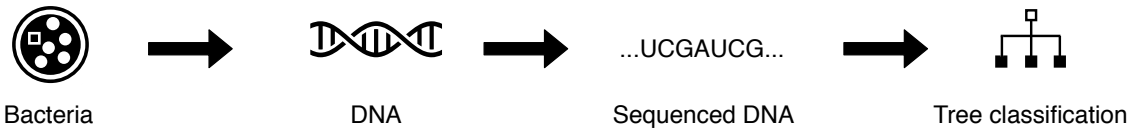
BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2019
http://excel.fit.vutbr.cz

Bioinformatic Tool for Classification of Bacteria into Taxonomic Categories Based on the Sequence of 16S rRNA Gene

Bc. Nikola Valešová*



Abstract
This work deals with the problem of automated classification and recognition of bacteria after obtaining their DNA by the sequencing process. In the scope of this paper, a new classification method based on the 16S rRNA gene segment is designed and described. The presented principle is based on the tree structure of taxonomic categories and uses well-known machine learning algorithms to classify bacteria into one of the connected classes at a given taxonomic level. a part of this work is also dedicated to implementation of the described algorithm and evaluation of its prediction accuracy. The performance of various classifier types and their settings is examined and the setting with the best accuracy is determined. Accuracy of the implemented algorithm is also compared to an existing method based on BLAST local alignment algorithm available in the QIIME microbiome analysis toolkit.

Keywords: Machine learning — Metagenomics — Bacteria classification — Phylogenetic tree — 16S rRNA — DNA sequencing — scikit-learn

Supplementary Material: Table of Results

*xvales02@stud.fit.vutbr.cz, *Faculty of Information Technology, Brno University of Technology*

1. Introduction

For a long time, it was possible to analyse bacteria only by their cultivation. However, many bacteria species are unculturable and therefore were not detectable at all. Thanks to recent advance in high-throughput sequencing, it is now achievable to efficiently investigate microbial communities and analyse the bacteria species found in them. With the obtained knowledge, attention of many scientists has been drawn to research of the relationship between the human microbiome and human health. Dysfunctional human microbiome has been already linked to many diseases, such as diabetes, inflammatory bowel disease, and antibiotic-resistant infection. [1], [2]

The aim of this article is to design and describe a new bacteria classification method, which is based on the 16S rRNA gene segment. In the scope of this work, the described method is implemented and its accuracy[1] is evaluated and compared with another existing bacteria classification tool.

Some methods of bacteria classification, which have already been implemented, are described in more detail in section 3. Most of the techniques are built

[1] Accuracy represents the true positive rate and can be computed using the following formula: $accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions}$ [3].

on the basis of k-NN classifier and apply various approaches to search for nearest neighbours.

The principle presented within this work is built on the basis of tree structure of taxonomic categories. The whole classifier consists of a tree of partial classifiers with topology respecting the taxonomic tree. Classification of an input specimen starts in the top classifier distinguishing between bacteria and archaea and the input sequence descends through the tree according to the predicted labels. The partial classifiers are well-known machine learning methods (such as SVM, decision tree, and nearest centroid) and their aim is to classify the given bacteria and assign it a label at lower taxonomic level.

Thanks to the use of taxonomic tree structure and the concept of successive classification it should be possible to decrease the overall classification error in comparison to direct classification on the lowest taxonomic level. This approach also offers the possibility of presenting the whole taxonomic classification from domain down to genus with values of reliability for predicted labels on every taxonomic level.

In section 2, some introductory terms in the field of molecular biology, which are essential for understanding the core of this work, will be introduced and explained. Section 3 is dedicated to the description of some other existing methods solving the bacteria classification problem. Section 4 contains detailed description of the proposed method. In section 5, there is a brief introduction to used tools and libraries. The aim of section 6 is to present some already achieved results and a comparison with one of the existing methods.

2. Definition of Important Terms

Metagenomics is a field of study focused on the microbial world. Its main characteristic is investigation of bacteria, viruses and fungi in complex communities, in which they usually exist, irrespective of whether they are culturable of not. Metagenomics tries to examine the genome of an entire organism concerning also the microbes existing within it. [4], [5]

2.1 DNA and Its Sequencing

Deoxyribonucleic acid is a material with hereditary information encoded in it. It can be found in almost all known organisms. Most cells in a human body share the same DNA. In a cell, DNA can be found in its nucleus, or in the mitochondria. [6]

The DNA can be represented as a code consisting of four chemical bases – adenine (A), guanine (G), cytosine (C), and thymine (T). The information in the DNA is encoded by combining these bases into long sequences. [6]

DNA sequencing is the process of determining the nucleotide sequence of DNA. The obtained sequence gives the most fundamental knowledge of a gene or a genome[2]. [8]

2.2 RNA, rRNA and 16S rRNA

Ribonucleic acid is a nucleic acid consisting of a long strand of nucleotides. Similarly to DNA, each nucleotide contains a nitrogenous base, a sugar, and a phosphate. [9]

rRNA is one type of RNA called ribosomal ribonucleic acid. It is located in ribosomes, which are the catalysts of protein synthesis. Over sixty percent of the ribosome consists of the ribosomal RNA which is a necessary part of all functions of a ribosome. [10]

16S rRNA is a sequence of DNA which encodes the RNA component of the smaller subunit of the bacterial ribosome. It can be found in the genome of all bacteria species and a related form can be found in all cells. In the 16S rRNA, two types of regions can be determined – slowly changing sections and variable parts, which undergo rapid genetic changes, therefore they are suitable for differentiating species. It has been proved to have the most information regarding examination of evolutionary relatedness. [11]

2.3 K-mer Spectrum

K-mers refer to subsequences of length k found in an input sequence. The term k-mer then represents all length k subsequences of a given sequence. In the field of computational genomics, the sequences, that are being processed, are often composed of nucleotides.

One of the biggest advantages of using k-mer spectrum is that the final spectrum is of the same length, regardless of the input sequence size. That is important when applying machine learning algorithms as they require their input data to have the same number of dimensions. On the other hand, extracting k-mer spectrum from a sequence leads to loss of the positional information of the subsequences in the original sequence.

3. Previous Works

One approach of solving bacteria classification has been introduced by Wang et al. in their article [12]. In this work, they described the RDP classifier, which extracts k-mer spectra from the input sequences and then applies the naïve Bayesian classifier to assign the

[2] A genome is the complete set of DNA of an organism, which includes all of its genes. All needed information on how to build and maintain the organism can be found in its genome. [7]

unknown specimen its taxonomic classification from domain to genus. Regarding k-mer size, they achieved the best accuracy with k-mers 8 and 9 and decided to use k-mer 8 to reduce memory requirements.

Some other existing solutions are implemented as a part of QIIME, a bioinformatic pipeline designed for analysing microbiome from raw DNA sequencing data. [13]

One of the methods implemented within QIIME is called USEARCH LCA. This method is based on the k-NN classifier. It uses the USEARCH [14] algorithm for finding k sequences, which are nearest to the given sequence and whose taxonomy is known. Then, on their taxonomic classifications, the LCA [15] algorithm is applied to obtain the taxonomy of the unknown sequence. Another approach is implemented in QIIME 2 and it is named BLAST LCA. The principle of this method is the same as in the previous algorithm, with the only change in the search algorithm. In this method, the BLAST [16] search algorithm is used. Last mentioned method is QIIME BLAST TOP HIT, which basically represents the k-NN algorithm with k set to 1. It uses the BLAST algorithm for finding the nearest neighbour and assigns the unknown sequence the taxonomy of the nearest sequence. [17]

Another solution microclass is available as an R package and while it has a standard R interface, its computational core is implemented in C++ to reduce time consumption. After experimenting with various k-mer based methods, the authors decided to use the multinomial method. The package offers two functions for training and applying a custom model, and it also offers a ready-to-use pre-trained classification tool, which uses k-mer size 8 and has been trained on full-length 16S rRNA sequences. K-mer of length 8 has been chosen since its increase to 9 or 10 results in high cost in memory consumption and computation time while the gain in accuracy is small. [18]

16S Classifier is based on a random forest classification model and uses only the hypervariable regions of the 16S rRNA in order to increase speed and prediction accuracy. To obtain the random forest model with the highest accuracy, parameter optimisation was applied. For the optimisations, hypervariable regions of V3 and k-mer sizes from 2 to 6 were experimented with. After performance testing, the authors came to conclusion, that performances of 2-mer and 3-mer models offered the lowest accuracy. 5-mer and 6-mer models gave results with the lowest error, however, the 4-mer model needed significantly less time and smaller size of training data. Therefore, the authors decided to use 4 as the k-mer size. [19]

4. Proposed Bacteria Classification Method Specification

This section contains detailed description of the proposed bacteria classification method, starting with k-mer spectra extraction from the input 16S rRNA sequences and continuing to classification tree creation, training and evaluation.

4.1 K-mer Spectra

This method is designed to classify bacteria according to the sequence of their 16S rRNA gene. The 16S rRNA sequence is different for every genus and can contain multiple mutations, insertions and deletions, therefore various 16S rRNA sequences can be of different length. This could cause inconvenience as machine learning algorithms require their input vectors to be of equal dimensions. To overcome these difficulties, a k-mer spectrum is extracted from the input sequence and used for classification afterwards. With the use of k-mer spectra, it is possible to transform every 16S rRNA sequence, which is in the form of a string, into a numeric vector, where each value represents the number of occurrences of the corresponding substring in the original sequence.

K-mer spectra extracted from the 16S rRNA sequences of all bacteria species have the same predetermined length, which can be computed using the formula n^k [20], where n is the number of possible characters (size of the alphabet) and k is the k-mer size. There are four nucleotides, which are being found in an RNA of a bacteria – adenine (A), cytosine (C), guanine (G), and uracil (U).

However, other characters are also frequently present in the rRNA sequences in various databases. These characters (similarly to regular expressions) represent two or more bases, e.g. Y, which can mean either cytosine or uracil [21]. To include also these characters and avoid unnecessary loss of information, the proposed classifier deals with substrings, which contain one or more non nucleotide characters, by transforming them into all possible real nucleotides they represent (created according to substitution table defined by the IUPAC federation [21]) and incrementing their count of occurrences.

4.2 Classification Tree

This classification tool is designed to assign unknown bacteria to their most probable genera. In order to minimise the classification error, the presented method is based on a tree structure of the taxonomic tree. The whole classifier is decomposed into multiple classifiers on all levels of taxonomy from domain down to genus.

This way it could be possible to obtain higher prediction accuracy since every classifier distinguishes only among the few classes belonging to it on the lower level of taxonomy.

The tree structure of component classifiers is shown in figure 1. Every rectangle represents a single classifier (of one of the well-known classifier types, such as SVM or nearest centroid) and its label shows the classification of the input sequence on the corresponding level of taxonomy. An example classification of *Escherichia coli* is given. The component classifiers are connected with arrows indicating the order of sequence classification through the whole tree of classifiers.

With this method, it is possible to obtain a complete taxonomic classification from domain to genus and, in case of an unsuccessful classification, determine the exact partial classifier, which caused the incorrect classification.

The training phase can be divided into two parts. First, all types of classifiers with various settings compete against each other so that it is possible to determine the best performing classifier settings. In order to execute the competition, the entire process of tree creation, training and validation is wrapped inside a cross-validation. In every iteration, multiple types of classifiers in various configurations are trained on the current training data set and their accuracy is then evaluated on the validation data set. The outcome of every iteration is the sum of accurate predictions obtained during validation.

After the entire cross-validation process, the best performing classifier is determined, trained on all available data and stored as the final model. With this approach it is possible to obtain the best performing classifier for every data set used.

The process of rRNA gene sequence classification is initiated by presenting the input sequence to the top classifier and categorising it as either a bacteria or an archaea. Afterwards, the chosen one of the two classifiers on the domain level is used to assign the input sequence its phylum. Then, the classifier belonging to the assigned phylum is used to classify the input into one of the connected classes and this process repeats itself until the final classification, genus, is assigned.

4.3 Used Classifier Settings

During the process of best classifier determination, eight types of classifiers altogether in forty-five configurations, which have been determined using grid search, are trained and validated. The overview of all used classifier types and their settings can be seen in table 1.

Table 1. Used classifier types and their settings

Classifier Type	Classifier Settings
SVM	kernel = {linear, rbf, sigmoid}
Nearest Centroid	metric = {euclidean, manhattan, chebyshev, minkowski, seuclidean, correlation}
k-NN	n_neighbors = {1, 2, … , 10}
Decision Tree	max_depth = {2, 3, … , 10}
Random Forest	max_depth = {7, 10} estimator_count = {10, 15} max_features = {auto, sqrt, log2}
MLP	penalty = {0.01, 0.1, 0.15} max_iterations = {300, 500}
AdaBoost	-
GaussianNB	-

5. Implementation of the Classifier

The described bacteria classification system is implemented in Python3 and compiled using Cython static compiler[3] to increase its performance. It uses the *numpy* library providing efficient implementation of multi-dimensional array object and advanced broadcasted operations[4], and classes of the mentioned classifiers imported from *scikit-learn*, which is an open source library implementing simple and efficient tools for data analysis[5].

6. Achieved Results

The BLAST 16S dataset[6] was used for evaluation of the implemented method. The dataset contains more than 7.500 region 4 16S rRNA sequences of well-known and examined bacteria in FASTA format. Every sequence is a string composed of four nucleotides – *A*, *C*, *G*, *T*, and for every sequence, there is its complete taxonomic classification.

Accuracy has been evaluated using the leave-one-out cross-validation. During every run, entire dataset except one specimen has been used for training and

[3] Cython developers and contributors. *Cython: C-Extensions for Python*. https://cython.org/, February 2019. [Online; visited 13. 3. 2019].

[4] NumPy developers. NumPy – NumPy. http://www.numpy.org, October 2018. [Online; visited 13. 3. 2019].

[5] F. Pedregosa et al. *scikit-learn: machine learning in Python – scikit-learn 0.20.3 documentation*. https://scikit-learn.org, March 2019.

[6] The dataset is available on site: https://drive5.com/taxxi/doc/fasta_index.html

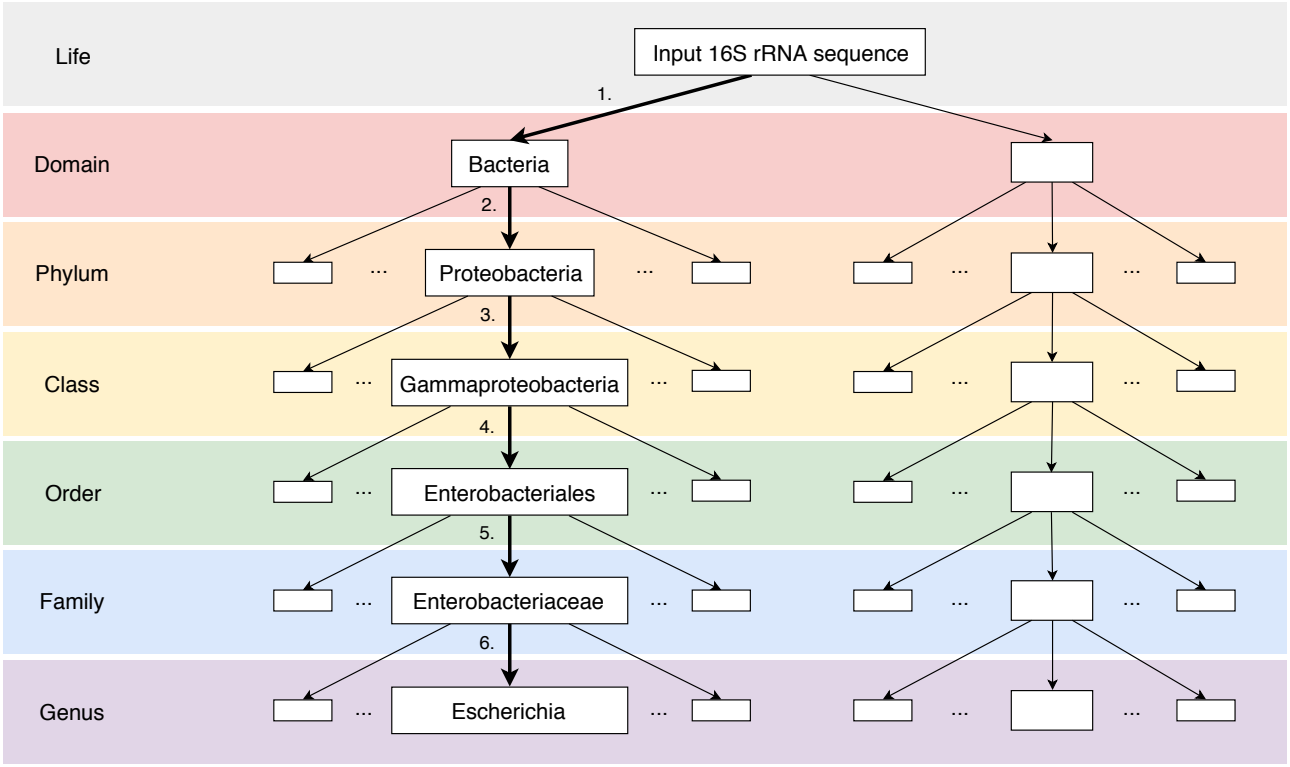


Figure 1. Tree structure of classifiers created according to the taxonomic tree with order of classification emphasised

the left out sample has been then used for validation. By utilizing this principle it was possible to obtain the overall accuracy of the presented algorithm on this dataset and also to get average accuracies for all classifier settings listed in table 1. For all results listed in this section, k-mer 5 has been used as it proved to be time and memory efficient and offer quite good accuracy at the same time.

The comparison of the best results every classifier type has reached can be seen in figure 2. The SVM classifier has obtained the best results using linear kernel, Nearest Centroid worked the best with correlation metrics, for k-NN algorithm it was setting the count of neighbours to 1, Decision Tree with maximal depth set to 10, Random Forest with maximum depth of 10, consisting of 10 randomly created decision trees and the number of features considered during looking for the best split set to the square root of number of features, and the MLP classifier with penalty 0.01 and a maximum of 300 iterations.

From the graph, the AdaBoost classifier has reached the worst performance. Neither Gaussian Naive Bayes gave satisfactory results. Slightly better accuracies were reached by Nearest Centroid and Decision Tree classifiers. Random Forest performed quite good, but especially on phylum, class and genus levels worse than the best ones. SVM, k-NN and MLP classifiers gave the best results, almost indistinguishable on all

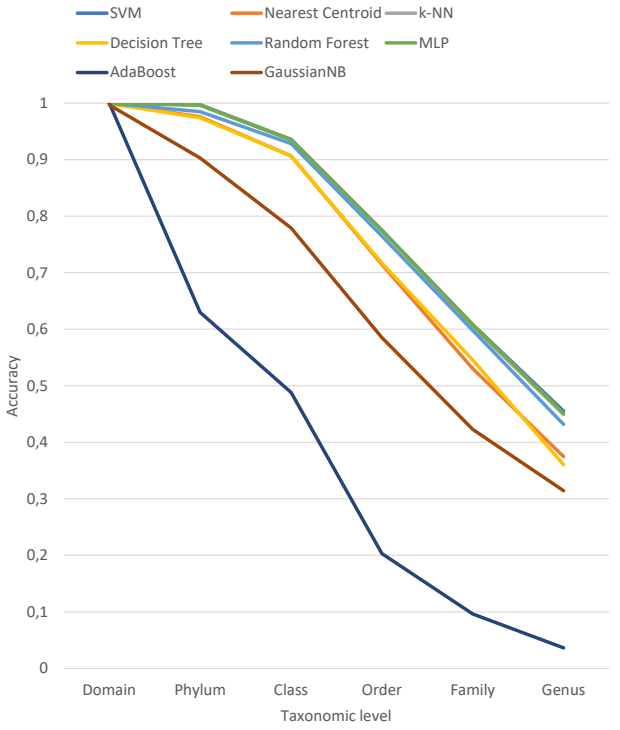


Figure 2. Comparison of the best results obtained by individual classifier types

levels except genus, where SVM slightly stands out.

For every classifier, there is a decrease in accuracy with increasing taxonomic level, as expected. While the domains bacteria and archaea separated millions of years ago (and their 16S rRNA sequences differ

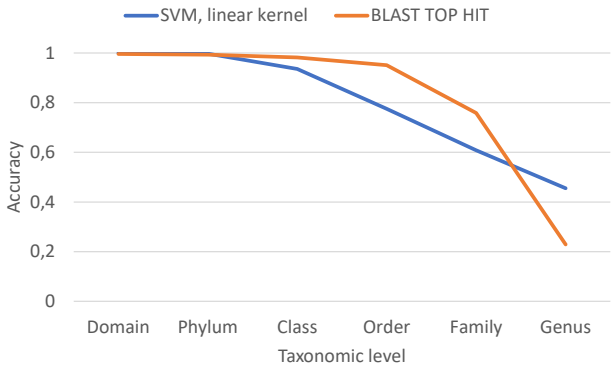


Figure 3. Comparison of results of the presented method (SVM classifier with linear kernel) and BLAST TOP HIT algorithm

significantly), two genera could have been split no longer than a decade ago.

In graph 3, there is a comparison of the results obtained by the presented method with BLAST TOP HIT algorithm described in section 3. The presented method has reached the best overall results on the used dataset when using SVM classifier with linear kernel. BLAST TOP HIT has been chosen for comparison since the method was fast and easy to implement in order to evaluate it on the chosen dataset.

On domain and phylum level, the accuracy of both compared methods is equally good. The class, order and family levels are predicted with better accuracy by the BLAST TOP HIT algorithm. However, on the genus level, the implemented method works better than BLAST TOP HIT.

Table in csv format with results, which were used to create these statistics, can be seen in the link in supplementary material.

7. Conclusions

In this work, the tree structured approach of bacteria classification has been introduced and described. The proposed algorithm was based on the use of the 16S rRNA gene sequences and applied well-known machine learning algorithms to solve the classification problem.

During examination of performance of the presented method, the best working classifier type and setting was SVM with linear kernel. The average accuracy reached on domain and phylum levels was above 99 %, around 78 % on order level, 61 % for family level and 46 % on genus level. When compared to another already existing method, the implemented algorithm gave competitive results.

Thanks to the presented principle, it was possible to train multiple classifiers with various settings and use the one, which performs best on the current dataset. With this approach, the final classification method can be chosen specifically for the dataset used.

In the future, further optimisations resulting in an increase in performance can be applied. Other approaches can focus on improving the prediction accuracy. One way this could be achieved is to let the classifiers compete on each level, or in every node, and then obtain a classifier tree composed of multiple classifier types and configurations.

Acknowledgements

I would like to express my sincere gratitude to my advisor, Ing. Stanislav Smatana, for the continuous support of my study, his motivation and valuable advice.

References

[1] Xochitl C. Morgan and Curtis Huttenhower. *Chapter 12: Human Microbiome Analysis*. In *PLoS Computational Biology*, 2012.

[2] Andreas Hiergeist, Udo Reischl, and Andrè Gessner. *Multicenter quality assessment of 16S ribosomal DNA-sequencing for microbiome analyses reveals high inter-center variability*. *International Journal of Medical Microbiology*, 306(5):334–342, Aug 2016.

[3] Google Developers. *Machine Learning Glossary — Google Developers*. https://developers.google.com/machine-learning/glossary, January 2019. [Online; visited 2. 4. 2019].

[4] Diana Marco. *Metagenomics: Theory, Methods and Applications*. Caister Academic Press, 2010.

[5] Yolanda Smith. *What is Metagenomics?* https://www.news-medical.net/life-sciences/What-is-Metagenomics.aspx, August 2018. [Online; visited 16. 1. 2019].

[6] Genetics Home Reference. *What is DNA? – Genetics Home Reference – NIH*. https://ghr.nlm.nih.gov/primer/basics/dna, January 2019. [Online; visited 16. 1. 2019].

[7] Genetics Home Reference. *What is a genome? – Genetics Home Reference – NIH*. https://ghr.nlm.nih.gov/primer/hgp/genome, January 2019. [Online; visited 16. 1. 2019].

[8] yourgenome. *What is DNA sequencing? — Stories — yourgenome.org*.

https://www.yourgenome.org/stories/what-is-dna-sequencing, June 2016. [Online; visited 16. 1. 2019].

[9] Ananya Mandal. *What is RNA?* https://www.news-medical.net/life-sciences/What-is-RNA.aspx, August 2018. [Online; visited 16. 1. 2019].

[10] Larry Li. *Ribosomal RNA (rRNA) – Definition and Functions — Biology Dictionary*. https://biologydictionary.net/ribosomal-rna/, April 2017. [Online; visited 16. 1. 2019].

[11] J. M. Janda and S. L. Abbott. 16s rRNA gene sequencing for bacterial identification in the diagnostic laboratory: Pluses, perils, and pitfalls. *Journal of Clinical Microbiology*, 45(9):2761–2764, Jul 2007.

[12] Q. Wang, G. M. Garrity, J. M. Tiedje, and J. R. Cole. *Naive Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy*. *Applied and Environmental Microbiology*, 73(16):5261–5267, Jun 2007.

[13] QIIME development team. *QIIME*. http://qiime.org, January 2018. [Online; visited 19. 3. 2019].

[14] Robert C. Edgar. *USEARCH*. https://www.drive5.com/usearch/, March 2019. [Online; visited 21. 3. 2019].

[15] Chris Lewis. *Microsoft PowerPoint – Lowest Common Ancestor (LCA) techniques.ppt – lowest_common_ancestor.pdf*. http://homepage.usask.ca/~ctl271/810/lowest_common_ancestor.pdf, September 2004. [Online; visited 19. 3. 2019].

[16] National Center for Biotechnology Information. *BLAST: Basic Local Alignment Search Tool*. https://blast.ncbi.nlm.nih.gov/Blast.cgi, March 2019. [Online; visited 21. 3. 2019].

[17] J Gregory Caporaso, Justin Kuczynski, and Jesse Stombaugh et al. *QIIME allows analysis of high-throughput community sequencing data*. *Nature Methods*, 7(5):335–336, Apr 2010.

[18] Kristian Hovde Liland, Hilde Vinje, and Lars Snipen. *microclass: an R-package for 16S taxonomy classification*. *BMC Bioinformatics*, 18(1), Mar 2017.

[19] Nikhil Chaudhary, Ashok K. Sharma, Piyush Agarwal, Ankit Gupta, and Vineet K. Sharma. *16S Classifier: A Tool for Fast and Accurate Taxonomic Classification of 16S rRNA Hypervariable Regions in Metagenomic Datasets*. *PLOS ONE*, 10(2):e0116106, Feb 2015.

[20] Benjamin Lee. *K-mer – PLoSWiki*. http://compbiolwiki.plos.org/wiki/K-mer, August 2018. [Online; visited 11. 3. 2019].

[21] Paul Stothard. *IUPAC Codes*. https://www.bioinformatics.org/sms/iupac.html, May 2000. [Online; visited 14. 3. 2019].