogue Reveals Earth's Multiscale Microbial Diversity." *Nature* 551 (7681): 457–63.
- Uddin, Shahadat, Arif Khan, Md Ekramul Hossain, and Mohammad Ali Moni. 2019. "Comparing Different Supervised Machine Learning Algorithms for Disease Prediction." *BMC Medical Informatics and Decision Making* 19 (1): 1–16.
- Ursell, Luke K, Jessica L Metcalf, Laura Wegener Parfrey, and Rob Knight. 2012. "Defining the Human Microbiome." *Nutrition Reviews* 70 (suppl_1): S38–44.
- Vapnik, Vladimir, and Alexey Chervonenkis. 1974. "Theory of Pattern Recognition." Nauka, Moscow.
- Venables, William N, and Brian D Ripley. 2013. *Modern Applied Statistics with s-PLUS*. Springer Science & Business Media.
- Wang, DeLiang. 2001. "Book Review Unsupervised Learning." Citeseer.
- Xue, Zhang, Li Xing'an, Su Qinzhi, Cao Qina, Li Chenyi, Niu Qingsheng, and Zheng Hao. 2019. "A Curated 16s rRNA Reference Database for the Classification of Honeybee and Bumblebee Gut Microbiota." *Biodiversity Science* 27 (5): 557.
- Yan, Qiulong, Yifang Gu, Xiangchun Li, Wei Yang, Liqiu Jia, Changming Chen, Xiuyan Han, et al. 2017. "Alterations of the Gut Microbiome in Hypertension." *Frontiers in Cellular and Infection Microbiology* 7: 381.
- Yang, Bo, Yong Wang, and Pei-Yuan Qian. 2016. "Sensitivity and Correlation of Hypervariable Regions in 16s rRNA Genes in Phylogenetic Analysis." *BMC Bioinformatics* 17 (1): 1–8.
- Yilmaz, Pelin, Renzo Kottmann, Dawn Field, Rob Knight, James R Cole, Linda Amaral-Zettler, Jack A Gilbert, et al. 2011. "Minimum Information about a Marker Gene Sequence (MIMARKS) and Minimum Information about Any (x) Sequence (MIXS) Specifications." *Nature Biotechnology* 29 (5): 415–20.
- Yoon, Seok-Hwan, Sung-Min Ha, Soonjae Kwon, Jeongmin Lim, Yeseul Kim, Hyungseok Seo, and Jongsik Chun. 2017. "Introducing EzBioCloud: A Taxonomically United Database of 16s rRNA Gene Sequences and Whole-Genome Assemblies." *International Journal of Systematic and Evolutionary Microbiology* 67 (5): 1613.
- Yu, Zhongtang, Ruben Garcia-Gonzalez, Floyd L Schanbacher, and Mark Morrison. 2008. "Evaluations of Different Hypervariable Regions of Archaeal 16s rRNA Genes in Profiling of Methanogens by Archaea-Specific PCR and Denaturing Gradient Gel Electrophoresis." *Applied and Environmental Microbiology* 74 (3): 889–93.
- Zhu, Baoli, Xin Wang, and Lanjuan Li. 2010. "Human Gut Microbiome: The Second Genome of Human Body." *Protein & Cell* 1 (8): 718–25.
- Zuckerkandl, Emile, and Linus Pauling. 1965. "Evolutionary Divergence and Convergence in Proteins." In *Evolving Genes and Proteins*, 97–166. Elsevier.