PHYLOGENETIC APPROACHES TO MICROBIAL COMMUNITY CLASSIFICATION

by

Jie Ning

Submitted in partial fulfillment of the
requirements for the degree of
Master of Computer Science

at

Dalhousie University
Halifax, Nova Scotia
August 2015

Table of Contents

List of Tables

List of Figures

Abstract

The microorganisms associated with our body, collectively known as the microbiome, have profound impacts on biological processes including human health and disease. Different body sites are dominated by different major groups of microbes, but the variations within a body site, such as the mouth, can be more subtle. High-throughput DNA sequencing allows the assessment of the microbiome at an unprecedented scale, but creates new computational challenges. Machine-learning procedures can serve as useful tools for distinguishing microbes from similar body sites, understanding key organisms and their roles can highlight deviations from expected distributions of microbes.

We focused our attention on the classification of nine oral sites, and dental plaque in particular, using data collected from the Human Microbiome Project. A key focus of our representations was the use of phylogenetic information, both as the basis for custom kernels and as a way to represent sets of microbes to the classifier. We also used the PICRUSt software, which draws on phylogenetic relationships to predict molecular functions, to generate additional features for the classifier. Custom kernels based on the UniFrac measure of community dissimilarity did not improve performance. However, feature representation was vital to classification accuracy, with microbial clade and functional representations providing useful information to the classifier. However, these two types of information were correlated rather than complementary, and combining the two types of features did not yield increased prediction accuracy. Many of the best-performing clades and functions had clear associations with the oral microbiome.

The classification of oral microbiota remains a challenging problem; our best accuracy on the plaque dataset was approximately 81%. Perfect accuracy may be unattainable due to the close proximity of the sites and intra-individual variation. However, further exploration of the space of both classifiers and feature representations is likely to increase the accuracy of predictive models.

List of Abbreviations and Symbols Used

16S rRNA	16S ribosomal RNA
CSS	Cumulative Sum Scaling
EDA	Exploratory Data Analysis
GG	GreenGenes
HMP	Human Microbiome Project
HMPDACC	Human Microbiome Project Data Analysis and Coordination Center
KEGG	Kypoto Encyclopedia of Genes and Genomes
NCBI	National Center of Biotechnology Information
NSTI	Nearest Sequenced Taxon Index
OTU	Operational Taxonomic Unit
PCA	Principal Component Analysis
PCoA	Principal Coordinate Analysis
PICRUSt	Phylogenetic Investigation of Communities by Reconstruction of Unob- served States
QC	Quality Control
QIIME	Quantitative Insights Into Microbial Ecology
RBF	Radial Basis Function
RDP	Ribosomal Database Project

RF	Random Forest
SVM	Support Vector Machine
TSS	Total Sum Scaling
UPGMA	Unweighted Pair Group Method with Arithmetic Mean

Acknowledgements

First of all, I would like to express my deepest gratitude to my supervisor Dr. Robert Beiko. Thanks for taking a chance on me when I came without any bioinformatics experience but only a heart full of interest. His far-sighted perspectives and enlightening suggestions on research make me learn a lot during the last two years. Most importantly, thank him for his detailed and patient guidance with his students.

Thanks are given to my thesis committee: Dr. Hong Gu and Dr. Thomas Trappenberg, for Dr. Gu's important suggestions on statistical analysis, for Dr. Trappenberg's diligent teaching in machine learning course I took. I am grateful for their detailed comments and questions on my thesis.

I want to thank all people in the lab, especially Dennis Wong, Michael Hall and Alex Keddy lfor their invaluable help. Thanks also go to people in Blouin Lab, Dr. Christian Blouin, Michelle Lu and Sergio Hleap, for providing me confidence during my master study.

Finally, I owe many thanks to my father and mother, for their endless love and patience with me.

Chapter 1

Introduction

The microorganisms that coexist within the human host are referred to as the human microbiome. The human body harbors a tremendous number of microbial cells that can outnumber human cells by a factor of 10 [6, 7, 8, 9, 10]. Interacting closely with their host, these microorganisms play an important role in human biological processes and disease states [11, 12, 13, 14, 15, 16]. For example, the intestinal microbes are capable of producing some required vitamins that the human body cannot synthesize, such as vitamin B12 and vitamin K [17]. Iron absorption [18, 19, 17] and the formation of antioxidants [20] in the human body also rely heavily on microbes. Microbes are also linked to different diseases, such as inflammatory bowel disease [21, 22, 11], periodontal disease [23, 24, 25, 15, 26, 27] and skin disease [28, 29, 30, 31].

The human oral cavity is one of the most diverse and complex microbial habitats to analyze, for several reasons. First, many ecologically distinct sites including different types of plaque, different oral surfaces, and saliva are found in close proximity to one another [12, 32, 33], which makes it easier for microbes to migrate among these sites. Second, the oral habitat is highly variable with frequent inputs of nutrients, often followed by mechanical removal of the biofilm (e.g., via tooth brushing). Third, the oral microbiome is also implicated in a number of diseases, including dental caries, periodontal disease and even infections in heart and liver [34, 35, 15].

Traditional microbiome studies were performed using laboratory culture methods [6, 36, 37, 38], which identify the microorganisms by plating samples on different artificial media. Culture-based approaches are slow and limit the detectable organisms to the minority that could be grown in a laboratory environment. DNA sequencing refers to the process of determining the order of nucleotides in a specific molecule of DNA. In a microbiome study, these sequences can reveal genetic information about the microorganisms. Recent developments in high-throughput DNA sequencing techniques allow a large number of microbial sequences to be identified

in a short time [39, 40]. With so much microbiome data available, interpretation of these huge datasets is a daunting challenge. Machine-learning approaches can be used to identify critical information to classify or distinguish microbiome samples. Several studies have constructed microbial features from DNA sequence and used machine-learning algorithms for classification [41, 42, 43, 44]. Classifying the samples between major body sites is relatively "easy" and commonly performed, however, differentiating samples within sites is challenging.

In our work, we tackled the problem of classifying human oral cavity samples, especially those associated with hard plaque. Integrating phylogenetic information among these microbes improved classifier performance. The work in this thesis is geared towards improving classifiers in order to better characterize microbial communities in the oral cavity. Important features for the classifiers also reveal the discriminative microorganisms and key functions within the microbial community.

1.1 Human Microbiome

1.1.1 Microbiome Interactions and Human Host

The term "human microbiome" was first proposed by Joshua Lederberg: "Microbiome is the ecological community of commensal, symbiotic, and pathogenic microorganisms that literally share our body space" [45]. Taking different parts of our body as their habitats, the microbes have substantial effects on human biological processes and disease states.

Microbial communities often interact with the host in a non-disease-inducing way. Members of these communities can have mutualistic relationships with the human host, where both partners derive some benefit from the association. Intestinal microbes inhabit in the human body and rely on nutrients from the host to survive. At the same time, these bacteria also produce vitamins and other substances which are vital elements for human health. Not all commensal microbes provide products that benefit the human body, some of them are associated with the enhancement of health. For example, immunological studies found probiotic bacteria, such as lactic acid bacteria in human milk contribute to the maturation of the baby's immune system [46].

Disease-associated microorganisms are also associated with the human body. Many medical conditions are associated with the breakdown in microbial balance in the human body, termed dysbiosis, which is readily detectible through changes in the diversity or composition of human-associated microbes [13]. For example, the abundance shifts in intestinal microbes have been identified as a vital factor of Crohn's disease [47, 48]. Some conditions such as psoriasis and acne are caused by an inappropriate immune response on the skin; this dysregulation was found to result from a change in skin microbes [31].

Some microorganisms play an indeterminate role in the human body. They are harmless to humans most of the time, but can switch to pathogenic status under some conditions. An example of such an "opportunistic pathogen" is *Staphylococcus epidermidis*. It is a permanent and commensal colonizer on human skin, but was found to be one of the most important causes of infections. Normally, they keep a benign relationship with their hosts and do not cause disease. However, when foreign bodies intrude into our body, *S.epidermidis* can cause infections. Because *S.epidermidis* can form a biofilm around foreign bodies, where biofilm is a number of densely-stick microbial cells growing on a surface. Due to the protection of this biofilm, our immune system cannot eliminate the infection [49, 50].

The oral cavity plays host to many complex microbial communities. Periodontal disease, one of the most common inflammatory and bone lytic diseases, is caused by abnormal composition of microorganisms in the gums [51, 15]. Besides the direct impacts of oral disease, periodontitis has been associated with systemic diseases such as cardiovascular disease and diabetes [25, 52]. Food is chewed and mixed in the oral cavity before reaching stomach and intestinal tract. If there was gingival crevice or other oral injury, some bacteria may follow the bloodstream to reach other body sites and cause infections. As one of the main communities of oral microbes, dental plaque refers to a condensed layer of bacteria on the teeth [53, 54]. Plaque is commonly associated with a number of diseases, including tooth decay and periodontal disease [55, 26]. While tooth brushing is an important mechanical control for plaque, many people fail to clear away all the plaque with regular tooth brushing, leading to very high prevalence of dental disease. The role of dental plaque in oral health and disease makes the oral cavity, and plaque in particular, worthwhile targets for

microbial community profiling and classification.

1.1.2 Microbial Communities of the Human Body

Microbial community ecology draws on concepts from traditional ecology to generate insights into the human microbiome: key aspects of microbial ecology include analysis of taxonomic proportions, functions, and interactions between microorganisms and environment [56, 57, 58].

Taxonomy provides approaches to define groups of organisms based on their common physiological or genetic characteristics. A taxon (plural: taxa) consists of a number of related organisms that share certain similarities. Recursively aggregating taxa forms higher-level groups, which creates a hierarchical taxonomic classification system. Organisms are typically assigned names at each rank of a taxonomic hierarchy comprising the ranks of kingdom, phylum, class, order, family, genus and species. Major body sites show very distinctive composition at all taxonomic levels; for example, healthy human gut samples are dominated by members of phyla *Bacteroidetes* and *Firmicutes*, while skin samples tend to be much richer in *Actinobacteria* and other groups [59, 60, 61]. Variation in environmental conditions results in the growth of various sets of bacteria. For example, hard plaque is subdivided into *subgingival plaque* below the gumline, and *supragingival plaque* above the gumline. Because of the oxygen-free environment, subgingival plaque consists mainly of anaerobic organisms, such as *Clostridia*, *Fusobacterium* and *Prevotella*, while aerobic bacteria, such as *Bacilli* and *Betaproteobacteria* are dominant members of the supragingival plaque [7, 32]. Taxa in human body is easy to be affected by factors, such as age, lifestyle, ethnicity and living environment. However, related studies found functional profiles within one site varies little between individuals. Functional profiles summarize metabolic or other traits existing in the samples, which is often a more stable measurement of microbial communities than taxonomic composition. Since functional traits characterize the microbial community from a different aspect, we might expect function to provide powerful features for microbial classification problem.

1.1.3 Microbiome Analysis

To characterize the microbiome requires the choice of a marker to identify and analyze the microorganisms. Genetic markers allow us to identify individuals, populations or species within the community. Although marker genes are a key tool for microbial study nowadays, the identification of microorganisms has undergone several decades of developments.

Culture-based approaches were applied to identify the microbes, but this method limited the range of microbes to those that can grow in the laboratory environment [62, 36, 63]: some authors have claimed that less than 10% of microbes can be cultured, thus preventing most microorganisms from having taxonomic names assigned [64]. The first use of molecular (i.e., DNA or protein) sequences for evolutionary analysis took place in the 1960s [65, 66, 67]. If all organisms possess similar genes that evolve relatively slowly, these sequences can be used to infer the evolutionary history of the organisms. The 16S ribosomal RNA (16S rRNA) gene emerged as the standard for identification of microorganisms. This gene is a constituent of ribosomes, which are responsible for synthesising proteins in the cell; since this is a universal function, all living organisms have ribosomes and the gene that encodes this ribosomal RNA. Several reasons justify the 16S rRNA gene (referred to hereafter as 16S) as a genetic marker for microbial diversity [68, 6, 37, 69]:

First, ribosomal RNA is present in all microorganisms, which makes it a universal target.

Second, the 16S rRNA sequence is a stable genetic marker. Many regions of the gene change slowly and 16S has a lesser chance of gene loss, mutation or genetic exchange between organisms (e.g. lateral gene transfer [70]). The evolutionary relatedness of the organisms can be inferred from these sequences.

Third, the 16S sequences include highly conserved and variable regions (Figure 1.1). The variable regions are different from species to species, which allows us to identify the taxa in the community. The conserved regions work as start-end marks on sequences to locate the variable regions.

Identifying the microbial diversity based on the 16S marker gene is widely used today, but there is still a large number of microbes that have not been characterized. Since we still have 16S sequences that cannot be reliably classified in taxonomy.

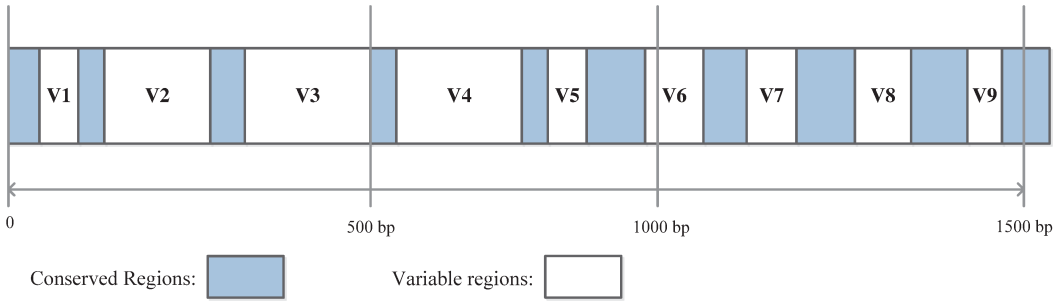


Figure 1.1: **Illusration of conserved (blue) and variable (white) regions of the 16S rRNA gene.**

The 16S sequences of many bacteria and archaea can be accessed via those public databases, such as National Center of Biotechnology Information (NCBI) and Genbank. However, quality of the sequences in those databases is not very authoritative [71]. So several secondary 16S database address this problem; the three most widely used are the Ribosomal Database Project (RDP) [72], SILVA [73] and GreenGenes (GG) [74]. RDP provides aligned and annotated sequences from not only bacteria and archaea but fungi as well. Sequences can also be aligned and derived phylogeny via RDP. SILVA has numerous types of rRNA sequences. Operations on sequences, such as searching and aligning are also available in SILVA. GG is a 16S rRNA gene database, which provides annotated, classified and full-length aligned sequences. The taxonomy in GG is based on a *de novo* phylogenetic tree created from the whole reference sequences in its database.

16S is a valuable tool to characterize microbial communities, however, it gives no direct information about the function of organisms in a given habitat. Metagenomics refers to the sequencing of fragments of DNA from given environment such as the human body. Metagenomics uses a "Shotgun" sequencing approach, which cuts total DNA from all microbes in an environment into small pieces. Sequencing a random sample of these fragments yields information about functional genes, and can potentially produce whole genomes based on the assembly of overlapping regions of these sequence fragments [75, 76]. However, due to the cost of metagenomics, the number of available metagenomic samples is smaller than that of 16S rRNA.

1.1.4 Human Microbiome Project

The recognition that microorganisms in the human body may play a more central role in health and disease that previously thought, motivated the development of large-scale projects to assess the microbiome in many individuals. To facilitate the understanding of the human microbiome, the National Institutes of Health launched the Human Microbiome Project (HMP) in 2008 [77, 78]. The goals of HMP included:

1) collecting samples from multiple body sites to produce an overall characterization of the microbial communities;

2) exploring the relationships between health state and changes in the microbiome;

3) providing researchers with a standard dataset and technology for further studies of the microbiome.

The HMP collected samples from 242 North American volunteers (129 males and 113 females) aged between 18 and 40. Microorganism samples were collected from the five major sites of greatest interest: the oral cavity, the nasal cavity, the skin, the gastrointestinal tract and the urogenital tract. In many cases several sub-sites were sampled from each site: for example, a total of nine different locations in the mouth were sampled from most study participants. A diagram of the oral cavity drawn by Sitepainter can be found in Figure 1.2. All generated sequences, developed software and related standard operating protocols in HMP have been released to the public, which can be accessed from HMP Data Analysis and Coordination Center (DACC) [Link: www.hmpdacc.org].

Works studying the HMP data found that *Firmicutes* is the prevalent phylum in the oral cavity, and *Streptococcus*, a genus of *Firmicutes*, has the most abundance at the genus level [7, 32]. However, through the digestive tract from mouse to gut, the abundance of *Firmicutes* typically decrease while the amount of *Bacteroidetes* increases. Following *Streptococcus*, other abundant species are different at sites: *Haemophilus* were found in the buccal mucosa, *Actinomyces* and *Prevotella* were in Supragingival plaque and Subgingival plaque [52]. Comparing to other sites, communities in the oral cavity and stool are diverse in microbial memberships and the taxonomic composition varies a lot between each individual. In addition, a number of metabolic processes were widely distributed on digestive tract, such as carbohydrate metabolism and the synthesis of energy molecules(e.g., adenosine triphospate

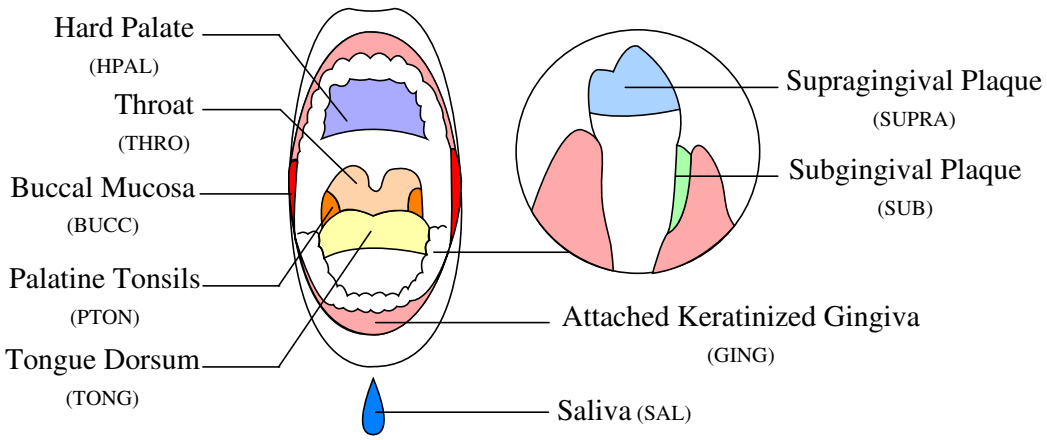


Figure 1.2: **Human oral cavity diagram drawn by SitePainter [1].**

(ATP)) [32].

1.2 Machine-learning Methods

Machine learning is the science of designing algorithms to recognize patterns in data, and making predictions based on these discovered patterns. Machine learning has been widely applied in many fields, such as text categorization, image recognition and intelligent robot control [79, 80, 81]. Recently developed "next-generation sequencing" technologies can produce a huge number of DNA sequences in a short time and low cost. To deal with such large amount of sequences, powerful computation tools and algorithms are required. Machine learning has been applied to many problems in bioinformatics [82, 83, 84, 85]; applying it to microbiome data may give better characterizations of the microbial community [44, 86, 41].

1.2.1 Overview

For the sake of consistency, this section introduces the definitions used throughout the thesis. The dataset that is used to build and evaluate the model is a collection of *samples*, also called instances or examples. Each sample is described by a number of *features*, which can also be referred to as *attributes*, variables or dimensions. Features can be assigned with values in continuous, categorical, or other data types. *Training* is the process of building a predictive model by learning from a subset of the entire

labeled dataset. To see how well the model performs on new samples, *testing* is performed using samples that were not used in the training process. The complete set of samples can be separated into training and testing set to evaluate the performance of the model. K-fold cross-validation is a common strategy that extends the idea of training and test sets by dividing the samples into k equal sized subsets. For each cross-validation process, the i_{th} (i=1,2,...,k) set is used for testing while the other k-1 sets are for training iteratively. Repeating the process k times and averaging the k results produces the final estimation.

Machine-learning methods are often split into two major categories. *Supervised learning* uses prior defined labels (for example, the body location associated with a given sample) and tries to build a mapping function between them according to the category they belong to. *Unsupervised learning* attempts to associate samples based on measures of between-sample similarity, without reference to any previously defined categories. Commonly used methods such as ordination and clustering are able to find out the associations from the most salient patterns, however, sometimes the achieved patterns may not reveal features with much interest [29]. Since supervised classification approaches use knowledge of features to train models that can draw on any pattern of co-variation in the data, it may give patterns with higher relevance to categories of interest than unsupervised approaches.

Due to the large variety of prediction algorithms, different evaluation measures have been proposed. For supervised learning algorithms, widely used evaluation methods assess the performance mainly from the proportion of correct predictions, sum of error and correlation coefficients [87]. The proportion of correctness reflects the percentage of samples that are correctly predicted, which can also be derived from the number of True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) predictions (description of TP, TN, FP and FN in Chapter 3.3). Sum of error approaches calculate the distance between the prediction and true label of the sample, while correlation coefficients measure the amount of agreement between them.

Typically, the performance of an algorithm is evaluated after repeating the prediction for a number of times. Although the performance can fluctuate during each

time, the deviation should always be within a reasonable scale. To verify the predicting results, statistical methods are usually used. Basic statistical descriptions include central tendency, consistency and range of a sample. In addition, a statistical hypothesis test can also be used to prove that results are not achieved by chance, which increases the soundness of the algorithm.

1.2.2 Important Considerations in Classification

The goal of classification is to generate a model that can maximize the accuracy under a given criterion. However, no single classifier can give optimal results on every dataset. Several factors may lead to the failure of classification, from the sample initialization to final evaluation [2]. Successful classification depends on a large number of factors, including:

- Creating an appropriate feature set is essential. Features are like the bricks that will be used to construct a building: if the bricks are poor, even the best architecture cannot make it strong. Raw data are not usually in a form that classifiers can use directly: for example, microbiome samples are based on DNA sequencing, models based directly on nucleotide sequences are unlikely to give good accuracy. To solve such problems, different feature construction strategies are required. For example, in sequence classification, a k-mer approach is used to create features, whose values represent the frequency of all possible k-length subsequences appearing in the target sequence [88]. For document categorization, features are constructed via a bag-of-words model, which counts the occurrences of each word in the document. Features can be constructed using various strategies [89].

- Biological datasets are often high dimensional, with many more features than there are samples. This is often described as *the curse of dimensionality* [90]. Many datasets in biology are of this type: gene microarray data records the expression level of thousands of genes under different conditions, but typically for under 100 samples [91]. Biological sequence data such as DNA and protein are often converted to features using a k-mer approach, which generates a large number of features (4^k for DNA and 20^k for protein) [92, 93, 88]. An

obvious problem caused by high-dimensional feature set are the high cost in running time and memory space. Moreover, overfitting may occur if the classifier is trained using a large number of features [94, 2]. Since feature space of large-dimension usually results in a complex model. Several strategies are available to counteract the curse of dimensionality. One way is to select an algorithm that can handle high-dimensional input features, such as Support Vector Machine (SVM) [95] and Random Forest (RF) [96]. Dimensionality reduction can also be used to reduce the size of the feature space. Feature selection and extraction are two commonly used types of approach. Feature selection (e.g. Information Gain [97], Feature Permutation [98]) chooses a subset of features highly relevant to the labels with different strategies. Feature extraction (e.g. Principal Component Analysis(PCA)) transforms all the features into a new feature space, where the first few dimensions captures most of variance about the dataset. These features may contain more useful information to the classifiers than those from feature selection, but the transformed features cannot reflect their biological meaning directly. Last, taking advantage of the properties of dataset and algorithms, for example, SVM gives predictions based on the similarity scores between pairs of samples. If information in the original feature space can be transformed into similarity scores, the high-dimensional problem will be avoided. Moreover, meaningful correlations often exist in biological data. For instance, several genes together may affect the same characteristic, a process known as epistasis [99]; and it may be a group of microorganisms, rather than a single one, that differentiates the communities [32]. If one representative member of this set can be picked out, the number of features would be reduced.

- *Information leakage* usually happens when the training data gets information about the labels in testing test beforehand, resulting in unrealistically good predictions. A typical example of a leaked model would be that it gives predictions based on the target label itself. This type of leakage is analogous to saying it is sunny on sunny days [100]. Although the predictions from such models look very satisfactory, they are not reliable. Information leakage can occur in most steps of the machine learning process, including feature construction, feature reduction, training and testing. An instance of leakage in feature construction

was seen in the KDD-Cup competition in 2008 [101]. One challenge required methods to predict if a patient had breast cancer based on her mammography data. The patient ID feature showed an extremely powerful ability to predict the label, because the ID string is encoded with the patients' health condition [102]. Information leakage can also occur when applying feature selection [103]. Features are first ranked on all the dataset using information gain or other criteria. Then a number of important features are selected, people build the model with these features and use cross-validation to measure the performance. But actually, all the data has already been known before testing, meaning the selected features leak the labels of all samples. So in this case, feature selection should be performed during cross-validation, that is divide the samples into training and testing set, rank the features based on training set and evaluate the model with testing set. Repeat this process for k times to get the result.

- The accuracy of machine-learning classifiers is rarely perfect on anything. People usually try to decrease the error by optimizing the training model. Bias and variance are two indicators of measuring the models [104]. Assume a training set (x_i, y_i) fit an hypothesis $H(x)$, bias describes the average error made by $H(x)$, while variance measures the amount of consistency of the predictions. Figure 1.3 visualizes the concept of bias and variance [105]. When a model is too simple, it may make consistently incorrect predictions. Although the consistency keeps a low variance, the value of bias will be very high. This situation is described as under-fitting. In contrast, if a model is too complex, even if it works well on training set, the prediction accuracy on the test set may not be that precise. This scenario leads to low bias and high variance, which is known as over-fitting. Since a decrease in bias implies an increase in variance and vice versa, there must be one point where the trade-off of bias and variance is optimal with respect to classification accuracy. A bias-variance trade-off curve graphically describes this idea (in Figure 1.4).

  Over-fitting can be avoided by reducing the complexity of the model.. Some algorithms reduce it via tuning specific parameters, others by controlling the number of parameters. Regularization is another way to control model complexity. From Figure 1.5 we find that the boundary generated by an over-fitted

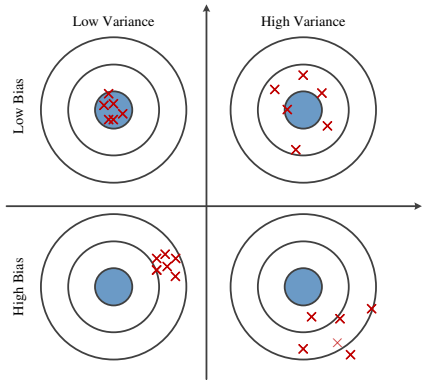


Figure 1.3: **Graphical definition of bias and variance.** Imitate the prediction as hits on the target. Each hit indicates a prediction of one sample. The figure was adapted from [2].

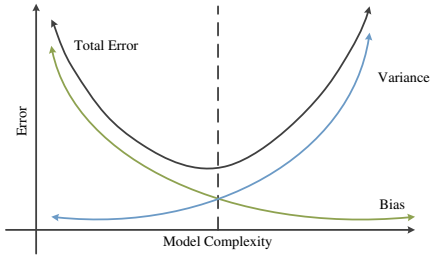


Figure 1.4: **The relationship between model complexity and total error, bias, variance.** The figure was adapted from [3]

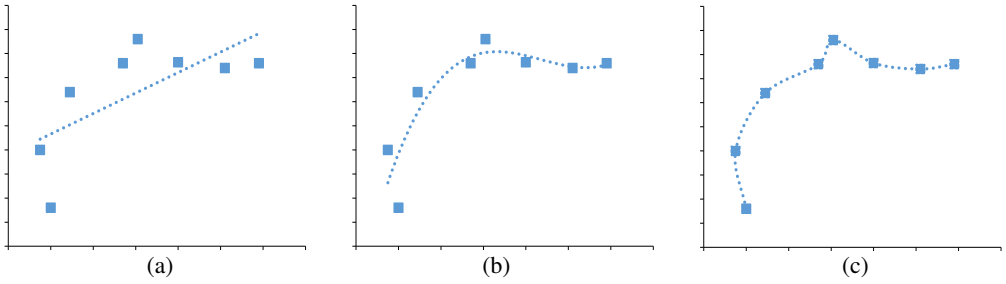


Figure 1.5: **An example of the models in different scenario.** (a) Under fitting: low variance and high bias; (b) good fitting: low variance and low bias; (c) over fitting: high variance and low bias.

model is usually highly curved, since it tries to get close to every training point. However, regularization can smoothe the curve functions by using different techniques [106].

1.3 Related Work

Supervised approaches have been used to classify the human microbiome [41, 107, 43, 108] with the abundance of different taxa (e.g., species) or Operational Taxonomic Units (OTUs) serving as feature vectors. OTUs were defined based on the similarity sharing between marker genes [109, 110]. For OTUs in microbiome, they are groups of sequences, typically 16S sequences, with a specific amount of similarity.

Knights *et al* performed one of the first supervised classification studies of the human microbiome, applying five classification algorithms [41], including RF and SVM. These classifiers were applied to five microbial datasets, comprising samples from healthy human volunteers, that were originally described in Costello *et al* [29, 111]. RF achieved the highest accuracy on three of the datasets when all features were used. However, SVM showed similar accuracy after being given a selected subset of features. Samples labeled with major body sites were easy to separate: for example, a classifier trained to distinguish external auditory canal (e.g., the ear), gut, hair, nostril, oral cavity and skin samples yielded an accuracy of 88.8%. Sub-sites were more difficult to distinguish: for example, a classifier trained on twelve different skin sites got an accuracy of 56.8% (*s.d.* 6.7%) on average.

Supervised classification has also been used to distinguish samples labeled with disease states. Galimanas *et al* found that microbial communities from supragingival plaque and the tongue dorsum can serve as alternative biomarkers for Chronic Periodontitis (CP), a disease of the subgingiva [12]. Subgingival, supragingival and tongue plaque samples from 11 healthy and 13 diseased subjects were analyzed in their study. In tongue dorsum, they found a group of disease-indicating OTUs, including *Treponema denticola* and *Treponema forsythia*, have a small proportion in healthy samples but are present in high abundance in people with CP. The amounts of *Porphyromonas gingivalis* and *Filifactor alocis* in supragingival plaque were found at significantly increased levels in people suffering from CP.

Statnikov *et al* tested the performance of several major classifiers and feature selection strategies [43]. They found that classification tasks of separating body sites or subjects yielded relatively high accuracy, while distinguishing healthy from diseased states was more difficult: for example, when classifying the samples from people with and without psoriasis (labels: control, psoriasis normal and psoriasis lesion), the performance of all classifiers were poor, whose accuracy is about 10% to 60% lower than classifying healthy body sites. When comparing the classification methods, their result was consistent with Knights *et al*'s: RF and SVM were two most effective machine-learning methods, followed by Kernel Ridge Regression and Bayesian Logistic Regression. Studies done previously recognized that the 16S rRNA data is of high sparsity and redundancy when expressed as input features. Wang *et al* present a feature reduction algorithm called Feature Merging and Selection (FMS), which integrated the Linear Discriminant Analysis [108]. FMS was able to reduce the feature space without losing original accuracy, and the relationships between features can also be preserved. They tested the pneumonia data (binary classification) with SVM and k-nearest neighbor (kNN) models. Results showed that features selected by FMS gained better performance than some popular feature selection methods: 5.5% (SVM) to 13.9% (kNN) improvement in accuracy.

For the high-dimension and sparsity of the feature space, an efficient machine learning algorithm for microbial classification is also in demand, especially multi-class classification. Liu *et al* integrated the SVM and KNN learning methods, and proposed a sparse distance-based learning algorithm for classifying 16S metagenomic data [42]. In their algorithm, the predictions were made by a kNN model. However, the distances between samples in the kNN model were given different weights, which is optimized via an efficient quadratic SVM method. They showed its efficiency in classifying 16S rRNA data and the suitability to unbalanced datasets.

Microbiome data is typically high dimensional, with potentially thousands of OTUs observed in each sample. Feature selection aims to identify a subset of all features that are most promising for classification, thereby eliminating uninformative features and decreasing the running time for the classifier [112]. Even when the accuracy of a classifier is not substantially improved, feature selection can still reveal key species or molecular functions of particular biological interest, because only the

set of features that are most useful to classification (typically a very small subset of all features) is retained.

1.4 Contributions

Supervised methods are effective for many classification problems; however, few studies have tackled the classification of the oral microbiome specifically. An important objective of this project is to augment standard representations of microbial communities (for example, OTUs) with additional biological and evolutionary information. For example, support vector machines (SVMs) can base their classifications on customized similarity values between samples from the same or different body sites; distances such as UniFrac [113, 114, 115] can be informed by phylogenetic relationships amongst species or OTUs.

Similarly, the use of OTUs in classification builds on an assumption that groups of closely related organisms can be treated as units sharing key similarities. This assumption may be violated by strain-level variation, and conversely may apply to aggregations of phylogenetic groups (i.e., *clades* that encompass all OTUs descended from a common ancestor) that comprise many OTUs, which again suggests a phylogenetic approach.

Finally, while taxonomic representations can contain a great deal of information, different microorganisms have different functional sets of genes involved in processes such as biosynthesis of important compounds, environmental adaptation, and antibiotic resistance. Information of functional genes are typically obtained by sequencing metagenomic data described in Chapter 1.1.3, however, this approach is costly. The recently developed Phylogenetic Investigation of Communities by Reconstruction of Unobserved States (PICRUSt) [116] (see Chapter 4.3) algorithm can map taxonomic samples to functional profiles, based on known gene repertoires of closely related organisms: these functional approaches may provide complementary information to taxonomic features. Functions may be similar between distantly related lineages and PICRUSt can potentially identify sets of clades whose similarities are functional rather than phylogenetic. Some of these approaches yield significant increases in classification accuracy, while feature selection highlights key phylogenetic and functional features. We have implemented these ideas in a machine-learning framework,

and used oral microbiome samples from the Human Microbiome Project [78, 77] as a challenging test case.

1.5 Thesis Outline

The remainder of this thesis is organized into five chapters. A brief description of the 16S rRNA dataset used in the thesis and standard data preprocessing are given in Chapter 2. Chapter 3 shows a preliminary experiment on the whole dataset and lists all the challenges encountered. Our exploration of this classification task, with a focus on samples from subgingival and supragingival plaque, is described in Chapter 4, which includes the prediction results, the key phylogenetic and functional features and classification with ensemble method. The conclusion of this thesis is in Chapter 5, which summarizes the results we achieved so far and gives ideas for future work.

Chapter 2

Data Preparation and Exploration

The 16S sequences used in the thesis came from HMP and all the sequences were processed via standard microbiome analysis software. The analysis of microbial community involves the following: quality filtering on 16S sequences, picking OTUs, representative sequence alignment, phylogenetic tree construction, diversity analysis and sample visualization. As a feature file for our classification problem, the OTU abundance table was analyzed from some basic statistical points of view. Preliminary interpretations of the microbial communities on different body sites were also given, including the comparison of taxonomic composition and phylogeny within the OTUs.

2.1 Microbiome Data

We retrieved the oral microbiome marker-gene dataset from the HMP DACC [77] in February 2014. There are nine sampled sites within the oral cavity: saliva, supragingival plaque and subgingival plaque (plaque above and below the gingival margin), tongue dorsum (top surface of the tongue), hard palate (roof of the mouth), buccal mucosa (inside lining of the cheek), attached keratinized gingiva (gums covering the jaw bones), and palatine tonsils (sides at the back of the throat) (see Figure 1.2). Samples in HMP were collected up to three times per site from each individual in a non-invasive manner. The process of sample collection obeyed the strict procedures; details were described in Manual of Procedures [117, 118]. Sequences in this dataset included amplified V1-V3 and V3-V5 regions of 16S rRNA gene, although there were more sequences associated with the V3-V5 region (see Table 2.1 for summary statistics).

2.2 Sequence Analysis

Many microbiome studies follow a standard protocol for sequence analysis and OTU construction. In human microbiome collection, related information about the sample donors such as sex, age and ethnicity is a vital factor for subsequent association studies. DNA sequence was then extracted from the collected samples. Sequences can be extracted using different experimental protocols, but it is essential that a single protocol be employed consistently in one study.

2.2.1 Sequence Quality Control

The small amount of sequence initially extracted from each sample was not enough for further study. So DNA amplification is performed to obtain a larger quantity of sequences. However, because of the employed technique, contaminant sequences are introduced in this process. So it is inevitable that microbial dataset usually comes with noise. The quality of raw sequences needs to be assessed in Quality Control (QC) steps.

There is not a gold standard to assess all sequences; QC strategies vary among sequencing techniques. For example, sequences from Sanger and Illumina sequencing machines come together with a quality file, which records the scores for each nucleotide directly [119]. For 454 or SOLid methods, each quality score is given as the probability of this nucleotide being wrong [120]. After assessing the quality, sequences with a

Table 2.1: **Details of human oral cavity samples from HMP, with associated abbreviations.**

Sub-sites	**Acronym**	***Samples***	***OTUs***	***Seqs/sample***	***OTUs/sample***
Saliva	SAL	281	6166	8596 ± 6034	521 ± 183
Attached keratinized gingiva	GING	304	3741	8998 ± 5756	313 ± 105
Buccal mucosa	BUCC	301	5370	9465 ± 10268	447 ± 166
Hard palate	HPAL	300	5848	8935 ± 6575	441 ± 154
Palatine tonsils	PTON	304	5339	9586 ± 7247	448 ± 146
Throat	THRO	301	6278	9053 ± 7233	422 ± 147
Tongue dorsum	TONG	305	4400	10351 ± 10450	398 ± 129
Subgingival plaque	SUB	301	6782	9877 ± 5926	495 ± 147
Supragingival plaque	SUPRA	305	5277	10413 ± 6564	497 ± 152

higher error rate than previously defined will be removed.

2.2.2 Building Operational Taxonomic Units

The definition of unique sequences has large effects on the identification of microorganisms [121]. Several degree of sequence divergences exists within the organism to define the organisms, the concept of OTU was proposed [109, 110]: based on the similarity shared among marker gene, sequences can be binned into groups. Different thresholds of similarity cutoffs denote different taxonomic levels, such as 97% for species level, 95% for genus level [122]. The assessment of sequence similarity can be done either by examining pairs of sequences within each sample, or by comparing each collected sequence to standard reference sequences from a reference database such as GG. With OTU abundance available, characterisics of the microbial communities such as microbial diversity and phylogenetic relationships can be inferred.

The approaches of clustering sequences into OTUs or OTU picking can be divided into three categories: *de novo*, closed-reference and open-reference (in Figure 2.1). In the *de novo* approach, sequences are compared against one another just within the samples; no external database is used as a reference. In the closed-reference OTU picking process, each sequence is compared against the sequences in a reference database such as GG. The sequence will be binned into the OTU centered on one reference database sequence, if their similarity is larger than the cutoff. As a compromise of *de novo* and closed-reference picking strategy, open-reference can be used. Sequences that succeed in finding a hit to the reference database are assigned as in the closed-reference approach, while the remainders are clustered using a *de novo* approach.

Closed-reference strategy discards sampled sequences that do not match any reference sequence at the specified threshold, which limits the identified OTUs. However, closed-reference OTU picking is fast since the implementation can be fully parallelizable. Moreover, the taxonomic assignment and the phylogenetic trees generated in the next steps are more reliable because all the OTUs are defined based on the well-constructed reference database. OTUs in our work were picked via a closed-reference strategy. For the reference database, although the RDP [72] and SILVA [73] have their own advantages and disadvantages, we adopted GG (gg_13_08) [74] as our reference database, which is consistent with our adopted pipeline's default setting.

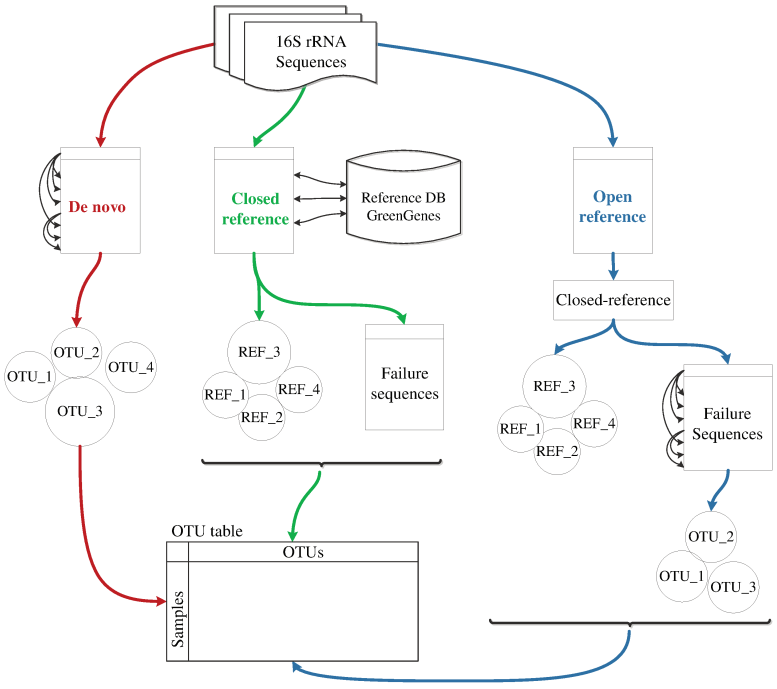


Figure 2.1: **Diagram of OTU picking strategies.** *De novo* (red stream), closed-reference (green stream) and open-reference (blue stream)

2.2.3 Sequence Alignment and Phylogenetic Tree Construction

The taxonomic assignment of OTUs can be used to express the similarity between samples, but it has drawbacks. The taxonomic annotation in the reference database usually cannot cover all sequences in our sample, leading to incomplete taxonomic summaries of the microbial community. The taxonomic hierarchy cannot always reflect the evolutionary distance precisely, which depends on the quality of annotations in reference database. Besides the count or proportion of different OTUs and species, the evolutionary distance between OTUs is also an important characteristic of microbial communities. Mapping OTUs takes advantage of intrinsic structure of a phylogenetic tree [123, 124]. Phylogeny refers to the evolutionary relatedness between different species, while a phylogenetic tree displays these relationships in a binary tree structure.

Phylogenetic trees are constructed after aligning the sequences of the target taxa.

Sequence alignment is a process trying to allow maximum number of identical nucleotides or amino acid be aligned against each other [125, 126]. By allowing substitutions and gaps, similar regions located of sequences can be identified. If a pair of sequences shares significant sequence similarity, there is a large chance that they evolved from a common ancestor, which refers to the parent node in the phylogenetic tree.

Branches and nodes are two main features in a phylogenetic tree. For each node in the tree, external nodes (leaves) represent the living individual in samples and inner nodes are their common ancestors. Branches in the tree represent the evolutionary relationships among a subset of the whole species and the degree of divergence between pairs of species. Phylogenetic tree can either be *rooted*, with a unique node that is ancestral to all other nodes, or *unrooted*, in which no common ancestor is explicitly defined (in Figure 2.2). These two types of tree can be interconverted. Simply removing the root of a rooted tree results in an unrooted tree, while giving a root to an unrooted tree needs more information. It can be done by adding a known outgroup sequence or finding an uncontroversial criterion that can split the whole species into two groups, such as bacteria and archaea.

Phylogenetic trees are mainly constructed with three different approaches: distance matrix, maximum parsimony and maximum likelihood. The distance method relies on a distance matrix that records the amount of mismatches or gaps between each pair of sequences in the set. Commonly used distance-based algorithms are Unweighted Pair Group Method with Arithmetic Mean (UPGMA) [127, 128] and Neighbor Joining (NJ) [129]. Their main procedure is similar to sample clustering: two nodes with the smallest distance merges together and then forms a new node; this process is repeated until all species are assigned to the leaves. Most distance methods are pretty fast, but building tree with only a distance matrix loses much information of the sequences themselves [130]. Maximum parsimony is a character-based algorithm that tries to construct a tree with minimum number of substitutions over all sequences, that is the smallest number of steps to map characters to reach the phylogenetic state [131]. This method is avoids providing sequences as a single distance, but searching the tree space takes much time. Moreover, parsimony assumes all substitutions in sequences happen with equal chance, but actually nucleotides change at

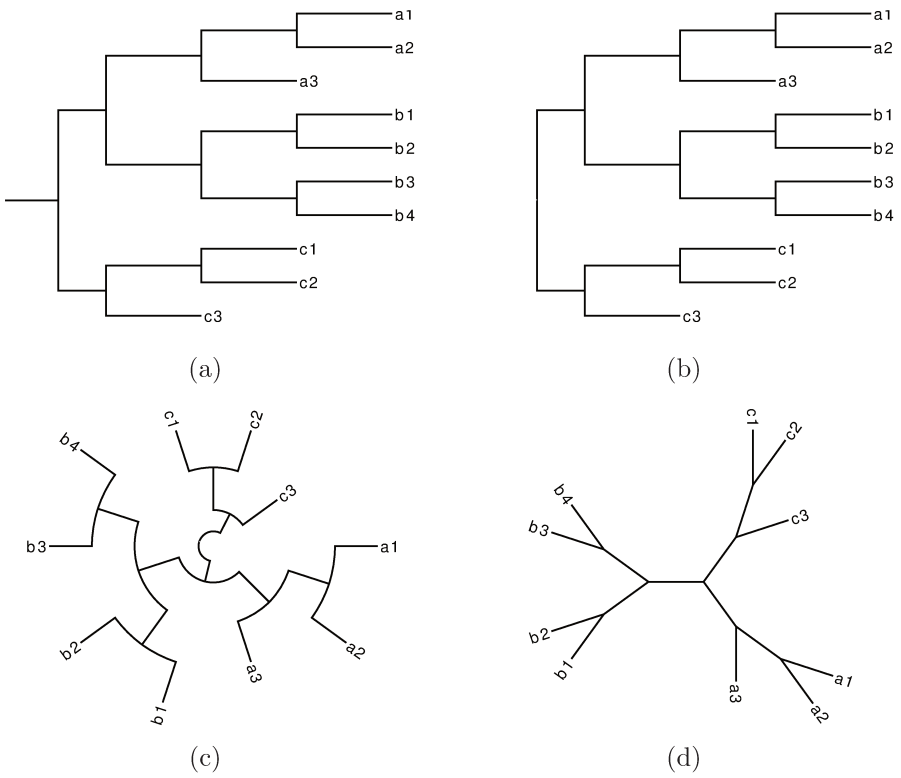


Figure 2.2: **Different ways to display a phylogenetic tree.** Diagrams of (a) rooted and (b) unrooted phylogenetic trees. The trees can also be layout as (a and b) rectangular, (c) circular, (d) radial phylograms.

different frequency rates. Maximum Likelihood approaches address this limitation by taking the probability of substitutions into consideration [132]. After assuming an empirical substitution matrix and a starting parsimony tree, the probability of different tree shapes can be calculated. The tree with highest likelihood will be selected at last. Although maximum likelihood is expensive in computation, it tends to create an accurate and robust evolutionary model. In our thesis work, FastTree [133] algorithm was adopted to build the tree, which is an approximately-maximum-likelihood method. Since instead of using a distance matrix as standard maximum-likelihood method, FastTree stores the sequences information of internal nodes. Together with other fast tree pruning and likelihood estimating algorithms, FastTree is efficient in both computational time and memory cost.

2.2.4 Measures of Microbial Diversity

Microbial diversity measures the variability among all types of microorganisms living in the community. High diversity can allow a community to cope with a changing and unpredictable environment, which increases their chances of survival [134]. The diversity of microorganisms in the oral cavity is higher than many other sites, since the wide variety of intaken food, changes in temperature and oxygen and saliva mixture make human mouth an unstable habitat. For human health, a number of diseases were found to correlate with variation of microbiome diversity [11], which also makes microbial diversity a potential indicator of disease detection, and possibly other conditions as well.

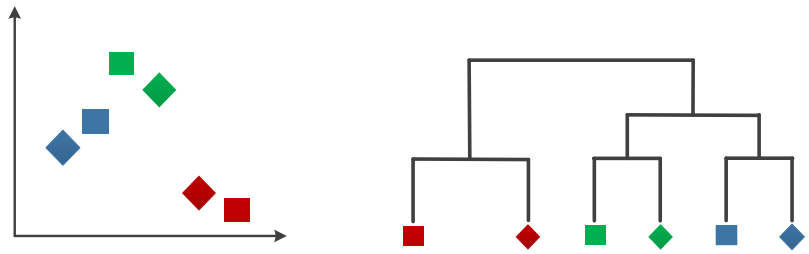
Alpha and beta diversity are two major categories of diversity measurements [6]. Alpha diversity represents the richness of taxa within a single community [135, 136]. This diversity criterion can vary a great deal between body sites: the study found that the alpha-diversity of microorganisms inhabiting in female is larger than that in males [137]. Beta-diversity quantifies the degree to which pairs of samples differ. The dissimilarity is usually expressed as a distance between communities [135, 138]. Simple distance measures such as those based on Euclidean or Manhattan distance use only information about the presence and absence (*qualitative* beta diversity) or the abundance (*quantitative* beta diversity) of OTUs in samples to calculate the distance. Those non-phylogenetic measures implicitly assume all organisms are equally

related in evolutionary history. However, phylogenetic information is lost under this assumption. Phylogenetic measurements can take different degrees of evolutionary relatedness into account. Two phylogenetic beta-diversity measures commonly used in microbial ecology are the weighted and unweighted UniFrac distances [113, 114, 115], which are described in Chapter 4.1.2.

However, because of the sequencing technique, the number of sequences in each sample is different. This variation can affect the estimation of microbial diversity: the more sequences, the more species will be found. Now people usually address this problem through two strategies: rarefaction and normalization. In rarefaction, N sequences will be randomly selected from each sample. The diversity is only calculated from those sampling sequences [139]. If the number of sequences in a sample is less than N, the entire sample will be omitted. Rarefaction eliminates low-quality samples with few sequences, but too many sequences may be excluded if N is given a very large value. Moreover, rarefaction is a random sampling process, there is no guarantee for a global optimal answer. In contrast to rarefaction, normalization attempts to adjust the sequence number to a common scale. The most straightforward approach to normalization, total-sum scaling (TSS) [140, 141], divides the number of sequence in each OTU by the total amount of sequences in that sample. In addition, other normalization methods were also proposed, for example scaling the number of sequence by the 75_{th} percentile of the non-zero abundance in each sample, which can normalize the dataset based on the sequence-count distribution. Paulson *et al* extended this idea [86], so that their cumulative sum scaling (CSS) is better suited for marker gene dataset.

2.2.5 Visualization of Microbial Community Structure

A number of different approaches can be used to visualize the beta diversity within a set of samples; two widely used approaches are Principal Coordinates Analysis (PCoA) plots [142] and hierarchical clustering [143]. PCoA is a scaling method that tries to represent the dissimilarity between samples in a low-dimensional space. PCoA transforms the distance matrix into a set of uncorrelated axes containing the maximum amount of dissimilarity information [144]. The axes are ranked by their importance in a descending order. The importance refers to the amount of variation



(a) (b)

Figure 2.3: **Two visualization methods (a) PCoA and (b) hierarchical clustering show the distances between each pairs of samples.**

in the matrix that can be captured by this axis. PCoA sometimes is confused with Principal Component Analysis (PCA) that also tries to display much information in a low-dimensional space [145]. However, PCA calculates from an initial variable matrix, whereas PCoA uses a distance matrix as its input. PCoA can accept distance matrices generated from any distance measurement such as weighted or unweighted UniFrac distance, which makes it possible for us to compare the samples based on different expressions of beta diversity. An example to explain PCoA can be found in Figure 2.3(a).

Hierarchical clustering expresses relationships among samples by grouping them into a tree (in Figure 2.3(b)). Process starts with finding the pair of samples with shortest distance, and then merging them into a common node. Repeating this process until all samples are clustered in the tree. A rooted tree will be generated in the end, reflecting the distances among samples.

2.3 Sequence Processing and Initial Data Analysis

Raw 16S sequences were processed and built into OTUs. A phylogenetic tree was constructed based on the representative sequences in each OTU. Different beta diversity measures were also calculated, reflecting the dissimilarity between samples. A preliminary exploration of the dataset helps us to better understand its main characteristics. Exploratory Data Analysis (EDA) is an approach for data analysis that uses visualization and basic statistical techniques [146]. EDA has become a critical step

before experiments: it allows an investigator to detect missing values and mistakes; know the range and distribution of the dataset; understand the biological meaning of features. We examined several important properties of our oral dataset via EDA. Comparing the OTU proportions at different ranks, and the precision to which different OTUs were classified (for example, at the phylum, genus or species level) is useful since we will use abundance information to classify the samples. The phylogenetic tree visualizes the relationships between OTUs, which gives us ideas to create new features.

2.3.1 Processing Workflow

All samples were processed using the Quantitative Insights Into Microbial Ecology (QIIME) software, version 1.8.0 [147], which is an open-source software pipeline to analyze and visualize microbial communities. HMP reviewed the data for quality and published necessary quality assurance report to declare the sequences are relatively complete and clean. So the sequences downloaded from DACC have already passed QC. All 16S sequences were clustered into OTUs at 97% similarity using UCLUST version 1.2.22q [148], using a closed-reference OTU-picking strategy with GreenGenes (gg_13_08) as our reference database. Representative sequences were aligned using QIIME's default alignment method Python Nearest Alignment Space Termination (PyNAST) version 1.2.2q [149], which implements the NAST alignment algorithm in Python. We used the default settings of PyNAST, which removes sequences with alignment length <150 nucleotides or <75% identity with the reference dataset. A phylogenetic tree of OTUs was constructed from the sequence alignment using FastTree version 2.1.3q [133]. Trees were visualized with Python Environment for Tree Exploration (ETE) version 2.1 [150]. Four beta-diversity metrics were used to calculate the distance between each pair of samples with QIIME. To visualize the dissimilarity of the samples, Principal Coordinates Analysis (PCoA) was performed to observe the samples in a low-dimensional space. We also used UPGMA approach to build hierarchical clusters. The workflow can be found in Figure 2.4.

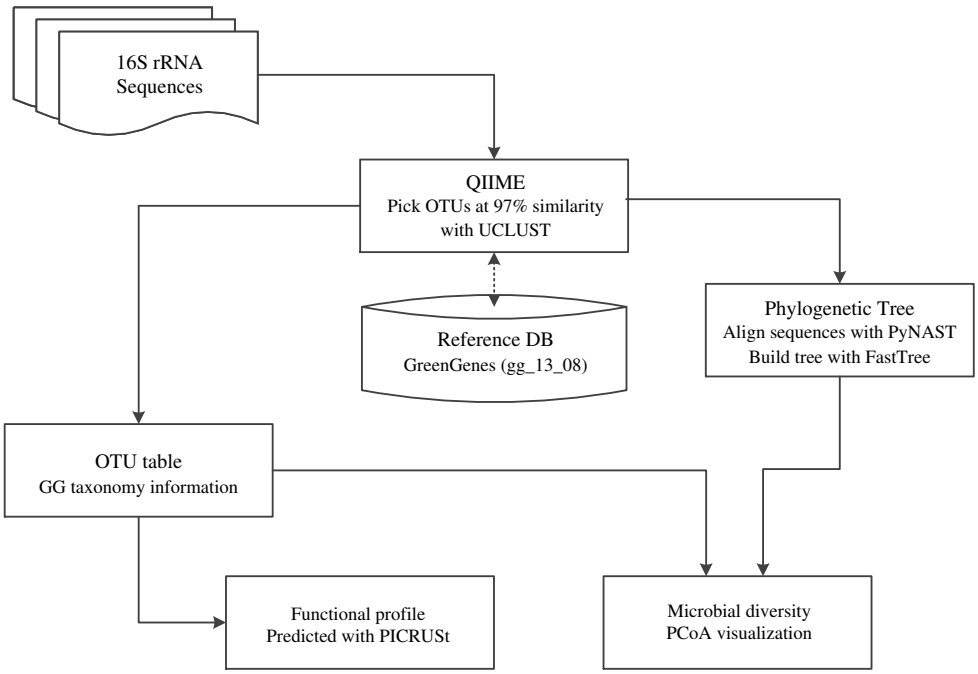


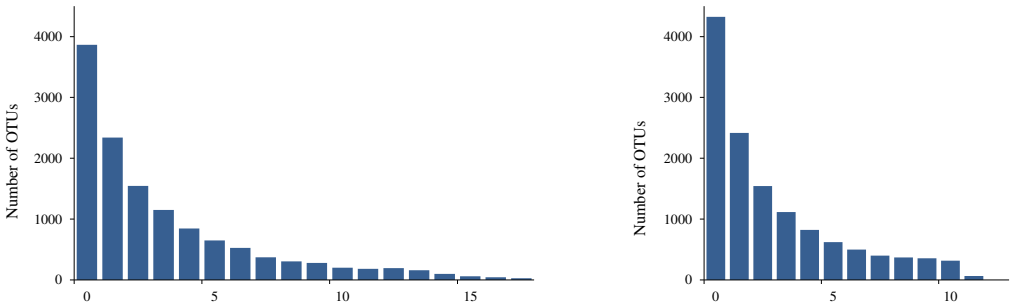
Figure 2.4: **Sequence data processing workflow.** Raw 16S sequences were obtained from HMP and put into QIIME to pick OTUs with a closed-reference strategy (GG as reference database). Steps also include building a phylogenetic tree of these OTUs, calculating the microbial diversity of each samples and visualization with PCoA. PICRUSt was used to predict the functional profiles based on the OTU table created.

2.3.2 Basic Statistical Description of Samples

A total of 2,706 human oral cavity samples from nine oral sites were collected from the HMP database. Some of the samples have sequences from both V1-V3 and V3-V5 regions, and some of samples have either one of them. We split each sample based on the variable regions, resulting in a total of 1,542 included sequence data from the V1-V3 region of the 16S rRNA gene, while 2,702 samples contained information from the V3-V5 region. Because of the disparity in data set size, and less accurate results obtained with the V1-V3 region (see Chapter 3.3), we focused on information retrieved from V3-V5.

The samples covered the V3-V5 region of the 16S rRNA gene (in Table 2.1). All sites had at least 281 associated samples. A total of 12,845 OTUs were generated by the closed-reference picking process, and OTU richness across all samples of a given site varied from a minimum of 3,741 (attached keratinized gingiva) to over 6,000 (saliva and throat). The average number of sequences per sample ranged from approximately 8,500 to 11,500, although the variation within each site was high. In terms of community members, the number of OTUs in a single sample varied between 313 (attached keratinized gingiva) and 521 (saliva).

A large number of identified OTUs were of low abundance (see Figure 2.5(a)). A number of 7,752 (60.4%) of the OTUs have fewer than 5 sequences, and 3,865 (30.1%) of them are singleton OTUs comprising only a single sequence. However, these low-abundance OTUs may nonetheless be useful for classification, so none of them was removed. Because of the high dimensionality of the data, the OTU table is very sparse. Figure 2.5(b) shows the number of OTUs presenting in different numbers of samples. Fewer than 1,105 (8.6%) OTUs are present in >10% of the samples, while 4,325 (33.7%) of them are present only in one sample. Rare and site-specific OTUs are common in microbial datasets, for several reasons. Rare OTUs can in fact be artifacts that arise from sequencing errors. Some microorganisms may be viable only in a subset of all sampled sites, while others may simply be rare. Specific human body sites typically comprise microbes from similar high-level groups such as phylum or family: for example, the gut microbiome is typically dominated by phyla *Firmicutes* and *Bacteroidetes*. However, when OTUs are clustered at very high levels of similarity, the overlap in composition tends to decrease dramatically. Different



(a) Number of sequences in OTU $\log_2 n$ (b) Number of samples $\log_2 n$

Figure 2.5: **Statistical summary of the input features.** Numbers on X-axis are displayed in logarithmic scale with base 2. (a) Each feature (OTU) has a different number of sequences. With the abundance of OTUs increasing, the number of corresponding features become less. (b) Features appear in different number of samples.

body sites support different kinds of microorganisms, and the body conditions also vary from person to person. Although some of the distinct features may not facilitate classification, we still kept all of them. Feature reduction strategies would be used to remove them alternatively.

2.3.3 Taxonomic composition of samples

The purpose of microbial community classification is definitely not only for a higher accuracy, identifying the discriminative taxa that differentiate the microbial communities is also important. Bacteria on the oral sites were compared at different taxonomic levels. Based on the GG taxonomic assignments, more than 60 phyla were detected, but four of these phyla constituted nearly 99% of the entire set of characterized OTUs (in Figure 2.6(a)). They are: *Firmicutes* (43.0%), *Proteobacteria* (20.4%), *Bacteroidetes*(17.8%) and *Fusobacteria* (9.4%). More than 180 classes were found in samples, and the top four classes covered 69% of OTUs that were classified at this rank: *Bacilli* (30.6%), *Bacteroidia* (14.1%), *Gammaproteobacteria* (12.4%) and *Clostridia* (12.3%) (in Figure 2.6(b)).

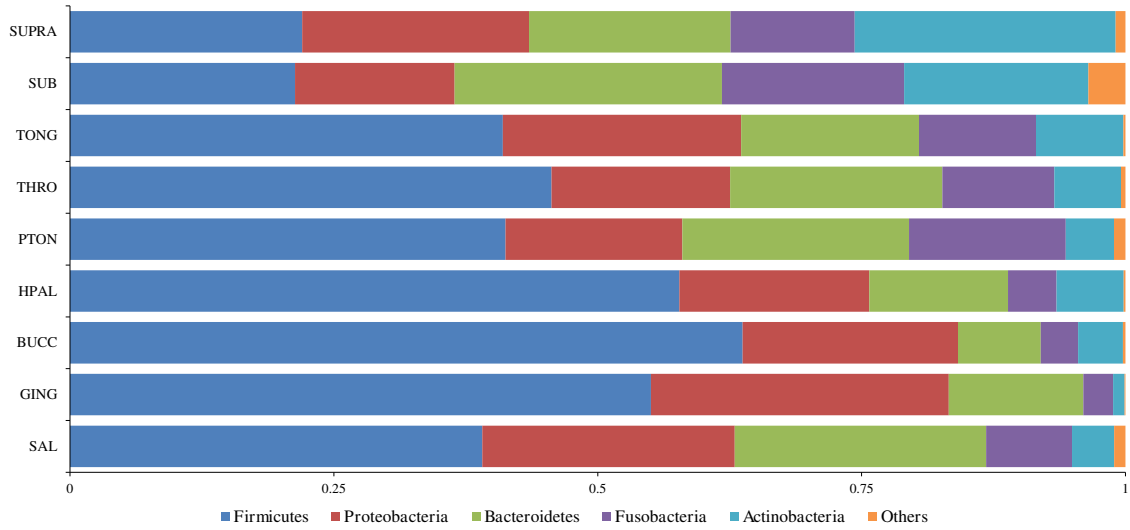
At the phylum level, *Proteobacteria* ($20.4 \pm 3.9\%$ *s.d.*) has an even distribution across all body sites, with a 3.9% standard deviation., while *Firmicutes* varies a lot ($43.0 \pm 13.9\%$ *s.d.*). Buccal mucosa, hard palate and attached keratinized gingiva are

the three sites with the most *Firmicutes*, while subgingival plaque and supragingival plaque have the smallest proportion. All sites have similar amount of *Bacteroidetes* except buccal mucosa. Although subgingival plaque and supragingival plaque consist of similar microorganisms, they are quite different from the other seven sites.

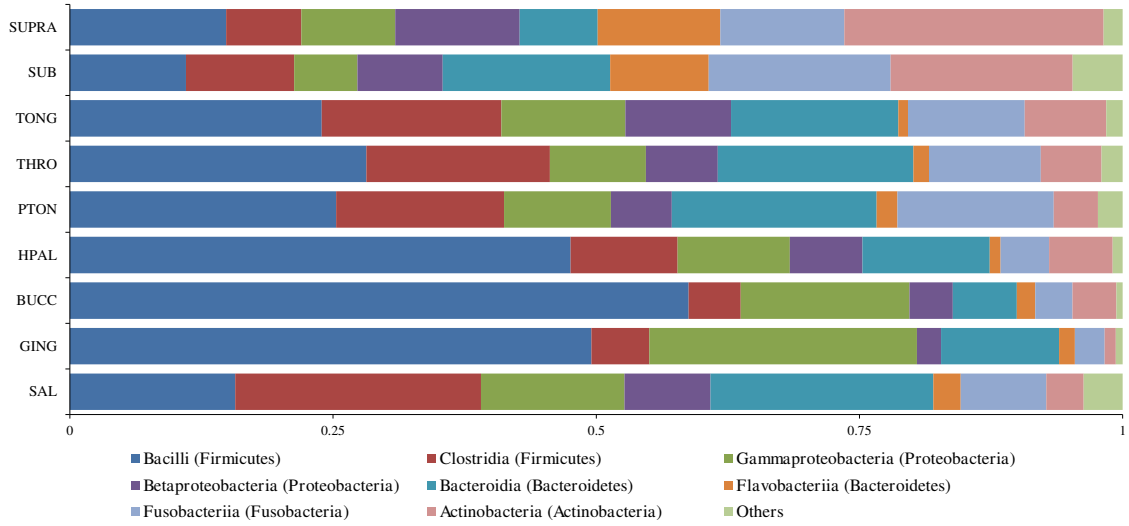
At the class level, the two major classes of *Firmicutes* together occupy more than 55.0% of the whole taxa on attached keratinized gingiva, buccal mucosa and hard palate. However, the amount of these two classes, *Bacilli* and *Clostridia* are extremely different. The ratios of *Bacilli* to *Clostridia* on these three sites are 9.0, 11.89 and 4.68. Since *Bacilli* are aerobic (i.e., they respire oxygen), these three sites are usually exposed to air. Palatine tonsils, throat and tongue dorsum are similar in proportion, 25.8% of *Bacilli* followed by 16.8% and 18.0% *Clostridia* and *Bacteroidia*. There is not a dominant class in saliva, since *Clostridia* and *Bacteroidia* have similar proportion. Although the taxa at phylum level on subgingival plaque and supragingival plaque are quite similar, they show some difference at the class level.

When coming to the genus level, *Streptococcus* ($26.0 \pm 13.8\% s.d.$) comprises a major population in the oral cavity. Two species of *Streptococcus* mainly appear in dental plaque of healthy human mouth, they are *S. sanguinis* and *S. mutans*. Both of them were found to be associated with dental caries. *Streptococcus salivarius* is fairly abundant in tongue dorsum, while *Streptococcus mitis* in other sites, such as buccal mucosa. *Haemophilus* also has a large population in the oral cavity, especially saliva samples. *Haemophilus parainfluenzae* is the biggest species of *Haemophilus* in oral cavity, which is reported to be highly associated with the pathogenicity of *Haemophilus parainfluenzae*.

OTUs were mapped into a phylogenetic tree whose root separates *Bacteria* and *Archaea*. The tree structure shows the evolutionary distance and relatedness of OTUs. The microbial communities are usually not distinguished by a single OTU, but a group of related members together. Phylogenetic tree put closely related species into a common branch. With the tree structure, the analysis did not need to be limited on the leaves (OTUs). Taking a branch of OTUs as a new taxonomic unit may give more information.

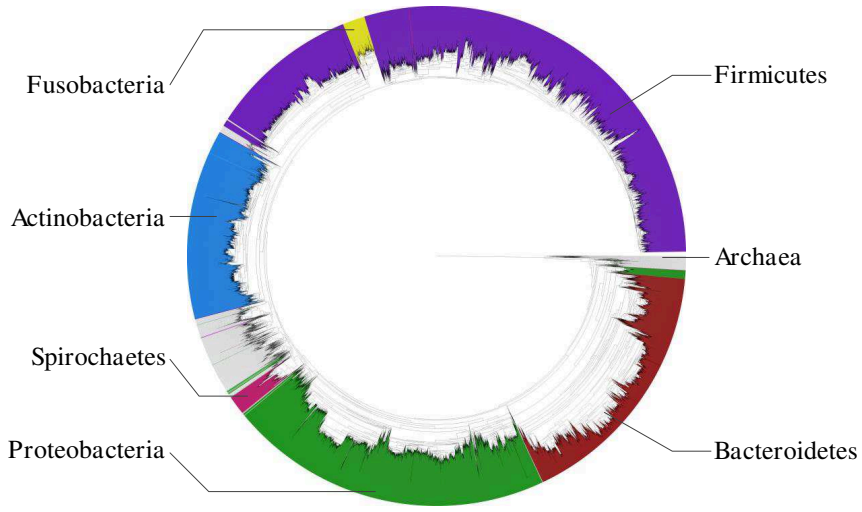


(a) Distribution of phyla in the nine oral cavity sites

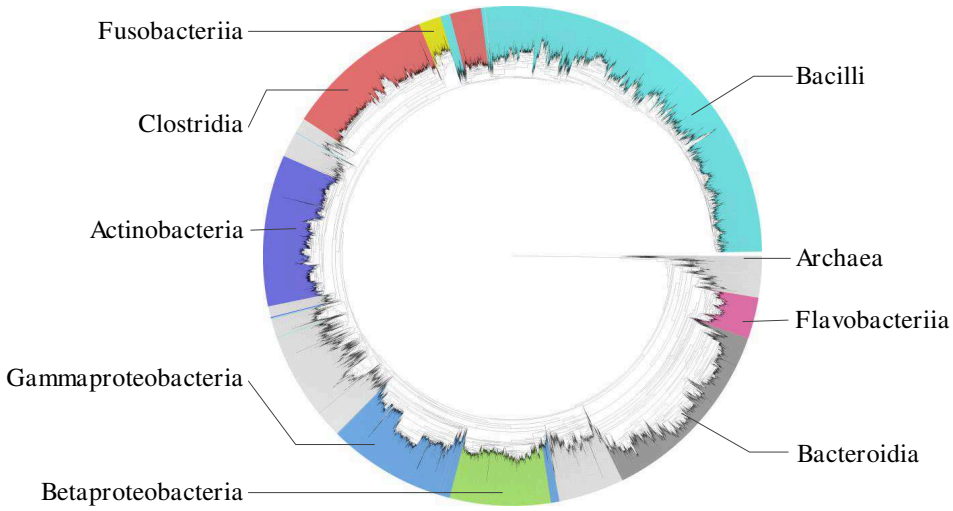


(b) Distribution of classes in the nine oral cavity sites

Figure 2.6: **Taxonomic composition of the microbes in nine oral cavity sites based on average relative abundance of 16S rRNA sequences.** Taxa from top 5 phyla (a), 8 classes and the remaining taxa are described as "Others". The full name of each abbreviations can be found in Table 2.1



(a) Abundant phyla in phylogenetic tree



(b) Abundant classes in phylogenetic tree

Figure 2.7: **Phylogenetic trees shows the relationships of OTUs from all oral samples.** OTUs are assigned to (a) phylum and (b) class level. The most abundant groups are highlighted in different colors.

Chapter 3

Classification of Oral Cavity Samples

In this thesis, 16S rRNA sequences were used as marker genes to identify the taxa in microbial communities. A supervised learning approach, SVM, was used to distinguish samples from nine different oral sites. Thousands of features were generated in our work, many of which are likely to be uninformative. With the help of feature selection, those uninformative features can be removed, resulting in an efficient training process. Moreover, a number of discriminative features were examined to uncover their biological relevance.

3.1 Feature Space

In our work, the OTU abundance calculated from the 16S rRNA samples acted as input features; the body site that the sample came from was the label attribute.

Assume the raw dataset is a sample-by-taxon abundance matrix $X(m, n)$, which can be displayed as follows:

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & & & \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{bmatrix}, \; y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_k \end{bmatrix}$$

where m and n indicate the number of samples and features. x_{ij} denotes the abundance of the i_{th} OTU in the j_{th} sample. Class label y indicates the body sites, $y_i \in C = \{c_1, c_2, \dots, c_g\}$. In our classification task, C= {saliva, attached keratinized gingiva, buccal mucosa, hard palate, palatine tonsils, throat, tongue dorsum, supragingival plaque, subgingival plaque}.

Data normalization or scaling converts the values of features into a specific range, which is usually performed as a data-reprocessing step. In SVM, scaling features

to $[-1, +1]$ or $[0, +1]$ is good for building the classifiers, since the algorithm may have less sensitivity to features with a small numeric range than a very large one. In addition, features in this numeric range also reduce the computational time in training step [151].

The number of sequences in each sample varied between approximately one and ten thousands, so we first converted raw abundance to proportions, or relative OTU abundance in each sample. The relative abundance was then scaled such that the largest value in each sample was set to 1.0.

3.2 Support Vector Machine

Support Vector Machines (SVMs) have been widely used in various applications since their introduction by Cortes and Vapnik in 1995 [95]. SVMs are model-based classification methods that try to maximize the width of a decision boundary between categories. This decision boundary or hyperplane is typically defined by a small number of boundary cases (the *support vectors*) with relatively small distances to cases of the other type [152]. A key attribute of SVMs is their ability to accept any similarity values that satisfy a set of constraints; the "kernel trick" allows mapping of cases into a higher-dimensional space where the linear SVM classifier can perform well [153].

3.2.1 Linear SVM

Suppose S is our training set containing n samples: $S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, x_i are the feature vectors of the samples and $y_i \in \{-1, +1\}$ represents the labels of the instances. To separate the positive and negative samples, SVM defines a hyperplane as:

$$f(\omega, x) = \omega^T x + b \tag{3.1}$$

which has the largest distance to the support vectors. This distance is called largest margin of the decision boundary. For all training samples on the right side of the margin, they should satisfy:

$$\omega^T x + b \begin{cases} \geq 1, when\ y_i = 1 \\ \leq -1, when\ y_i = -1 \end{cases} \tag{3.2}$$

Searching for the optimal hyperplane $f(\omega^T, b)$ is a process of solving a quadratic programming problem to minimize:

$$\frac{1}{2}|\omega|^2 \tag{3.3}$$

under the constraints:

$$\begin{gathered} \forall i = 1, 2, \ldots, n \\ s.t.\ y_i(\omega^T x_i + b) \geq 1 \end{gathered} \tag{3.4}$$

The Lagrange multiplier is introduced to solve this problem yielding the final decision function:

$$\begin{aligned} f(\omega, x) &= \omega^T x + b \\ &= \left(\sum_{i=1}^{n} a_i y_i x_i \right) x + b \\ &= \sum_{i=1}^{n} a_i yi \left\langle x_i, x \right\rangle + b \end{aligned} \tag{3.5}$$

This is the simplest SVM, which can only cope with a linearly separable training set. However, real-world datasets are typically more complex, with various amounts of noise and outliers. Samples cannot be correctly separated since noise and outlier cases can interfere with correct separation. The classifier will go worse if noise or outliers appear in the support vectors , since a small number of support vectors determines the decision boundary. The introduction of slack variable ξ_i blurs the decision boundary, so that the abnormal incorrectly separated points can be given less weight. So the optimization problem can be converted to minimize:

$$\frac{1}{2}|\omega|^2 + \sum_{i=1}^{n} \xi_i \tag{3.6}$$

under the constraints:

$$\begin{gathered} \forall i = 1, 2, \ldots, n \\ s.t.\ y_i(\omega^T x_i + b) \geq 1 - \xi_i \end{gathered} \tag{3.7}$$

Where the cost penalty C is used to control the model complexity of SVM, which allows the optimal trade off between bias and variance. When C is too large, the hyperplane will try to classify each training sample correctly while ignoring the test

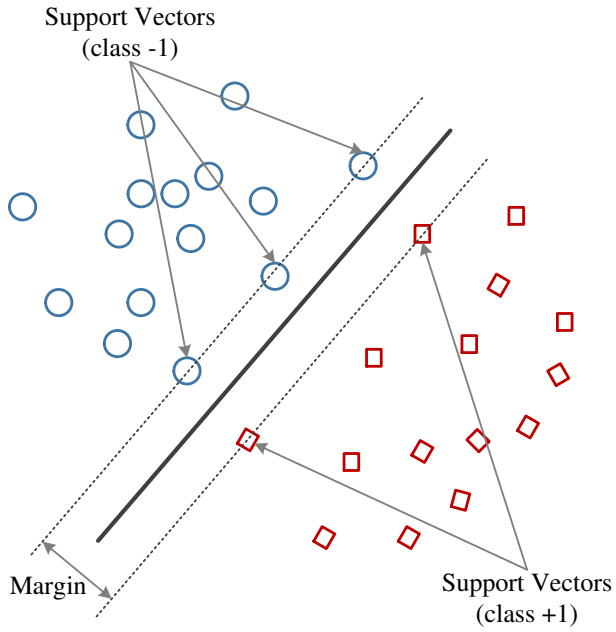


Figure 3.1: **Illustration of two labeled groups of samples in feature space separated by a hyperlane.**

set. This usually causes an over-fitting problem. However, if C is too small, all samples will be separated by a very large margin. It will hard to segregate the samples into their respective classes. However, there is no well-developed theory for determining an appropriate C value, and the typical approach to optimize C is to exhaustively try a number of values and chose the one with highest cross-validation accuracy.

3.2.2 Non-linear SVM

Finding a hyperplane on a linear separable dataset is straightforward given the formulation of the SVM, however, many real-world samples are distributed in a non-linear space. The kernel technique maps low-dimensional features into higher space and tries to separate the samples in this space. It also largely improves the computational efficiency of SVM, since the computation cost increases a lot with dimensionality. In SVM, kernel method works more like kernel trick. For the decision boundary is defined from the inner products of pair-variables and this inner product can be directly

replaced by kernel. So SVM does not actually compute the coordinates in that high and complex feature space, but gets this result directly from the kernel functions. If the input feature vector x_i can be expressed as:

$$x_i^{'} = \psi(x_i) \tag{3.8}$$

Then a kernel function can be introduced to satisfy:

$$K(x_1, x_2) = \langle \psi(x_1), \psi(x_2) \rangle \tag{3.9}$$

So the decision function becomes:

$$f(x) = \sum_{i=1}^{n} a_i y_i K(x_i, x) + b \tag{3.10}$$

An example of classifying nonlinear data with kernel function is given in Figure 3.2. The points form two curves on $X - Y$ space, so it is impossible to find a hyperplane that perfectly separates the two classes. However, when points are mapped to a third dimension, the boundary is obvious. There are a few commonly used kernels in SVM, including linear, polynomial, radial basis function (RBF) and sigmoid kernels. The RBF kernel was used as our baseline because of its reasonable number of parameters and widely applied in many problems. The formula can be given as:

$$K(x_1, x_2) = e^{-\frac{|x_1 - x_2|^2}{2\delta^2}} \tag{3.11}$$

A replacement of $\gamma = \frac{1}{2\delta^2}$ is usually used to simplify the equation as:

$$K(x_1, x_2) = e^{-\gamma |x_1 - x_2|^2} \tag{3.12}$$

where γ define the influence of each training sample can have. When γ is small, samples can have far-reaching influence. So the separation will be smooth. However, if γ is too large, the model will be very specific and highly sensitive to noise. Figure 3.3 explains how points were mapped into higher space using RBF kernel.

The kernel trick can convert a nonlinear separable problem into a linearly separable one. Because of the multiple choice of kernel functions, SVM works well on many different types of datasets.

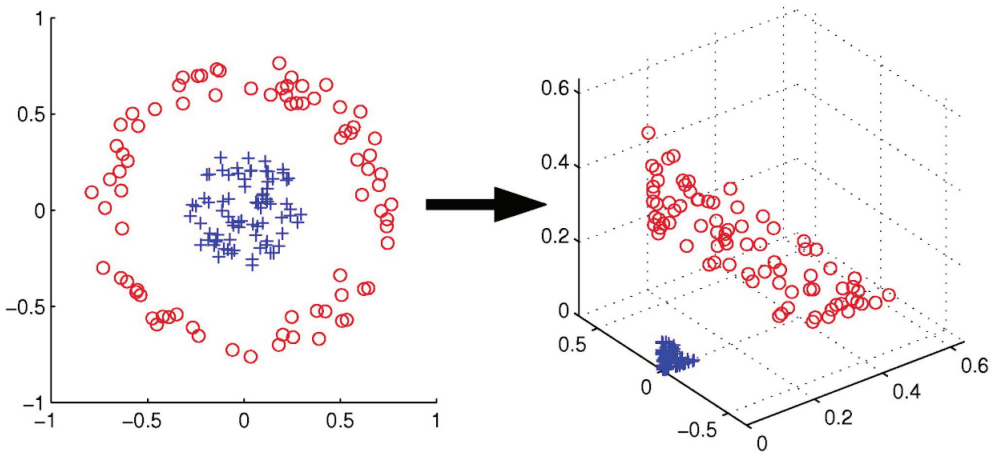


Figure 3.2: **An example of features that are not linear separable in the first two dimension, but becomes separable after mapping to a third dimension space [4].**

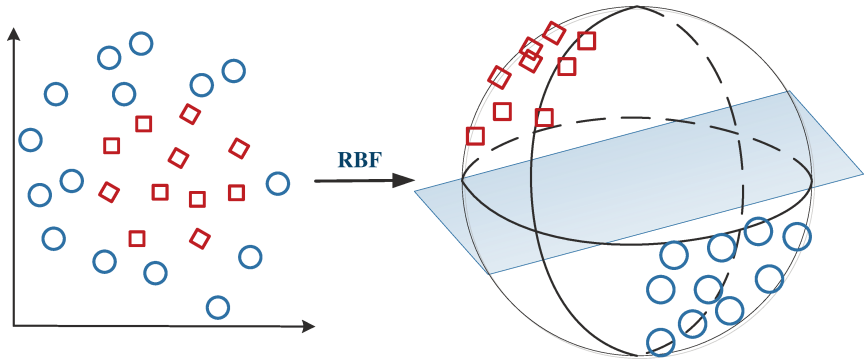


Figure 3.3: **Graphical description of how RBF kernel can map features into higher features space.**

3.2.3 SVM in Multi-class Classification

SVM itself is a binary classifier, however it can be applied to multiclass classification by decomposing the multiclass task into several binary ones. One-*vs*-all and one-*vs*-one are two commonly used strategies [154].

A one-*vs*-all classifier treats the samples from the i_{th} class as positive samples and the others as negative. It is computationally efficient, since only n classifiers are needed when there are n classes. But the sub-samples for each classifier are unbalanced; the number of negative samples outnumbers positive samples $n-1$ times. Such unbalanced dataset usually generates models with bias, which have preference to the class with more samples. The one-*vs*-one approach builds classifiers from each pair of classes. In the testing process, classifiers will vote for the class they preferred and the one with the most votes wins. The number of classifiers in one-*vs*-one quadratically grows with class.

For an n-class dataset, $\frac{n(n-1)}{2}$ classifiers will be constructed. Although one-*vs*-one requires $O(n^2)$ classifiers comparing to $O(n)$, the sample size for training each classifier is much smaller. It results in a faster and less memory-intensive training process. Importantly, each classifier is trained from a balanced subset as long as the original dataset is uniformly distributed.

3.3 Results Evaluation and Verification

3.3.1 Performance Evaluation

The performance of the prediction algorithm can be assessed by the accuracy, which is usually expressed as the percentage of correct predictions, quadratic error measures or correlation coefficients. Raw percentage can correctly reflect the performance when the number of samples in each label bin is similar [87]. However, classes can be imbalanced, with one group containing many more samples than the other: for example, in a dataset that compares diseased *vs* healthy individuals, there may be a much smaller diseased set. If all samples were roughly predicted as non-disease, there still would be a very high accuracy as the majority would be correct. To solve this problem, predictions are summarized into a confusion matrix with four numbers: True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). A confusion

		Predicted Class	
		Disease	Non-disease
Actual Class	Disease	**TP**	**FP**
	Non-disease	**FN**	**TN**

Figure 3.4: **Confusion matrix of a binary classification problem, disease vs non-disease.** TP is the number of samples that are disease and predicted as positive; FP is the number of samples that are disease but predicted as negative; FN is the number of samples that are non-disease but predicted as positive; TN is the number of samples that are non-disease and predicted as negative

matrix displays the number of actual and predicted samples in each class made by the classification algorithm. Explanation of the four numbers is in Figure 3.4.

Precision and recall evaluate different aspects of the performance. Precision is defined as the number of true positive samples over the number of all true samples, which tells you the percentage of the selected items that are correctly predicted. Recall provides the complementary information, which calculated as the number of true positive samples over the total number of positive predictions. The assessment of quadratic error methods is based on the distance between the true and predicted label, such as Hamming or Euclidean distance [155]. Correlation coefficient measures are frequently used in machine learning. One of the most commonly used is the Matthews Correlation Coefficient (MCC), proposed by Matthews in 1975 [156]. The value can be calculated as:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{3.13}$$

The result is ranged between -1 and +1, which reflects the degree to which correct assignments agree with the predictions. A coefficient of +1 indicates a perfect prediction, while -1 means a total disagreement between them. Measures mentioned above all can be extended to the case of multi-class classification problem. The confusion matrix displays the number of correctly predicted samples in each class. The rows

indicate the true label for each sample, while columns indicate the label predicted by the classifier. Detail extension of other measures can be found in [87]. We adopted the percentage of correct predictions and confusion matrix to evaluate the classifiers, since the dataset in the thesis is in balance.

3.3.2 Statistical Testing

Classification with different input features was repeated for 100 times with shuffling of the samples, and the final accuracy was expressed as their mean value. To exam the results, statistical methods were used. Basic statistical information was first given: the *mean* reflects the average performance under one type of classification model; *standard deviation* measures how consistent this group of results are; *standard error* estimates the likely difference between the mean and future data; *minimum* and *maximum* values give the range of the data; *confidence interval* establishes a range of values that within which a future data may fall with a specific probability. A group of accuracies was also shown with a histogram and a boxplot to examine its normal distribution.

Compare to the benchmark done, some attempts we did improved the classification performance, while some did not. For the significance of improvement, we did two-sample t-tests. The two-sample t-test is used for determining whether the means of two samples are significantly different. Assumptions behind this test include: both samples must follow normal distributions and they are independent. We established the null hypothesis as a lack of difference between two samples. T-value was calculated from:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}}} \tag{3.14}$$

where N_1 and N_2 are the sample sizes, X_1 and X_2 are the means, S_1^2 and S_1^2 the standard deviations. The p-value can determined then, indicating the probability that the null hypothesis can be accepted. A threshold of 5% was used in the thesis. If the p-value was lower than 0.05, we can declare that the improvement did not occur by chance.

3.4 Classification of Nine Sub-sites

Classification was performed using the LIBSVM package [157] with RBF kernel. To pick the best combination of kernel width γ and error penalty parameter C, a grid search using different combinations of C and γ was done as a pre-experiment (finite sets of attempt values for $C = [\log_2 -5, \log_2 15]$, $\gamma = [\log_2 -15, \log_2 3]$. An one *vs* one strategy was adopted to perform multi-class classification. A five-fold cross-validation approach was adopted to evaluate the classification models. This cross-validation procedure was repeated 100 times for each trial, each time using a different random number seed, in order to generate distributions of accuracy scores.

3.4.1 Performance Comparison

Samples from both the V1-V3 and V3-V5 variable regions of 16S were classified using SVM with RBF kernel. The model built from V1-V3 dataset achieved an accuracy of 64.4%, while the V3-V5 samples contributed to a better model whose accuracy reached 69.7%. A two-sample t-test was given to measure the significance of improvement. The performance of the model from V3-V5 samples yielded 4.70% (in Figure 3.5) higher accuracy than that from V1-V3 samples. Since V3-V5 regions have 1,160 more samples than V1-V3 regions, the increased accuracy may be due to a larger training set. Sequences from the V3-V5 regions may also be more powerful in identifying different microorganisms, regardless of sample size. Based on the result above, we chose to focus on samples from the V3-V5 regions only.

3.4.2 Grouping of Sites

We generated PCoA plots based on unweighted UniFrac distances between samples to visualize the separation of points between the nine sample types (in Figure 3.6(a)). A table containing the relative abundance of OTUs in each sample was used to calculate the distance. A phylogenetic tree was also passed as input to inform the evolutionary relationships between OTUs.

We used QIIME to calculate the principal coordinate axes for each sample. The first two principal coordinates explain 15.07% of the total variance in the data set, and do not provide clear separation of any of the nine sample types. Clustering patterns

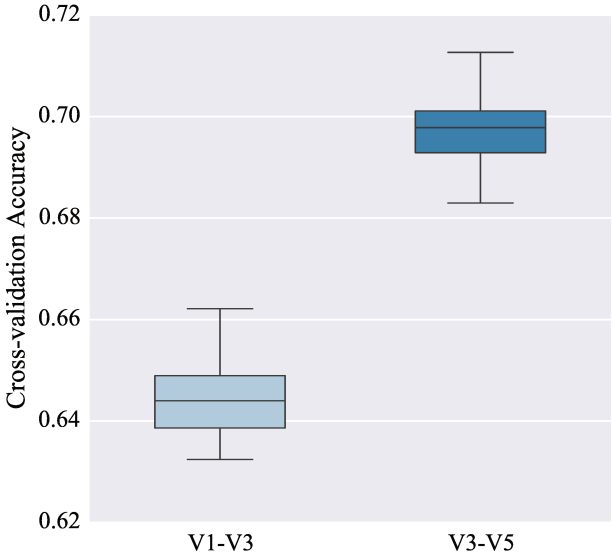
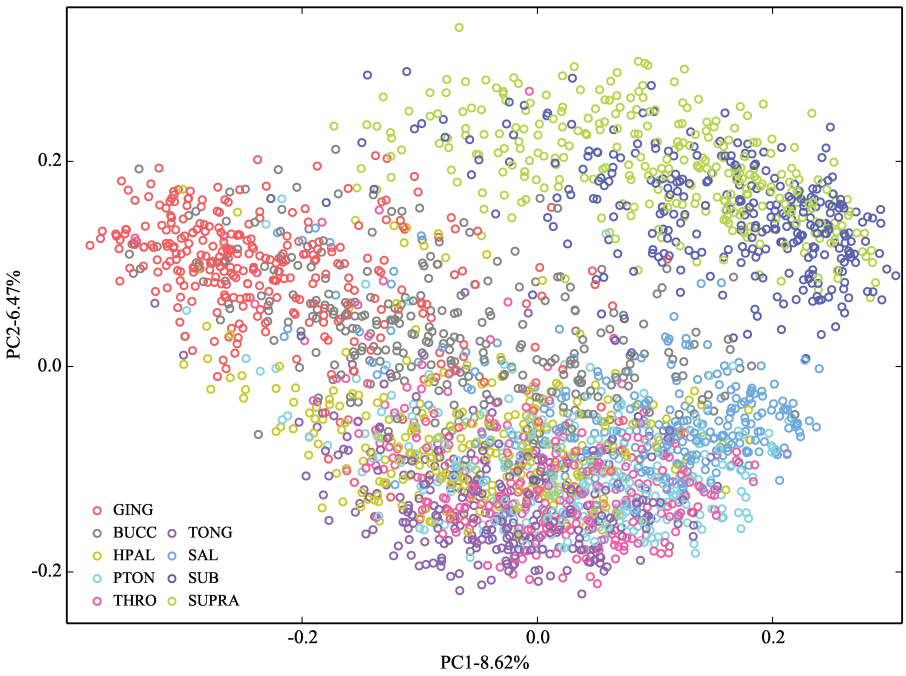


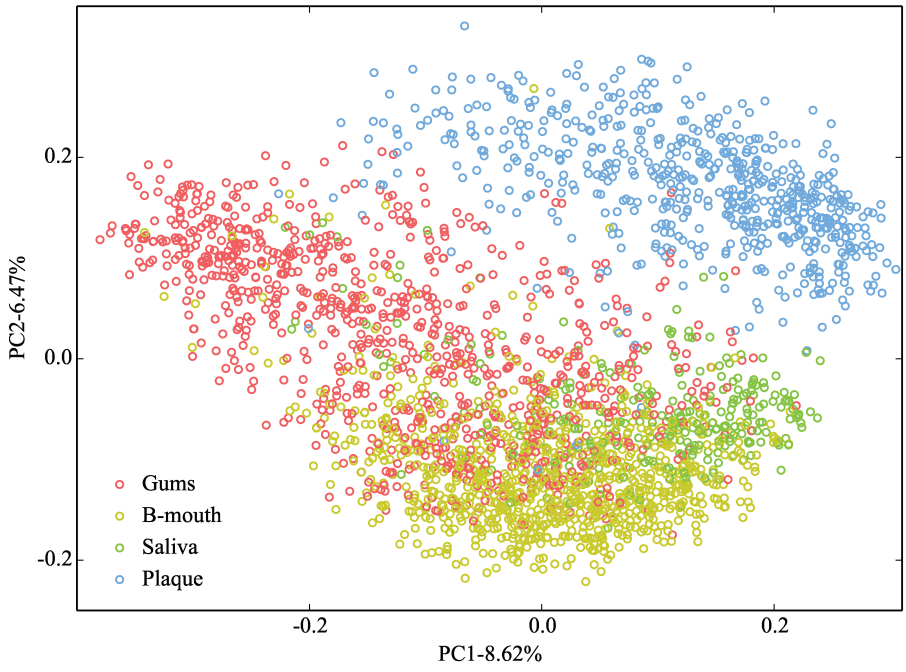
Figure 3.5: **Classification accuracy with features of sequences from different variable regions.** Significant difference between models built from V1-V3 and V3-V5 dataset was observed.

are nonetheless visible in the figure; in particular, the supragingival and subgingival plaque samples constitute a group that is largely separate from the other sample types. The other seven sample types occupy one large cluster, but none of these is uniformly distributed throughout the cluster: for example, attached keratinized gingival samples tend to have negative values in principal coordinate one and near-zero values in principal coordinate two, while both principal coordinates of buccal mucosa samples are near-zero values. The PCoA plot reflects the distance between these oral samples. Overlapping sets of samples share considerable amount of similarity, which may imply a challenging classificaton problem.

We performed SVM classification using an RBF kernel on all 2,702 oral cavity samples from the V3-V5 regions. The cross-validated classification accuracy with respect to the sample type label was 69.7%. The confusion matrix (in Figure 3.7(a)), which shows the frequency with which samples of a given type were correctly classified or misclassified to another category, shows a non-random pattern of misclassification. Of the nine oral cavity sites, saliva and tongue dorsum were classified with the highest accuracy (87.2% and 84.4%, respectively), while samples from the palatine tonsils and throat were correctly classified less than 45% of the time.



(a) Principal coordinates analysis of nine sites



(b) Principal coordinates analysis of four clusters

Figure 3.6: **Principal coordinates analysis of nine oral cavity sites.** The same data set is shown with all nine oral cavity sites (a) and four clustered groups (b) as labels. Distances were computed using the unweighted UniFrac distance.

		Predicted Class								
		GING	**BUCC**	**HPAL**	**SAL**	**SUB**	**SUPRA**	**PTON**	**THRO**	**TONG**
Actual Class	**GING**	**0.799**	**0.155**	**0.016**	0.000	0.017	0.007	0.002	0.004	0.000
	BUCC	**0.135**	**0.734**	**0.067**	0.008	0.014	0.014	0.008	0.011	0.009
	HPAL	**0.030**	**0.085**	**0.678**	0.011	0.007	0.006	0.020	0.125	0.037
	SAL	0.001	0.011	0.012	**0.872**	0.007	0.008	0.035	0.036	0.019
	SUB	0.003	0.008	0.003	0.005	**0.643**	**0.332**	0.003	0.004	0.000
	SUPRA	0.000	0.001	0.003	0.000	**0.215**	**0.781**	0.000	0.000	0.000
	PTON	0.017	0.056	0.063	0.018	0.021	0.005	**0.424**	**0.213**	**0.183**
	THRO	0.011	0.029	0.167	0.020	0.007	0.003	**0.179**	**0.443**	**0.140**
	TONG	0.000	0.002	0.013	0.006	0.000	0.000	**0.072**	**0.063**	**0.844**

(a) Confusion matrix of nine-way classification

		Predicted Class			
		Gum	**Saliva**	**Teeth**	**Mouth**
Actual Class	**Gum**	**0.870**	0.000	0.029	0.101
	Saliva	0.020	**0.732**	0.020	0.229
	Teeth	0.025	0.000	**0.969**	0.006
	Mouth	0.052	0.006	0.019	**0.923**

(b) Confusion matrix of four-way classification

Figure 3.7: **Confusion matrix of nine-way oral site classification without feature selection.** Rows indicate the correct label for each sample, while columns indicate the label predicted by the classifier. Each cell indicates the number of samples of a given type classified to each sample type. The classification patterns of all nine classes (a) and a recoding into four classes (b) are shown.

We identified four natural groupings of sites based on these patterns. 90.0% of samples from a group comprising the attached keratinized gingiva, buccal mucosa and hard palate samples were most often classified within the same group, which we define as "gums"; the misclassification of 12.5% of hard palate samples to the throat represents the only major confusion between this group and any other. Consistent with the separation seen in Figure 3.6(a), 98.6% of subgingival and supragingival plaque samples are classified as one of these two sites. Samples from throat, palatine tonsils and tongue dorsum constitute another group responsible for 85.4% of all classifications, although the throat and tonsils are also conflated with the hard palate and buccal mucosa. Finally, salivary samples are relatively better classified, with an accuracy of 87.2 %. In general, these four major groupings consist of sites that are proximal in the mouth, corresponding roughly to gums (attached keratinized gingiva, buccal mucosa and hard palate), plaque (supragingival and subgingival plaque), back of the mouth (throat, palatine tonsils and tongue dorsum), and saliva. Because of the gag reflex, collecting samples from throat is the most difficult work among the nine sites. Samples are easy to be contaminated during the depressor getting back from throat, so throat samples may be mixed with hard palate microbes [117].

Recoloring sample points in the original PCoA plot to reflect the four groups (in Figure 3.6(b)) shows a clearer distinction among sites, albeit still with a substantial amount of overlap among all but the plaque group. The nine sites were recoded into their four constituent categories, and once again classified using an SVM with the RBF kernel. The classification accuracy of plaque samples is 96.9%, as compared with 73.2% accuracy for saliva, 87.0% for gums, and 92.3% for the back of the mouth. In the four-way classification, the number of saliva samples is much smaller than that of gums or hard plaque, which makes the decision boundary prone to labels with large sample size. It can be one reason for the reduced accuracy of saliva samples. The plaque samples were well separated from the other groups, but difficult to distinguish based on the confusion matrix in Figure 3.7(b), we chose to focus on this two-class problem in order to try and improve the classification accuracy for a tractable subset of sites.

3.5 Challenges

The work done in this chapter used all samples from the oral cavity. Features were constructed from OTU abundance without any optimization. From this preliminary attempt, three major challenges in microbial community classification can be identified:

1. High dimensionality and sparseness. A total of 12,845 OTUs were generated at a 97% OTU identity threshold, but there are only 2,702 samples in our dataset. High-dimensional classification is expensive both in time and memory. Moreover, it may raise over-fitting problem. The taxon abundance feature is a sparse matrix. Among all the OTUs, as many as (4,325) 33.7% of them appear only once. Only (1105) 8.6% of the OTUs exist in >10% of the samples. Rare features cannot be proved to be useless for classification, so none of them were removed.

2. Feature dependence: many machine-learning methods and feature selection operators assume that all the input features are independent. However, many input OTUs have highly correlated patterns of abundance. In microbial classification problem, sometimes a number of OTUs cannot be useful features independently, but they may become powerful when combined together.

3. Limited information contained within OTUs: OTUs that clustered from 16S rRNA only contain taxonomic information within communities. However, other information such as phylogenetic distance, functional profiles can also be used to differentiate microbial communities. Building features space based on various information is another challenge.

Results from the 9-class classification show the correlations among oral sites. Saliva was mixed with microorganisms from various sources since it bathes several sites in the mouth. Samples from teeth are quite isolated, which can be found in the confusion matrix and the PCoA plot. However, subgingival plaque and supragingival plaque samples were very difficult to distinguish from one another. Since these sites provide a challenging binary classification problem, in the next chapter, we develop and test ideas to improve classification accuracy using these two sites.

Chapter 4

Classification of Hard Plaque Samples

In the previous chapter, the microbiome of the oral cavity was characterized and samples were classified using SVMs. However, the classifiers so far were informed only by the relative OTU abundance of each sample. Features represented in this way usually presume that OTUs are the appropriate unit of analysis, which may not be the case. Phylogenetic relationships among the organisms express their evolutionary distances; these relationships may be key attributes of microbial communities. Augmenting the classification methods with phylogenetic insights may yield better results. Representing organisms based on their phylogenetic groupings breaks the constraint of rigid OTU thresholds, which may provide additional information to the SVM.

In addition to taxonomic information, functions in the microbial community can also serve as useful features. Each body site differs remarkably in functions and the microbial pathways are highly associated with the body site functions. Intestinal microbes are mainly responsible for the production and absorption of nutrients, while microbes on skin are protective against pathogenic bacteria. Functional profiles characterize the microbial communities differently from taxonomic components. The functional difference between body sites may provide discriminative information to the classifiers.

Ensemble methods are algorithms that integrate a number of classifiers to give better predictions than a single classifier does. Ensemble methods suggest another approach, since we found the predictions from different classifiers are not always consistent.

4.1 Custom Kernels based on Phylogenetic Distances

4.1.1 Kernel Methods

Kernel methods are a set of algorithms that work on various types of dataset and find the general relationships in it. Datasets such as, sequences, documents, and images, all can be performed. In addition to their application in SVMs, kernel methods can be used in a range of techniques including Fisher Discriminant Analysis, Principal Components Analysis and Spectral Clustering. Problems are usually solved by kernel methods in two steps: map the original dataset into the feature space that the adopted algorithm can deal with; design a function to discover linear patterns in the mapped feature space.
From the formula 3.8 we induced in Chapter 3:

$$x_i^{'} = \psi(x_i) \tag{4.1}$$

Then a kernel function can be introduced to satisfy:

$$K(x_1, x_2) = \langle \psi(x_1), \psi(x_2) \rangle \tag{4.2}$$

where ψ maps the original features into a dot product space, or feature space. The inner product in the feature space is K, kernel function. Kernel function is the vital ingredient for kernel method, since it has to be designed depending on the types of the specific dataset and be efficient in computation. SVMs are capable of working on nonlinear datasets, since a hyperplane can always be found by mapping features into higher dimensions and the kernel function still works efficiently in infinite feature space.

Generic polynomial and RBF kernels are widely used, but custom kernels that incorporate biological insights can be useful as well. For example, alignment-based kernels improved SVM performance in predicting protein subcellular localization, which is a vital aspect of protein function [88]. Since phylogenetic distance is an effective measure in the comparison of microbial communities, custom kernels based on this property may be effective in discriminating microbiome samples.

4.1.2 Four Distance Measures

Parks *et al* examined 39 measures of beta diversity and analyzed their relative similarity [5]. The authors identified several groups of measures with very high correlations among predictions. We chose single representatives from four of these groups. The custom kernels were developed based on two phylogenetic (weighted and unweighted UniFrac distances) and two non-phylogenetic (Euclidean and Canberra distances) measures. Non-phylogenetic measures assess community differences based only on OTU presence and abundance, while phylogenetic measures are also informed by evolutionary relationships between these OTUs in the phylogenetic tree.

The Euclidean distance is one of the most popular and straightforward distance measures. It can be expressed as the length of the path connecting two points $(x_1, x_2, \ldots, x_n)$ and $(y_1, y_2, \ldots, y_n)$ in n-space:

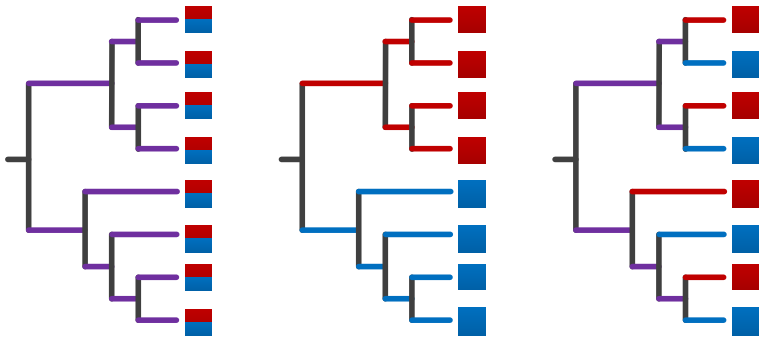
$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{4.3}$$

The Canberra distance is a non-phylogenetic measurement introduced by Lance et al in 1966 and the modified form mainly used today was suggested a year later [158, 159]. The dissimilarity sums the results calculated from the absolute difference between the pair of variables divided by their total value. The Canberra distance between $(x_1, x_2, \ldots, x_n)$ and $(y_1, y_2, \ldots, y_n)$ in n-space can be calculated as:

$$d(x, y) = \sum_{i=1}^{n} \frac{|x_i - y_i|}{|x_i| + |y_i|} \tag{4.4}$$

This metric is more sensitive to quantitative (abundance) than binary (presence-absence) differences between samples.

Some studies found the Canberra distance works well in separating community samples and had a better performance than the Euclidean distances [160]. Possible reasons for these differences were proposed: the Euclidean distance does not scale the values during calculation, while the Canberra distance uses the sum of the variables in each dimension as a scaling factor. The Canberra distance was initially proposed as a software metric and performed well in detecting intrusions in networks and information systems, but it has been readily adopted in ecology and genomics. Jurman *et al* used the Canberra distance as an indicator to measure the stability of ranked



(a) (b) (c)

Figure 4.1: **Calculation of the UniFrac distance.** Blocks in blue and red colors indicate sequences from each of two communities. Branches in purple means taxa from these two samples are mutually shared. (a) A tree of taxa from two similar communities, where all the braches are shared. A minimum UniFrac distance value of 0.0. (b) A tree of two very different communities, sequences in red and sequences in blue appear in disjoint sets of branches. A maximum UniFrac distance value of 1.0. (c) A tree shows that parts of sequences from these two communities share branches on the tree, while some of them do not.

biomarkers from functional genomics, so that a reliable set of genes can be selected for classification or annotation [161].

UniFrac distance, as a method of estimating microbial distance based on phylogenetic information, was proposed by Lozupone *et al* [113, 114, 115]. UniFrac expresses the phylogenetic dissimilarity between each pair of samples after all taxa in these samples have been placed into a rooted phylogenetic tree. Any branch that has taxa from both samples as children is called a "shared branch", whereas a branch whose children are from one sample only are "unique branches". The UniFrac distance of these two samples can be calculated as the sum of lengths of all unique branches divided by the sum of all branch lengths in the tree..

Figure 4.1 gives examples of three UniFrac distances. The two communities in Figure 4.1(a) have similar phylogeny (minimum UniFrac distance 0.0), since all the taxa in red and blue samples simultaneously appear in each branch. However, Figure 4.1(b) shows taxa from these two samples are in distinct braches, which leads to a maximum UniFrac distance 1.0. Between these two extreme cases, Figure 4.1(c) shows a UniFrac distance of 0.5, whose total branch length from shared branches and unshared branches are equal. The formula of calculating the distance between two

communities, x and y, is consistent with [162]:

$$d\left(x,y\right)=\sum_{i=1}^{n}\frac{b_i\left|I(P_i^x>0)-I(P_i^y>0)\right|}{\sum_{i=1}^{n}b_i} \tag{4.5}$$

where n is the number of braches in the phylogenetic tree, and b_i corresponds to the length of branch i. P_i^x and P_i^y are the taxa proportions in the i_{th} branch from community x and y respectively. Function $I(.)$ indicates the presence or absence of species within the branch.

The results from this measurement have an assumption that all taxa in the phylogenetic tree come with similar abundance, which is called unweighted UniFrac distance. However, the difference in abundance can be critical for distinguishing communities, so a weighted UniFrac distance was developed. Weighted UniFrac distance gives a weight to each branch based on the amount of divergence in the taxa abundance. The equation can be given as:

$$d\left(x,y\right)=\frac{\sum_{i=1}^{n}b_i(P_i^x-P_i^y)}{\sum_{i=1}^{n}b_i(P_i^x+P_i^y)} \tag{4.6}$$

These two measurements yield different and complementary dissimilarity score between communities. Unweighted UniFrac is sensitive only to the presence and absence of different OTUs, which may amplify the effect of rare lineages. However, not all rare members are important. Weighted UniFrac considers the relative abundance, but sometimes the most abundant lineages are not the discriminative members. Generalized UniFrac distances which offer different tradeoffs between presence and abundance have been developed [162]. However, here we focus on the widely used unweighted and weighted UniFrac distance.

A distance matrix is used to store the distance between each pair of samples, which is square and symmetric. It can be represented as an n-by-n matrix:

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \dots & \dots & d(n,n) \end{bmatrix}$$

where $d(i, j)$ measures the dissimilarity between sample i and j. All the diagonal values are defined as zero, which indicates that there is no difference between the element and itself. A value of 1.0 denotes the maximum dissimilarity.

All four beta-diversity measures in the thesis were calculated with QIIME. Converting the data to a common range before applying distance calculation is necessary, so that all the attributes can be given an equal weight. To account for disparities in OTU counts in different samples, these similarity scores were combined with several different OTU table preprocessing approaches, including raw OTU count, relative abundance, rarified counts from 500 to 3,000 per samples and cumulative sum scaling (CSS) normalization [86]. Since beta diversity expresses the dissimilarity between each pair of samples, we subtracted each such value from 1.0 in order to generate similarity values for the SVM classifier. The classifiers with custom kernel were performed using Libsvm package [157].

4.1.3 Performance Comparison

Using the four beta-diversity measures above, we developed custom kernels that express the similarity between all pairs of samples. The hypothesis underlying the use of these kernels is that similarity scores based on ecological similarity measures will outperform a naïve RBF kernel, especially when these measures are based on information not available to the classifier (for example, phylogenetic information in the case of UniFrac). The performance of SVMs with different custom kernels is given in Figure 4.2. Colors are consistent with Parks *et al*'s clusters on beta diversity measurements.

Phylogenetic measures did not work better than non-phylogenetic measures: for example, the widely used unweighted and weighted UniFrac measures yielded 74.4% and 73.7% accuracy. The Canberra distance obtained an accuracy score of 76.5%, which is better than the UniFrac distance, but still worse than using OTU abundance with an RBF kernel. Although many types of microbial samples cluster well based on beta-diversity measures such as UniFrac, this is clearly not the case with the two types of plaque. A possible reason for the discrepancy between RBF and our custom kernels is the optimization of the gamma parameter, since none of the four beta-diversity measures have such process. Another reason may be that the semi-defined distance function makes samples not satisfy the KarushKuhnTucker (KKT)

distance_metric	Canberra	unweighted UniFrac	weighted UniFrac	Euclidean	RBF
otu_count	0.762	0.729	0.734	0.695	/
otu_abundance	0.762	0.729	0.736	0.736	0.762
rarefaction	0.766	0.770	0.740	0.622	/
css	0.769	0.729	0.738	0.754	/
mean ± s.d	0.765 ± 0.003	0.739 ± 0.017	0.737 ±0.002	0.702 ± 0.051	0.762

Figure 4.2: **The performance of SVMs with different custom kernels.** The distance metrics are ranked by their mean values and highlighted with colors consistent to Parks et al's cluster result. Highly correlated and prominent measures are grouped in one color set, the calculation of correlation can be found in [5].

conditions, which is a generalized method Lagrange multipliers . So we cannot ensure the Lagrange function is still convex, resulting that the hyperplane may not be global optimal.

4.2 Clade Features based on Phylogenetic Relationships Among OTUs

A *clade* refers to a group containing a common ancestor and all its descendants, which is a grouping of lineages based on phylogenetic relationships. The term was first proposed by Huxley in 1957 [163]. In the phylogenetic tree, clades are nested within one another (in Figure 4.3). A clade can have thousands of organisms or only a few of them.

By using the phylogenetic tree generated in the sequence-processing step, sets of closely related OTUs can be grouped into clades. Since OTUs can only be identified after assuming a fixed similarity, OTUs cannot go into deep lineage if the similarity were set very low. However, if the microbial communities were differentiated by taxa in big families, a very high similarity cannot detect such groups [164]. In fact, although many studies pick OTUs at 97% similarity, this percentage is an empirical value. So it is reasonable to argue that OTUs at lower or higher similarity, such as 90% or 99% may better characterize the communities. To support the argument, Knights *et al* identified OTUs at different similarity between 50% and 95%, besides 97% and 99% [41]. Features were constructed from these OTUs and put into RF classifiers.

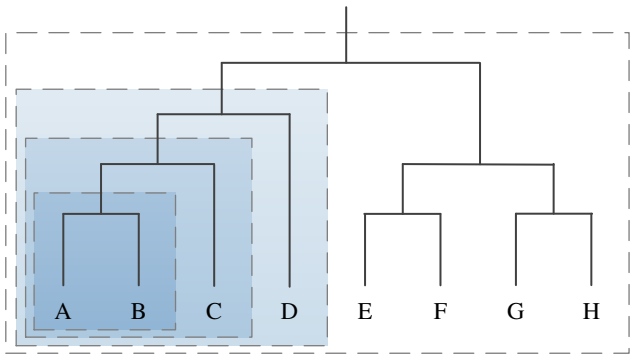


Figure 4.3: **Generation of clade-based features.** Each clade in the tree corresponds to a feature in the data set; for example, the darkest box encompasses OTUs A and B.

Results show the accuracy did not change much when the similarity became larger than 65%. However, the accuracy got worse at high levels of similarity, for example, in addition to the more widely used 97% and 99%. Their results indicate that OTUs of high similarity cannot always provide discriminative features for classifiers. However, it is difficult to determine a perfect similarity before classification. To solve this problem, we proposed the idea of clade features. The aggregation of clades breaks the limitation by strain-level variation, so that discriminative taxa at different levels can be found.

4.2.1 Clade Features

We constructed clades using the reference phylogenetic tree as described above, and added each clade to the existing OTU feature space. The abundance of a clade was calculated as the sum of abundances of all its descendants. Clade abundance reduces the sparsity of the dataset and removes the need for a single, universal similarity threshold. However, this method increases the number of features relative to the original OTU table. Since the number of non-leaf clades is equal to the number of internal nodes in the phylogenetic tree, this clade-based approach can generate a total of $l-2$ features, where l is the number of leaves in the tree, if the uninformative root clade that includes all OTUs is ignored. To solve this problem, we applied different feature selection strategies.

4.2.2 Feature Selection

The clade approach generated thousands of features, many of which are likely to be uninformative, and in aggregate can reduce the speed and accuracy of SVM training. Although some species appear in only a small number of samples, rare features may nonetheless be useful for classification and should not be removed by default. We used feature selection to accelerate learning by removing uninformative OTUs. Among the multitude of available feature selection techniques, we used two types of approach: filter methods, which consider the usefulness of features based on their apparent relevance to the classification problem, and wrapper methods, which assess features by quantifying their effect on the accuracy of a trained model.

One of the filter methods used was information gain, which ranks the features based on the amount of predictive information obtained from the presence or absence of a term [165]. To measure the amount of information, we introduce the concept of entropy, which tells the expected amount of information in the content. Let $Y = \{y_1, y_2, \ldots, y_k\}$ denote the set of values in the space and fit the probability function $P(Y)$. So the entropy $H(Y)$ can be defined as:

$$H(Y) = -\sum_{y \in Y} p(y) \log_2 p(y) \tag{4.7}$$

where the log is usually to the base of 2, meaning the entropy is measured in bits. A simple example is tossing a fair coin, where the probability of heads and tails are both 0.5. So the entropy is $-0.5 \times \log 0.5 - 0.5 \times \log 0.5 = 1$.

The definition of entropy can be extended to a pair of variables, which is called joint entropy $H(X, Y)$. If the variables $(X, Y) \sim p(x, y)$, the conditional entropy $H(Y|X)$ can be defined as:

$$H(Y|X) = -\sum_{x \in X, y \in Y} p(y, x) log \frac{p(y, x)}{p(x)} \tag{4.8}$$

Conditional entropy can be regarded as the uncertainty of Y after given the information of X. Mutual information is the difference between entropy $H(Y)$ and the conditional entropy $H(YX)$:

$$MI(Y; X) = H(Y) - H(Y|X) \tag{4.9}$$

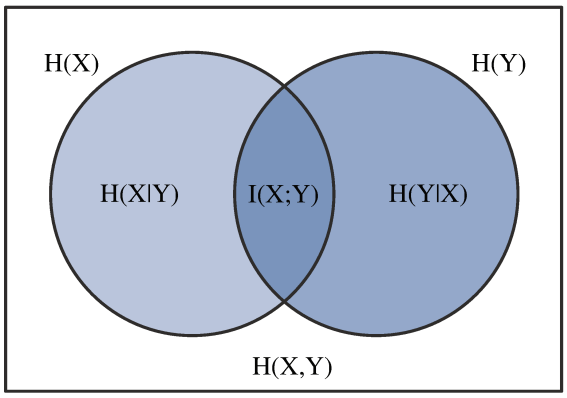


Figure 4.4: **Illustration of relationship between entroy and mutual information.** For the two variables X and Y, area of the rectangle is joint entropy H(X,Y). The full area of the circles on the left denotes individual entropy H(X), the light blue area is the conditional entropy H(X|Y). The circle on the right shows for variable Y. The area overlopped by these two circles is the mutual information I(X;Y).

which gives the reduction in entropy if X is known. A Venn diagram in Figure 4.4 shows the relationship between entropy and mutual information. However, information gain is different from mutual information, since it calculates the expected reduction in entropy:

$$\begin{aligned} IG(Y;X) &= \sum \frac{S_i}{S} MI(Y;X_i) \\ &= \sum \frac{S_i}{S} H(Y) - \sum \frac{S_i}{S} H(Y|X_i) \\ &= H(Y) - \sum \frac{S_i}{S} H(Y|X_i) \end{aligned} \tag{4.10}$$

where S denotes the whole number of samples, $\frac{S_i}{S}$ is the proportion of samples with the i_{th} value in feature X. We computed the information gain of each feature and selected the top N of them to build the classifier.

The Chi-square (χ^2) test is used to identify the difference between the sampling distribution and the expected distribution given by the hypothesis [97, 166]. Suppose there are N samples in total, $Y = \{y_1, y_2, \ldots, y_k\}$ enumerates the classes in target space and $X = \{x_1, x_2, \ldots, x_k\}$ is one variable in the feature set. A null hypothesis H_0 states that variables X and Y are independent. The probability of rejecting this hypothesis can be calculated from the χ^2 value under the χ^2 distribution with one

degree of freedom. Formulas are referred from:

$$\chi^2(X,Y) = \sum_{i=1}^{n}\sum_{j=1}^{k}\frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} \tag{4.11}$$

$$E_{i,j} = \frac{M_i B_i}{N} \tag{4.12}$$

where $O_{i,j}$ is the number of samples has the i_{th} feature from the j_{th} class. M_i indicates the number of samples has the i_{th} feature, whatever the class is. B_j is the amount of samples from the j_{th} class. So the $E_{i,j}$ is the expected number of samples with the i_{th} feature and from the j_{th} class.

By searching the χ^2 distribution table, the corresponding p-value can be found. A small probability allows us to reject the null hypothesis, which indicates this feature has high correlation with the class label.

Wrapper methods were also employed in the thesis. These approaches select features according to their performance when used by a learning algorithm, as the evaluation function. We considered Random Forest (RF) feature permutation as a wrapper method, which is very popular in feature ranking [98]. It is easy to use and only few parameters need to be tuned. Each feature value would be shuffled randomly and test the performance of the model trained by those permutated features. Variables were ranked based on the effect of randomizing their values between the categories to be predicted. In the context of a trained RF classifier, randomizing a useful variable would lead to a significant drop in accuracy, whereas a similar procedure on an uninformative variable would have no effect. Measurements such as prediction accuracy, precision or MCC can be used to evaluate the performance.

Filter methods are fast and suitable for problems of high dimensionality. Since these approaches are independent of the classification algorithm and typically consider features one at a time, features only need to be ranked once. However, filter methods lose the interaction with the classifiers. This means the selected features may not be the subset that are of greatest utility to the classifier. What is worse, filter methods treat all features as independent from one another, which is not true in fact. By contrast, wrapper methods interact well with the learning algorithm and may take feature dependencies into account. Features selected by wrapper methods are usually more powerful than those by filter methods [112], but the searching process require

long time and high computational cost. Although OTUs with strong marginal effects (i.e., those that have good predictive power independent of any other variables) should be identified by all three of our chosen approaches, useful combinations of variables might be highlighted by the RF approach.

4.2.3 Performance Comparison

We augmented the OTU table with relative abundance information about clades that contain multiple OTUs, to determine whether explicit specification of relationships amongst OTUs might lead to better prediction accuracy. Fifty-two OTUs were lost because their corresponding sequences failed the PyNast quality control filters, leaving a total of 6,996 OTUs. To this set we added 6,994 clades, corresponding to all internal nodes in the reference tree, minus the uninformative root node which always has a relative abundance of 1.0. The classification accuracy obtained without feature selection was less than that obtained from the OTU table without clade information (73.8% vs. 76.2%). While the OTU+clade table has almost twice as many features

Table 4.1: **Maximum accuracy of SVM classifiers trained with different combinations of input features.** The initial numbers show the accuracy score, with numbers in parentheses indicating the total number of features used to train and test the classifier. The four types of input features used were (i) OTUs only; (ii) OTUs and clades comprising related sets of OTUs; (iii) Functional predictions made using PICRUSt; and (iv) a dataset comprising all generated features. Feature selection techniques used were the filter methods, information gain and Chi-square; and the feature permutation wrapper method.

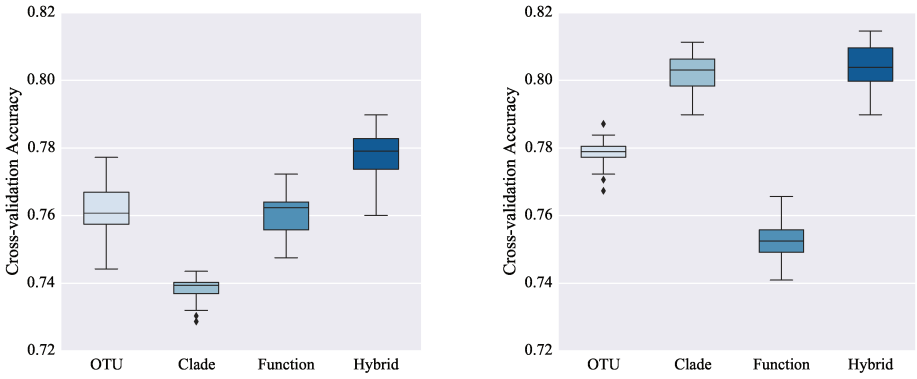
Cross-Validation Accuracy (number of features)				
Features	***Without Feature Selection***	***With Feature Selection***		
		Info_Gain	***Chi_Square***	***Feat_Perm***
(i) OTU	0.762 (7048)	0.779 (60)	0.777 (50)	0.798 (20)
(ii) Clade	0.738 (14402)	0.802 (110)	0.800 (170)	0.802 (100)
(iii) Function	0.761 (6191)	0.762 (120)	0.754 (100)	0.761 (60)
(iv) Hybrid	0.777 (1556/1518)	0.804 (92/78)	0.805 (68/62)	0.805 (28/23)

as the OTU abundance table alone and includes over 99% of the original OTUs, it appears that the higher dimensionality of the data confounds the SVM classifier, making it more difficult to build an accurate model. However, applying feature selection as above gave at least 80% accuracy (results in Table 4.1, Figure 4.5). A ten-fold cross-validation was performed and we recorded the accuracy from each fold and the overall dispersion.. The statistical descriptions are displayed in Table 4.2.

As was observed previously with the OTU table, the filter methods required more features to achieve their maximum classification accuracy (110 and 170 for information gain and Chi-square versus 100 features for the RF approach). When analyzing the selected features, all three methods selected more clades than OTUs (106 clades for information gain, 159 clades for Chi-square and 19 clades for the RF feature permutation). Figure 4.6 shows the performance of classifiers with different numbers of features. The information gain (in Figure 4.6(a)) and Chi-square (in Figure 4.6(b)) approaches had similar performance: the accuracy of OTU abundance varied between 76% and 78% with different numbers of features. However, clade abundance gave accuracy scores that were often in excess of 80%. Both OTU and clade abundance can classify samples well with a small number of RF-ranked features (in Figure 4.6(c)), but with the number of features increasing, the performance of OTU abundance worsened

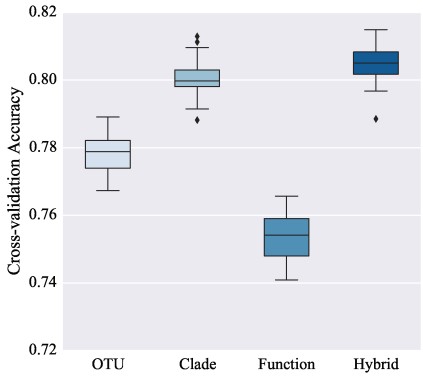
Table 4.2: **Statistical summary of the accuracies from each of ten cross-validation folds with different features.**

Features		*Mean*	*Std.Deviation*	*Std.Error*	*95% Confidence intervals*		*Min*	*Max*
					Lower	*Upper*		
OTU	Info_Gain	0.776	0.069	0.005	0.648	0.847	0.645	0.850
	Chi_Square	0.778	0.053	0.003	0.684	0.856	0.677	0.867
	Feat_Perm	0.783	0.064	0.004	0.664	0.863	0.656	0.867
Clade	Info_Gain	0.806	0.025	0.001	0.778	0.853	0.777	0.860
	Chi_Square	0.792	0.053	0.003	0.698	0.870	0.691	0.881
	Feat_Perm	0.800	0.035	0.001	0.743	0.850	0.742	0.854
Function	Info_Gain	0.756	0.057	0.003	0.659	0.816	0.653	0.817
	Chi_Square	0.753	0.054	0.003	0.659	0.816	0.653	0.817
	Feat_Perm	0.751	0.058	0.003	0.655	0.833	0.653	0.833
Hybrid	Info_Gain	0.804	0.035	0.001	0.756	0.872	0.753	0.884
	Chi_Square	0.795	0.043	0.002	0.746	0.888	0.739	0.903
	Feat_Perm	0.805	0.045	0.002	0.744	0.884	0.740	0.892

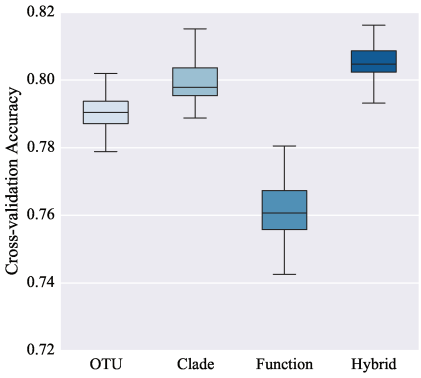


(a) Without Feature Selection

(b) Information gain



(c) Chi-square



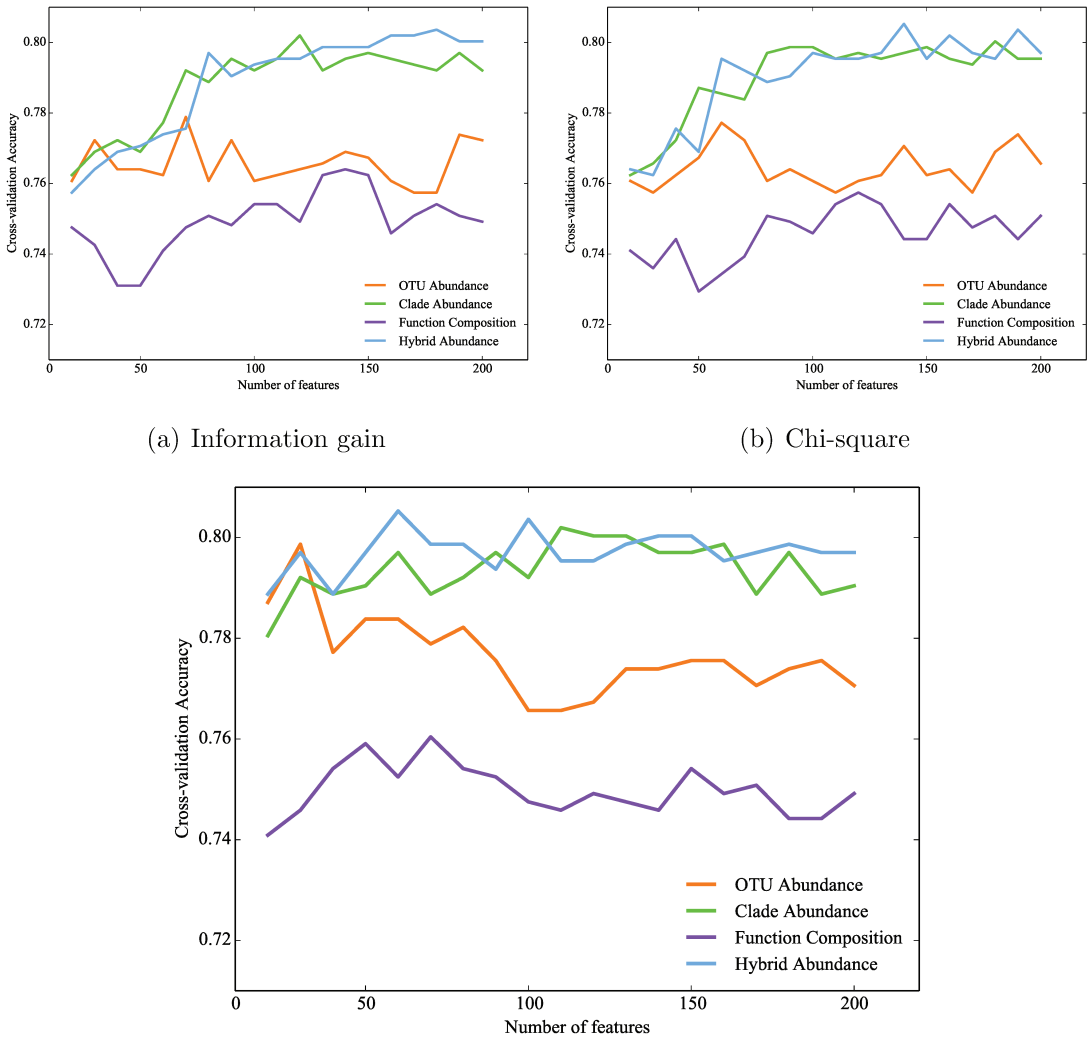
(d) RF feature permutation

Figure 4.5: **Boxplots show the distribution of 100 times cross-validation accuracies with different input features.** Features (a) without feature selection and with feature selection: (b) information gain, (c) Chi-square, (d) RF feature permutation criteria.

whereas clade abundance kept working well. Clade abundance gained improvement in accuracy, which were tested by two-sample t-tests (results in Table 4.3). It appears that explicitly modeling the phylogenetic correlations between OTUs allows the filter methods to exploit the interactions that were previously accessible only to the wrapper method.

4.2.4 Phylogenetic Distribution of Selected Features

The phylogenetic mappings and corresponding phylum-level GreenGenes taxonomic classifications of OTUs are shown in Figure 4.7(a). Subgingival plaque samples tended



(a) Information gain

(b) Chi-square

(c) RF feature permutation

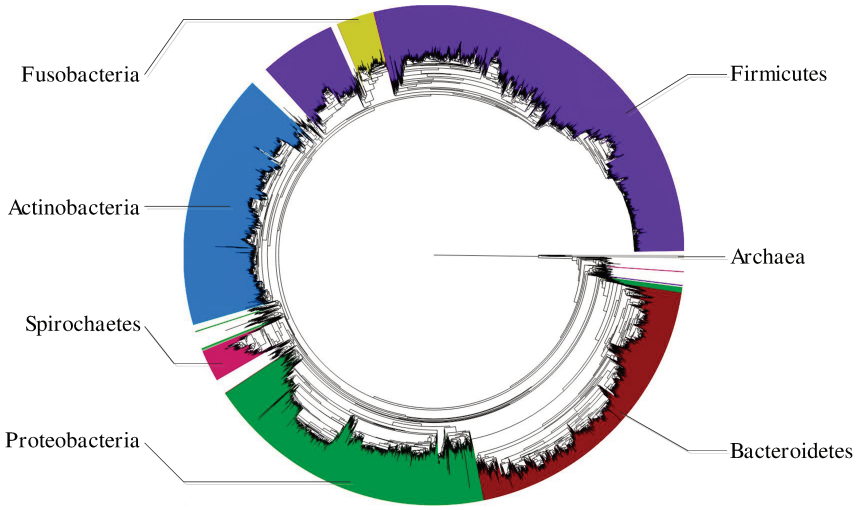
Figure 4.6: **Classification accuracy with different sets of input features.** The classification accuracy is shown for sets of 10 to 200 of the top-ranked features according to the information gain (a), Chi-square (b), and RF feature permutation (c) criteria. The four types of input features used were OTUs only (orange markers); OTUs and clades (green markers); Functional predictions made using PICRUSt (purple markers); and all generated features (blue markers).

to have higher proportions of *Bacteriodetes* (sub: 0.254 *vs* supra: 0.191), *Fusobacteria* (0.172 *vs* 0.118) and *Spirochaetes* (0.029 *vs* 0.006), whereas *Actinobacteria* (0.175 *vs* 0.247) and *Proteobacteria* (0.151 *vs* 0.215) are more abundant in supragingival plaque. *Firmicutes* had similar abundance in both types of site (0.213 *vs* 0.220), however, at the class level, *Bacilli* (0.110 *vs* 0.148) and *Clostridia* (0.103 *vs* 0.071) showed larger deviations.

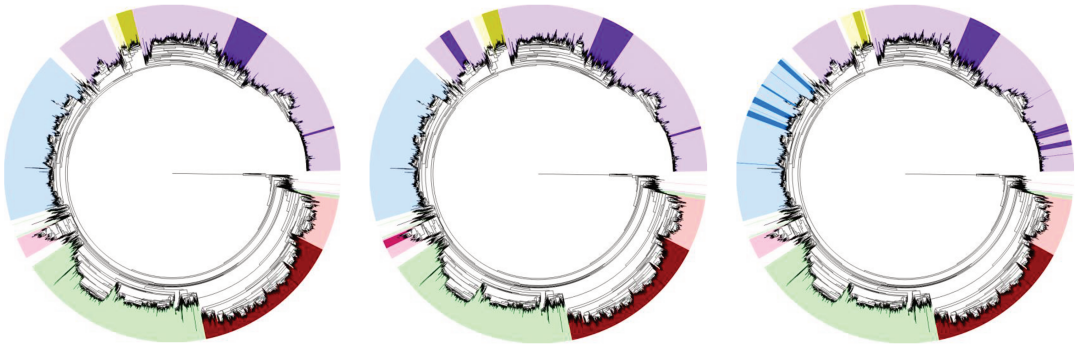
We then highlighted the optimal features that were selected by each method in the phylogenetic tree. The filter methods, information gain (in Figure 4.7(a)) and Chi-square (in Figure 4.7(c)) chose similar clades including a large clade within *Bacteroidetes* and smaller groupings within *Firmicutes* and *Fusobacteria*. The Chi-square approach chose the largest number of features, including *Spirochaetes* and *Clostridia* clades that were not chosen by the information gain criterion. By contrast, the RF feature-permutation approach, which included the fewest features in its optimal set, selected a greater diversity of features (in Figure 4.7(d)). This set of features included unique clades of *Firmicutes* and *Actinobacteria* that were not identified by the information gain or Chi-square approaches. For all the three feature-selection methods, near-optimal classification accuracy was obtained for many different numbers of selected features, suggesting that some of the highlighted clades in Figure 4.7 may not

Table 4.3: **Improved accuracy of SVM classifiers trained with different combinations of input features.** The initial numbers show the improvement of accuracy score, with numbers in parentheses indicating the p-value and t-value from t-test. Three pairs of features were compared: (i) Clade *vs* OTU; (ii) Hybrid *vs* OTU; (iii) Hybrid *vs* Clade. Feature selection techniques used were the filter methods, information gain and Chi-square; and the feature permutation wrapper method.

Improvement of Accuracy (p-value, t-value)				
Features	***Without F.S.***	***With Feature Selection***		
		Info_Gain	***Chi_Square***	***Feat_Perm***
(i)Clade *vs* OTU	-0.024 (p<2.2e-16;t=-34.6)	0.023 (p<2.2e-16;t=26.5)	0.024 (p<2.2e-16;t=24.8)	0.009 (p<1.7e-13;t=8.6)
(ii)Hybrid *vs* OTU	0.015 (p<2.2e-16;t=18.1)	0.025 (p<2.2e-16;t=25.9)	0.028 (p<2.2e-16;t=29.7)	0.015 (p<2.2e-16;t=15.6)
(iii)Hybrid *vs* Clade	0.039 (p<2.2e-16;t=56.9)	0.002 (p<1.1e-1;t=1.6)	0.005 (p<3.6e-5;t=4.3)	0.003 (p<7.3e-3;t=3.5)



(a)



(b) Information gain (c) Chi-square (d) RF feature permutation

Figure 4.7: **Phylogenetic mapping of top-ranked clade and OTU features.** (a) Reference tree comprising all observed oral site OTUs, with branch lengths proportional to substitutions per site. Key phyla are highlighted with different colors. (b-d) mapping of highest-ranked clade and OTU features according to information gain (b: 110 features), Chi-square (c: 170 features) and RF feature permutation (d: 100 features).

be important for classification purposes. The selected clades are nested; one big clade may contain a number of small ones. Because of that, although 110 features were selected by information gain, there are only 5 big clades highlighted in the phylogenetic tree. Nonetheless, the higher variety of features selected by the RF feature permutation approach shows the value of testing combinations of features during the selection process.

4.3 Functional Encondings

Taxonomic diversity is an important characteristic of microbial communities, however, essential knowledge of functional capabilities helps understand the role they play. The functional capabilities of a microbiome sample are often assessed using metagenomics (see Chapter 1.4), but the HMP collected only 764 metagenomic samples as compared to 2,702 16S samples. Since large numbers of samples are desirable for model training, we used a method that predicts functions within the community from 16S rRNA gene sequences.

4.3.1 Functional Features

The PICRUSt software [116] allows the prediction of functional gene complements in microbial samples that have been characterized with marker genes such as 16S. We used these predictions as the basis for classification; if the functional capacity of microbes in a system is more important than their specific taxonomic affiliations, then a function-based approach to classification may yield higher accuracy. PICRUSt uses phylogenetic information to make its predictions, and thus functional information will be highly correlated with the OTU and clade data. However, since phylogenetically distant lineages can share common functional features, the predictions made by PICRUSt may identify functional similarities between OTUs that belong to different high-level taxonomic groups such as classes and phyla. Thus the predictions made by PICRUSt are not completely redundant with the OTU and clade features considered in this work.

The functional profile predicted by PICRUSt is expressed as a table containing the count of functional genes in each sample. Kyoto Encyclopedia of Genes and Genomes (KEGG) orthologs were the gene family profiles we adopted [167]. KEGG

is a database containing genomic information and high-level functions. The functional assignments in KEGG connect genes into different molecular interaction and reaction networks, such as metabolism and genetic information processing. KEGG Ortholog (KO) groups were manually defined based on the hierarchy of pathways and used as identifiers to map each gene function. New features were constructed from the KO abundance in each sample, where maximum values were also scale to 1.0 to eliminate the disparity.

To measure the reliability of the functional predictions, we calculated the Nearest Sequenced Taxon Index (NSTI) values for each sample. The value of NSTI was calculated from average length of the braches that can separate the OTUs from the reference genome, given a weighted by the normalized abundance of the OTU. So the lower the value is, the more reliable the prediction is expected to be. A 0.04 ± 0.01 *s.d.* was obtained in the thesis. It is similar to the values reported for HMP samples (mean NSTI $= 0.03 \pm 0.02$ *s.d.*), which were generally well predicted by PICRUSt, as compared with 0.23 ± 0.07 *s.d.* for a less well-predicted hypersaline community [116].

4.3.2 Hybrid Features

Since both function and taxonomy can potentially characterize the community, combinations of the two types of feature may provide complementary information. We combined clade and functional abundances and created new hybrid features. Results show that classifiers were able to distinguish samples with a small number of features, which allows us to focus on the key attributes that distinguish the two types of plaque.

4.3.3 Performance Comparison

A total of 6,191 KEGG orthologs, which incorporate functional predictions in addition to homology information, were used as input features to an SVM with an RBF kernel as performed above. The cross-validated accuracy of the model trained with all features was 76.1%, almost the same with the corresponding OTU abundance model. These observations are consistent with those of Xu *et al* [168], who found that taxonomy alone was sufficient to model microbial community structure. Functional features are still useful for predictive purposes, but their failure to improve

classification accuracy may be attributable to several factors. It may be that the crucial functions are not well annotated by KEGG, because of misannotations or a failure to assign to any meaningful functional category. The granularity of KEGG functional attributions and the presence of irrelevant features may also impede the discovery of important predictive attributes.

To assess the performance of classifiers based on combined clade and functional information, we performed feature selection on a hybrid data set containing features of both types. The results of feature selection and classification are shown in Table 4.1 and statistical descriptions in Table 4.2. The accuracy obtained from all three types of feature selection was 80.4%-80.5%, and the RF feature permutation approach yielded a maximum accuracy score with only 28 clade-based and 22 functional features. The small improvement in accuracy of the hybrid approach relative to clade-based classification alone (in Table 4.3) suggests that the functional features do not provide much useful complementary information to taxonomy: the increase of 0.3% relative to previous wrapper-based results corresponds to only a few additional correctly classified cases.

4.3.4 Biological Meaning of Selected Features

The selected clade-based features are mainly from *Streptococcus*, with several of them restricted to the opportunistic pathogen *S.anginosus*. There are other clades of *Streptococcus*, underscoring the importance of different members of this genus in the oral cavity. Although *Streptococcus* is typically a more significant component of supragingival plaque, consistent with its facultative anaerobic lifestyle, three of the *Streptococcus*-containing groups were overrepresented in subgingival plaque, while the fourth was 50% more abundant in supragingival plaque. This finding suggests that the most common types of *Streptococcus* may not be the best discriminators between the two types of plaque. Some selected features were broader in their taxonomic distributions, including genera such as *Prevotella*, *Fusobacterium* and *Dialister*.

One of the selected functional features is *sag*A, which encodes the basic structural unit of Streptolysin S (SLS). Bacteria such as *S.pyogenes* use SLS to lyse host cells and acquire iron [169, 170]. This function appears to be strongly associated with subgingival plaque. High correlated functional features also include a beta-lactam

resistance protein, overrepresented in subgingival plaque; streptokinase, which can aid the spread of *Streptococcus* infection through cleavage of fibrils [171]; proteins involved in resistance to tellurium and vancomycin; and a Type IV secretion system component. Although many of the implicated functions relate to host-microbial interactions, we found no clear, strong connections to aerobic or anaerobic lifestyles.

4.4 Combining Information from Multiple Classifiers

Empirical studies shows that no algorithm can outperform all others on all possible datasets. Instead of trying to optimize a single learning model, combing several different trained predictors may yield better results [2]. Ensemble methods train a group of base classifiers on the same dataset and make decisions based on the predictions from all of them. SVM, kNN and decision trees are all commonly used base classifier.

4.4.1 Different Predictions from Various Classifiers

Although our focus was on SVMs, we also considered two other supervised classification methods, SourceTracker [172] and RF [96]. SourceTracker is a Bayesian approach that assigns probabilities that a given sample is derived from each of a set of environment types, which characterizes the microbial community in another aspect. We used SourceTracker version 0.9.5 software as implemented in QIIME with default settings. Analogous to five-fold cross validation, the set of samples was divided into 5 subsets: one subset was sink samples for testing while the other four were source samples training. We repeated this process five times with different cross-validation subsampling.

RFs, first introduced in 2001, are an ensemble method merging decision trees with voting schemes. Each decision tree is constructed based on M (mtry) randomly chosen features from the input dataset. The prediction of every sample is determined by the majority vote of all these decision trees. RF classifiers are popular both for feature selection and classification, and were found by Knights *et al* to perform well on several test datasets. RF classification was implemented with scikit-learn 0.15 [173].

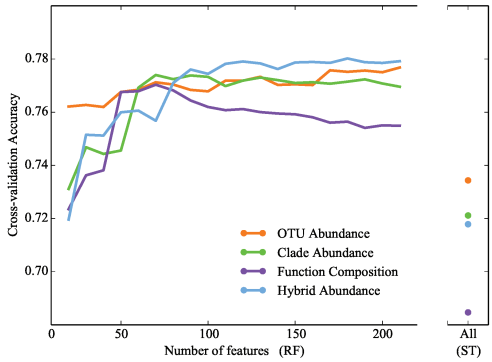
RF trials were carried out in an analogous manner to SVM, using sets of 10 to 200 features ranked by the three feature selection methods. Since SourceTracker has

a much longer running time, we only used the whole set of features to test its performance. For the RF model, the lowest and highest accuracies were respectively achieved by functional abundance and the hybrid feature set, as was observed with the SVM classifiers (in Figure 4.8). OTUs and clades had similar performance when features were ranked by information gain (in Figure 4.8(a)) and Chi-square (in Figure 4.8(b)), but clade abundance improved under RF feature permutation (in Figure 4.8(c)) ranked features. SourceTracker estimates the posterior probability scores for each possible source of a sample; we used the source with the highest posterior probability as the final prediction. The clade and hybrid feature sets did not perform as well as OTUs, likely due to the large number of highly similar clade features that were not removed with a feature selection process.

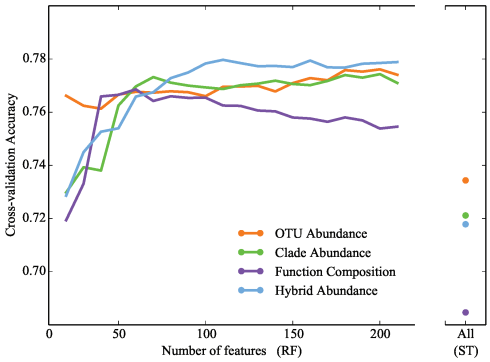
Both SourceTracker and RF had similar performance in distinguishing the two hard plaque sites, with classification accuracy between 75% to 78% with OTU abundance features. However, the predictions on each sample were different between methods. Figure 4.9 contrasts the predictions made by each pair of methods on each sample. All three methods had consistent predictions on most samples, since the majority of samples are either perfectly classified or perfectly misclassified by each pair of methods. However, off-diagonal samples show differences between two methods, and some samples are classified 100% correctly by one approach and 0% correctly by the other.

4.4.2 Design of Ensemble Algorithm

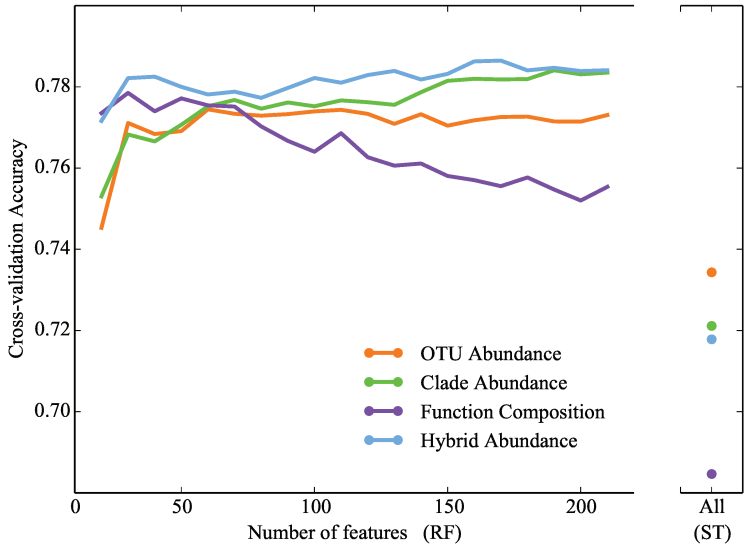
Since the predictions from SVM, RandomForest and SourceTracker were sometimes divergent, we used all three methods as base classifiers for an ensemble model. All the samples were divided into five subsets as 5-fold cross-validation (in Figure 4.10). Training set was then trained with SVM, RandomForest and SourceTracker respectively. During this training process, another 5-fold cross-validation was adopted, that separated the training samples into five subsets. After this inner cross-validation, each sample was assigned a label. As was done with the individual SVM and RF classifiers, this process was repeated 100 times with random shuffling of the cross-validation sets. For SourceTracker, the classification was performed for one time due to the speed of SourceTracker. However, SourceTracker are able to give predictions in



(a) Information gain



(b) Chi-square



(c) RF feature permutation

Figure 4.8: **Classification accuracy with different sets of input features.** Each plot is split into two portions; (Left) the random forests classification accuracy with sets of 10 to 200 of the top-ranked features according to the information gain (A), Chi-square (B), and RF feature permutation (C) criteria, (Right) the Source Trackers classification accuracy with only top 200 and the whole features. The four types of input features used were OTUs only (orange markers); OTUs and clades (green markers); Functional predictions made using PICRUSt (purple markers); and all generated features (blue markers).

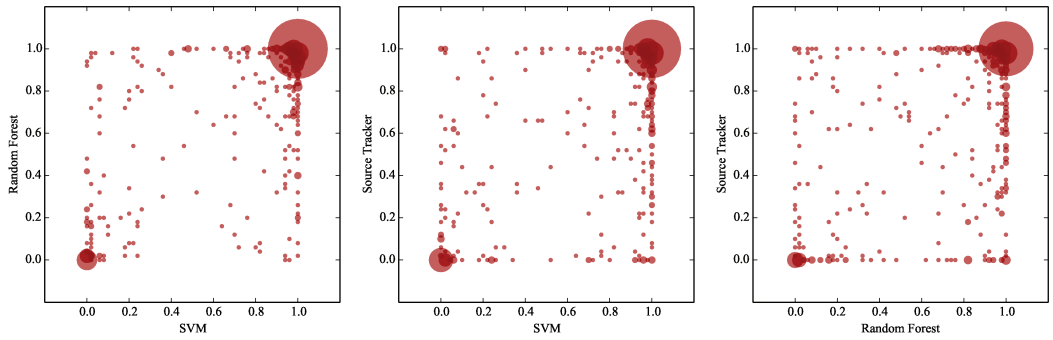


Figure 4.9: **The predictions on all samples: (a) SVM *vs* random forests, (b) SVM *vs* SourceTracker, (c) random forests *vs* SourceTracker.** The values on x and y axis indicate the frequency of samples were correctly predicted. The size of the nodes reflects the number of samples that were classified with the indicated accuracy, from 0% by both classifiers in the lower left-hand corner to 100% in the upper right-hand corner.

probability (possibility of the sample from Subgingival plaque, Supragingival plaque and Unknown) contrasting to the discrete values (+1, -1) from SVM and RF.

To construct the new feature space, we combined the corresponding probabilities and the top N (from 10 to 200 at a step of 10) features from RF feature permutation. A RF classifier was adopted as our final classifier with default settings. The classifiers were trained by the new features and evaluated using the testing set.

4.4.3 Performance Comparison

Figure 4.11 compares the performance of classifiers from ensemble method and with original OTU abundance. Unfortunately, the ensemble method did not yield a substantial increase in accuracy. The best accuracy obtained was 78.5% with 120 features. Two possible reasons may account for the failure of ensemble method: the performance of base classifiers themselves and the design of final classifiers. The performances of the three base classifiers are not perfect, so we cannot expect the samples that are wrongly predicted by all base members can be correctly labeled by the final classifier. The ensemble methods follow the philosophy that errors made by one base classifier can be compensated by another, so that the final result would ideally have less errors than either base classifiers. However, our result failed to achieve this goal, though predictions from SVM, RF and SourceTrack did have some amount

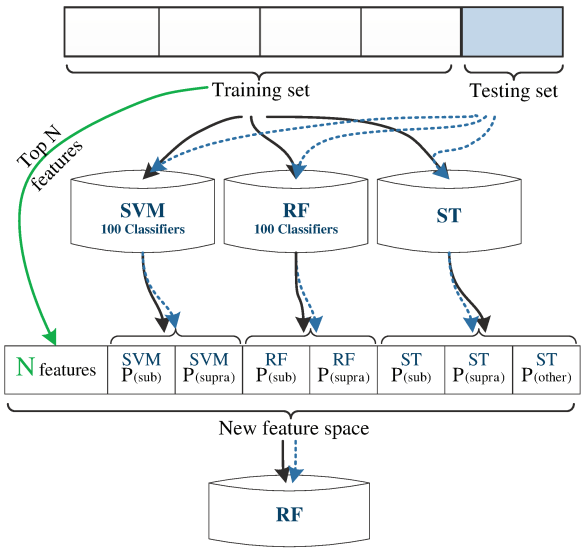


Figure 4.10: **Illustration of ensemble method.** The streams in black indicate the data flow of training set, while streams in blue are from the testing set. The top N features adding to the final classifier were highlighted in green.

of disagreement. Possibly the design of final classifier needs to be improved. Instead of using random forest to combining all classifiers, we can try some other algorithms, such as artificial neural network.

Figure 4.11: **Ensemble classification with different sets of input features.** The accuracy is shown for sets of 10 to 200 of the top-ranked features based on RF feature permutation. Two types of classifiers were compared: ensemble classifier (green markers) and classifier with original OTU abundance (blue markers).

Chapter 5

Conclusion and Future Work

Hard plaque is one of the four groups we defined in Chapter 3. With the exception of the saliva samples which were somewhat conflated with the back of the mouth, distinguishing samples between groups was relatively easy (accuracy $\geq$ 87%). The differences between samples of the same group are more subtle. Using hard plaque samples as a test case, we proposed several improvements to the sample classification problem. The best test-set accuracy scores obtained were in the range 80%-81%, which demonstrates useful learning but is of little value for diagnostic applications. Clade and clade-function abundance encodings did yield improvements in this problem, while functional abundance alone and custom-kernel strategies failed to give better performance. The ensemble method was expected to yield higher accuracy, but failed to improve upon the accuracy obtained from the SVM method alone.

5.1 Nine-site classification with clade and functional abundance

The principal object of this thesis is to develop new, biologically informed strategies for microbial community classification, with the oral microbiome as a test case, so our final analysis was to apply the successful modifications demonstrated in Chapter 4 to the 9-way classification problem first explored in Chapter 3. Although some of the modifications are useful for the binary classification of hard plaque samples, the whole sample set offers several other challenging cases. Most notably, classification accuracy within the "back of the mouth" group, which comprised samples from the throat, palatine tonsils and tongue dorsum, was very poor: over 50% of throat and palatine tonsil samples were incorrectly assigned to other sites by the classifier.

Since none of the four custom kernels yielded better performance, we focused on using different abundance measures on all nine oral cavity sub-sites. For clade abundance, we used the same approach as described in Chapter 4. Ninety-one OTUs were eliminated because of the PyNast quality control filters, yielding a total of 12,754

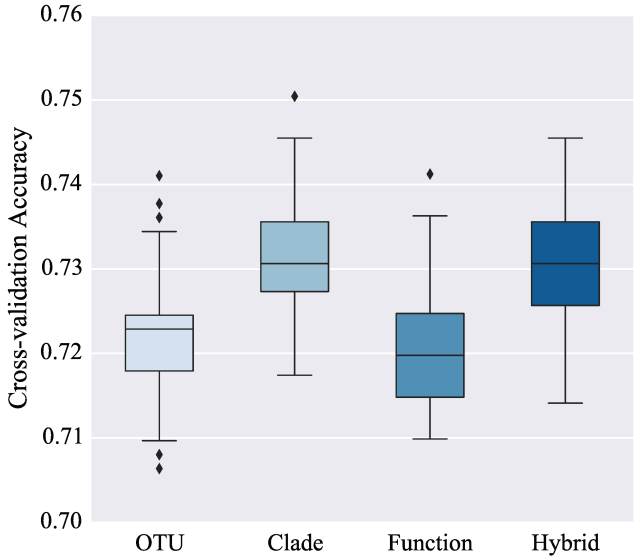


Figure 5.1: **Boxplots show the distribution of 100 times cross-validation accuracies with different input features.**

OTUs. Because of its performance in classifying hard plaque samples, RF feature permutation was employed to select the important features. Classifiers with 10 to 200 features were trained as a pre-experiment, however, these classifiers performed poorly, probably since the number of selected features was insufficient to distinguish all nine classes. So we increased the feature dimension, from 100 to 2,000 at an increasing step of 100. The accuracy of the default OTU abundance approach was similar to that carried out in Chapter 3, varying between 71% and 72% with different numbers of

Table 5.1: **Statistical summary of the accuracies from each of ten cross-validation folds with different features.**

Features	*Mean*	*Std.Deviation*	*Std.Error*	*95% Confidence Intervals*		*Min*	*Max*
				Lower	*Upper*		
OTU	0.721	0.024	0.001	0.671	0.755	0.665	0.760
Clade	0.730	0.025	0.001	0.685	0.763	0.677	0.765
Function	0.716	0.023	0.001	0.687	0.763	0.687	0.769
Hybrid	0.731	0.027	0.001	0.696	0.781	0.692	0.790

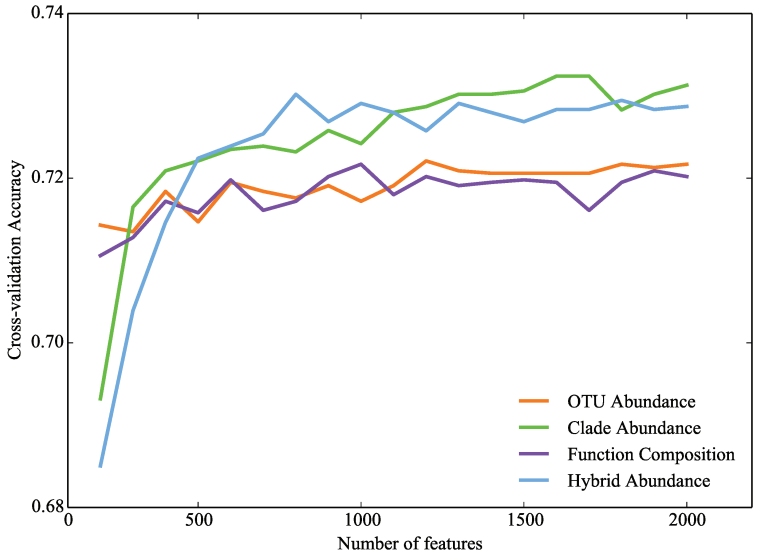


Figure 5.2: **Nine-way classification accuracy with different sets of input features.** The classification accuracy is shown for sets of 100 to 2,000 of the top-ranked features according to RF feature permutation criteria. The four types of input features used were OTUs only (orange markers); OTUs and clades (green markers); Functional predictions made using PICRUSt (purple markers); and all generated features (blue markers).

features (in Figure 5.1, Figure 5.2). However, clade abundance gave accuracy scores that were often in excess of 73%. A ten-fold cross-validation was also performed and the statistical descriptions of each fold's accuracy and the overall dispersion are displayed in Table 5.1. Figure 5.3 compares the performance of OTU abundance and clade abundance in confusion matrices. Four natural groups can still be found in clade confusion matrix (in Figure 5.3(b)), especially the dental plaque group. Seven of the nine sites improved predictions with clade abundance as features. Accuracies increased from 1.0% (subgingival plaque) to 6.0% (hard palate). Although accuracy of supragingival plaque and tongue dorsum dropped, both their the decrement was only 0.6%.

Functional and hybrid clade-function abundance measures were also considered. A total of 6,348 KEGG orthologs were predicted by PICRUSt. The cross-validated accuracies of functional profiles was around 72% with different numbers of features, essentially the same as the accuracy obtained with the OTU abundance model. However, functional abundance required fewer features for optimal classification than did

the measures of OTU and clade abundance, reaching at least 70% accuracy with fewer than 100 features. Since the predictions made by PICRUSt were based on the functional similarities between numerous OTUs that belong to different high-level taxonomic groups, a small number of functional features may contain enough information to distinguish some communities on the oral cavity.

The accuracy obtained from the hybrid clade-function abundance encoding is consistent with the classification on hard plaque samples, obtaining accuracy around 73%. The performance of hybrid features and clade features had very similar performance, indicating that functional features failed to provide much complementary information to taxonomy in the 9-way classification as well.

5.2 Summary and Conclusion

The microbiome of human oral cavity is associated with both health and disease. The analysis of HMP dataset provided a detailed view of the healthy oral cavity microbiome, which can serve as a baseline for further studies that consider variation in disease states and therapeutic responses [7, 25, 15, 27]. A primary objective of machine learning is to train models that can distinguish classes of entities, in this case microbial samples encoded as OTU tables, with high accuracy. Previous authors have tested many different machine-learning algorithms on reference data sets [41, 172, 43, 42]; our principal focus here on SVMs allowed us to consider different encodings of the input data. The SVM is well-suited to our microbial classification problems. It is a robust algorithm that can deal with high-dimensional and mutually-dependent features, such as clade and functional abundance. Various kinds of kernels have been developed to solve different real-world problems [88]; biologically inspired kernels were considered in the thesis. The performances of models were evaluated via prediction accuracy and displayed in confusion matrix. Results were also described and tested with statistical methods. Because of the limited number of accuracies in the 10-fold cross-validation, the statistical power may not be adequate. We repeated the classification for 100 times and used two-sample t-test to examine the difference between models.

A preliminary classification attempt was performed on the 9-way oral cavity samples. By representing the results to a confusion matrix, we identified four distinct

		Predicted Class								
		GING	**BUCC**	**HPAL**	**SAL**	**SUB**	**SUPRA**	**PTON**	**THRO**	**TONG**
Actual Class	**GING**	**0.826**	0.120	0.010	0.000	0.013	0.000	0.003	0.010	0.000
	BUCC	0.145	**0.748**	0.087	0.007	0.013	0.013	0.063	0.043	0.000
	HPAL	0.013	0.083	**0.697**	0.011	0.000	0.003	0.059	0.163	0.020
	SAL	0.000	0.007	0.020	**0.879**	0.007	0.003	0.026	0.017	0.010
	SUB	0.007	0.010	0.007	0.004	**0.631**	0.151	0.016	0.007	0.000
	SUPRA	0.007	0.007	0.003	0.007	0.326	**0.826**	0.007	0.007	0.000
	PTON	0.003	0.010	0.027	0.057	0.003	0.003	**0.477**	0.150	0.036
	THRO	0.000	0.010	0.107	0.028	0.007	0.000	0.191	**0.439**	0.049
	TONG	0.000	0.007	0.043	0.007	0.000	0.000	0.158	0.166	**0.885**

(a) Confusion matrix with OTU features

		Predicted Class								
		GING	**BUCC**	**HPAL**	**SAL**	**SUB**	**SUPRA**	**PTON**	**THRO**	**TONG**
Actual Class	**GING**	**0.859**	0.105	0.013	0.003	0.010	0.003	0.003	0.003	0.000
	BUCC	0.126	**0.771**	0.053	0.007	0.003	0.007	0.013	0.020	0.000
	HPAL	0.003	0.073	**0.757**	0.017	0.000	0.007	0.013	0.097	0.033
	SAL	0.004	0.007	0.004	**0.897**	0.000	0.004	0.039	0.036	0.011
	SUB	0.007	0.013	0.000	0.010	**0.641**	0.302	0.020	0.007	0.000
	SUPRA	0.003	0.007	0.007	0.003	0.157	**0.820**	0.003	0.000	0.000
	PTON	0.007	0.069	0.063	0.016	0.016	0.003	**0.490**	0.207	0.128
	THRO	0.007	0.027	0.176	0.030	0.000	0.003	0.143	**0.482**	0.133
	TONG	0.000	0.003	0.016	0.013	0.003	0.000	0.052	0.033	**0.879**

(b) Confusion matrix with clade features

Figure 5.3: **Confusion matrix of nine-way oral site classification with feature selection.** Results of classifiers with different features (a) OTU abundance and (b) clade abundance are shown.

groups of body sites: gums, saliva, hard plaque and back of mouth. These patterns were also confirmed by mapping the samples to a 2D PCoA plot, which showed a clearer distinction among sites after recoloring the points. We focused on samples from hard plaque to develop new approaches: the phylogenetic relationships and functions provided more information about the similarity of communities than OTU abundance. Therefore, our improvements on the microbial classification problem were based on the following aspects:

- *Phylogenetic distance.* Four different custom similarity kernels were developed based on ecological beta-diversity measures, with and without phylogenetic distances. However, the performance of the classifiers did not improve with the custom kernels, possibly because the RBF kernel has the optimization of parameters in the SVM grid search, whereas the custom kernels have no such optimization process.

- *Phylogenetic relationships.* Closely related OTUs were grouped into clades based on their relationships in a phylogenetic tree. Of the modifications we tried, clade-based representations gave the largest increase in performance. Although the combinations of OTUs that constituted clades could in principle be discovered by the classifier, it is clear that explicit clade representations yielded some advantage in both feature selection and classification. Selected clades contained genera known to be important in the human oral cavity, in particular *Streptococcus*.

- *Function.* Our predictive approach to function did not improve the accuracy of our classifiers, in spite of the potential for PICRUSt to identify functional as well as phylogenetic connections between OTUs and clades. It may be that shotgun metagenome sequencing, which generates accurate information about even those genes that are frequently transferred, may yield higher predictive accuracy.

In the case of oral samples, and hard plaque samples in particular, complete separation (i.e., 100% classification accuracy) may not be achievable, for several reasons. Chief amongst these is the physical proximity of the supragingival and subgingival

plaque. Although the two sites are different in terms of nutrient and oxygen availability, the formation of plaque indicates that the migration of microorganisms from supragingival plaque to subgingival plaque does exist. We can infer that communities on these two sites are highly overlapped, so sample misidentification may appear, which also contribute to diminished classification; indeed this was one motivation for the development of SourceTracker. However, we expect misidentified samples will have a minimal impact on classification, for two reasons: first, the HMP followed very strict protocols regarding the collection and handling of samples; second, the overlapping of sample types we see in Figure 3.6(a) suggests a gradient of diversity from one sample type to others, rather than a few scattered outliers that might be indicative of misclassified samples. It is also unlikely that there is a single type of healthy subgingival and supragingival microbial community, which would impede the ability of a classifier to learn a single, general model of classification.

5.3 Future Work

Although complete separation is not achievable on oral samples, there still may exist room for improvement. Improvement in microbial classification is more than seeking higher accuracy; identifying discriminative OTUs or functions is also an important step towards understanding key processes in the microbiome. Based on the work done in the thesis, several promising approaches can be explored:

- Use concordance of classifiers may give better prediction accuracy on a subset of the data. Previous work suggests that a different choice of classifier may yield higher classification accuracy; clearly further work is needed to explore this question, and there is a multitude of different approaches that can be applied to the data. Chapter 4 uncovered the inconsistency in predictions from different methods and built an ensemble classifier combing SVM, Random Forest and SourceTracker. No improvements were seen, which may reflect either a lack of complementary information from the combination of classifiers, or insufficient optimization of our ensemble approach. So future work trying to build a well-designed ensemble classifier is strongly recommended.

- Changing the definition and inference of OTUs may improve performance as

well: in particular, changing the OTU threshold from 97% to 99% would highlight finer-scale differences in abundance, for example, differences that may manifest only at or below the species level. In this work we used closed-reference OTU picking because it maps sampled sequences to reference groups that are defined prior to the analysis. However, closed-reference picking discards any sequences that do not map to existing OTUs at the required level of sequence similarity, a phenomenon that is especially acute at higher thresholds such as 99%. An approach that combines closed-reference and *de novo* OTU generation would likely be ideal, but requires that new OTUs be comparable between samples and across studies.

- Other measurements of microbial community structure could serve as input features as well. In additional to taxonomic abundance, functional profiles were also constructed in Chapter 4. However, it did not obtain as much improvement as clade abundance, possibly in part because the functions were predicted from 16S rRNA gene samples rather than sampled directly from the community. Future work could attempt to identify key functions from shotgun sequencing, which generates direct estimates of functional abundance from the community without the prediction step of PICRUSt. Features constructed from other measurements, such as divergence-based methods described in [38] may reveal different characteristics of the microbial communities, since feature selection did find out features of biological meaning.

Bibliography

[1] Antonio Gonzalez, Jesse Stombaugh, Christian L Lauber, Noah Fierer, and Rob Knight. Sitepainter: a tool for exploring biogeographical patterns. *Bioinformatics*, 28(3):436–438, 2012.

[2] Pedro Domingos. A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87, 2012.

[3] Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):38, 2005.

[4] Jordan Michael. Advanced topics in learning and decision making, January 2004.

[5] Donovan H Parks and Robert G Beiko. Measures of phylogenetic differentiation provide robust and complementary insights into microbial communities. *The ISME journal*, 7(1):173–183, 2013.

[6] Xochitl C Morgan and Curtis Huttenhower. Chapter 12: human microbiome analysis. *PLoS Comput Biol*, 8(12):e1002808, 2012.

[7] Human Microbiome Project Consortium et al. Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–214, 2012.

[8] Dwayne C Savage. Microbial ecology of the gastrointestinal tract. *Annual Reviews in Microbiology*, 31(1):107–133, 1977.

[9] Rodney D Berg. The indigenous gastrointestinal microflora. *Trends in microbiology*, 4(11):430–435, 1996.

[10] Eva Bianconi, Allison Piovesan, Federica Facchin, Alina Beraudi, Raffaella Casadei, Flavia Frabetti, Lorenza Vitale, Maria Chiara Pelleri, Simone Tassani, Francesco Piva, et al. An estimation of the number of cells in the human body. *Annals of human biology*, 40(6):463–471, 2013.

[11] Ilseung Cho and Martin J Blaser. The human microbiome: at the interface of health and disease. *Nature Reviews Genetics*, 13(4):260–270, 2012.

[12] Vaia Galimanas, Michael W Hall, Natasha Singh, Michael David J Lynch, Michael Goldberg, Howard Tenenbaum, Dennis G Cvitkovitch, Josh D Neufeld, and Dilani B Senadheera. Bacterial community composition of chronic periodontitis and novel oral sampling sites for detecting disease indicators. *Microbiome*, 2(1):32, 2014.

[13] Peter J Turnbaugh, Micah Hamady, Tanya Yatsunenko, Brandi L Cantarel, Alexis Duncan, Ruth E Ley, Mitchell L Sogin, William J Jones, Bruce A Roe, Jason P Affourtit, et al. A core gut microbiome in obese and lean twins. *nature*, 457(7228):480–484, 2009.

[14] Brian L Schmidt, Justin Kuczynski, Aditi Bhattacharya, Bing Huey, Patricia M Corby, Erica LS Queiroz, Kira Nightingale, A Ross Kerr, Mark D DeLacure, Ratna Veeramachaneni, et al. Changes in abundance of oral microbiota associated with oral cancer. 2014.

[15] William G Wade. The oral microbiome in health and disease. *Pharmacological Research*, 69(1):137–143, 2013.

[16] Elizabeth A Grice, Heidi H Kong, Sean Conlan, Clayton B Deming, Joie Davis, Alice C Young, Gerard G Bouffard, Robert W Blakesley, Patrick R Murray, Eric D Green, et al. Topographical and temporal diversity of the human skin microbiome. *science*, 324(5931):1190–1192, 2009.

[17] Andrew L Kau, Philip P Ahern, Nicholas W Griffin, Andrew L Goodman, and Jeffrey I Gordon. Human nutrition, the gut microbiome and the immune system. *Nature*, 474(7351):327–336, 2011.

[18] Tanja Jaeggi, Guus AM Kortman, Diego Moretti, Christophe Chassard, Penny Holding, Alexandra Dostal, Jos Boekhorst, Harro M Timmerman, Dorine W Swinkels, Harold Tjalsma, et al. Iron fortification adversely affects the gut microbiome, increases pathogen abundance and induces intestinal inflammation in kenyan infants. *Gut*, 64(5):731–742, 2015.

[19] Michael B Zimmermann, Christophe Chassard, Fabian Rohner, Eliézer K N'goran, Charlemagne Nindjin, Alexandra Dostal, Jürg Utzinger, Hala Ghattas, Christophe Lacroix, and Richard F Hurrell. The effects of iron fortification on the gut microbiota in african children: a randomized controlled trial in cote d'ivoire. *The American journal of clinical nutrition*, 92(6):1406–1415, 2010.

[20] Barry Halliwell, John Gutteridge, and Carroll E Cross. Free radicals, antioxidants, and human disease: where are we now? *The Journal of laboratory and clinical medicine*, 119(6):598–620, 1992.

[21] Daniel K Podolsky. Inflammatory bowel disease. *New England Journal of Medicine*, 325(13):928–937, 1991.

[22] Noora Ottman, Hauke Smidt, Willem M De Vos, and Clara Belzer. The function of our microbiota: who is out there and what do they do? *Frontiers in cellular and infection microbiology*, 2, 2012.

[23] Bruce L Pihlstrom, Bryan S Michalowicz, and Newell W Johnson. Periodontal diseases. *The Lancet*, 366(9499):1809–1820, 2005.

[24] MF Zarco, TJ Vess, and GS Ginsburg. The oral microbiome in health and disease and the potential impact on personalized dental medicine. *Oral diseases*, 18(2):109–120, 2012.

[25] Jinzhi He, Yan Li, Yangpei Cao, Jin Xue, and Xuedong Zhou. The oral microbiome diversity and its relation to human diseases. *Folia microbiologica*, 60(1):69–80, 2015.

[26] T Daniluk, G Tokajuk, D Cylwik-Rokicka, D Rozkiewicz, ML Zaremba, and W Stokowska. Aerobic and anaerobic bacteria in subgingival and supragingival plaques of adult patients with periodontal disease. *Advances in medical sciences*, 51:81–85, 2005.

[27] Laurie Ann Ximénez-Fyvie, Anne D Haffajee, and Sigmund S Socransky. Comparison of the microbiota of supra-and subgingival plaque in health and periodontitis. *Journal of clinical periodontology*, 27(9):648–657, 2000.

[28] Elizabeth A Grice and Julia A Segre. The skin microbiome. *Nature Reviews Microbiology*, 9(4):244–253, 2011.

[29] Elizabeth K Costello, Christian L Lauber, Micah Hamady, Noah Fierer, Jeffrey I Gordon, and Rob Knight. Bacterial community variation in human body habitats across space and time. *Science*, 326(5960):1694–1697, 2009.

[30] Sorel Fitz-Gibbon, Shuta Tomida, Bor-Han Chiu, Lin Nguyen, Christine Du, Minghsun Liu, David Elashoff, Marie C Erfe, Anya Loncaric, Jenny Kim, et al. Propionibacterium acnes strain populations in the human skin microbiome associated with acne. *Journal of Investigative Dermatology*, 133(9):2152–2160, 2013.

[31] Alexander V Alekseyenko, Guillermo I Perez-Perez, Aieska De Souza, Bruce Strober, Zhan Gao, Monika Bihan, Kelvin Li, Barbara A Methé, and Martin J Blaser. Community differentiation of the cutaneous microbiota in psoriasis. *Microbiome*, 1(1):31, 2013.

[32] Nicola Segata, Susan Kinder Haake, Peter Mannon, Katherine P Lemon, Levi Waldron, Dirk Gevers, Curtis Huttenhower, and Jacques Izard. Composition of the adult digestive tract bacterial microbiome based on seven mouth surfaces, tonsils, throat and stool samples. *Genome Biology*, 13(6):R42, 2012.

[33] Karoline Faust, J Fah Sathirapongsasuti, Jacques Izard, Nicola Segata, Dirk Gevers, Jeroen Raes, and Curtis Huttenhower. Microbial co-occurrence relationships in the human microbiome. *PLoS Compute Biology*, 8(7):e1002606–e1002606, 2012.

[34] Frank A Scannapieco, Renee B Bush, and Susanna Paju. Associations between periodontal disease and risk for atherosclerosis, cardiovascular disease, and stroke. a systematic review. *Annals of Periodontology*, 8(1):38–53, 2003.

[35] Jill E Clarridge, Silvia Attorri, Daniel M Musher, Jeff Hebert, and Sherry Dunbar. Streptococcus intermedius, streptococcus constellatus, and streptococcus anginosus (streptococcus milleri group) are of different clinical importance and are not equally associated with abscess. *Clinical infectious diseases*, 32(10):1511–1515, 2001.

[36] Rudolf I Amann, Wolfgang Ludwig, and Karl-Heinz Schleifer. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiological reviews*, 59(1):143–169, 1995.

[37] IM Head, JR Saunders, and RW kup. Microbial evolution, diversity, and ecology: a decade of ribosomal rna analysis of uncultivated microorganisms. *Microbial ecology*, 35(1):1–21, 1998.

[38] Catherine A Lozupone and Rob Knight. Species divergence and the measurement of microbial diversity. *FEMS microbiology reviews*, 32(4):557–578, 2008.

[39] Jay Shendure and Hanlee Ji. Next-generation dna sequencing. *Nature biotechnology*, 26(10):1135–1145, 2008.

[40] Martin Kircher and Janet Kelso. High-throughput dna sequencing–concepts and limitations. *Bioessays*, 32(6):524–536, 2010.

[41] Dan Knights, Elizabeth K Costello, and Rob Knight. Supervised classification of human microbiota. *FEMS microbiology reviews*, 35(2):343–359, 2011.

[42] Zhenqiu Liu, William Hsiao, Brandi L Cantarel, Elliott Franco Drábek, and Claire Fraser-Liggett. Sparse distance-based learning for simultaneous multiclass classification and feature selection of metagenomic data. *Bioinformatics*, 27(23):3242–3249, 2011.

[43] Alexander Statnikov, Mikael Henaff, Varun Narendra, Kranti Konganti, Zhiguo Li, Liying Yang, Zhiheng Pei, Martin J Blaser, Constantin F Aliferis, and Alexander V Alekseyenko. A comprehensive evaluation of multicategory classification methods for microbiomic data. *Microbiome*, 1(1):11, 2013.

[44] Olga Tanaseichuk, James Borneman, and Tao Jiang. Phylogeny-based classification of microbial communities. *Bioinformatics*, page btt700, 2013.

[45] Joshua Lederberg and Alexa Mccray. The scientist:'ome sweet'omics–a genealogical treasury of words. *The Scientist*, 17(7), 2001.

[46] Leónides Fernández, Susana Langa, Virginia Martín, Antonio Maldonado, Esther Jiménez, Rocío Martín, and Juan M Rodríguez. The human milk microbiota: origin and potential roles in health and disease. *Pharmacological Research*, 69(1):1–10, 2013.

[47] Daniel C Baumgart and William J Sandborn. Crohn's disease. *The Lancet*, 380(9853):1590–1605, 2012.

[48] Stanislas Mondot, S Kang, J-Pierre Furet, Daniel Aguirre de Cárcer, C McSweeney, M Morrison, P Marteau, Joel Dore, and Marion Leclerc. Highlighting new phylogenetic specificities of crohn's disease microbiota. *Inflammatory bowel diseases*, 17(1):185–192, 2011.

[49] ErnestD Gray, Marjorie Verstegen, Georg Peters, and WarrenE Regelmann. Effect of extracellular slime substance from staphylococcus epidermidis on the human cellular immune response. *The Lancet*, 323(8373):365–367, 1984.

[50] Holger Rohde, Christoph Burdelski, Katrin Bartscht, Muzaffar Hussain, Friedrich Buck, Matthias A Horstkotte, Johannes K-M Knobloch, Christine Heilmann, Mathias Herrmann, and Dietrich Mack. Induction of staphylococcus epidermidis biofilm formation via proteolytic processing of the accumulation-associated protein by staphylococcal and host proteases. *Molecular microbiology*, 55(6):1883–1895, 2005.

[51] DR Haynes. Bone lysis and inflammation. *Inflammation research*, 53(11):596–600, 2004.

[52] Floyd E Dewhirst, Tuste Chen, Jacques Izard, Bruce J Paster, Anne CR Tanner, Wen-Han Yu, Abirami Lakshmanan, and William G Wade. The human oral microbiome. *Journal of bacteriology*, 192(19):5002–5017, 2010.

[53] Burton Rosan and Richard J Lamont. Dental plaque formation. *Microbes and infection*, 2(13):1599–1607, 2000.

[54] PD Marsh. Sugar, fluoride, ph and microbial homeostasis in dental plaque. *Proceedings of the Finnish Dental Society. Suomen Hammaslaakariseuran toimituksia*, 87(4):515–525, 1990.

[55] DA Spratt and J Pratten. Biofilms and the oral cavity. *Reviews in environmental science and biotechnology*, 2(2-4):109–120, 2003.

[56] Eva Boon, Conor J Meehan, Chris Whidden, Dennis H-J Wong, Morgan GI Langille, and Robert G Beiko. Interactions in the microbiome: communities of organisms and communities of genes. *FEMS microbiology reviews*, 38(1):90–118, 2014.

[57] James I Prosser, Brendan JM Bohannan, Tom P Curtis, Richard J Ellis, Mary K Firestone, Rob P Freckleton, Jessica L Green, Laura E Green, Ken Killham, Jack J Lennon, et al. The role of ecological theory in microbial ecology. *Nature Reviews Microbiology*, 5(5):384–392, 2007.

[58] Jennifer B Hughes Martiny, Brendan JM Bohannan, James H Brown, Robert K Colwell, Jed A Fuhrman, Jessica L Green, M Claire Horner-Devine, Matthew Kane, Jennifer Adams Krumins, Cheryl R Kuske, et al. Microbial biogeography: putting microorganisms on the map. *Nature Reviews Microbiology*, 4(2):102–112, 2006.

[59] Yanjiao Zhou, Hongyu Gao, Kathie A Mihindukulasuriya, Patricio S La Rosa, Kristine M Wylie, Tatiana Vishnivetskaya, Mircea Podar, Barb Warner, Phillip I Tarr, David E Nelson, et al. Biogeography of the ecosystems of the healthy human body. *Genome Biol*, 14(1):R1, 2013.

[60] Heidi H Kong and Julia A Segre. Skin microbiome: looking back to move forward. *Journal of Investigative Dermatology*, 132:933–939, 2012.

[61] Patrick D Schloss. Microbiology: An integrated view of the skin microbiome. *Nature*, 514(7520):44–45, 2014.

[62] Robert S Breed and HJ Conn. The status of the generic term bacterium ehrenberg 1828. *Journal of bacteriology*, 31(5):517, 1936.

[63] Jo Handelsman. Metagenomics: application of genomics to uncultured microorganisms. *Microbiology and molecular biology reviews*, 68(4):669–685, 2004.

[64] Aharon Oren and George M Garrity. Then and now: a systematic review of the systematics of prokaryotes in the last 80 years. *Antonie Van Leeuwenhoek*, 106(1):43–56, 2014.

[65] Emile Zuckerkandl and Linus Pauling. Molecules as documents of evolutionary history. *Journal of theoretical biology*, 8(2):357–366, 1965.

[66] Carl R Woese and George E Fox. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences*, 74(11):5088–5090, 1977.

[67] James R Brown, Christophe J Douady, Michael J Italia, William E Marshall, and Michael J Stanhope. Universal trees based on large combined protein sequence data sets. *Nature genetics*, 28(3):281–285, 2001.

[68] M Faveri, MPA Mayer, M Feres, LC De Figueiredo, FE Dewhirst, and BJ Paster. Microbiological diversity of generalized aggressive periodontitis by 16s rrna clonal analysis. *Oral microbiology and immunology*, 23(2):112–118, 2008.

[69] Justin Kuczynski, Christian L Lauber, William A Walters, Laura Wegener Parfrey, José C Clemente, Dirk Gevers, and Rob Knight. Experimental and analytical tools for studying the human microbiome. *Nature Reviews Genetics*, 13(1):47–58, 2012.

[70] C Gyles and P Boerlin. Horizontally transferred genetic elements and their role in pathogenesis of bacterial disease. *Veterinary Pathology Online*, page 0300985813511131, 2013.

[71] Daniel McDonald, Morgan N Price, Julia Goodrich, Eric P Nawrocki, Todd Z DeSantis, Alexander Probst, Gary L Andersen, Rob Knight, and Philip Hugenholtz. An improved greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *The ISME journal*, 6(3):610–618, 2012.

[72] James R Cole, Qiong Wang, Jordan A Fish, Benli Chai, Donna M McGarrell, Yanni Sun, C Titus Brown, Andrea Porras-Alfaro, Cheryl R Kuske, and James M Tiedje. Ribosomal database project: data and tools for high throughput rrna analysis. *Nucleic acids research*, page gkt1244, 2013.

[73] Christian Quast, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. The silva ribosomal rna gene database project: improved data processing and web-based tools. *Nucleic acids research*, page gks1219, 2012.

[74] Todd Z DeSantis, Philip Hugenholtz, Neils Larsen, Mark Rojas, Eoin L Brodie, Keith Keller, Thomas Huber, Daniel Dalevi, Ping Hu, and Gary L Andersen. Greengenes, a chimera-checked 16s rrna gene database and workbench compatible with arb. *Applied and environmental microbiology*, 72(7):5069–5072, 2006.

[75] R Staden. A strategy of dna sequencing employing computer programs. *Nucleic acids research*, 6(7):2601–2610, 1979.

[76] Stephen Anderson. Shotgun dna sequencing using cloned dnase i-generated fragments. *Nucleic Acids Research*, 9(13):3015–3027, 1981.

[77] Jane Peterson, Susan Garges, Maria Giovanni, Pamela McInnes, Lu Wang, Jeffery A Schloss, Vivien Bonazzi, Jean E McEwen, Kris A Wetterstrand, Carolyn Deal, et al. The nih human microbiome project. *Genome research*, 19(12):2317–2323, 2009.

[78] Peter J Turnbaugh, Ruth E Ley, Micah Hamady, Claire Fraser-Liggett, Rob Knight, and Jeffrey I Gordon. The human microbiome project: exploring the microbial part of ourselves in a changing world. *Nature*, 449(7164):804, 2007.

[79] David E Goldberg and John H Holland. Genetic algorithms and machine learning. *Machine learning*, 3(2):95–99, 1988.

[80] Donald Michie, David J Spiegelhalter, and Charles C Taylor. Machine learning, neural and statistical classification. 1994.

[81] SB Kotsiantis. Supervised machine learning: A review of classification techniques. *Informatica*, 31:249–268, 2007.

[82] Yonghong Peng. A novel ensemble machine learning for robust microarray data classification. *Computers in Biology and Medicine*, 36(6):553–573, 2006.

[83] Zafer Barutcuoglu, Robert E Schapire, and Olga G Troyanskaya. Hierarchical multi-label prediction of gene function. *Bioinformatics*, 22(7):830–836, 2006.

[84] Jean-Luc Bouchot, William L Trimble, Gregory Ditzler, Yemin Lan, Steve Essinger, and Gail Rosen. Advances in machine learning for processing and comparison of metagenomic data. *Computational Systems Biology: From Molecular Mechanisms to Disease: Second Edition.: Elsevier Inc*, pages 295–329, 2013.

[85] Wei Yu, Tiebin Liu, Rodolfo Valdez, Marta Gwinn, and Muin J Khoury. Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. *BMC Medical Informatics and Decision Making*, 10(1):16, 2010.

[86] Joseph N Paulson, O Colin Stine, Héctor Corrada Bravo, and Mihai Pop. Differential abundance analysis for microbial marker-gene surveys. *Nature methods*, 10(12):1200–1202, 2013.

[87] Pierre Baldi, Søren Brunak, Yves Chauvin, Claus AF Andersen, and Henrik Nielsen. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16(5):412–424, 2000.

[88] Lynne Davis, John Hawkins, Stefan Maetschke, and Mikael Bodén. Comparing svm sequence kernels: A protein subcellular localization theme. In *Proceedings of the 2006 workshop on Intelligent systems for bioinformatics-Volume 73*, pages 39–47. Australian Computer Society, Inc., 2006.

[89] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47, 2002.

[90] Mario Köppen. The curse of dimensionality. In *5th Online World Conference on Soft Computing in Industrial Applications (WSC5)*, pages 4–8, 2000.

[91] John Quackenbush. Microarray data normalization and transformation. *Nature genetics*, 32:496–501, 2002.

[92] Eric P Xing, Michael I Jordan, Richard M Karp, et al. Feature selection for high-dimensional genomic microarray data. In *ICML*, volume 1, pages 601–608. Citeseer, 2001.

[93] Y Wang, DJ Miller, and R Clarke. Approaches to working in high-dimensional data spaces: gene expression microarrays. *British journal of cancer*, 98(6):1023–1028, 2008.

[94] Rich Caruana, Nikos Karampatziakis, and Ainur Yessenalina. An empirical evaluation of supervised learning in high dimensions. In *Proceedings of the 25th international conference on Machine learning*, pages 96–103. ACM, 2008.

[95] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[96] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[97] Yiming Yang and Jan O Pedersen. A comparative study on feature selection in text categorization. In *ICML*, volume 97, pages 412–420, 1997.

[98] André Altmann, Laura Toloşi, Oliver Sander, and Thomas Lengauer. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347, 2010.

[99] Makio Tamura and Patrik D'haeseleer. Microbial genotype–phenotype mapping by class association rule mining. *Bioinformatics*, 24(13):1523–1529, 2008.

[100] Gary Miner, Robert Nisbet, and John Elder IV. *Handbook of statistical analysis and data mining applications*. Academic Press, 2009.

[101] Shachar Kaufman, Saharon Rosset, Claudia Perlich, and Ori Stitelman. Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6(4):15, 2012.

[102] Ron Kohavi, Carla E Brodley, Brian Frasca, Llew Mason, and Zijian Zheng. Kdd-cup 2000 organizers' report: Peeling the onion. *ACM SIGKDD Explorations Newsletter*, 2(2):86–93, 2000.

[103] Rong She, Ke Wang, Yabo Xu, and S Yu Philip. Pushing feature selection ahead of join. *transfer*, 100:2, 2005.

[104] Pedro Domingos. A unified bias-variance decomposition. In *Proceedings of 17th International Conference on Machine Learning. Stanford CA Morgan Kaufmann*, pages 231–238, 2000.

[105] DS Moore and GP McCabe. Introduction to the practice of statistics. 1989.

[106] Mark JL Orr. Regularization in the selection of radial basis function centers. *Neural computation*, 7(3):606–623, 1995.

[107] Nuttachat Wisittipanit. *Machine learning approach for profiling human microbiome*. PhD thesis, GEORGE MASON UNIVERSITY, 2012.

[108] Yin Wang, Yuhua Zhou, Yixue Li, Zongxin Ling, Yan Zhu, Xiaokui Guo, and Hong Sun. An improved dimensionality reduction method for meta-transcriptome indexing based diseases classification. *BMC systems biology*, 6(Suppl 3):S12, 2012.

[109] Paul R Ehrlich and Richard W Holm. Patterns and populations basic problems of population biology transcend artificial disciplinary boundaries. *Science*, 137(3531):652–657, 1962.

[110] Robert R Sokal, Peter HA Sneath, et al. Principles of numerical taxonomy. *Principles of numerical taxonomy.*, 1963.

[111] Noah Fierer, Christian L Lauber, Nick Zhou, Daniel McDonald, Elizabeth K Costello, and Rob Knight. Forensic identification using skin bacterial communities. *Proceedings of the National Academy of Sciences*, 107(14):6477–6481, 2010.

[112] Yvan Saeys, Iñaki Inza, and Pedro Larrañaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517, 2007.

[113] Catherine Lozupone and Rob Knight. Unifrac: a new phylogenetic method for comparing microbial communities. *Applied and environmental microbiology*, 71(12):8228–8235, 2005.

[114] Catherine Lozupone, Micah Hamady, and Rob Knight. Unifrac–an online tool for comparing microbial community diversity in a phylogenetic context. *BMC bioinformatics*, 7(1):371, 2006.

[115] Catherine Lozupone, Manuel E Lladser, Dan Knights, Jesse Stombaugh, and Rob Knight. Unifrac: an effective distance metric for microbial community comparison. *The ISME journal*, 5(2):169, 2011.

[116] Morgan GI Langille, Jesse Zaneveld, J Gregory Caporaso, Daniel McDonald, Dan Knights, Joshua A Reyes, Jose C Clemente, Deron E Burkepile, Rebecca L Vega Thurber, Rob Knight, et al. Predictive functional profiling of microbial communities using 16s rrna marker gene sequences. *Nature biotechnology*, 31(9):814–821, 2013.

[117] Human Microbiome Project Consortium et al. Manual of procedures for human microbiome project core microbiome sampling protocol.

[118] Kjersti Aagaard, Joseph Petrosino, Wendy Keitel, Mark Watson, James Katancik, Nathalia Garcia, Shital Patel, Mary Cutting, Tessa Madden, Holli Hamilton, et al. The human microbiome project strategy for comprehensive sampling of the human microbiome and why it matters. *The FASEB Journal*, 27(3):1012–1022, 2013.

[119] Peter JA Cock, Christopher J Fields, Naohisa Goto, Michael L Heuer, and Peter M Rice. The sanger fastq file format for sequences with quality scores, and the solexa/illumina fastq variants. *Nucleic acids research*, 38(6):1767–1771, 2010.

[120] Brent Ewing and Phil Green. Base-calling of automated sequencer traces using phred. ii. error probabilities. *Genome research*, 8(3):186–194, 1998.

[121] Mark Achtman and Michael Wagner. Microbial diversity and the genetic nature of microbial species. *Nature Reviews Microbiology*, 6(6):431–440, 2008.

[122] Patrick D Schloss. The effects of alignment quality, distance calculation method, sequence filtering, and region on the analysis of 16s rrna gene-based studies. *PLoS Compute Biology*, 6(7):e1000844, 2010.

[123] Robert G Beiko, W Ford Doolittle, and Robert L Charlebois. The impact of reticulate evolution on genome phylogeny. *Systematic biology*, 57(6):844–856, 2008.

[124] Weilong Hao and Jeffrey Palmer. Hgt turbulence: Confounding phylogenetic influence of duplicative horizontal transfer and differential gene conversion. *Mobile genetic elements*, 1(4):256–304, 2011.

[125] DavidF Robinson and Leslie R Foulds. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1):131–147, 1981.

[126] Walter M Fitch, Emanuel Margoliash, et al. Construction of phylogenetic trees. *Science*, 155(3760):279–284, 1967.

[127] Pierre Legendre and Loic FJ Legendre. *Numerical ecology*, volume 24. Elsevier, 2012.

[128] Robert R Sokal. A statistical method for evaluating systematic relationships. *Univ Kans Sci Bull*, 38:1409–1438, 1958.

[129] Naruya Saitou and Masatoshi Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, 4(4):406–425, 1987.

[130] Manfred J Sippl. Biological sequence analysis. probabilistic models of proteins and nucleic acids, edited by r. durbin, s. eddy, a. krogh, and g. mitchinson. 1998. cambridge: Cambridge university press. 356 pp.£ 55.00 (80.00)(*hardcover*); £19.95(34.95)(paper). *Protein Science*, 8(3):695–695, 1999.

[131] Walter M Fitch. Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic zoology*, pages 406–416, 1971.

[132] Joseph Felsenstein. Evolutionary trees from dna sequences: a maximum likelihood approach. *Journal of molecular evolution*, 17(6):368–376, 1981.

[133] Morgan N Price, Paramvir S Dehal, and Adam P Arkin. Fasttree: computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular biology and evolution*, 26(7):1641–1650, 2009.

[134] G Booy, RJJ Hendriks, MJM Smulders, JM van Groenendael, and B Vosman. Genetic diversity and the survival of populations. *Plant biology*, 2(4):379–395, 2000.

[135] Robert Harding Whittaker. Vegetation of the siskiyou mountains, oregon and california. *Ecological monographs*, 30(3):279–338, 1960.

[136] Robert H Whittaker. Evolution and measurement of species diversity. *Taxon*, pages 213–251, 1972.

[137] Noah Fierer, Micah Hamady, Christian L Lauber, and Rob Knight. The influence of sex, handedness, and washing on the diversity of hand surface bacteria. *Proceedings of the National Academy of Sciences*, 105(46):17994–17999, 2008.

[138] Anne E Magurran. *Measuring biological diversity*. John Wiley & Sons, 2013.

[139] Daniel Simberloff. Properties of the rarefaction diversity measurement. *American Naturalist*, pages 414–418, 1972.

[140] James H Bullard, Elizabeth Purdom, Kasper D Hansen, and Sandrine Dudoit. Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments. *BMC bioinformatics*, 11(1):94, 2010.

[141] Marie-Agnès Dillies, Andrea Rau, Julie Aubert, Christelle Hennequet-Antier, Marine Jeanmougin, Nicolas Servant, Céline Keime, Guillemette Marot, David Castel, Jordi Estelle, et al. A comprehensive evaluation of normalization methods for illumina high-throughput rna sequencing data analysis. *Briefings in bioinformatics*, 14(6):671–683, 2013.

[142] Kantilal Varichand Mardia, John T Kent, and John M Bibby. *Multivariate analysis*. Academic press, 1979.

[143] Benjamin S Duran and Patrick L Odell. *Cluster analysis: a survey*, volume 100. Springer Science & Business Media, 2013.

[144] John C Gower. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53(3-4):325–338, 1966.

[145] Ian Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.

[146] John W Tukey. Exploratory data analysis. *age*, 1(68):2–70.

[147] J Gregory Caporaso, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D Bushman, Elizabeth K Costello, Noah Fierer, Antonio Gonzalez Pena, Julia K Goodrich, Jeffrey I Gordon, et al. Qiime allows analysis of high-throughput community sequencing data. *Nature methods*, 7(5):335–336, 2010.

[148] Robert C Edgar. Search and clustering orders of magnitude faster than blast. *Bioinformatics*, 26(19):2460–2461, 2010.

[149] J Gregory Caporaso, Kyle Bittinger, Frederic D Bushman, Todd Z DeSantis, Gary L Andersen, and Rob Knight. Pynast: a flexible tool for aligning sequences to a template alignment. *Bioinformatics*, 26(2):266–267, 2010.

[150] Jaime Huerta-Cepas, Joaquin Dopazo, and Toni Gabaldón. Ete: a python environment for tree exploration. *BMC bioinformatics*, 11(1):24, 2010.

[151] Chih-Wei Hsu, Chih-Chung Chang, Chih-Jen Lin, et al. A practical guide to support vector classification, 2003.

[152] Vladimir N Vapnik. An overview of statistical learning theory. *Neural Networks, IEEE Transactions on*, 10(5):988–999, 1999.

[153] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM, 1992.

[154] Chih-Wei Hsu and Chih-Jen Lin. A comparison of methods for multiclass support vector machines. *Neural Networks, IEEE Transactions on*, 13(2):415–425, 2002.

[155] Erich Leo Lehmann and George Casella. *Theory of point estimation*, volume 31. Springer Science & Business Media, 1998.

[156] Brian W Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451, 1975.

[157] Chih-Chung Chang and Chih-Jen Lin. Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[158] GN Lance and WT Williams. Computer programs for hierarchical polythetic classification (similarity analyses). *The Computer Journal*, 9(1):60–64, 1966.

[159] Godfrey N Lance and William T Williams. Mixed-data classificatory programs i - agglomerative systems. *Australian Computer Journal*, 1(1):15–20, 1967.

[160] Justin Kuczynski, Zongzhi Liu, Catherine Lozupone, Daniel McDonald, Noah Fierer, and Rob Knight. Microbial community resemblance methods differ in their ability to detect biologically relevant patterns. *Nature methods*, 7(10):813–819, 2010.

[161] Syed Masum Emran and Nong Ye. Robustness of chi-square and canberra distance metrics for computer intrusion detection. *Quality and Reliability Engineering International*, 18(1):19–28, 2002.

[162] Qin Chang, Yihui Luan, and Fengzhu Sun. Variance adjusted weighted unifrac: a powerful beta diversity measure for comparing communities based on phylogeny. *BMC bioinformatics*, 12(1):118, 2011.

[163] Julian Huxley. The three types of evolutionary process. *Nature*, 180:454–455, 1957.

[164] Frederick A Matsen, Robin B Kodner, and E Virginia Armbrust. pplacer: linear time maximum-likelihood and bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC bioinformatics*, 11(1):538, 2010.

[165] John T Kent. Information gain and a general measure of correlation. *Biometrika*, 70(1):163–173, 1983.

[166] Frank Yates. Contingency tables involving small numbers and the χ 2 test. *Supplement to the Journal of the Royal Statistical Society*, pages 217–235, 1934.

[167] Minoru Kanehisa, Susumu Goto, Yoko Sato, Miho Furumichi, and Mao Tanabe. Kegg for integration and interpretation of large-scale molecular data sets. *Nucleic acids research*, page gkr988, 2011.

[168] Zhenjiang Xu, Daniel Malmer, Morgan GI Langille, Samuel F Way, and Rob Knight. Which is more important for classifying microbial communities: whos there or what they can do&quest. *The ISME journal*, 2014.

[169] Kowthar Y Salim, Joyce C De Azavedo, Darrin J Bast, and Dennis G Cvitkovitch. Role for saga and siaa in quorum sensing and iron regulation in streptococcus pyogenes. *Infection and immunity*, 75(10):5011–5017, 2007.

[170] Christopher S Bates, Griselle E Montanez, Charles R Woods, Rebecca M Vincent, and Zehava Eichenbaum. Identification and characterization of a streptococcus pyogenes operon involved in binding of hemoproteins and acquisition of iron. *Infection and immunity*, 71(3):1042–1055, 2003.

[171] S Bergmann and S Hammerschmidt. Fibrinolysis and host response in bacterial infections. *Thrombosis and haemostasis*, 98(3):512, 2007.

[172] Dan Knights, Justin Kuczynski, Emily S Charlson, Jesse Zaneveld, Michael C Mozer, Ronald G Collman, Frederic D Bushman, Rob Knight, and Scott T Kelley. Bayesian community-wide culture-independent microbial source tracking. *Nature methods*, 8(9):761–763, 2011.

[173] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12:2825–2830, 2011.