Prokaryote classification - method development and novel insight in 16S ribosomal RNA-based classification

Prokaryote klassifisering - metodeutvikling og
ny innsikt i 16S ribosomal RNA-basert klassifisering

Philosophiae Doctor (PhD) Thesis

Hilde Vinje

Dept. of Chemistry, Biotechnology and Food Science
Norwegian University of Life Sciences

Ås 2016

Thesis number 2016:95
ISSN 1894-6402
ISBN 978-82-575-1409-9

Acknowledgements

The work presented in this thesis has been carried out in the Biostatistic group at the department of Chemistry, Biotechnology and Food Science at the Norwegian University of life Sciences (NMBU). Reaching this point in my education would be impossible had it not been for great support from several people. Specially three persons need to be highlighted.

First of all I want to thank my main supervisor Lars Snipen. He has been nothing else than exceptional during my time as a PhD student. I could not wish for a better supervisor. His door has always been open and he has answered all my questions almost exclusively with a smile. His knowledge in the field of bioinformatics and also his endless curiosity have inspired me for five years now. And this is the reason that my knowledge now have exceeded the level *"I have heard about bacteria"* which was my clever answer when he asked me about bacteria at the job interview.

The next person I want to thank is Kristian Hovde Liland. Kristian is my computer guru. He has always been available for questions; office, mail or Facebook. He always answers at a glance and uses happily (I think) the time watching the evening news to make our R-program run ten times faster. Then I would like to express my special appreciation and thanks to Trygve Almøy. If it had not been for Trygve I would never even considered statistics. Since the first course I took at NMBU in 2008 he has inspired me to move forward. It was he who made me consider a PhD after my master and I will be forever grateful for his support all these years. These three men are, without doubt, a supervising dream-team. With their different knowledge and perspectives, they have made my job as a PhD student truly enjoyable.

Thanks to the Biostatistic group for lots of laugher and memories during these years. I will never forget our movie-making, parties, coffee breaks, and ski trips. There has always been help and support from this group and I feel lucky to have been a part of it. A special thanks to my parents, for all the time they have used helping me with my homework and always made me work harder (than everybody else) during my education. To my sister and my friends for support, especially Sebastian and Niklas for proofreading my writing and of course, Eivind, for coping with my mood swings and watching over our two beautiful daughters whenever I needed to work.

Ås, October 2016
Hilde Vinje

List of papers

I. Vinje H, Almøy T, Liland KH, Snipen L (2014) *A systematic search for discriminating sites in the 16S ribosomal RNA gene*, Microbial Informatics and Experimentation **4**:2

II. Vinje H, Liland KH, Almøy T, Snipen L (2015) *Comparing K-mer based methods for improved classification of 16S sequences*, BMC Bioinformatics **16(1)**:205

III. Vinje H, Liland KH, Snipen L (2016) *The ConTax data: Improved supervised learning of prokaryotic taxonomy*, submitted manuscript

IV. Liland KH, Vinje H, Snipen L (2016) *`microclass`: An R-package for 16S taxonomy classification*, submitted manuscript

Summary

The main objective of this thesis is the improvement of prokaryotic classification based on the 16S ribosomal RNA. As a result of the shift in sequencing technology, generating enormous amounts of sequencing data and the rise of cultivation-independent methods, the need for reliable, fast and memory efficient methods has been revealed. The 16S rRNA is used for building the existing taxonomy of prokaryotes and map it into the phylogenetic tree of life, as well as for the exploration of microbial communities, which has become a major focus in microbiology. It is a common belief that the discriminant power of the 16S marker lies within nine variable regions located along the gene. We began our work challenging this assumption by searching systematically for discriminating sites that contributes to a correct classification. 50 discriminating sites were found when classifying down to phylum level and for genus identification, over 80% of all sites were important, they were all scattered throughout the gene.

We further present a systematic comparison of five K-mer based classification methods for the 16S rRNA gene. Classification methods based on counting K-mer are popular because they are fast, consider the whole sequence and will not suffer from the same uncertainties as evolutionary models and alignments. The five methods differ both in data usage and modelling strategies. Preprocessed nearest-neighbour (PLSNN) performed best on full-length sequences, but overall, for both full and fragmented sequences, the multinomial method outperformed the others. It was significantly better than the RDP-classifier, which today works as a gold standard classification method.

There is no official taxonomy of prokaryotes and any classification method will suffer from the lack of consensus in training data. The *ConTax* database, presented in this thesis, is a seed-set of the most accurately classified sequences from which we can continue to explore the prokaryotic taxonomy and train new classification methods. A major feature of the new dataset is that a sequence is included only if three primary 16S databases agree on its assigned taxonomy down to genus level. The results are combined and presented in an R-package, `microclass`, which provide classification tools down to genus level. Efforts have been made to make the tools both fast and memory-efficient. All methods can be trained on new data, but a ready-to-use tool, the `taxMachine`, is also presented. The `taxMachine` has been trained with the multinomial method on full-length 16S sequences to recognize full or fragmented sequences, using the designed and optimized trimmed `ConTax` dataset for training.

Sammendrag

Hovedmålet med avhandlingen er å forbedre klassifikasjon av prokaryoter basert på 16S ribosomalt RNA. Som et resultat av skiftet i sekvenseringsteknologien, som nå genererer enorme mengder med sekvensdata, og fremveksten av kultiverings-uavhengige metoder, er det blitt avdekket et behovet for stabile, kjappe og minne-effektive metoder. 16S rRNA er blitt brukt for å bygge den eksisterende taxonomien av prokaryoter og kartelgge de i det fylogentiske livstreet. Samt, utforske mikrobielle samfunn, som har blitt et hovedfokus i mikrobiologi. Det er en vanlig oppfatning at den diskriminante evnen til 16S markøren ligger innenfor ni variable regioner lokalisert langs genet. Vi begynte vårt arbeid med å utfordre denne antagelsen ved å søke systematisk for posisjoner med diskriminerende evner som bidro til korrekt klassifikasjon. 50 diskriminerende posisjoner ble funnet ved klassifisering ned til phylum nivå og for genus identifisering var over 80% av alle posisjoner viktige, de var alle spredt over hele genet.

Videre presenterer vi en systematisk sammenligning av fem K-mer baserte klassifiseringsmetoder for 16S rRNA genet. Klassifiseringsmetoder basert på å telle K-merer er populære fordi de er raske, tar for seg hele sekvensen og lider ikke av usikkerhetene som evolusjonære modeller og sammenstillinger gjør. De fem metodene er forskjellige både i databruken og modellerings-strategien. Den forbehandlede nærmeste-nabo metoden (PLSNN) gjorde det best for full lengde sekvenser, men generelt, for både full lengde og fragmenterte sekvenser, gjorde mutinomial metoden det bedre enn de andre. Den var signifikant bedre enn RDP klassifikatoren, som idag fungerer som en gullstandard av klassifiseringsmetoder.

Det finnes ingen offisiell taxonomi av prokaryoter og enhver klassifiseringsmetode vil lide av mangelen på konsensus i treningsdata. *ConTax* datasettene, presentert i denne avhandlingen, er en samling av de mest nøyaktig klassifiserte sekvensene som vi kan fortsette å utforske den prokaryote taksonomien utifra og trene nye methoder med. Den viktigste egenskapen til datasettet er at en sekvens bare blir inkludert hvis tre hoved 16S databaser er enige om den tildelte taksonomien ned til genus nivå. Resultatene er kombinert og presentert i en R-pakke, `microclass`, som inneholder klassifiseringsverktøy ned til genus nivå. En innsats har blitt gjort i å gjøre redskapene både kjappe og minne-effektive. Alle metoder kan bli trent med nye data, men en klar-til-bruk metode, `taxMachine`, er også presentert. `taxMachine` er blitt trent med multinomial metoden på full lengde 16S sekvenser for å kjenne igjen fulle eller fragmenterte sekvenser, ved å bruke det konstruerte og optimerte trimmede `ConTax` datasettet for trening.

Contents

Abbreviations and Explanations

DNA	Deoxyribonucleic acid
RNA	Ribonucleic acid
16S rRNA	16S ribosomal RNA
PCR	Polymerase chain reaction
Taxon/taxa	A group of one or more organisms that categorize to form a unit
Reads	Short DNA sequences, product of sequencing
K-mers	All possible words of length K in the DNA alphabet
KNN	K nearest neighbour
PLS	Partial least squares
MSA	Multiple sequence alignments
OTU	Operational taxonomic unit

Errors using inadequate data are much less than those using no data at all.

CHARLES BABBAGE

1 Introduction

1.1 Background

Microbes such as archaea and bacteria, together referred to as prokaryotes, are single-cell organisms found everywhere on earth, in the air, soil and water and inside (and on) living creatures. They play a major role of biological functions. Without them we would have been unable to digest food, plants would not grow and there would have been considerably less oxygen in the air. Microbes maintain the ecological balance of our planet and are indispensable for the survival of most species and essential for every part of human life [12]. Simply put: Microbes run the world!

DNA sequencing is a highly prioritized area in modern biology as it provides us with the most basic information of all: The DNA sequence of nucleotides. Over the past ten years several new sequencing technologies have revolutionized the field; They have gone from high-costs, low throughput to high-throughput. The costs have been reduced significantly and the sequencing has become considerably more efficient, generating an enormous amount of data. Thousands of sequences are now available in public repositories and today there are tens of thousands of sequencing projects in progress. For a long time it has been difficult to study microbes in their own environment; the microbiologists studied individual species one by one in the laboratory and these so called pure-cultures were the standard for microbiology. In the middle of the $20th$ century, in the pure-culture paradigm, microbes that grew well as single cells suspended in a liquid medium became the model for much of modern biology [46]. Today we know that these microbes only covered 1% of the diversity of all species [27].
As a result of the shift in sequence technology, and the rise of cultivation-independent methods, it has become clear that microbes actually live in communities and interact in complex systems. In these communities they are in physical contact with microbes of their own kind and with other species, as well as the environment. In fact, bacterial monocultures hardly ever occur in nature. The composition of microbial communities is far from accidental and it has opened a whole new spectre of understanding the (microbial) world. "Who is out there?", "what are they doing there?" and "how are they doing it?" are questions that now can be investigated by studying a microbial community from a specific habitat.

1.2 The study of complex microbial communities

There are two common ways to obtain data and study complex microbial communities.

I) **Amplicon sequencing for building a 16S-profile:** The diversity of a microbial community is often investigated by sequencing a selected genomic marker which is able to identify the microbes to a certain level of taxonomy. For prokaryotes the choice of marker usually falls on the variable regions of the 16S ribosomal RNA (see section 1.3). Amplicon sequencing is the most widely used method for identifying the composition of microbial communities. A sample is taken from the environment (for example water, soil, gut) and the DNA is extracted from all the sampled cells. An informative genetic marker (here the 16S rRNA) is targeted with universal primers. Primers are short sequences matching and enclosing key parts of the marker and are universal in the way that they span all taxa. The regions enclosed by the primers are called amplicon-sequences and are subtracted from the DNA by a polymerase chain reaction (PCR). The PCR products are sequenced by using technology such as Sanger, but more recently, high-throughput sequencing platforms like Illumina, Ion Torrent and PacBio. These generated short amplicon-sequences can be bioinformatically characterized to determine which microbes are present in the sample and at what relative abundance, hence building a 16S-profile. The study of microbial communities has been revolutionized after ribosomal RNA profiling methods were implemented [69, 47, 18, 59]. It can also be used to ascertain the similarity of two or more communities (hence, communities that share a greater degree of identical taxa are more similar) [56]. The method is widespread, it has for instance been used to map the diversity of the human gut [13, 72], *Arabidopsis thaliana* roots [36], ocean thermal vents [38] and Antarctic volcano mineral soils [61].

Although this method is used extensively to study community constituents, amplicon sequencing is not without limitations. It may, for instance, fail to identify the true diversity of the sample because of biases associated with the primers chosen, or because of failure during the PCR process [24, 57, 33]. Sequencing errors and formation of chimeric sequences during PCR produces artificial sequences, which are difficult to identify. A chimera is an incorrect pairing of two different sequences. Because the 16S gene can be transferred horizontally between distantly related taxa the 16S analysis can result in overestimations of the diversity in the community [3]. Also, the amplicon sequencing and construction of the 16S-profile is unable to resolve the biological function associated with the community and is also limited to the analysis of taxa that contain known genetic markers.

II) Some of these limitations can be overcome by **Shotgun metagenomic DNA sequencing**. Shotgun metagenomic DNA sequencing is a relatively new and powerful environmental sequencing approach that provides insight into the diversity and functions of a community. The DNA is again extracted from all the cells in a sample, but instead of targeting a genomic marker, random sequencing of short sub-sequences across the entire genomic content in the sample is considered, hence the name shotgun sequencing. The fragments are independently sequenced and results in short read sequences. Some of the reads are sampled from genomic markers, such as the 16S rRNA, and some are from coding sequences and as a result, metagenomics then allows us to explore: "Which genes are present in the sample?" and "what are they capable of doing?". For example, studies of metagenomic samples from the human body have discovered a relationship between the human microbiome and human health [13, 12] an approach used to study antibiotic-resistance genes [55]. But metagenomic data is complex and vast, and there are limitations connected to the retrieval of information for this approach as well. The short reads need to reassemble into longer contigs or, if possible, whole genomes. This is not trivial. For instance, because of the diversity of the community, some genomes are not completely represented by reads. If two reads do not overlap it is impossible to rebuild the genome through an assembly. And, if they do overlap it is not hard evidence that they are from the same genome. Therefore, a large volume of sequences is required to obtain meaningful results from a microbial community.

The preferred approach to study a microbial community depends on what you want to know and how much you are willing to spend. The metagenomic approach has the advantage of providing more information of the community and its function. However, it is more expensive than amplicon sequencing to reach the detailed level of sequencing needed to carry out the analysis. It also requires significantly more computational analysis to obtain the information. As sequencing has become cheaper and faster, and as new methods of extracting information from random sequencing metagenomics have been developed, researchers have speculated whether the amplicon sequencing for creating a 16S-profile will soon be obsolete. Recent studies, however, have shown that this is not the case. In many studies, the 16S-profiles provide exactly the information needed, for less expenses and computational effort. An approach is to use 16S-profiles in forensics, where these are used to investigate which part of the body a biological trace originates from. This cannot be determined by the "regular" DNA-trace, because human DNA is identical wherever it is retrieved from in the body. Instead one must withdraw the DNA from the microbes in the sample, and determine the relative abundance of the various taxa, hence building a 16S-profile. The abundance can then easily be compared to the 16S-profiles for all known body sites. Body sites show very distinctive profiles [73, 53, 9] and it has been demonstrated that marker-gene profiles of human microbi-

ota can provide a view of microbial diversity across body sites [14].

Both amplicon sequencing and shotgun metagenomic sequencing are still computationally challenged due to the rapid advances in sequencing technologies which now generate billions of reads in a few days. It is therefore crucial for software tools to minimize computational resources (for example time, memory and I/O) and obtaining high accuracy when analysing sequencing data.

1.3 The 16S ribosomal RNA gene

In the 1980's, it was demonstrated that phylogenetic relationships of bacteria and all other life-forms, could be well approximated by comparing stable parts of the genome [70, 68]. Today, the most commonly used marker for phylogenetic studies and studies of microbial communities is the 16S ribosomal RNA. The 16S rRNA is a structural component of the prokaryotic ribosome and present in all bacteria and archaea. Its common use in metagenomics and phylogenetic analysis is both due to its presence in all prokaryotes and because of its high level of conservation [47]. The 16S sequence is around 1550 base pair long, and the pattern of sequence conservation is assumed to be because of its essential role in cell function. It is an efficient tool for all sequencing platforms and its structure of both variable and conserved regions makes it very convenient for use as a molecular chronometer in evolutionary studies. Previous mapping of bacterial 16S genes show nine hypervariable regions [7]. These regions, denoted V1-V9, show distinct sequence diversity among different bacteria and have commonly been used for bacterial identification and taxonomic studies [63, 8]. Since the sequenced rRNA genes are frequently used as markers in metagenomics and phylogenetic reconstruction, the need for systematic databases has increased. The Silva comprehensive ribosomal RNA database (http://www.arb-silva.de/, [48]), the Greengenes 16S rRNA gene database (GG, http://greengenes.lbl.gov/cgi-bin/nph-index.cgi, [19]) and the Ribosomal Database Project (RDP, http://rdp.cme.msu.edu/, [11]) provide access to sets of rRNA sequences and tools for analysis useful in microbial studies. Due to the 16S rRNA's properties it can be used for bacterial identification and classification down to genus level [10].

1.4 Taxonomic classification

Following on from Darwin's work in the book "On The Origin Of Species" (1859), which discussed that all life forms arose from a common ancestor, biologists have attempted to classify life according to evolution. Taxonomy (sometimes known as "systematics") is the science of classifying organisms. It is a rank-based classification built on a hierarchical system. Each organism belongs to a series of ranked taxonomic categories where the broadest is by domain and kingdom, and the most specific classification is by genus and species. At any level in the hierarchy an organism belongs to only one taxon, or taxonomic group.

Classification is the arrangement of an organism into these predetermined groups, the taxa. The classification of 16S sequences obtained from some samples is a classical pattern recognition problem, i.e. recognizing patterns in a sequence, and assign it to one out of several predetermined categories, based on how similar these sequences are to the known sequences from the respective category. Figure 2 is an illustration of this concept. When a new sequence (i) is obtained, one of the first questions that arises is "What are we looking at?". For a microbial community study, you are dependent upon identifying the DNA sequence regardless of whether you are interested in exactly this bacteria alone, or whether you are facing the problem of identification of community constituents. You need a classifier (ii); a method that inputs one, or a couple of million, sequences and returns a qualified guess of which predetermined class (iii) it belongs to.

The retrieval of information from bacterial sequencing data is not trivial. The conditions for scientific success and progress depends mainly on two aspects; the quality of the methods, the classifier, and the quality of the data available, which the classifier is trained on. The amount of sequence data, the skewness in the data obtained and, perhaps the fishy problem, that there is no universal taxonomy makes the retrieval of information from the data we have today challenging.

In his paper from 1971 Cowan presented it as follows: *"Taxonomy is written by taxonomists for taxonomists; in this form the subject is so dull that few, if any, non-taxonomists are tempted to read it, presumably even fewer try their hand at it"*. Cowan followed up with: *"It is the most subjective branch of any biological discipline, and in many ways is more an art than a science"*[15]. Although this is an exaggerated assertion it is important to realize that considerable subjectivity has been allowed when designing the predetermined classes that today's taxonomy depends on. The taxonomy has, to a certain extent, been built on the intuition of individual researchers rather than fact based knowledge of an organism [6, 51].

Some of the branches in the bacterial tree of life are widely accepted and will most likely

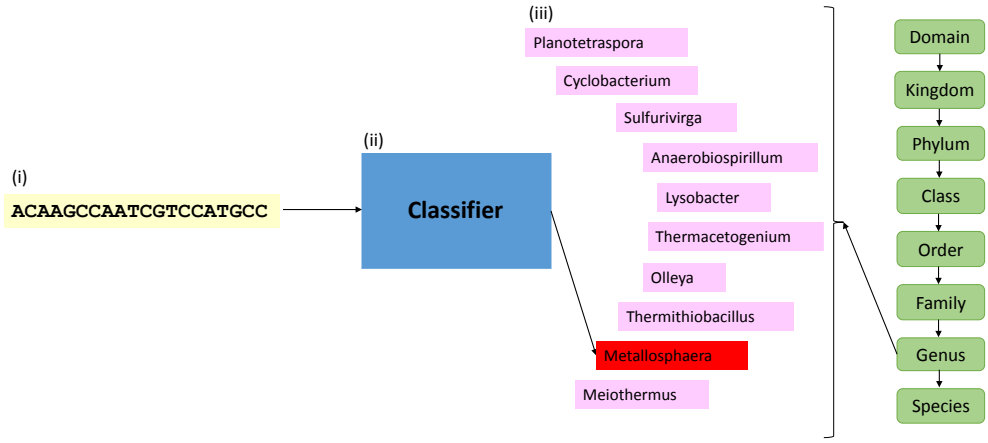


Figure 1: A visualization of the classification concept. A new DNA sequence (i) is used as input in a classifier (ii). This classifier will use a classification method to try to place the new sequence into one of many predetermined classes (iii), which again is divided into a hierarchic system: the taxonomy

never change, but a significant part of the taxonomy is still evolving. The culture-independent techniques highlighted our fragmented view of the phylogenetic tree of life by crudely outlining its borders. This was achieved, as explained in section 1.1, by sequencing 16S genes from DNA extracted directly from environmental sources as microbial communities. This known lack of biological insight is often referred to as the *microbial dark matter*. It is a daunting fact that the coverage of the real microbial diversity is estimated conservatively to represent hundreds of thousands of species [16]. The improvements in sequencing technologies have highlighted this issue [58, 39, 42], which again has resulted in many new taxa [64] and an ever increasing diversity in data available [45, 26]. Figure 2 visualizes a proposal of the phylogenetic tree of life from April 2016. The bluish, top-right branch of the tree is a new addition as a result of the culture-independent methods. The branch was completely unknown until a few years ago! The question remains; *How much of the diversity in the microbial world have we not yet seen?*

Thus, there is no comprehensive gold standard dataset which gives a clear picture of the microbial world as we know it; the classification system represented by Bergey's Manual of Systematic Bacteriology [67] is widely accepted and is therefore often considered the best ap-

proximation to an official classification [6]. The Bergey's classification system is based on the phylogenetic analysis of the 16S gene, together with classical microscopic and biochemical observations of relations between organisms, and it has been valuable in describing the width in prokaryotic diversity and setting the framework for the study of relationships between taxa [34, 6, 28].

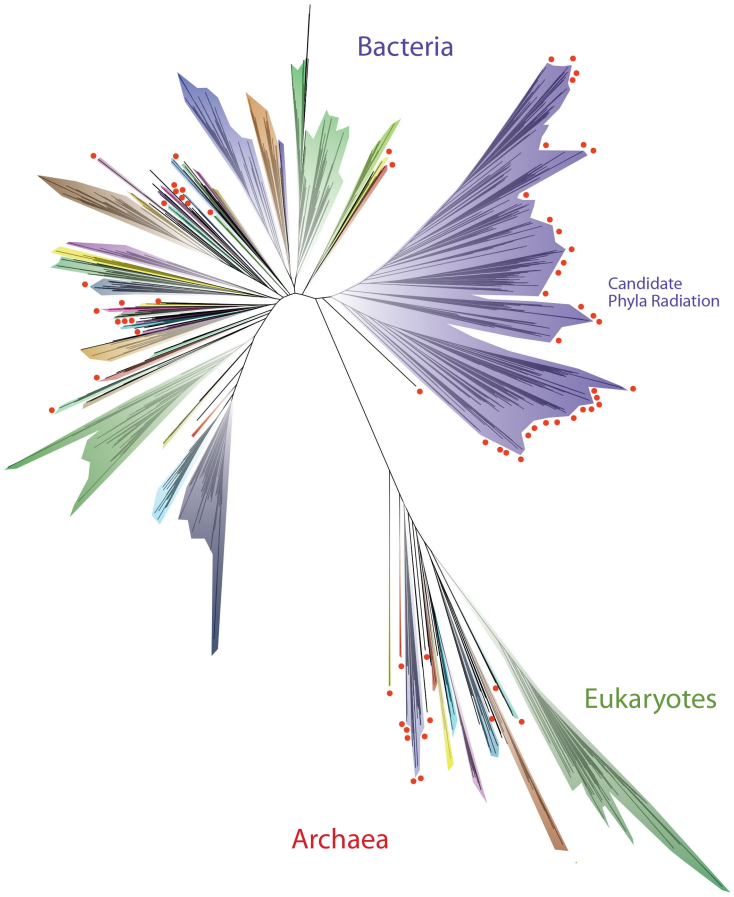


Figure 2: A current view of the phylogenetic tree of life, encompassing the total diversity represented by sequenced genomes from April 2016, source: Hug et al. 2016 [25]

Classification error

Despite the numerous new classification tools and the approximately $600 - 700$ new descriptions of bacteria and archaea published each year, we still come up short of finding an adequate solution to the classification problems. The fact that this is so hard to achieve is due to several reasons. 1) DNA sequences are complex. We try to categorize nature with classification-based approaches that make taxonomic assignments to help us understand the whole aspect of life. For further discussion on this issue see section 3.3. 2) Sequencing and sequencing error in the sequenced data. There are numerous quality-filtering techniques to correct or eliminate reads that contain sequencing errors [54, 66, 29, 49]. However, there is a long way left to obtain error-free data. 3) Databases and classification methods are constructed considering the existing diffuse taxonomy. As mentioned, the taxonomy for every organism has evolved over years and there is no consensus.

2 Outline and aim of thesis

The amount of sequence data available today gives us a solid foundation to understand the phenomena of biology, and has led to an increased need for computer-efficient and time-saving methods to extract information from data such as this. The overall aim of this PhD project has been to **improve classification of Prokaryotes**, both by presenting new insight concerning the amount of data we are exposed to and by rigorously testing and presenting different classification methods. The project also gives an indication of where the current sequencing data falls short, and points out future improvements.

Two important aspects of why it is important to improve classification of Prokaryotes are:
I) *Taxonomy*: This is what it is all about! It is the mapping and understanding of the relationships of all living organisms which allows us insight in how life on earth collaborates and functions. Figure 2 shows a suggestion for this relationship referred to as the phylogenetic tree of life, where the prokaryotes are represented by the bacteria, the big left branch pointing west and the archaea, located together with the Eukaryotes pointing south.

II) *Microbial communities*: The exploration of microbial communities has recently become a major focus in microbiology as mentioned in section 1.2. A main goal is to model the constituents of microbes within complex communities. This can be used as markers for diagnosing disease, identify body fluids in forensics and describe environments.

The 16S gene performs well in phylogenetic studies and studies of microbial communities. Due to the properties/qualities of the 16S gene, the research presented in this thesis is build solely on 16S sequencing data. The data is obtained from the three main databases: The Silva comprehensive ribosomal RNA database, the Greengenes 16S rRNA gene database and the Ribosomal Database Project.

3 Methods

After emphasizing the breakthrough in sequencing technology closely followed by the age of sequencing data, we must look at how all this data can be beneficial for understanding the phenomena of biology. How can we extract information from the hundreds of millions of A's, C's, G's and T's that we now have so easy access to? A bottleneck has arisen when it comes to method development. Computer memory and capacity are bottlenecks considering millions of sequences. Even with an excellent classification method, most computers today do not have the ability to take as input millions of full length genomes, 16S sequences, or even short reads. We need an approach with respect to both speed, memory and accuracy. Instead it is very often a question of a trade-off between these. In metagenomics, classifying sequences more efficiently is a key since the number of sequences to classify may be vast. On the other hand, since the 16S marker is used to build the entire prokaryotic taxonomy one should make all possible efforts to have the absolute best classification available, meaning accuracy trumps time-efficiency. Ultimately, a method that can input millions of sequences and rapidly return them to their correct classes is a goal.

3.1 Sequences to numbers

In order to apply classification methods to sequencing data it is crucial to convert it into meaningful numeric variables, to maintain and identify relevant information and useful signals. There are several ways to translate letters into numbers [30, 37, 22]. Two different approaches have been used in this thesis, which have proven satisfactory in conserving important information. One by considering a multiple sequence alignment (MSA) and dummy-code each symbol in the alignment into a row-vector of five binary values. The symbol `A` is coded as $(1,0,0,0,0)$, `C` as $(0,1,0,0,0)$, `G` as $(0,0,1,0,0)$, `T` as $(0,0,0,1,0)$ and the indel `-` as $(0,0,0,0,1)$. Hence, each site gives rise to five numerical (binary) variables (paper I). The other approach by consider the frequency of a short nucleotide word, known as a K-mer (paper II-IV).
After converting sequences into a matrix consisting of numerical variables there are numerous statistical tools that can be applied to gain insight and information from the data.

Multiple Sequence Alignments

Multiple sequence alignments (MSA) have been widely used in all areas of DNA and protein sequence analysis. It has several purposes, including finding interesting patterns, detecting

homology between new sequences and existing families, building phylogenetic trees and/or calculate the evolutionary distances between sequences.

MSA is generally the alignment of three or more biological sequences (protein or nucleic acid) in such a way that the sequences obtain the best match of each other. Most procedures for taxonomic studies have actually been based on alignments and reconstruction of phylogenetic trees, making use of predefined evolutionary models and relevant algorithms [35, 68, 4]. However, considering the huge amount of data available for the 16S sequences, sequence analysis based on MSA can pose problems. Firstly, the time and computational load required to align DNA sequences increases exponentially with the number of taxa analysed. Secondly, it is not trivial to fit a new sequence into the already existing alignment. Hence, the entire alignment must be redone for each new sequence. Thirdly, different algorithms give different alignment results. For 16S data the alignments will be fairly similar in the conserved regions, but show dramatic differences in the variable regions, causing the multiple alignments to be very software dependent. Another problem is that today's shotgun sequencing technology only provides fragments (short reads) of the 16S gene, making the result of the alignment methods dependent upon good assembling methods to obtain the full length gene.

In paper I three different MSA datasets with 16S sequences were downloaded from the databases Greengene, Silva, and RDP. All three databased have their unique alignment methods, making the alignments based on the same sequences differ in a distinct way.

K-mer

The limitations of MSA have resulted in that methods based on counting K-mers by sliding windows now are considered the most interesting classification approach with respect to both speed and accuracy [52, 32, 60, 65]. Every sequence is converted to a numeric vector by counting overlapping "words" of length K in the sequence. Hence, there are $D = 4^K$ possible words of length K in the DNA alphabet. Wang et al. [65] developed the RDP classifier, based on the naïve Bayes principle (see section 3.3) and a word-length of $K = 8$. The RDP classifier is close to a standard in 16S based classification. In 2011, Essential Science Indicators found Wang et. al. to be the most-cited research paper in microbiology [1]. K-mer methods are fast and do not suffer from the same uncertainties as the procedures based on evolutionary models and alignments. This way of converting sequences to numerical data is not as intuitive as evolutionary models and lack the obvious interpretation given by evolutionary distances. Positional information will also get lost with this method, but it is

very objective in its mechanism and has proven to work impeccably on obtaining information from sequencing data.

Because of its strengths we use K-mer counts as numerical variables in paper I-IV.

3.2 Supervised learning methods

The aim of a supervised machine learning method, is to build a model that makes predictions and classifications based on known information. As adaptive algorithms the methods identify patterns in data and "learns" from the observations. When you input new observations, the methods consider them and improve its predictive performance.

Specifically, a supervised learning algorithm takes a set of explanatory variables and their known response classes, and trains a model to generate reasonable predictions for the response to new data. Supervised learning methods can be divided into two broad categories:
Classification: The goal is to assign an observation to one out of a distinct set of classes. As such, the response variable is a categorical variable as in the case of taxonomic classification.
Regression: The response variables are real numbers and the goal is to predict a continuous measurement of an observation. Applications for the method include prediction of apartment prices based on size, age and location, or forecasting and financial analysis.

3.3 Classification

The main goal for a classifier is to come up with well-defined rules, that can be used for assigning new objects. Let k_i denote the classes of random feature vectors $\mathbf{x}$. The classification task is to predict k_i after observing a new $\mathbf{x}$. The quantitative outputs k_i, are disjunct categorical variables and the assumption is made that one observation belongs to only one of the classes. Therefore, a prediction based on a new feature will either be correct or wrong in a classification perspective. One of the simplest and most efficient classifiers is the *Bayes Classifier* also known as the *Minimum Misclassification Rule*. This is an easy and preferred method, but it is dependent on known class densities as well as class prior probabilities $Pr(k_i)$, which, in most cases, are not possible to obtain. The Bayes Classifier says that we classify to the most probable class, using the conditional probabilities $Pr(k_i|\mathbf{x})$.

For each class the probability for the class k_i given the new $\mathbf{x}$ is calculated in the following way:

$$Pr(k_i|\mathbf{x}) = \frac{Pr(\mathbf{x}|k_i)Pr(k_i)}{Pr(\mathbf{x})} \tag{1}$$

The probability on the left hand side in (1) is the *posterior probability* of class k_i given the observed $\mathbf{x}$, and we classify to the corresponding group k_i that $max[Pr(k_1|\mathbf{x}), Pr(k_2|\mathbf{x}), \ldots, Pr(k_i|\mathbf{x})]$. Notice that the denominator $Pr(\mathbf{x})$ does not depend on the classes k_i. Hence, the k that maximizes $Pr(k_i|\mathbf{x})$ is exactly the same k that maximizes $Pr(\mathbf{x}|k_i)Pr(k_i)$, and we can ignore $Pr(\mathbf{x})$ altogether. If the prior probabilities of every $Pr(k_i)$ are identical for all classes, we get the simple relation $Pr(k_i|\mathbf{x}) \propto Pr(\mathbf{x}|k_i)$. A way to calculate the joint probability, $Pr(\mathbf{x}|k_i)$, is to write each element of the feature vector $\mathbf{x}$ as a product of their marginal probabilities:

$$Pr(\mathbf{x}|k_i) = \prod_{j=1}^{r} Pr(x_j|k_i) \tag{2}$$

where r is the number of elements in $\mathbf{x}$.
This assumption is correct only if the elements are independent, which in most cases is a naïve assumption, but has shown to often work satisfactory [20]. This approach is known as the *naïve Bayes approach*.
Figure 3 is a visualization of the Bayes decision boundary. The line is the threshold of two classes (blue and yellow). A new observation located at the upper side of the line is classified as yellow and vice versa; all new observations located in the lower part of the plot, under the line, are now classified as blue. As one can see, this is not perfect and emphasizes the fact that the world, with all its biology, is too complex to be categorized into disjunct classes. Even though we give it our best try, will we repeatedly fail at some level.

There are a vast number of methods with various rules associated with them. In the coming sections some of the main classification methods considered in this project are presented.

K-nearest-neighbour

K-nearest-neighbour (KNN) classification is widely used to classify objects based on distances between them. The idea is simple, by finding the K training points closest in distance to the new observation, it is classified using a majority vote among this K neighbours [23]. It is widely used for all classification purposes because of its simplicity and that it in many cases has shown to work satisfactory.

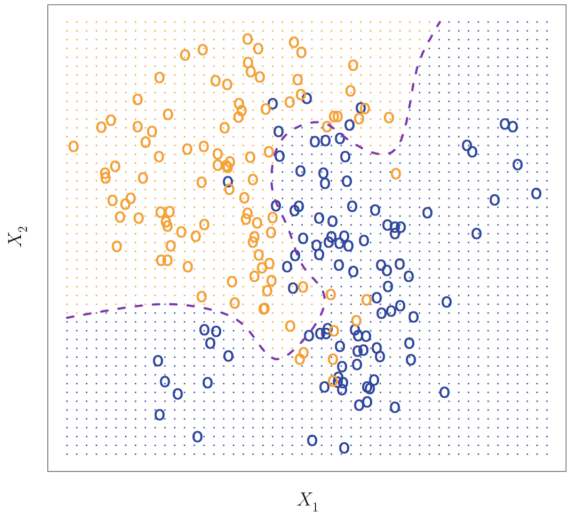


Figure 3: Example picture of the Bayes decision boundary from *"The elements of Statistical learning"* by Hastie et al. [23]

PLS

Partial Least Squares (PLS) [71] is a well-established classification method that has been used in many bioinformatic applications, including analysis of sequence data [44, 43, 2, 41, 40]. In practice PLS finds linear combinations of the explanatory (known here as the numerical) variables which gives the minimum classification error. These combinations are referred to as PLS components. In principle, all explanatory variables are included, and given more or less weight in the components. Unlike the very general KNN method, PLS needs numerical responses and is especially applicable when there are many and correlated explanatory variables. This will typically be the case for K-mer data, as K increases. In K-mer space every sequence has $D = 4^K$ numerical variables (words), and every one of these variables carries equal weight. However, it is more than likely that some of these will be more, or less, important for recognizing a particular class. Replacing the original D dimensional space by a smaller number of variables, with more emphasis on the important numerical variables, can be beneficial for classification.

The RDP-classifier and the Multinomial method, presented below, are more specific classification methods, made for the purpose of classifying nucleotide sequences by considering K-mer data.

RDP-classifier

The RDP-classifier [65] considers the presence/absence of a K-mer in a sequence. All words of length K are ordered alphabetically as $w_1, w_2, ..., w_D$. For every sequence, we create a vector of D elements where element j is 1 if K-mer w_j is present in the sequence, and 0 if not. For a new sequence we construct a similar vector and compute the posterior probabilities for each taxon using the trained model and a naïve Bayes assumption. The predicted taxon is the one producing the maximum posterior probability.

Multinomial method

The Multinomial method differs from the RDP method by considering the relative frequency of every K-mer instead of only presence/absence. The calculation steps are similar as described for the RDP-classifier above. This approach has also been tested by Lui and Wong in [31].

For a more detailed explanation of the two latter methods see paper II.

3.4 Validation

In practice most classification methods will be used to classify a new sequence not included in the training data. To obtain a realistic impression of the accuracy of classification, and to avoid over-fitting to the data, some kind of validation of the methods is needed. Over-fitting is the risk of making the trained models too complex, too "fit" for the respective training data. In this case the accuracy measure will be unnaturally high, because your model is customized to describe exactly the trained data. Classification methods built on such models will not be able to classify new sequences in a meaningful way and therefore not serve as a decent and general classification method. There are several ways to validate classification methods. Two popular ways are by external validation and cross-validation. In the case of external validation the performance of a method is measured in its prediction capability on an independent test

dataset. This is solved by splitting the dataset into one training- and one test set. Cross-Validation [62] is also frequently used. In a K-fold cross-validation the dataset is split into K segments, where $K \leq n$ and n is the number of samples. Each segment is set aside once as a validation set. The model is then trained on the remaining $K - 1$ segments and the accuracy is calculated based on the models ability to predict each K validation set. What is the optimal K is an ongoing discussion [23]. In the case where $K = 1$, one observation is removed, the model is trained on every other observation and the class/value for the excluded observation is predicted with the trained model.

4 Paper summaries

Paper I – A systematic search for discriminating sites in the 16S ribosomal RNA gene

The 16S sequence is usually divided into conserved and variable regions of the sequence dependent on similarity of the genes. Nine variable regions have been detected, named $V1$-$V9$, and in earlier research it has been claimed that the discriminatory power of the 16S marker lies within these variable regions. The location of the variable regions, and implicitly the conserved parts flanking them, has been based on multiple alignments of full-length 16S sequences. Three multiple alignments were considered, comprising the same 16S sequences, but differs in the way they were conducted. A supervised learning method was used to search systematically for sites that contribute to correct classifications at either phylum or genus level. After dummy coding each site along the alignment (including gaps), PLS was run accompanied by the variable selection algorithm, Selectivity Ratio, to detect the sites in the alignments that were significant with respect to correct classification. The site selection algorithm located 50 discriminative sites when considering classification on phylum level. Instead of being exclusively located in the variable regions, these were scattered across most of the alignments. Therefore, while variable regions are important, they are not more important than any other region. The discriminative sites are also typically sites with high entropy (variability) located among neighbouring sites of much lower entropy. Regions of lower entropy imply some degree of conservation, and alignments tend to be more accurate in such regions. When classifying genera the site selection algorithm needed around 80% of the sites in the 16S sequence before the classification error reached a minimum. This means that all variation in the 16S sequence, in both variable and conserved regions, is needed in order to separate the prokaryotes down to genus level in an optimal way.

As a result of the findings in this paper, we continued our pursuit for the ultimate classification method by considering full length 16S sequences, and fragmented sequences obtained from the entire sequence.

Paper II – Comparing K-mer based methods for improved classification of 16S sequences

There has been a shift in approaches considering classification methods. Previously, alignment-based methods were seen as the most applicable tools. Now however, methods based on counting K-mers by sliding windows are preferred due to improved speed and accuracy. In this paper we presented a systematic comparison of five different K-mer based classification methods for the 16S sequence. The study is based on the commonly known and well-used RDP-classifier, which only registers whether or whether not a K-mer is observed in a sequence. Four other methods were considered: The Multinomial method, this is quite similar to the RDP-classifier, except here we consider frequencies of K-mer rather than presence/absence, Nearest-neighbour classification (NN), the Preprocessed nearest-neighbour (PLSNN) and an ordinary Markov model. They were all implemented and tested on two different datasets, and differed from each other in data usage and modelling strategies. Due to the results obtained in paper I, all five methods were tested on full-length sequences and on fragments (of typical read-length) coming from the whole sequence. The difference in classification error generated by the methods seemed to be small, but they were stable and present for both datasets tested. The PLSNN method performed best for full-length 16S rRNA sequences, and significantly better than the RDP-classifier. On fragmented sequences the Multinomial method performed significantly better than all other methods. For the two datasets explored, and both on full-length and fragmented sequences, all five methods reached an error-plateau. The error-plateau differed between the different datasets, indicating that we need better training data to further improve classifications and classification models. Classification errors occur most frequently for genera that consists of few sequences.

We concluded that for further improvements of the taxonomy and testing of new classification methods, the need for a better and more universal training dataset is crucial.

Paper III – The ConTax data: Improved supervised learning of prokaryotic taxonomy

There is no official taxonomy for prokaryotes, which is a great hindrance when it comes to classification and the construction of reliable and stable classification methods. In paper II we concluded that in order to achieve classification improvement for prokaryotes, we must take a step back and focus on obtaining stable and correct training data before we continue the development of classification methods. We used to our advantage that there are some branches in the phylogenetic tree of life that are widely accepted - and will most likely never change - and that there are several different databases containing hundreds of thousands of 16S sequences. From three of the most popular databases; the Silva comprehensive ribosomal RNA database, the Greengenes 16S rRNA gene database, and the Ribosomal Database Project, we downloaded all the high quality, full-length 16S sequences. From this enormous amount of data we filtered out the sequences where the databases agreed on its assigned taxonomy down to genus level; where there was a taxonomy consensus. Together these sequences forms a new and improved dataset with consensus taxonomy: The `ConTax` set. The `ConTax` set comprises $664,199$ 16S rRNA sequences. From a method training perspective, the enormous number of sequences in some of the genera provides no additional information for classification purposes. Therefore, the largest genera in the `ConTax` set were trimmed in such a way that the main information is retained. The trimmed `ConTax` set consists of $38,784$ sequences. The results presented in this paper confirmed that the performance of different classification methods improves once trained and tested on the `ConTax` set, compared to other established datasets. The `ConTax` set can be seen as a seed-set of the most accurately classified sequences from which we can continue to investigate the prokaryotic taxonomy. We observed that most of the mis-classifications are caused by small taxa, drawing the conclusion that for real classification improvements to take place, an effort must be made to investigate the rarely explored and poorly represented taxa.

Paper IV – `microclass`: An R-package for 16S taxonomy classification

In paper II we found that the algorithm for the multinomial method obtained the best overall results for classification of full-length and fragmented 16S sequences at the shortest amount of time. In paper III we corrected some of the uncertainty related to diffuse taxonomy, and created a stable and most likely correct training set for taxonomic classification. Based on these findings we have constructed an R-package, `microclass`, with tools addressing classification down to genus level. Efforts have been made to make the tools both fast and memory-efficient. All methods in the package can be trained on new data, but considering our previous results we also developed a ready-to-use tool, the `taxMachine`. The `taxMachine` has been trained with the multinomial method for full-length 16S sequences to recognize full or partial (reads) sequences at the genus level, using the designed and optimized trimmed `ConTax` set, from paper III, for training. It has been optimized to produce the most accurate classifications at genus level, without consuming too much memory. One of the major benefits with this machinery is that it, together with an accurate classification for every sequence or read (even short ones), also provides some quantified uncertainties indicating if the input sequences are difficult to recognize. Based on input sequences of varying length and quality, we demonstrated how the output from the classifications can be used to obtain high quality taxonomic assignments from 16S sequences within the R computing environment. The `microclass` R-package, as well as its symbiotic data package `microcontax`, are freely available at the Comprehensive R Archive Network (CRAN, [50]).

5 Discussion and concluding remarks

The need for precise and stable taxonomic classification is crucial in modern microbiology and in this project we have made an effort to improve taxonomic classification by accuracy, computer capability and run time. We started this project by considering the nine variable regions in the 16S rRNA gene, since it is a common belief that the discriminatory power of the 16S marker lies within these nine regions and because amplicon sequencing techniques provide targeted sequence reads as output from the much discussed regions. Alignments of the 16S sequences were investigated and, by only considering the nine variable regions, we got a considerably higher error rate for our distance tests than by calculating simple p-distances across the full length 16S alignments. The p-distance is one of the simplest ways to calculate evolutionary distances, it is the proportion in nucleotides that differs between two sequences. Some discriminant information got lost in the attempt to only consider the variable regions. With this as the premise, in paper I we made a detailed examination of the input data, the 16S sequences ((i) in figure 1). We used a supervised learning method (PLS) to search systematically for sites that contribute to the correct classification and found that discriminating sites were scattered all over the sequence and not at all located solely in the variable regions, as often presumed. Another conclusion drawn from these results was that methods built on multiple sequence alignments are not compatible with the large amount of sequencing data we face today. A natural continuation was to consider methods that did not discriminate against some parts of the 16S sequence and that circumvented the alignment step. Our choice fell on one of the most popular pattern-recognition method in the literature: counting K-mers by sliding windows. The K-mer methods are fast and all parts of the sequence are included, it also evades the uncertainties and computational drawback by producing an alignment. In paper II, we presented a thorough overview of five different classification methods ((ii) in figure 1). All methods reached an error plateau indicating that in order to improving taxonomy, and test new classification methods, the need for a better and more universal training dataset is crucial. In short, it is impossible to make the ultimate classification tool without the ultimate training data. The fact that there is no universal consensus of the taxonomy makes the classification problem highly vulnerable; the same DNA sequence can have different taxonomy when considering different databases and classification methods trained on uncertain data will lead to a growing number of false taxonomy assignments. Therefore, a stable dataset where we are confident about the sequences taxonomy will be of greatly beneficial. A consensus of the taxonomical system is fundamental also for precise communication between scientists [21]. Paper III is again a concrete result of the conclusion drawn in paper II, and concerns the predetermined

taxonomic classes ((iii) in figure 1). In paper III we discussed the problem of unreliable training data more thoroughly and came up with a solution on a consensus taxonomic dataset based on the amount of data we now are exposed to. We presented the new and improved training dataset with consensus taxonomy: The `ConTax` set. It consists of sequences from three main 16S databases that are agreeing on the assigned taxonomy down to genus level. We also presented a trimmed down, more usable version of this, were the largest genera are trimmed in such a way that the main information is retained. The `ConTax` set can be seen as a seed-set or a gold standard set of the most accurately classified sequences from which we can continue to improve classification methods, as well as explore and map the prokaryotic taxonomy. Paper IV sums up our detailed work in the previous papers and present the `microclass` package which provides optimized tools for taxonomic classification of 16S sequence data in the R computing environment. In this package a ready-to-use improved classifier ((ii) in figure 1) is presented: The `taxMachine`. The `taxMachine` has been trained with the multinomial method and $K = 8$, which was the method found to conduct the best overall results in paper II. Efforts have been made to make the R-function both **speed and memory usage efficient**. It is found to be significantly better in classification accuracy and also faster than today's gold standard method: The RDP-classifier. One new, and highly important improvement of this tool is its quantifies uncertainties measures, which indicates if a new input sequence is difficult to recognize. As such, `taxMachine` **improves classification**, and is superior with respect to time and computer memory, compared with other highly used methods.

There is a growing concern about the reproducibility of scientific work. Science builds upon itself and in an attempt to investigate if there really is a reproducibility crisis, Monya Baker investigated this in Nature News Feature [5]. For the **usability and reproducibility** of some of the work executed in this project the two R-packages, `microcontax` and `microclass`, are conducted and available for free at The Comprehensive R Archive Network [50].

The work here is solely based on 16S data, making the results, for now, limited to prokaryotes. 16S is, as repeatedly mentioned, a conserved marker which can be used to identify prokaryotes down to genus level. Accurate classifications down to species level is not possible using the 16S rRNA due to the small sequence differences between species within the same genus, which creates another boundary for research based on this genomic marker. The classification methods presented in `microclass`, can be trained on other genomic markers, coding genes as well as full-length genomes. However, it is important to stress that we have optimized it for the 16S marker. If it were to be used on other sequencing data, a systematic investigation,

as in paper II, should be conducted to, for instance, optimize the word length K.

Another interesting observation worth mentioning, is the robustness of `taxMachine` with respect to short reads. It is trained on full-length 16S sequences, so the K-mer frequency for the whole sequence is considered. However, it works astonishingly well identifying shorter fragments (reads). The fact that these short fragments of 16S, which for the most part so easily can be classified, are scattered all over the gene is interesting. Amplicon-sequences obtained from region $V4$ were also tested in paper III-IV. These are of approximately 292 bases long. In paper IV the number of misclassified was reported for these versus for randomly scattered $270 - 300$ bases long fragments and the classification accuracy differs with as little as 0.021. In fact, random scattered fragments with $450 - 500$ bases outperformed the amplicon-sequences from $V4$, indicating that the length of the fragment has a greater impact on classification than the position along the 16S gene. These results support the fact that interesting and discriminating sites are located all over the 16S gene. The longer the fragment considered the more of these sites it will enclose which again highlights the findings in paper I.

16S studies go through a series of in vitro and in silico steps that can greatly influence their outcomes. For an overview of the steps regarding amplicon sequencing see figure 4. The raw sequences obtained from the PCR process are processed using bioinformatic pipelines that try their best to remove low-quality reads and detect and remove chimeric sequences. Two common ways to continue the analysis of the sequence content are: *a)* Cluster the curated sequences into operational taxonomic units (OTUs). From here one can compare OTU content of several microbial communities, or at different states, for the same community, and/or do a taxonomic classification based on a core sequence in each OTU. b) By straightforward taxonomic labelling of the sequences by rigorously comparing them against a reference database and assigning them to the taxon with the best match. Regardless of the angle of approach, it will often be a request of classification of the sequence data and as such, the need of a stable, fast and reliable classifier is crucial.

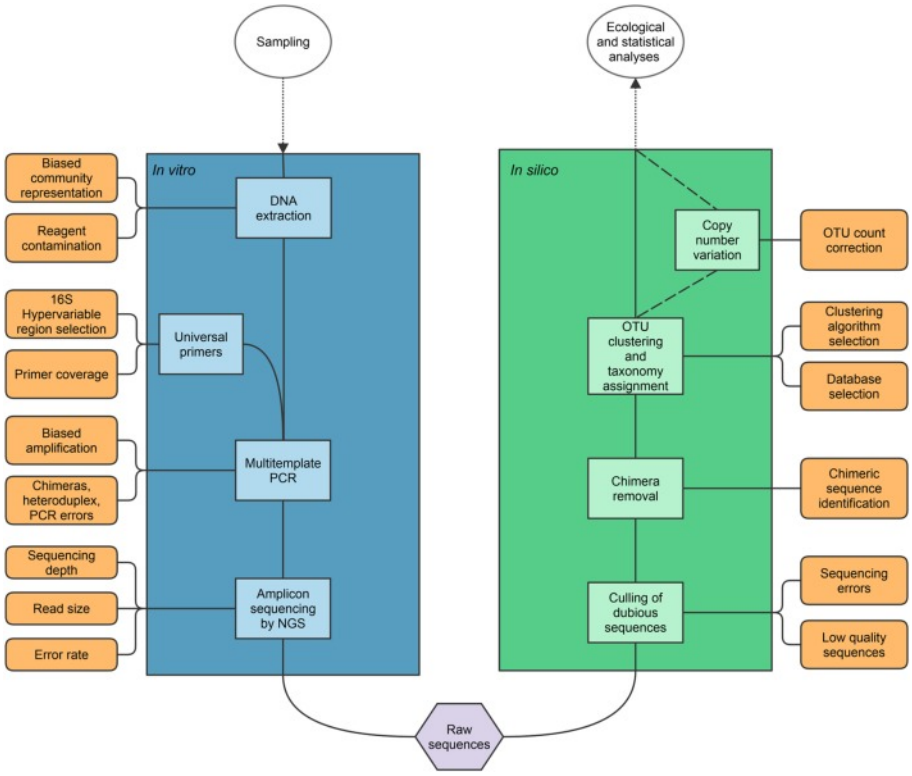


Figure 4: A flowchart of a usual work-flow for a 16S study from sampling to the beginning of the data analysis. The orange boxes highlights the problems associated with each step. (Adapted from de la Cuesta-Zuluaga et al. [17])

5.1 Future perspectives

For some parts of the taxonomy a huge amount of data is available, from other branches in the phylogenetic tree of life we have very few observations (sometimes as few as one) and some branches have we, most likely, not even seen yet. We observed that most of the misclassifications are made for small taxa. The problem of singleton taxa haunted us throughout this project and even with various attempts to smooth out the skewness we could not overcome this problem. A further challenge and request from a bioinformatics point of view, to really improve classifications from here, is to make an effort to investigate the rarely explored and poorly represented taxa.

The classification of 16S is the most basic approach to profiling a microbial community, and due to the explosion in metagenomic research activities, tools for recognizing taxa from 16S

sequences (reads) should be tuned to their optimal performance. A questionable result of K-mer methods is that position information is getting lost and an attempt to include such information can be beneficial.

Future perspectives of our new classification method, the `taxMachine`, lie in the investigation of the elusive classification, those sequences that can not easily be recognized and assigned to a class. Are they elusive because they are something new, infected with sequencing error or because they lie in the middle of two genera? The fact that the uncertainty measures, defined in paper IV, gives us an indication of whether a new sequence read is coming from the 16S gene or not, combined with the fact that K-mer frequency can be used to identify fragmented sequences scattered over the whole 16S gene, opens up a possibility to identify and classify the 16S reads using `taxMachine` directly on shotgun metagenomic DNA sequences. In the very end of this project we did a brief test on the ability the `taxMachine` has to determine or not if a short sequencing read is coming from the 16S gene. For the *Escherichia coli* genome we sampled at random 150 bases long reads inside and outside the 16S regions in the genome. Every read was used as an input in `taxMachine`. The r-score and the p-recognize, indication if the sequence is recognized in the test data or not, were investigated and a distinct separation of the 16S from "the others" was observed. Figure 5 is an illustration in how the scores distributed. It shows a distinct separation, both in r-score density, in the upper panel, and also in the lower panel where the r-score is plotted against the d-score. The horizontal and vertical lines are the boundary lines of a certain classification suggested in paper IV. If a classified sequence (the points) are located in the upper right box bounded by these lines it is defined as a certain classification. The mis-classified sequences (the red points in figure 5) are all identified by `taxMachine` and given a small d-value. The same test was tried out on several other genomes (e.g. *Acaryochloris marina*, *Carnobacterium sp.* and *Pseudomonas aeruginosa*) and they all showed the same total disunity between 16S regions and not 16S. This is a good indication that `taxMachine` may be used as a tool in shotgun metagenomic samples taking as input short reads from shotgun sequencing, recognizing the once that are 16S reads and in a fast and accurate way classify them. Considering the advantage in speed, memory and accuracy that the `taxMachine` provides, these indications are worth an extended investigation.

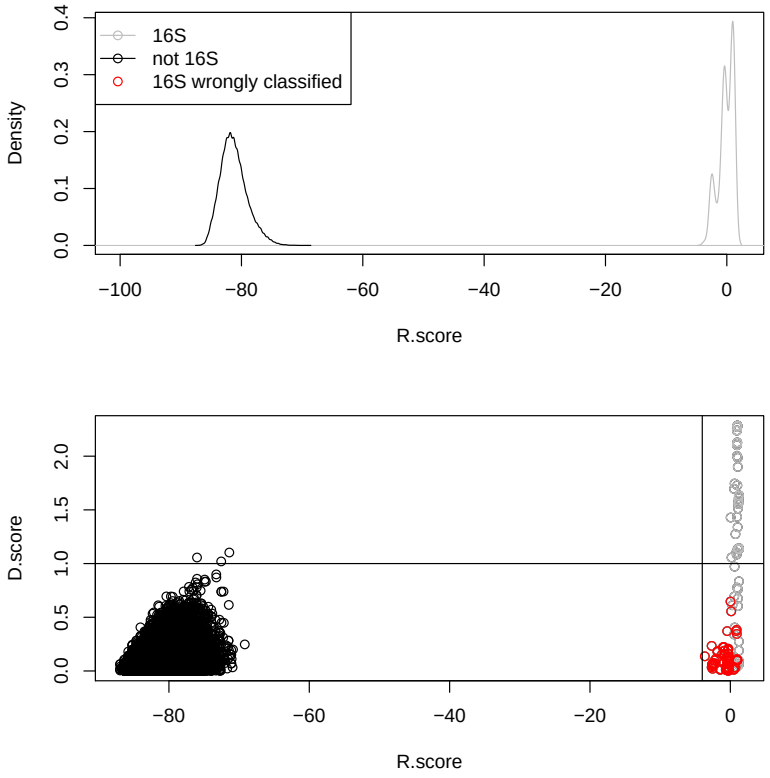


Figure 5: Randomly shotgun sequences, 150 bases long, were sampled from 16S regions and outside the 16S regions on the *Escherichia coli* genome. The short reads were used as input in the `taxMachine`. The top panel is the r-score density for both regions, and the lower panel displays the r-scores plotted against the d-score. The lines are the boundary lines of a certain classification suggested in paper IV

6 References

[1] Sience watch. `http://archive.sciencewatch.com/dr/erf/2011/11decerf/11decerfCole/`. Accessed: 2016.10.17.

[2] J Aarøe, T Lindahl, V Dumeaux, S Sæbø, D Tobin, N Hagen, P Skaane, A Lönneborg, P Sharma, and A Børresen-Dale. Gene expression profiling of peripheral blood cells for early detection of breast cancer. *Breast Cancer Research*, 12:R7, 2010.

[3] S G Acinas, L A Marcelino, V Klepac-Ceraj, and M F Polz. Divergence and redundancy of 16S rRNA sequences in genomes with multiple *rrn* operons. *Journal of bacteriology*, 186(9):2629–2635, 2004.

[4] Kolaczkowski B and J W Thornton. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature*, 431:980–984, 2004.

[5] M Baker. 1,500 scientists lift the lid on reproducibility. Nature News Feature, May 2016.

[6] D J Brenner, J T Staley, and Krieg N R. *Bergey's Manual of Systematic Bacteriology*, chapter Classification of Procaryotic Organisms and the Concept of Bacterial Speciation, pages 27–32. The Williams & Wilkins Co, 2005.

[7] J Brosius, M L Palmer, P J Kennedy, and H F Noller. Complete nucleotide sequence of a 16S ribosomal RNA gene from Escherichia coli. *Proceedings of the National Academy of Science of the United States of America*, 75:4801–4805, 1978.

[8] S Chakravorty, D Helb, M Burday, N Connell, and D Alland. A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *Journal of Microbiol Methods*, 69(2):330–339, 2007.

[9] I Cho and M J Blaser. The human microbiome: at the interface of health and disease. *nature Reviews Genetics*, 13:260–270, 2012.

[10] J E Clarridge. Impact of 16S rRNA Gene Sequence Analysis for Identification of Bacteria on Clinical Microbiology and Infectious Diseases. *Clinical Microbiology*, 17:840–862, 2004.

[11] J R Cole, Q Wang, E Cardenas, J Fish, B Chai, R J Farris, A S Kulam-Syed-Mohideen, D M McGarrell, T Marsh, G M Garrity, and J M Tiedje. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Research*, 37:D141–D145, 2008.

[12] Human Microbiome Project Consortium. A framework for human microbiome research. *Nature*, 486(7402):215–21, 2012.

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

[14] E K Costello, C L Lauber, M Hamady, N Fierer, J I Gordon, and R Knight. Bacterial community variation in human body habitats across space and time. *Science*, 326:1694–1697, 2009.

[15] S T Cowan. Sense and nonsense in bacterial taxonomy. *Journal of general microbiology*, 67:1–8, 1971.

[16] T P Curtis and W T Sloan. Microbiology. Exploring microbial diversity–a vast below. *Science*, 26(309(5739):1331–1333, 2005.

[17] J de la Cuesta-Zuluaga and J S Escobar. Considerations For Optimizing Microbiome Analysis Using a Marker Gene. *Frontiers in Nutrition*, 3:26, 2016.

[18] E F DeLong and N R Pace. Environmental diversity of bacteria and archaea. *Systematic Biology*, 50(4):470–278, 2001.

[19] T Z DeSantis, P Hugenholtz, N Larsen, M Rojas, E L Brodie, K Keller, T Huber, D Dalevi, P Hu, and G L Andersen. Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. *Appl Environ Microbiol*, 72:5069–5072, 2006.

[20] P Domingos and M Pazzani. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29:103–130, 1997.

[21] H C J Godfray. Challenges for taxonomy. *Nature*, 417:17–19, 2002.

[22] J Han, M Kamber, and J Pei. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2011.

[23] T Hastie, R Tibshirani, and Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science+Business Media, 2009.

[24] S Hong, Bunge J, C Leslin, S Jeon, and S S Epstein. Polymerase chain reaction primers miss half of rRNA microbial diversity. *ISME Journal*, 3(12):1365–1373, 2009.

[25] L A Hug, B J Baker, K Anantharaman, C T Brown, A J Probst, C J Castelle, C N Butterfield, A W Hernsdorf, Y Amano, K Ise, Y Suzuki, N Dudek, D A Relman, K M Finstad, R Amundson, B C Thomas, and J F Banfield. A new view of the tree of life. *Nature Microbiology*, 1:16048, 2016.

[26] P Hugenholtz, A Skarshewski, and D H Parks. Genome-Based Microbial Taxonomy Coming of Age. *Cold Spring Harbor Perspectives in Biology*, 8(6):a018085, 2016.

[27] E Kellenberger. Exploring the unknown - The silent revolution of microbiolog. *EMBO reports*, 2(1):5–7, 2001.

[28] K T Konstantinidis and J M Tiedje. Towards a Genome-Based Taxonomy for Prokaryotes. *Journal of Bacteriology*, 187(18):6258–6264, 2005.

[29] J J Kozich, S L Westcott, N T Baxter, S K Highlander, and P D Schloss. Development of a Dual-Index Sequencing Strategy and Curation Pipeline for Analyzing Amplicon Sequence Data on the MiSeq Illumina Sequencing Platform. *Applied and Environmental Microbiology*, 79(17):5112–5120, 2013.

[30] D Kudenko and H Hirsh. Feature Generation for Sequence Categorization. *In Proceedings of the Fifteenth National Conference on Artifical Intelligence*, 1:733–738, 1998.

[31] KL Liu and TT Wong. Naïve Bayesian Classifiers with Multinomial Models for rRNA Taxonomic Assignment. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 10(5):1334–9, 2013.

[32] Z Liu, GL DeSantis TZ, Andersen, and Knight R. Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. *Nucleic Acids Research*, 36:e120, 2008.

[33] R Logares, S Sunagawa, G Salazar, F M Cornejo-Castillo, I Ferrera, H Sarmento, P Hingamp, H Ogata, C de Vargas, G Lima-Mendez, J Raes, J Poulain, O Jaillon, P Wincker, S Kandels-Lewis, E Karsenti, P Bork, and S G Acinas. Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. *Environmental Microbiology*, 16(9):2659–2671, 2014.

[34] W Ludwig and H Klenk. Overview: a phylogenetic backbone and taxonomic framework for prokaryotic systematics. *Bergey's Manual of Systematic Bacteriology*, 1:49–65, 2000.

[35] W Ludwig, O Strunk, S Klugbauer, N Klugbauer, M Weizenegger, J Neumaier, M Bachleitner, and K H Schleifer. Bacterial phylogeny based on comparative sequence analysis. *Electrophoresis*, 19(4):554–68, 1998.

[36] D S Lundberg, S L Lebeis, S H Paredes, S Yourstone, J Gehring, S Malfatti, J Tremblay, A Engelbrektson, V Kunin, T G del Rio, R C Edgar, T Eickhorst, R E Ley, P Hugenholtz, S G Tringe, and J L Dangl. Defining the core Arabidopsis thaliana root microbiome. *Nature*, 488(7409):86–90, 2012.

[37] B G Ma. How to describe genes: enlightenment from the quaternary number system. *Biosystems*, 90(1):20–27, 2007.

[38] E A McCliment, K M Voglesonger, P A O'Day, E E Dunn, J R Holloway, and S C Cary. Colonization of nascent, deep-sea hydrothermal vents by a novel Archaeal and Nanoarchaeal assemblage. *Environmental Microbiology*, 8(1):114–125, 2006.

[39] J A McLean, M Lombardo, J H Badger, A Edlund, M Novotny, J Yee-Greenbaum, N Vyahhi, A P Hall, Y Yang, C L Dupont, M G Ziegler, H Chitsaz, A E Allen, S Yooseph, G Tesler, P A Pevzner, R M Friedman, K H Nealson, J C Venter, and R S Lasken. Candidate phylum TM6 genome recovered from a hospital sink biofilm provides genomic insights into this uncultivated phylum. *Proceedings of the National Academy of Science*, 110(26):E2390–E2399, 2013.

[40] T Mehmood, J Bohlin, A B Kristoffersen, S Sæbø, J Warringer, and L Snipen. Exploration of multivariate analysis in microbial coding sequence modeling. *BMC Bioinformatics*, 13:97, 2012.

[41] T Mehmood, H Martens, S Sæbø, J Warringer, and L Snipen. Mining for genotype-phenotype relations in Saccharomyces using partial least squares. *BMC Bioinformatics*, 12(318):318, 2011.

[42] B Mole. Microbiome research goes without a home. *Nature*, 500:16–17, 2013.

[43] D V Nguyen and D M Rocke. Multi-class cancer classification via partial least squares with gene expression profiles. *Bioinformatics*, 18:1216–1226, 2002.

[44] D V Nguyen and D M Rocke. Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, 18:39–50, 2002.

[45] S Nikolaki and G Tsiamis. Microbial Diversity in the Era of Omic Technologies. *BioMed Research International*, 2013:958719, 2013.

[46] Committee on Metagenomics: Challenges, Functional Applications; Board on Life Sciences; Division on Earth, and Life Studies; National Research Council. *The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet.* National Academies Press (US), 2007.

[47] N R Pace. A molecular view of microbial diversity and the biosphere. *Science*, 276:734–740, 1997.

[48] E Pruesse, C Quast, K Knittel, B Fuchs, W. Ludwig, J Peplies, and F O Glöckner. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research*, 35:7188–7196, 2007.

[49] F Puente-Sánchez, J Aguirre, and V Parro. A novel conceptual approach to read-filtering in high-throughput amplicon sequencing studies. *Nucleic Acid Research*, 44(4):e40, 2015.

[50] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

[51] R Roselló-Móra. Towards a taxonomy of Bacteria and Archaea based on interactive and cumulative data repositories. *Environmetnal Microbiology*, 14:318–334, 2011.

[52] K Rudi, M Zimonja, and T Næs. Alignment-independent bilinear multivariate modelling (AIBIMM) for global analyses of 16S rRNA gene phylogeny. *International Journal of Systematic and Evolutionary Microbiology*, 56:1565–1575, 2006.

[53] P D Schloss. Microbiology: an integrated view of the skin microbiome. *Nature*, 514:44–45, 2014.

[54] P D Schloss and S L Westcott. Assessing and Improving Methods Used in Operational Taxonomic Unit-Based Approaches for 16S rRNA Gene Sequence Analysis. *Appl. Environ. Microbiol.*, 77(10):3219–3226, 2011.

[55] R Schmieder and R Edwards. Insights into antibiotic resistance through metagenomic approaches. *Future Microbiology*, 7(1):73–89, 2012.

[56] T J Sharpton. An introduction to the analysis of shotgun metagenomic data. *Frontiers in Plant Science*, 5:209, 2014.

[57] Kembel SW Ladau J O'Dwyer JP Green JL Eisen JA Pollard KS. Sharpton TJ1, Riesenfeld SJ. PhylOTU: a high-throughput procedure quantifies microbial community diversity and resolves novel taxa from metagenomic data. *Plos Computational Biology*, 7(1):e1001061, 2011.

[58] A Siegl, J Kamke, T Hochmuth, J Piel, M Richter, C Liang, T Dandekar, and U Hentschel. Single-cell genomics reveals the lifestyle of Poribacteria, a candidate phylum symbiotically associated with marine sponges. *The ISME Journal*, 5:61–70, 2011.

[59] Z Smith, A E McCaig, J R Stephen, T M Embley, and J I Prosser. Species diversity of uncultured and cultured populations of soil and marine ammonia-oxidizing bacteria. *Microb Ecol*, 42:228–237, 2001.

[60] DA Soergel, N Dey, R Knight, and SE Brenner. Selection of primers for optimal taxonomic classification of environmental 16S rRNA gene sequences. *The ISME Journal*, 6:1440–4, 2012.

[61] R M Soo, S A Wood, J J Grzymski, I R McDonald, and S C Cary. Microbial biodiversity of thermophilic communities in hot mineral soils of Tramway Ridge, Mount Erebus, Antarctica. *Environmental Microbiology*, 11(3):715–728, 2009.

[62] M Stone. Cross-validatory choice and assesment of statistical predictions. *Journal of the Royal Statistical Society, Serie B-Methodological*, 36:111–147, 1974.

[63] Y Van de Peer, S Chapelle, and R De Wachter. A quantitative map of nucleotide substitution rates in bacterial rRNA. *Nucleic Acids Research*, 24:3381–3391, 1996.

[64] J C Venter, K Remington, J F Heidelberg, A L Halpern, D Rusch, J A Eisen, D Wu, I Paulsen, K E Nelson, W Nelson, D E Fouts, A Levy, A H Knap, M W Lomas, K Nealson, O White, J Peterson, J Hoffman, R Parsons, H Baden-Tillson, C Pfannkoch, Y Rogers, and H O Smith. Environmental Genome ShotguncSequencing of the Sargasso Sea. *Science*, 304(5667):66–74, 2004.

[65] Q Wang, G M Garrity, J M Tiedje, and J R Cole. Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. *Applied and enviromental Microbiology*, 73:5261–5267, 2007.

[66] X V Wang, N Blades, J Ding, R Sultana, and G Parmigiani. Estimation of sequencing error rates in short reads. *BMC Bioinformatics*, 13(1):1, 2012.

[67] WB Whitman. Bergey's manual of systematics of archaea and bacteria. Online ISBN: 9781118960608, April 2015.

[68] C R Woese. Bacterial evolution. *Syst Appl Microbiol*, 51:221–271, 1987.

[69] C R Woese and G E Fox. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci U S A*, 74(11):5088–90, 1977.

[70] C R Woese, E Stackebrand, T J Macke, and G E Fox. A phylogenetic definition of the major eubacterial taxa. *Syst Appl Microbiol*, 6:143–151, 1985.

[71] S Wold, H Martens, and H Wold. The Multivariate Calibration Problem in Chemistry solved by the PLS Method. *Lecture Notes in Mathematics*, 973:286–293, 1983.

[72] T Yatsunenko, F E Rey, M J Manary, I Trehan, M G Dominguez-Bello, M Contreras, M Magris, G Hidalgo, R N Baldassano, A P Anokhin, A C Heath, B Warner, J Reeder, J Kuczynski, J G Caporaso, C A Lozupone, C Lauber, J C Clemente, D Knights, R Knight, and J I Gordon. Human gut microbiome viewed across age and geography. *Nature*, 486(7402):222–227, 2012.

[73] Y Zhou, H Gao, K A Mihindukulasuriya, P S La Rosa, K M Wylie, T Vishnivetskaya, M Podar, B Warner, P I Tarr, D E Nelson, J D Fortenberry, M J Holland, S E Burr, W D Shannon, E Sodergren, and G M Weinstock. Biogeography of the ecosystems of the healthy human body. *Genome Biology*, 14:R1, 2013.

Paper I

RESEARCH Open Access

A systematic search for discriminating sites in the 16S ribosomal RNA gene

Hilde Vinje[1*], Trygve Almøy[1], Kristian Hovde Liland[1,2] and Lars Snipen[1]

Abstract

Background: The 16S rRNA is by far the most common genomic marker used for prokaryotic classification, and has been used extensively in metagenomic studies over recent years. Along the 16S gene there are regions with more or less variation across the kingdom of bacteria. Nine variable regions have been identified, flanked by more conserved parts of the sequence. It has been stated that the discriminatory power of the 16S marker lies in these variable regions. In the present study we wanted to examine this more closely, and used a supervised learning method to search systematically for sites that contribute to correct classification at either the phylum or genus level.

Results: When classifying phyla the site selection algorithm located 50 discriminative sites. These were scattered over most of the alignments and only around half of them were located in the variable regions. The selected sites did, however, have an entropy significantly larger than expected, meaning they are sites of large variation. We found that the discriminative sites typically have a large entropy compared to their closest neighbours along the alignments. When classifying genera the site selection algorithm needed around 80% of the sites in the 16S gene before the classification error reached a minimum. This means that all variation, in both variable and conserved regions, is needed in order to separate genera.

Conclusions: Our findings does not support the statement that the discriminative power of the 16S gene is located only in the variable regions. Variable regions are important, but just as many discriminative sites are found in the more conserved parts. The discriminative power is typically found in sites of large variation located inside shorter regions of higher conservation.

Background

The use of stable parts of the genomic content as an evolutionary marker was a breakthrough for microbial studies in the 1980s [1,2]. The 16S small ribosomal subunit gene (16S rRNA) is today considered the gold standard for phylogenetic studies of microbial communities and for assigning taxonomic names to bacteria [3-5]. There are several properties of the 16S gene that has made it useful as a taxonomic target. First, the 16S gene is present in all bacteria. Second, it contains regions resistant to prokaryotic evolution [2]. This has made it possible to recognize the 16S without too much problems in most genomes. Third, and most important to this study, the 16S gene also includes some variable regions in between the more conserved parts. Nine such regions were once identified and named V1-V9 [6] from the sequence data available at that time. Based on the data sets of those days, it was concluded that the conserved regions are too conserved to be useful for discriminating between taxa, and that the variable regions are the key to classification of prokaryotes. Some later studies [7,8] have also confirmed these results, establishing a dogma in the use of 16S sequence data: The information separating taxa is found in the variable regions of the 16S gene.

The location of the variable regions, and implicitly the conserved parts flanking them, has been based on some multiple alignment of more or less full-length 16S genes. Van de Peer et al. [6] used distances between sequences together with the specific nucleotide substitution rate for each position to identify the variable regions. Another approach is to compute the entropy for each position in the alignment [9], and conserved/variable regions correspond to low/high entropy.

*Correspondence: hilde.vinje@nmbu.no
[1]Department of Chemistry, Biotechnology and Food Sciences, Norwegian University of Life Sciences, Ås N-1432, Norway
Full list of author information is available at the end of the article

The conserved parts are used to locate the marker gene, either *in silico* in a sequence of genomic DNA, or more commonly, *in situ* by polymerase chain reaction (PCR) amplification [10] based on primers matching these conserved parts. The first sets of primers were named according to their positions on *Escherichia coli* 16S rRNA [11]. Over the years many publications have been devoted to improving these primers [12,13].

In recent years it has been discovered that the conserved parts are not in fact as conserved as once conceived, and that there are really no such thing as universal PCR-primers that will sample equally well in all branches of the tree-of-life [14-16]. A recent study by Mizrahi-Man *et al.* [17] consider, among other things, how well the various variable regions are suited for classification. Still, these investigations all have in common that they first fix a set of primers, and then look at the regions between the primer-matching sites to see if the corresponding sub-sequences discriminate well or not. In this article we want to examine the whole length of the 16S gene, and to see if mining in the huge set of available 16S sequences can tell us something about where the discriminating sites are located, without any constraints with respect to primer matching sites.

We approach this problem by classifying the 16S sequences using a multivariate method and data consisting of multiple alignments. We conduct a systematic search for the best discriminative sites along the alignments. We use high-quality data from the databases Greengenes [18], the Ribosomal Database Project (RDP) [19] and SILVA [20]. The aim of this study is to investigate where the most discriminative sites in the 16S marker gene are located, more specifically if they correspond to variable or conserved regions.

Methods

Data

Data were downloaded from three databases; Greengenes [21], RDP [22] and SILVA [23]. From Greengenes we downloaded the alignment of isolated named strains, containing 117 101 sequences over 7682 positions. From RDP we downloaded all bacterial sequences marked as good quality and with at least 1200 bases which resulted in an alignment containing 1 151 913 sequences over 22 721 positions. From SILVA we downloaded the archived alignment named `SSURef_111_NR_tax_silva_trunc_aligned` containing 286 858 sequences over 45 984 positions.

From all alignments we discarded sequences less than 1200 bases long, sequences having alien characters (not `A,C,G,T` or `-`) and sequences not classified to one of the 2074 bacterial genera listed in the List of Prokaryotic names with Standing in Nomenclature (LPSN, http://www.bacterio.cict.fr/). We also discarded duplicated sequences. This resulted in a reduced alignment of high-quality data from each database, see Table 1.

Finally, we focused on the subset of sequences found in all three databases, i.e. the intersection between the databases. In order to obtain a consensus-based class label for all sequences, we also discarded sequences assigned to different genera in the three databases. We were then left with 12362 sequences found in all three databases, see Table 1. For each of these sequences both the assigned phylum and genus were recorded as two alternative class labels. Figure 1 shows the distribution of phyla in this data set.

When performing the systematic search for discriminating sites, phyla with less than 25 sequences were discarded, leaving us with data for 11 phyla and a total of 12270 sequences in the data set. When using genus as response, we required at least 10 sequences in each genus, resulting in 198 distinct genera (9948 sequences).

Entropy

To relate sites in the three alignments to each other, and to conserved/variable regions, we computed the entropy for each site in each alignment. This approach has also been used in previous studies (e.g. [9]). For all three alignments all sites consisting of less than 30 `A, C, G` and `T` were discarded as these provided too little data. At each remaining site k we computed the entropy

$$H_k = -\sum_{i=1}^{4} p_i \log(p_i) \tag{1}$$

where p_1, p_2, p_3 and p_4 are the empirical proportions of the four bases appearing at position k.

In order to visually identify the regions of high/low entropy, this entropy was smoothed across positions using a centered moving average of length 51. Figure 2 is a visualisation of this from the three different alignments. Note, the position specific entropy from (1) was used in the subsequent analysis, the smoothing was only used to illustrate.

Table 1 Overview of data

Database	Downloaded	Filtered	Intersection
Greengenes	117101 × 7682	74928 × 3664	12362 × 3166
RDP	1151913 × 22721	135120 × 16686	12362 × 4084
SILVA	286858 × 45984	111914 × 13172	12362 × 4230

Each cell shows the number of sequences × the number of positions of each alignment. Downloaded are the original alignments, Filtered means after filtering of high-quality data (see text) and Intersection is the subset of sequences common to all databases.

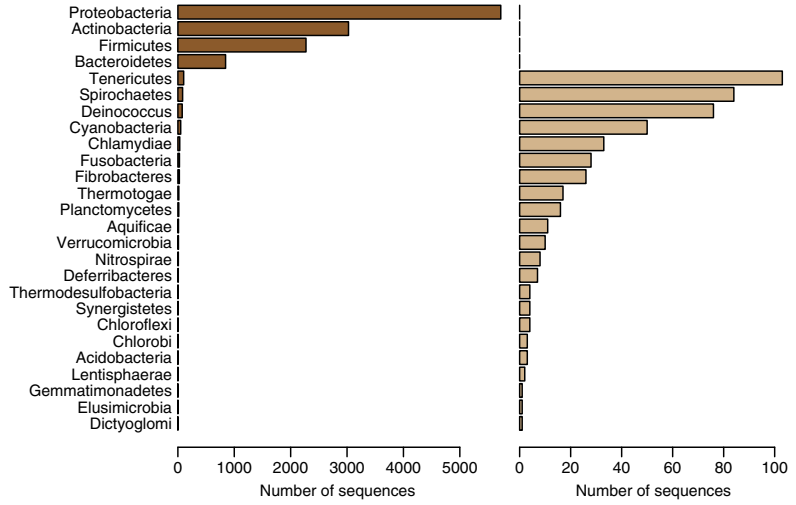


Figure 1 Distribution of phyla in data set. The final 12362 sequences are classified into 26 different phyla in the data set. The left panel shows the distribution among these phyla. The right panel gives a more detailed picture for the smaller numbers, with the four most common phyla ignored.

Site selection algorithm

In order to search for discriminating sites along the 16S alignments in a systematic way, we implemented a supervised learning approach. The input data to the supervised learning method are one of the three alignments previously described and the class-labels for each sequence in the alignment. We have used the Partial Least Squares (PLS) method [24], which is one in a long list of supervised learning methods. PLS is well established and has been used in many bioinformatics applications, also for the analysis of sequence data [25,26]. PLS is especially applicable when there are many correlated explanatory variables. This will typically be the case for the present data since the explanatory variables are in our case the sites in the alignments, and many sites along the alignment will have similar base compositions giving high correlations.

All three alignments were considered one at a time. Each site in the alignment contains a column with the symbols `A,C,G,T` or `-`. In order to use the supervised learning method we coded each symbol into a row-vector of five binary values. The symbol `A` was coded as $(1,0,0,0,0)$, `C` as $(0,1,0,0,0)$, `G` as $(0,0,1,0,0)$, `T` as $(0,0,0,1,0)$ and the indel `-` as $(0,0,0,0,1)$. Thus, each $N \times 1$ column of symbols in the alignment gives rise to a $N \times 5$ matrix of binary values to be used in the PLS-algorithm. Where N is the number of sequences. We use the term *variable* instead of *site* below, but each site actually gives rise to five numerical (binary) variables.

The response variable is in this case the class labels, and this was also coded in a similar way, using one bit for each class. As an example, when using phylum as response, the single $N \times 1$ column containing 11 different phyla was translated into an $N \times 11$ matrix of binary values, where Proteobacteria corresponds to $(1,0,0,0,0,0,0,0,0,0,0)$, Firmicutes to $(0,1,0,0,0,0,0,0,0,0,0)$ etc.

Being a multivariate method, PLS finds combinations of the explanatory variables giving the minimum classification error. These combinations are referred to as PLS components. In principle, all explanatory variables are included, and given more or less weight in the components. Variable selection means we intend to select only a subset of the original explanatory variables, and then combine these to achieve the best possible discrimination. There are many approaches to variable selection under the PLS paradigm [27], and for this application we have chosen the Selectivity Ratio (SR) score as the criterion. The SR-score is the ratio of explained variance to residual variance for each variable. This represents a measure of the ability to discriminate between the classes. High SR-score for a variable means it contains information about the classes and can discriminate between these in a good way [28].

The site-selection algorithm contained the following steps:

1. A 10-fold cross validation was first used to find the optimal number of PLS-components needed to classify the given response with the minimum obtainable error.
2. A PLS regression model was fitted to the full data set, with the fixed number of components from Step 1, to

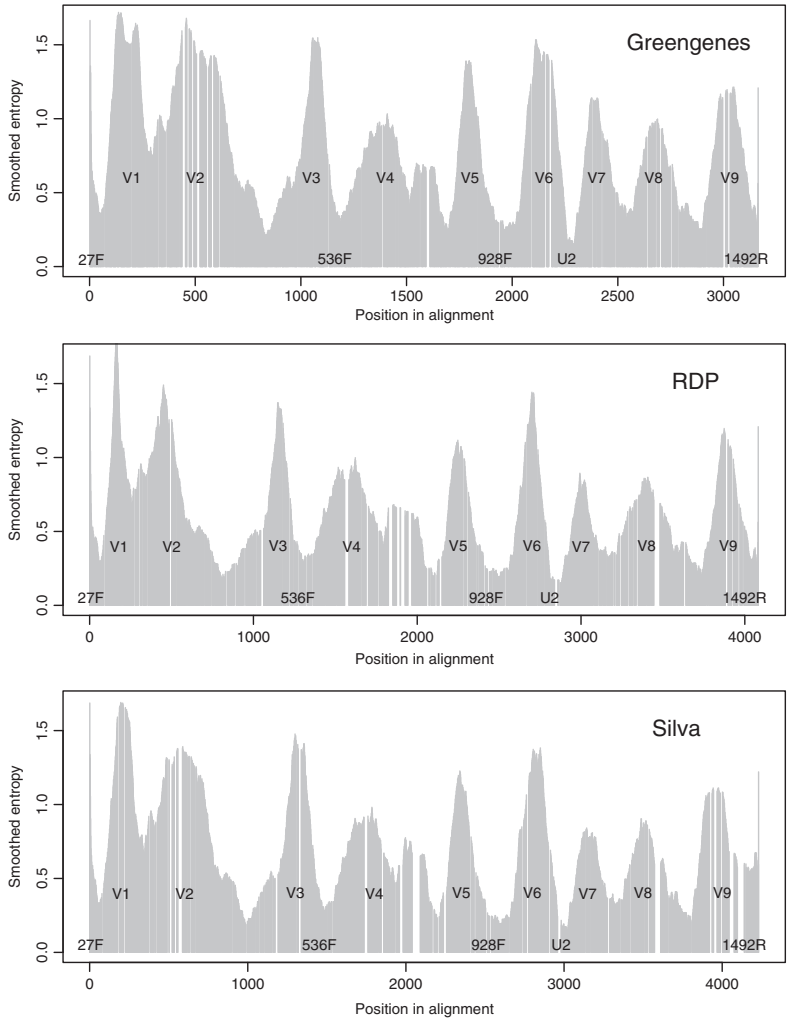


Figure 2 Smoothed entropy. The three panels show the smoothed entropy for the Greengenes, RDP and SILVA alignments covering the same 12362 sequences in this study. Positions with less than 30 bases have no entropy here, hence the 'holes' at some positions. Notice the difference in the number of positions, Greengenes being the shortest and SILVA the longest alignment. The nine variable regions V1,..., V9 are indicated for each alignment. Five examples of primers (27F, 536F, 928F, U2 and 1492R) used for PCR amplification of 16S are also marked along the position axis, indicating where they frequently match.

obtain regression coefficients for all explanatory variables. For every explanatory variable the selectivity ratio.

3. For every explanatory variable the selectivity ratio was calculated based on the regression coefficients from 2. Due to the coding, each site in the alignment corresponds to five SR-scores. The maximum of these five SR-scores was used as a site specific SR-score.
4. These site specific SR-scores were sorted in descending order; the largest SR-score corresponding to the most interesting sites. One by one the sites were included in the final model, and a 10-fold cross validation was again conducted to estimate a classification error. The final choice of how many sites to include was based on this classification error.

Results and discussion

We extracted 12362 unique sequences from the three databases Greengenes, RDP and SILVA, all having at least 1200 bases, no alien characters, found in all three

databases and with identical assignment to genus. This consensus data set must be considered a high-quality data set for 16S sequences, and an overview is given in Table 1. The three databases provide alignments of these sequences, and Figure 2 shows the smoothed entropy in each case. The three alignments differ, specifically the number of sites are different, which is due to a differing number of gaps introduced. However, the smoothed entropy shows a fairly similar pattern in all cases, and nine peaks can, with some good will, be identified. We emphasize that the grey bars in Figure 2 shows the *smoothed* entropy in order to display the regions. The actual entropy at the various sites fluctuates much more, as we will come back to below.

Instead of focusing on conserved or variable sites, we used the PLS supervised learning method to extract the sites giving the best possible discrimination regardless of where they may be along the alignment. First, we used phylum as response, i.e. there are 11 distinct classes, and for each of the three alignments (Greengenes, RDP and SILVA) we employed the site selection algorithm.

Figure 3 is an illustration of the selected discriminative sites together with the smoothed entropy from Figure 2. For all three alignments we ended up with 50 selected sites. The coloured bars indicate the selected sites. The height of a bar is the (log-transformed) SR-score, i.e. the tallest bars indicate the most discriminative sites. The color shows which symbol had the largest discriminatory power at the respective site. As an example, the leftmost bar is red, meaning the majority (but not necessarily all) of the information at this site is connected to whether a sequence has an `A` or not an `A` at this position. The three panels in Figure 3 are the results for the three different alignments. Despite the differences between the alignments, the selected sites are remarkably similar with respect to the variable and conserved regions. The largest single SR-score is the site indicated by the tallest blue bar. If we compare its location to the entropy in the background, we find it at the left hand side of region V4 in all three cases. Since both relative location and the colors of the selected sites are similar for the three panels, the results of the selection algorithm are stable with respect to the different alignments.

The first impression given by Figure 3 is that the selected sites are scattered across almost the entire alignment, there are no specific regions where they tend to cluster. As

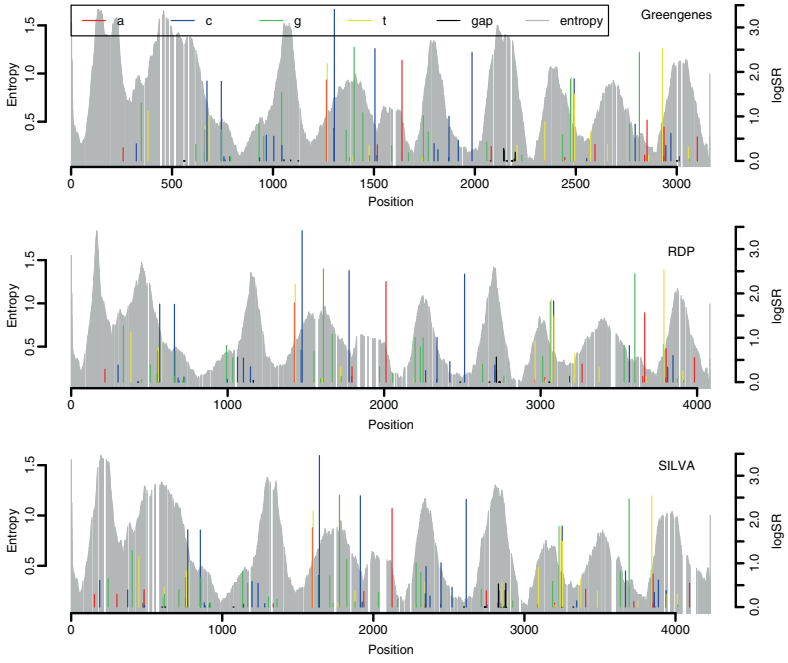


Figure 3 Selected discriminative sites. The selected sites for classification of phyla are plotted as coloured bars along the alignments. The height of a bar is the log-transformed SR-score (right hand vertical axis). The colors indicate which of the symbols `A`, `C`, `G`, `T` or `-` have the dominating discriminating power at the respective site, see legend. The grey bars in the background show the smoothed entropy values (left hand vertical axis) at each site as in Figure 2.

Table 2 Overview of the positions of the selected sites

Database	V1	V2	V3	V4	V5	V6	V7	V8	V9	Outside
Greengenes	0	7	2	6	3	1	3	0	2	26
RDP	0	7	1	6	3	3	1	0	1	28
SILVA	2	6	1	6	4	3	1	0	1	26

Each cell shows the number of selected sites for phylum classification found inside the variable regions V1-V9 for the three data sets. The rightmost column, named Outside, are the number of selected sites outside the variable regions. The total number of selected sites are 50 in each case.

shown in Figure 2 we can identify the nine variable regions in each of the three alignments. By manual inspection we found their boundaries, and Table 2 shows the number of selected sites in each. Most notably is that around half of the 50 selected sites are outside the variable regions. The variable regions cover roughly half of all the positions in the alignments, hence a selected discriminative site is just as likely to occur outside as inside of these regions. From Table 2 we also see that regions V2 and V4 contain many selected sites, while V8 has none in all three cases. Regions differ in width, and V4 has most selected sites per position.

Even if selected sites are both inside and outside of variable regions, their actual site-specific entropy from eq. (1) are in all cases significantly above the average entropy for the entire alignment. This was tested by a simple permutation test, and the results are displayed in the left panel of Figure 4. The histogram shows the average entropy for 50 randomly sampled sites (repeated 10 000 times) in the Greengenes alignment, and the red bar marks the average for the 50 sites selected by PLS. Clearly, the selected sites have a mean entropy (1.23) which is much larger than what we expect at random (histogram), giving a p-value $p < 0.0001$ here. The point is that selected sites have high entropy, but are not necessarily located in high-entropy regions. In fact, they tend to have much higher entropy than their surrounding sites, which is shown in the right panel of Figure 4. Here we computed the difference between the entropy of a selected site and its 10 neighbouring sites at each side. For the Greengenes data this resulted in the average difference 0.57 marked by the red bar. The histogram is again the result of a permutation test (10 000 permutations) where the same difference has been computed for randomly sampled sites. The results of Figure 4 were very similar for the RDP and Silva alignments, and are not shown here.

Figure 5 presents some detailed results for phylum classification based on the Greengenes alignment, again the results turned out similar for the RDP and SILVA alignment. Panel A (top left) shows how the number of mis-classifications decreases by including more selected variables, and converging at around 100 errors, giving an accuracy of over 99%. The other five panels visualize sequences in PLS-plots. Every point represents a sequence and the coordinate axes represent the optimal combinations selected by PLS (PLS components). Sequences located near each other are aligned similarly, at least in the discriminative sites. The colors represent the true classes (phyla). The first components separate the large classes, and it is not until the 10th component that smaller groups are separated. In panel B of Figure 4, we can see some obvious mis-classifications. Some black dots (supposedly Proteobacteria) are found in the center cloud of yellow (Firmicutes). This must be due to either alignment errors or sequences assigned to the wrong class from the beginning. In order to construct the huge alignments we use here, greedy algorithms of some kind are required. This means errors accumulate, and alignments of this size will most likely contain a substantial number of errors. Structure-based alignment methods should perform better for RNA-sequences. The RDP alignment we use here is based on the Infernal software

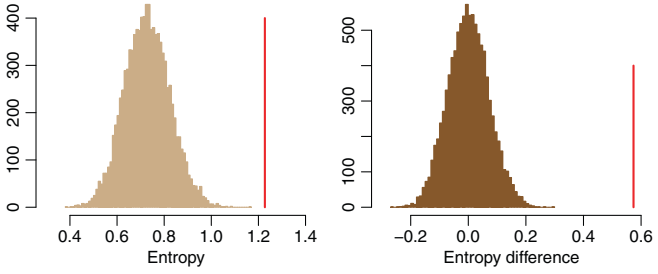


Figure 4 Entropy of selected sites. The left panel shows that mean entropy of the selected sites compared to random samples. The vertical red bar marks the mean entropy of the 50 selected sites, at 1.23. The histogram is constructed by sampling 50 random positions, computing their mean entropy, and repeating this 10 000 times. The right panel shows the mean difference between the entropy of a selected site and its 20 neighbors (10 on each side). Again the red bar marks this difference for the 50 selected sites and the histogram displays the same difference for 50 sites sampled at random, repeated 10 000 times. This figure is based on the Greengenes data, but the RDP and SILVA data gave similar results.

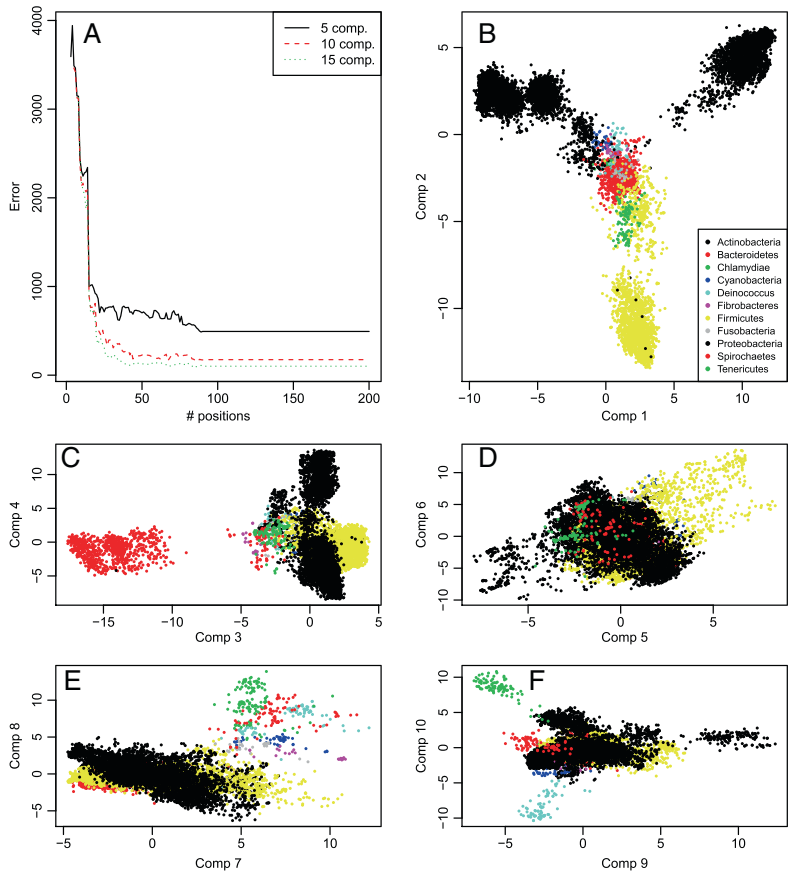


Figure 5 Details of phylum classification. Panel **A** shows how the number of mis-classifications drops as more and more sites are selected by the site selection algorithm. The error levels out at around 100 mis-classifications, and 50 selected sites seems to be enough to achieve this error rate. Panels **B-F** are PLS-plots of the sequence data, and the various panels show the same data from different perspectives. In panel **B** we plot the data in a coordinate system spanned by PLS-component 1 and 2, in panel **C** it is spanned by component 3 and 4 and so on. Each dot corresponds to a sequence, and the colors represent the true class label for each sequence, indicated by the legend in panel **B**. This figure is based on the Greengenes data, but the RDP and SILVA data gave similar results.

[29], but still we find a number of mis-classifications. These errors constitutes a significant source of the classification errors we observe. In fact, the methods most frequently used for classification are those based on word-frequencies instead of alignments, e.g. the RDP-classifier [30], indicating that huge, monolithic alignments are quite poor data for classification purposes. However, when linking the classification to the location of conserved and variable regions, the use of alignments seems unavoidable.

From Figure 5 we see how the separation of the larger classes is more important than the smaller classes, since the first PLS-components are devoted to this. Each mis-classification counts equally much, and separating larger classes will always reduce the total error more. This means the selected sites we find are those sites most important for separating the larger classes. The number of sequences in each class varies a lot in all available 16S data sets, e.g. see Figure 1. In this study we have only focused on the total error, and different results would be found if we focused only on the smaller classes.

Next, we repeated everything done so far, but using genus instead of phylum as class labels. This means we have 198 instead of 11 classes, making the separation much more difficult. In Figure 6 we show how the number

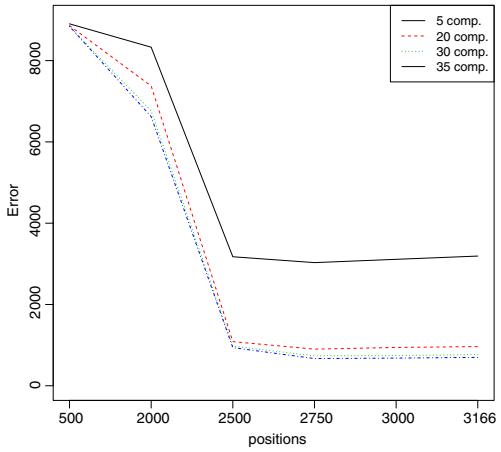


Figure 6 Site selection for genus classification. The figure illustrates how the number of mis-classifications drops by increased number of selected sites for genus classification. The error levels out at around 1000, and we need around 2500 sites to achieve this. This figure is based on the Greengenes data, but RDP and SILVA gave similar results.

of mis-classifications drops as we select more and more sites in the Greengenes alignment. We need to include many more sites than for phylum, and the classification error seems to level out after around 2500 selected sites, the remaining 600-700 sites do not provide further information about genus. Since around 80% of the sites are selected, it is obvious that the discriminating information in this case is not restricted to the variable regions. In fact, it tells us that in order to separate genera, we need to utilize almost every difference that can be found in the sequences regardless of where they are located. The error level we reach here, around 10% mis-classifications, is comparable to those reported by other studies on the genus level. This error rate and the number of selected sites indicates that a 16S based classification of genera means we are pushing the limit for how much information we can extract from the alignments of a single gene marker.

Conclusion

The aim of this study was to investigate the dogma of 16S based classification, stating that the key information for separating classes is harboured in the variable regions of this marker. By using three different multiple alignments of the same sequence data, we implemented a supervised learning method to systematically search for discriminative sites without any constraints with respect to conservation. The selected sites came out remarkably similar for the three data sets, a sign of a stable selection despite the obvious differences between the three alignments.

Our first major finding is that the discriminative sites are not exclusively located in the variable regions. In fact, the nine variable regions are not even enriched with sites selected by our algorithm. Variable regions are important, but not more important than any other region. The second major finding is that discriminative sites are typically sites with high entropy located among neighbouring sites of much lower entropy. This seems like a logical outcome. Regions of lower entropy means some degree of conservation, and alignments tend to be more accurate in such regions. If a site inside such regions show a much larger variation, it is more likely this is due to real biology, not alignment errors.

We believe these findings should be taken into consideration when it comes to improving methods for 16S based classification of bacteria.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

The project was initiated by LS and KHL. All authors have been involved in the development of the approach. KHL and HV did the programming. HV and LS drafted the manuscript. All authors have read and approved the final version.

Author details

[1]Department of Chemistry, Biotechnology and Food Sciences, Norwegian University of Life Sciences, Ås N-1432, Norway. [2]Nofima AS, Osloveien 1, Ås 1430, Norway.

Received: 16 May 2013 Accepted: 16 December 2013
Published: 27 January 2014

References

1. Woese CR, Stackebrand E, Macke TJ, Fox G E: **A phylogenetic definition of the major eubacterial taxa.** *Syst Appl Microbiol* 1985, **6:**143–151.
2. Woese CR: **Bacterial evolution.** *Syst Appl Microbiol* 1987, **51:**221–271.
3. Pace NR: **A molecular view of microbial diversity and the biosphere.** *Science* 1997, **276:**734–740.
4. Woese CR, Fox GE: **Phylogenetic structure of the prokaryotic domain: the primary kingdoms.** *Proc Natl Acad Sci USA* 1977, **74**(11):5088–90.
5. Harmsen D, Karch H: **16S rDNA for diagnosing pathogens: a living tree.** *ASM News* 2004, **70:**19–24.
6. Van de Peer Y, Chapelle S, De Wachter R: **A quantitative map of nucleotide substitution rates in bacterial rRNA.** *Nucleic Acids Res* 1996, **24:**3381–3391.
7. Clarridge JE: **Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases.** *Clin Microbiol* 2004, **17:**840–862.
8. Chakravorty S, Helb D, Burday M, Connell N, Alland D: **A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria.** *J Microbiol Methods* 2007, **69**(2):330–339.
9. Vasileiadis S, Puglisi E, Arena M, Cappa F, Cocconcelli PS, Trevisan M: **Soil bacterial diversity screening using single 16S rRNA gene V regions coupled with multi-million read generating sequencing technologies.** *PLoS One* 2012, **7**(8):e42671. doi: 10.1371/journal.pone.0042671.
10. Bartlett JMS, Stirling D: **A short history of the polymerase chain reaction.** *Methods Mol Biol* 2003, **226:**3–6.
11. Lane DJ, Pace B, Olsen GJ, Stahl DA, Sogin L, Pace NR: **Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses.** *Proc Nat Acad Sci* 1985, **82:**6955–6959.
12. Baker GC, Smith JJ, Cowan DA: **Review and re-analysis of domain-specific 16S primers.** *J Microbiol Methods* 2003, **55:**541–555.

13. Wang Y, Qian P: **Conservative fragments in bacterial 16S rRNA genes and primer design for 16S ribosomal DNA amplicons in metagenomic studies.** *PLoS ONE* 2009, **4**(10):e7401. doi:10.1371/journal.pone.0007401.
14. Mao D, Zhou Q, Chen C, Quan Z: **Coverage evaluation of universal bacterial primers using the metagenomic datasets.** *BMC Microbiol* 2012, **12:**66.
15. Winsley T, van Dorst JM, Brown MV, Ferrari BC: **Capturing greater 16S rRNA gene sequence diversity within the domain bacteria.** *Appl Environ Microbiol* 2012, **78:**5938–5941.
16. Cai L, Ye L, Tong AHY, Lok S, Zhang T: **Biased diversity metrics revealed by bacterial 16S Pyrotags derived from different primer sets.** *PLoS ONE* 2013, **8**(1):e53649. doi:10.1371/journal.pone.0053649.
17. Mizrahi-Man O, Davenport ER, Gilad Y: **Taxonomic classification of bacterial 16S rRNA genes using short sequencing reads: evaluation of effective study designs.** *PLoS ONE* 2013, **8**(1):e53608. doi:10.1371/journal.pone.0053608.
18. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen G L: **Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB.** *Appl Environ Microbiol* 2006, **72:**5069–5072.
19. Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ, Kulam-Syed-Mohideen AS, McGarrell DM, Marsh T, Garrity GM, Tiedje J M: **The ribosomal database project: improved alignments and new tools for rRNA analysis.** *Nucleic Acids Res* 2008, **37:**D141–D145.
20. Pruesse E, Quast C, Knittel K, Fuchs B, Ludwig W, Peplies J, Glöckner FO: **SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB.** *Nucleic Acids Res* 2007, **35:**7188–7196.
21. **Greengenes database.** [http://greengenes.lbl.gov/cgi-bin/nph-index.cgi].
22. **Ribosomal Database Project.** [http://rdp.cme.msu.edu/].
23. **SILVA database.** [http://www.arb-silva.de/].
24. Wold S, Martens H, Wold H: **The multivariate calibration problem in chemistry solved by the PLS method.** *Lect Notes Math* 1983, **973:**286–293.
25. Mehmood T, Martens H, Sæbø S, Warringer J, Snipen L: **Mining for genotype-phenotype relations in Saccharomyces using partial least squares.** *BMC Bioinformatics* 2011, **12:**318.
26. Mehmood T, Bohlin J, Kristoffersen AB, Sæbø S, Warringer J, Snipen L: **Exploration of multivariate analysis in microbial coding sequence modeling.** *BMC Bioinformatics* 2012, **13:**97.
27. Mehmood T, Liland KH, Snipen L, Sæbø S: **A review of variable selection methods in partial least squares regression.** *Chemometrics Intell Lab Syst* 2012, **118:**62–69.
28. Rajalahti T, Arneberg R, Kroksveen AC, Berle M, Myhr KM, Kvalheim OM: **Discriminating variable test and selectivity ratio plot: quantitative tools for interpretation and variable (biomarker) selection in complex spectral or chromatographic profiles.** *Anal Chem* 2009, **81**(7):2581–90.
29. Nawrocki EP, Kolbe DL: **Eddy SR: Infernal 1.0: inference of RNA alignments.** *Bioinformatics* 2009, **25**(10):1335–1337.
30. Wang Q, Garrity GM, Tiedje JM, Cole JR: **Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy.** *Appl Environ Microbiol* 2007, **73:**5261–5267.

doi:10.1186/2042-5783-4-2
Cite this article as: Vinje *et al.*: A systematic search for discriminating sites in the 16S ribosomal RNA gene. *Microbial Informatics and Experimentation* 2014 **4**:2.

Paper II

Vinje *et al. BMC Bioinformatics* (2015) 16:205
DOI 10.1186/s12859-015-0647-4

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Comparing K-mer based methods for improved classification of 16S sequences

Hilde Vinje[1*], Kristian Hovde Liland[1,2], Trygve Almøy[1] and Lars Snipen[1]

Abstract

Background: The need for precise and stable taxonomic classification is highly relevant in modern microbiology. Parallel to the explosion in the amount of sequence data accessible, there has also been a shift in focus for classification methods. Previously, alignment-based methods were the most applicable tools. Now, methods based on counting *K*-mers by sliding windows are the most interesting classification approach with respect to both speed and accuracy. Here, we present a systematic comparison on five different K-mer based classification methods for the 16S rRNA gene. The methods differ from each other both in data usage and modelling strategies. We have based our study on the commonly known and well-used naïve Bayes classifier from the RDP project, and four other methods were implemented and tested on two different data sets, on full-length sequences as well as fragments of typical read-length.

Results: The difference in classification error obtained by the methods seemed to be small, but they were stable and for both data sets tested. The Preprocessed nearest-neighbour (PLSNN) method performed best for full-length 16S rRNA sequences, significantly better than the naïve Bayes RDP method. On fragmented sequences the naïve Bayes Multinomial method performed best, significantly better than all other methods. For both data sets explored, and on both full-length and fragmented sequences, all the five methods reached an error-plateau.

Conclusions: We conclude that no *K*-mer based method is universally best for classifying both full-length sequences and fragments (reads). All methods approach an error plateau indicating improved training data is needed to improve classification from here. Classification errors occur most frequent for genera with few sequences present. For improving the taxonomy and testing new classification methods, the need for a better and more universal and robust training data set is crucial.

Background

The exploration of microbial communities is now a major focus in microbiology, opening new approaches to the study of microbiomes of humans and other organisms as well as the communities found in natural environments of air, water or soil [1]. Already in the 1980s Carl Woese introduced the rRNA-based phylogenetic comparisons of prokaryotes [2, 3], and the 16S rRNA gene is still the most useful genomic marker for the study of diversity and composition of metagenomes. The classification of 16S sequences obtained from some samples is a classical pattern recognition problem, i.e. recognizing some pattern in a sequence and assign it to one out of several predetermined categories. Whether the sequences are subjected to multiple alignments or, as in this paper, counting of short words, some assignment must be made based on how similar these sequences are to previously classified sequences. Naturally, the methods employed should give as accurate classifications as possible, but in metagenomics time-efficiency is also an issue since the number of sequences to classify may be vast. It should also be noted that with today's massively parallel sequencing technologies, shorter reads covering only a region of the gene are more accessible [4–6], making classification methods that perform well on sequence fragments essential.

However, classifications based on 16S rRNA sequences do not only have a practical use in metagenomics. In fact, this marker is used to build the entire prokaryotic taxonomy and is considered the gold standard for phylogenetic

*Correspondence: hilde.vinje@nmbu.no
[1]Department of Chemistry, Biotechnology and Food Sciences, Norwegian University of Life Sciences, Oslo N-1432 Ås, Norway
Full list of author information is available at the end of the article

studies [7–9]. In this perspective the classification of full-length 16S sequences is the issue. It should also be noted that in this context we should make all possible efforts to have the absolute best classifications available, and time-efficiency is no longer important.

A number of different procedures have been used to classify 16S sequences, and several different databases purposely designed as 16S rRNA repositories are available, e.g. Greengenes [10], RDP [11] and SILVA [12]. Most procedures for taxonomic studies have been based on alignments and reconstruction of phylogenetic trees, making use of some predefined evolutionary models and relevant algorithms [3, 13, 14]. However, with the enormous increase in data from next generation sequencing technology, these approaches suffer some problems. First, the computational time required to align a large set of sequences increases exponentially by its size. Secondly, greedy algorithms of some kind are required to construct these huge alignments and these sparse, monolithic alignments will most likely contain a substantial number of errors due to the heuristics employed. Finally, the lack of consensus, e.g. on evolutionary model assumptions, has made it impossible to arrive at an official taxonomy for prokaryotes, the most widely accepted taxonomy being the Bergey's Manual of Systematics of Archaea and Bacteria [15]. Thus, objective pattern recognition algorithms are likely to be valuable tools for building the prokaryotic taxonomy itself.

The most popular pattern recognition methods for 16S sequences are those based on counting K-mers, i.e. overlapping 'words' of length K in the sequences [16–19]. Wang et al. [19] developed the RDP classifier, based on the naïve Bayes principle and a word-length of $K = 8$. The RDP classifier is now close to being a standard in 16S based classification, and was in 2011 selected by Essential Science Indicators as the most-cited paper in a highlighted research area of microbiology [20]. K-mer methods are fast and will not suffer from the same uncertainties as the procedures based on evolutionary models and alignments. This way of converting sequences to numerical data is not as intuitive as evolutionary models, and lack the obvious interpretation given by evolutionary distances, but they are very objective in their mechanism. Also, in a previous study [21] we found that in order to obtain the best possible classification at the genus level, one has to consider more or less all positions along the full-length 16S sequences (around 1500 bases), not only hypervariable regions or other subsequences. This is another advantage of the K-mer methods; they use all data in a sequence.

However, K-mer based pattern recognition methods are not without model assumptions, and the RDP classifier uses the K-mer counts in one out of a number of alternative ways. Recent suggested improvements of this approach [22] have made it necessary to make a more systematic investigation on how well other K-mer based methods would perform, and possibly to reveal how and where efforts should be made to improve the objective classification of prokaryotes. In this paper we have compared different classification methods based on K-mer data for 16S sequences. We consider five different methods based on different machine-learning approaches, and we have compared their performance for full-length sequences as well as fragments. In addition to the method comparison, we also try to pinpoint where improvements should be made in order to give us better future methods for the important problem of identifying the majority of species on this planet.

Methods

Data

To compare methods we used two data sets. The `Trainingset9` is the data used to compare 16S classification methods in [19], and was downloaded from RDP [11]. It consists of 10032 16S rRNA sequences varying from 320 to 2210 bases in length, with the majority around 1400 bases. There are 37 phyla and 1943 genera represented in this set.

The `SilvaSet` is an extract from the SILVA database [12], where the largest genera have been 'pruned' by random sampling to contain fewer sequences. This set has 29520 sequences, covering 29 phyla and 1533 genera. The main reason for including this data set is that it is a manually curated data set different from `Trainingset9`, which was used during the development of the RDP-classifier.

In this paper we only consider classification to genus, i.e. the lowest taxonomic level of these data. This is the most challenging and also the most relevant problem for most studies where taxonomic classification is important.

The distributions of sequence abundance across genera are skewed for both `Trainingset9` and `SilvaSet`. Genera with only one sequence available are by far the most common in `Trainingset9` (Fig. 1). These singleton genera were included in the analysis, but will always be mis-classified by all methods, and all reported errors exclude these sequences. For `Trainingset9` few genera have more than 15 sequences, while some genera are considerably larger (not shown). The genus with most sequences is *Streptomyces*, which consists of 513 sequences. In the `SilvaSet` the difference in genus sizes is not as pronounced as in `Trainingset9`, but the majority of genera consists of 40 or less sequences. The genus with most sequences is *Pseudomonas* with 115 sequences.

To estimate the model performance we conducted a 10-fold cross validation [23] for all methods. The data

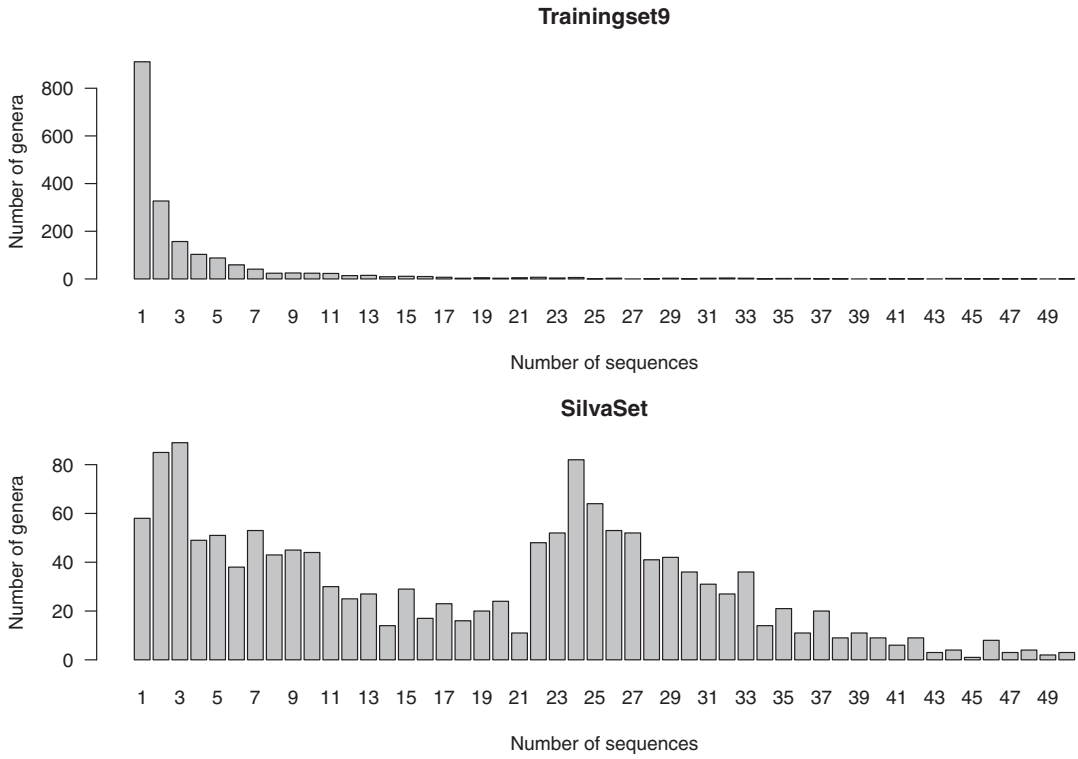


Fig. 1 The size distibution histogram of genera in `Trainingset9` and `SilvaSet`. The panels show the size distibution of genera in `Trainingset9` and `SilvaSet`. Genera with more than 50 sequences are not included here, but in `Trainingset9` there are 26 and in `SilvaSet` there are 40 such genera

were ordered alphabetically by genus name and split into ten different segments by enumeration from one to ten repeatedly, and then assigned to segments according to this number, i.e. every tenth sequence belongs to the same segment. This ensured a maximum spread of all genera across the segments. Each segment was set aside once as a test set, while the rest were used as training set in each cross-validation iteration.

K-mer based methods

All methods compared here represent a 16S sequence by its overlapping K-mers, i.e. words of length K. There are $D = 4^K$ possible words of length K in the DNA (RNA) alphabet, and in our study we tested word lengths from two to eight. The methods tested differ in the way they represent a sequence as K-mers and how this information is utilized in a statistical learning algorithm to achieve best possible classification.

All five methods were implemented in the software environment R [24]. Our implementation of the RDP classifier was tested against the original Java-implementation to ensure consistency. The PLS and nearest-neighbour methods already exist in the R-environment.

RDP

The RDP method considers only the presence/absence of a word in a sequence, not its frequency. All words of length K are ordered alphabetically as $w_1, w_2, \ldots, w_D$. For every sequence, we create a vector of D elements where element j is 1 if word w_j is present in the sequence, and 0 if not. We have chosen to describe the RDP method in detail below, even if this has been done in [19], because this method serves as a reference for the other methods described later.

Training

For each of the N sequences in the training set we get a vector of 1's and 0's, and these vectors are arranged as rows in the $N \times D$ matrix $\boldsymbol{A}^{rdp}$.

First, we estimate the unconditional probability: The probability of presence of each word regardless of genus. Summing the elements in each column of $\boldsymbol{A}^{rdp}$ produces the vector $n_1, n_2, \ldots, n_D$, i.e. n_j is the number of sequences

in the training set where word w_j is observed at least once. The probability that word w_j will be found present in any sequence is estimated by

$$Pr(w_j) = \frac{n_j + 0.5}{N + 1} \tag{1}$$

where the added 0.5 and 1 guarantees that no probability is zero or one.

Next, consider only sequences from genus g, i.e. we consider a sub-matrix $\boldsymbol{A}_g^{rdp}$ containing only the M_g rows corresponding to genus g. Again we can sum over the rows of $\boldsymbol{A}_g^{rdp}$, and we get the vector $m_{g,1}, m_{g,2}, \ldots, m_{g,D}$, i.e. $m_{g,j}$ is the number of sequences from genus g where we observe the word w_j at least once. The genus-specific or conditional probabilities are estimated by

$$q_{g,j} = Pr(w_j|g) = \frac{m_{g,j} + Pr(w_j)}{M_g + 1} \tag{2}$$

If the training set contains data for G genera, we can arrange the probabilities $q_{g,j}$ in a $G \times D$ matrix $\boldsymbol{Q}^{rdp}$ where the element in row g and column j is $q_{g,j}$, for $g = 1, \ldots, G$, $j = 1, \ldots, D$. This matrix $\boldsymbol{Q}^{rdp}$ is the trained model, with a set of probabilities (a row) for each genus.

Classification

Given a new sequence we construct the vector $\boldsymbol{a}$ corresponding to a row in the matrix $\boldsymbol{A}^{rdp}$ from above. Element j in $\boldsymbol{a}$ is 1 if word w_j is found in the new sequence, and 0 otherwise. The unconditional probability of $\boldsymbol{a}$ is found from (1) by

$$Pr(\boldsymbol{a}) = \prod_{j=1}^{D} Pr(w_j)^{a_j} \tag{3}$$

where a_j is element j in $\boldsymbol{a}$ and p_j is from (1). Notice that $Pr(\boldsymbol{a}$ is a joint probability of observing the words we see in this sequence. The naïve Bayes approach lies in the assumption that this joint probability can be written as a product of the marginal probabilities, as we have done on the right hand side above. This assumption is correct only if the elements of $\boldsymbol{a}$ are independent, which is a naïve assumption, but often still works in a satisfactory manner.

The conditional probability of $\boldsymbol{a}$ given some genus g is computed in a similar way from (2) by

$$Pr(\boldsymbol{a}|g) = \prod_{j=1}^{D} q_{g,j}^{a_j} \tag{4}$$

From the general relation between conditional and marginal probabilities it follows that

$$Pr(g|\boldsymbol{a}) = \frac{Pr(g)Pr(\boldsymbol{a}|g)}{Pr(\boldsymbol{a})} \tag{5}$$

where the probability on the left hand side is the criterion we use to classify. This is the *posterior probability* of genus g given the observed sequence $\boldsymbol{a}$, and we classify to the genus that maximizes this probability. On the right hand side we have the *prior probability* of genus g, $Pr(g)$, in addition to the two probabilities we computed in (3) and (4). It is customary to set the prior probability equal to the proportion of data from genus g in the training data set. In the RDP classifier the prior probabilities are assumed to be equal for all genera, and genera with few sequences are just as likely to be observed as those with many sequences in the training set. In our study we considered both flat priors (RDP) as well as priors proportional to genus abundances.

The posterior probability $Pr(g|\boldsymbol{a})$ is computed for every genus, and we assign the sequence to the genus where we get the largest probability. Notice that the denominator $Pr(\boldsymbol{a})$ in (5) does not depend on genus g. Hence, the g that maximizes $Pr(g|\boldsymbol{a})$ is exactly the same g that maximizes $Pr(\boldsymbol{a}|g)Pr(g)$, and we can ignore $Pr(\boldsymbol{a})$ altogether. Also, if the prior probabilities $Pr(g)$ are identical for all genera, we get the simple relation $Pr(g|\boldsymbol{a}) = Pr(\boldsymbol{a}|g)$.

From a computational perspective, we prefer the log-transformed version of (5) (ignoring $Pr(\boldsymbol{a})$), and using the relation in (4) we get

$$\log_2(Pr(g|\boldsymbol{a})) = \log_2(Pr(g)) + \sum_{j=1}^{D} a_j \log_2(q_{g,j}) \tag{6}$$

since this log-probability is maximized for the same g as the one in (4). If the matrix $\boldsymbol{Q}^{rdp}$ from the training step is log-transformed and called $\boldsymbol{L}^{rdp}$, and $\boldsymbol{p}$ is the column-vector of the G log-priors for all genera, we can compute the *score* vector

$$\boldsymbol{z} = \boldsymbol{p} + \boldsymbol{L}^{rdp} \cdot \boldsymbol{a}' \tag{7}$$

as the inner product of $\boldsymbol{L}^{rdp}$ and the column vector $\boldsymbol{a}'$. The score vector $\boldsymbol{z}$ has one element for each genus, and we assign to the genus where $\boldsymbol{z}$ has its maximum value. In case of two or more genera obtaining the same maximum value, the sequence is marked as unclassified.

Notice that with flat priors, the terms $\log_2(Pr(g))$ are identical for all g, i.e. all elements of $\boldsymbol{p}$ are identical, and it can be omitted from (7) since it will add the same to all genera.

Multinomial

The Multinomial method differs from the RDP method by considering the relative frequency of every word instead of presence/absence. The naïve Bayes principle is the same. A similar approach has also been tested by Lui and Wong in their work in [22].

Training

For each of the N sequences in the training set we get a vector of frequencies, i.e. element j is the number of times

we observe w_j in the sequence. These vectors are arranged as rows in the $N \times D$ matrix $\boldsymbol{A}^{frq}$.

As before we consider a sub-matrix $\boldsymbol{A}_g^{frq}$ containing the M_g rows corresponding to genus g. Summing over the columns of $\boldsymbol{A}_g^{frq}$ we get a vector $m_{g,1}, m_{g,2}, \ldots, m_{g,D}$. The genus-specific frequencies $F(w_j|g)$ are:

$$F(w_j|g) = \frac{m_{g,j}}{M_g} + \frac{1}{D} \tag{8}$$

where $\frac{1}{D}$ pseudo-counts are added to each frequency to avoid 0 counts. The multinomial probabilities for genus g is then calculated by dividing each $F(w_j|g)$ by their respective row sum, giving us a new set of multinomial probabilities $q_{g,i}$:

$$q_{g,j} = F(w_j|g) / \sum_{k=1}^{D} F(w_k|g) \tag{9}$$

The trained model consists of the $(G \times D)$ matrix $\boldsymbol{Q}^{mlt}$ where row g contains the multinomial probabilities $q_{g,j}$ for genus g.

Classification

From the new sequence we construct the frequency vector $\boldsymbol{a}$ corresponding to a row in the matrix $\boldsymbol{A}^{frq}$ above. Again we use the naïve Bayes approach to compute a score vector $\boldsymbol{z}$:

$$\boldsymbol{z} = \boldsymbol{p} + \boldsymbol{L}^{mlt} \cdot \boldsymbol{a}' \tag{10}$$

where $\boldsymbol{L}^{mlt}$ is the log-transformation of $\boldsymbol{Q}^{mlt}$ from the training step and $\boldsymbol{p}$ are the log-priors just as for the RDP-classifier. The score vector $\boldsymbol{z}$ has one element for each genus, and the sequence is assigned to the genus with maximum score in $\boldsymbol{z}$. In case of two or more genera obtaining the same maximum value, the sequence is marked as unclassified.

Markov

In the present context ordinary Markov models consider word frequencies, but differ from the naïve Bayes principle used by the previous two methods. Markov models have been tested on sequence data with the K-mer approach in earlier studies, e.g. by Davidsen et al. [25].

Training

The training step corresponds to estimating the transition probabilities of the Markov model. Any word of length K can be split into the *pretext* consisting of the first $K-1$ symbols, and the last letter, being A, C, G or T. The transition probabilities are the conditional probabilities of the last letter given the pretext. These probabilities are usually organized in a transition matrix with 4 columns (one for each letter) and one row for each pretext (4^{K-1} rows). However, these probabilities can equally well be organised in a single row-vector, where the conditional probabilities of A given the ordered pretexts is found at positions $I_A = (1, 5, 9, \ldots)$, for C given the ordered pretexts in positions $I_C = (2, 6, 10, \ldots)$ and so on. Note that this corresponds to the K-mers in alphabetical order. Each consecutive four positions corresponds to the same pretext, extended by A, C, G and T, respectively.

The matrices $\boldsymbol{A}^{frq}$ and $\boldsymbol{A}_g^{frq}$ are computed as for the Multinomial method. Summing over the columns of $\boldsymbol{A}_g^{frq}$ again produces genus-specific frequencies $F(w_j|g)$ as in (8). If K-mer w_j contains pretext h followed by, say, A, then the corresponding genus-specific transition probability is estimated by

$$q_{g,j} = F(w_j|g) / \sum_{k \in I_A} F(w_k|g) \tag{11}$$

and similar if the pretext is followed by C, G or T, I_A is replaced by the corresponding index set. If we had organized the transition probabilities in a matrix, this value would appear in cell $(h, 1)$ since we consider pretext h followed by A (column 1). Instead we arrange these probabilities in a row vector of D elements. Having the transition probabilities for each genus, we arrange the vectors as rows in a $(G \times D)$ matrix $\boldsymbol{Q}^{mrk}$. The latter organization of the transition probabilities is done only to have the same data structure as for the other methods; it does not affect the computations.

Classification

From the new sequence we count K-mers as for the Multinomial method, constructing the frequency vector $\boldsymbol{a}$ corresponding to a row in the matrix $\boldsymbol{A}^{frq}$. We compute scores for the sequence as

$$\boldsymbol{z} = \boldsymbol{L}^{mrk} \cdot \boldsymbol{a}' \tag{12}$$

where $\boldsymbol{L}^{mrk}$ is the log-transformation of $\boldsymbol{Q}^{mrk}$. Again we classify to the genus yielding maximum score. In case of a tie, the sequence is marked as unclassified.

Nearest-neighbour (NN)

In this method we use nearest-neighbour classification based on multinomial probabilities. Nearest-neighbour methods have no specific training step, but use the training data as a database and perform a lookup based on some characteristics of the query sequence. Another 16S nearest-neighbour method, called the Similarity Rank tool, was published by Maidak et al. [26] for use in The Ribosomal Database Project.

As before we compute the $(N \times D)$ matrix $\boldsymbol{A}^{frq}$ by word counting, where N is the number of sequences in the training set. Then we divide all elements in a row by its row-sum to obtain multinomial probabilities, and these are stored in the $(N \times D)$ matrix $\boldsymbol{A}^{mlt}$. Thus, each training

sequence, with its labelled genus, is represented as a row in this matrix.

For every new sequence we also count word frequencies and divide by the number of words in the sequence, producing a vector $\boldsymbol{a}$ similar to a row in $\boldsymbol{A}^{mlt}$. The Euclidean distance from $\boldsymbol{a}$ to all sequences (rows) in the training set is computed. The new sequence is assigned to the same genus as the nearest neighbour in the training set. In case of a tie, i.e. two or more genera are nearest neighbours, it is left unclassified.

Preprocessed nearest-neighbour (PLSNN)

In this method we extend the nearest-neighbour by combining it with the partial least squares (PLS) method [27]. This is a supervised learning method that has been used in many bioinformatics applications (e.g. [28–32]). A reason for the wide-spread use of PLS is that it is especially applicable when we have many correlated explanatory variables, which is typical for the present K-mer data, especially as K increases.

The idea is to compute a linear mapping from the K-mer frequency space to a much lower dimensional space, and then look for nearest-neighbours in this low-dimensional space. In K-mer space every sequence has $D = 4^K$ coordinates, and in the nearest-neighbour method above all coordinates (K-mers) have equal weight. However, it is more than likely that some of these will be more or less important for recognizing a particular genus. Replacing the original D dimensional space by a smaller number of combinations can be seen as a preprocessing of the data before the nearest-neighbour step, hopefully resulting in more 'correct' distances between sequences when seeking the nearest neighbour.

Training

From the training data we again compute the $(N \times D)$ matrix $\boldsymbol{A}^{mlt}$ as above. This is used as the matrix of explanatory variables in training the PLS-method. The response is the genus for each sequence. This is coded as a row-vector of G elements, with 1 in position g if the sequence comes from genus g and 0 in all other positions. This assembles into an $(N \times G)$ matrix $\boldsymbol{Y}$.

The PLS assumption is based on the linear model

$$E(\boldsymbol{Y}) = \boldsymbol{A}^{mlt}\boldsymbol{\beta} \tag{13}$$

where $\boldsymbol{\beta}$ is some $(D \times G)$ vector of regression coefficients. The algorithm will search for an orthogonal subspace by combining the variables (columns) of $\boldsymbol{A}^{mlt}$ and maximising the covariance between $\boldsymbol{Y}$ and $\boldsymbol{A}^{mlt}$. The algorithm first finds the 1-dimensional sub-space, then the 2-dimensional, etc. The main idea is to stop the search after C dimensions, where $C << D$ but still enough to have a good fit according to the model in (13). This means we end with

$$\boldsymbol{A}^{mlt} \approx \boldsymbol{S}\boldsymbol{R}' \tag{14}$$

where the $(N \times C)$ dimensional matrix $\boldsymbol{S}$ consist of linear combinations of the columns in $\boldsymbol{A}^{mlt}$, and $\boldsymbol{R}$ is some orthonormal projection matrix. The rows of $\boldsymbol{S}$ are the training sequences represented in the C-dimensional subspace with maximum covariance to genus information. In this representation we have filtered out less important variation in K-mer frequencies, e.g. variation within genera. Distances between sequences in this space should be more sensitive to between-genus variation and less sensitive to within-genus variation. For every word length K we tested 8 different dimensions C. The maximum was set to $C_{max} = \min(N-1, D-1, 2000)$, and we used $C = iC_{max}/8$ for $i = 1, 2 \ldots, 8$.

Classification

For every new sequence we compute a vector $\boldsymbol{a}$ similar to a row in $\boldsymbol{A}^{mlt}$. From (14) it follows that $\boldsymbol{A}mlt\boldsymbol{R} \approx \boldsymbol{S}$ since $\boldsymbol{R}$ is orthonormal, and thus we can compute $\boldsymbol{s} = \boldsymbol{a}\boldsymbol{R}$. The vector $\boldsymbol{s}$ is the representation of the new sequence in the subspace spanned by $\boldsymbol{S}$. The new sequence is finally classified with the nearest-neighbour method as before, where Euclidean distances from $\boldsymbol{s}$ to all rows of $\boldsymbol{S}$ are considered.

Results and discussion

We have tested five methods for K-mer based classification of 16S sequences, using a 10-fold cross validation, on two different data sets to compare their performance.

Figure 2 shows the classification error for full-length sequences for both `Trainingset9` and the `SilvaSet`. The Multinomial, the NN and the PLSNN method, all had a smooth, steady reduction in classification error from word length three, while the RDP method did not stabilize until word length five. The latter is due to the present/absent logic of this method: With too short word-length almost all words are present in most sequences. RDP and Multinomial had their minimum error at word length eight, NN and PLSNN at word length seven for `Trainingset9` and eight for the `SilvaSet`, while The Markov method reached the minimum error rate at word lengths of four and six, respectively, for the two data sets. The smallest error reached is fairly similar for all methods. The minimum error level was around 5 % for `Trainingset9`, and slightly higher for the `SilvaSet`.

The classification errors for the optimal word lengths are summarized in Table 1. For full-length sequences, using the optimal word length for each method, PLSNN performed best on both data sets with classification errors 4.2 % and 4.9 % respectively. The differences from the

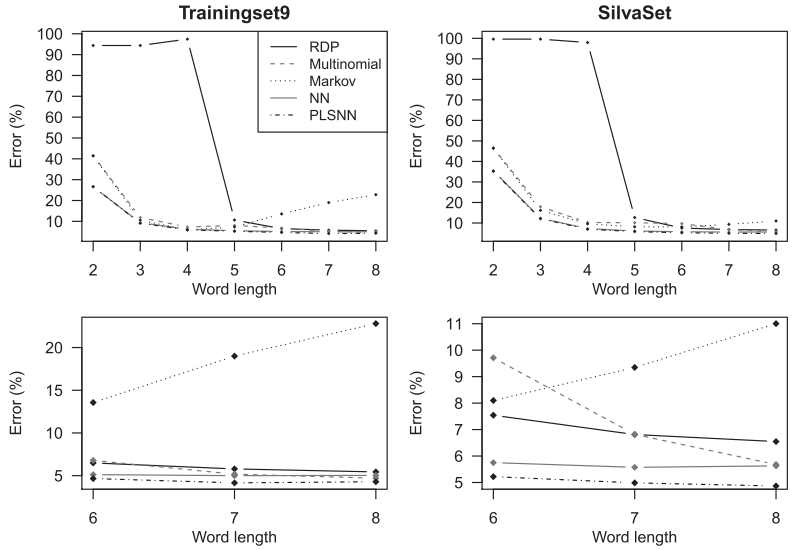


Fig. 2 Classification error for full-length 16S sequences. The top panels display the classification error for full-length sequences using all methods on word lengths $K2 - K8$. The bottom panels are the same results only zoomed at the last three word lengths ($K6 - K8$). Hence, the results are discrete values for every K-mer length and the connecting lines are merely to aid visual interpretation

other methods may seem small, but were stable. This is indicated by the error percentages in each of the ten cross-validation test-sets (Fig. 3). Each test set was a random subset of the full data set. The fact that methods behave consistent across subsets is an indication of a stable difference. From Fig. 3 we observed that not only was the PLSNN method overall best, but also best in nine out of ten sub-sets. We also noticed that the RDP method was not among the best methods in any sub-sets, and the Markov method produced the largest error in most cases. To test the effect of methods on the classification error, we employed a standard analysis-of-variance, using method as fixed effect (five levels) and test set as random effect (ten levels). Using the RDP method as a reference method, we made a pairwise comparison with Tukey's Honestly Significant test of the other four methods. The p-values are found in Table 2. The Markov method was significantly poorer and both the Multinomial and the PLSNN methods were significantly better ($p < 0.05$) than RDP on full-length sequences for both data sets.

Table 1 Results from a 10-fold cross validation. Classification errors (% misclassified) for the different methods at their optimal word length and for various data sets. Singleton genera errors are not included since they add the same to all methods

	Trainingset9		SilvaSet	
Method	Full-length	Fragments	Full-length	Fragments
PLSNN	4.15 (K7)	16.96 (K8)	4.87 (K8)	24.33 (K7)
Multinomial	4.70 (K8)	16.00 (K8)	5.68 (K8)	19.73 (K8)
NN	4.99 (K7)	16.54 (K8)	5.63 (K8)	24.02 (K8)
RDP	5.43 (K8)	16.42 (K8)	6.55 (K8)	20.49 (K8)
Markov	5.93 (K4)	21.78 (K6)	8.10 (K6)	22.98 (K7)

All methods were also tested on shorter fragments of 16S sequences. Present sequencing technologies provide high-quality reads up to a few hundred bases, and some kind of assembly is required to provide a full-length 16S sequence (minimum 1200 bases). Thus, classification based directly from the reads is desired. We divided the test sequences into ten partially overlapping fragments of 200 bases, and all fragments were classified. Figure 4 shows the classification error based on the fragment sequences for both `Trainingset9` and the `SilvaSet`. No method behaved well before word lengths of at least five or six, and again there was some error-plateau below which no method reached. Naturally, the errors were larger than for full sequences, since the information content of these shorter fragments must be smaller than the full sequence. Again, the ANOVA analysis was performed and we found that, compared to our control method RDP, the Multinomial was the only method that performed significantly better ($p < 0.05$) for both

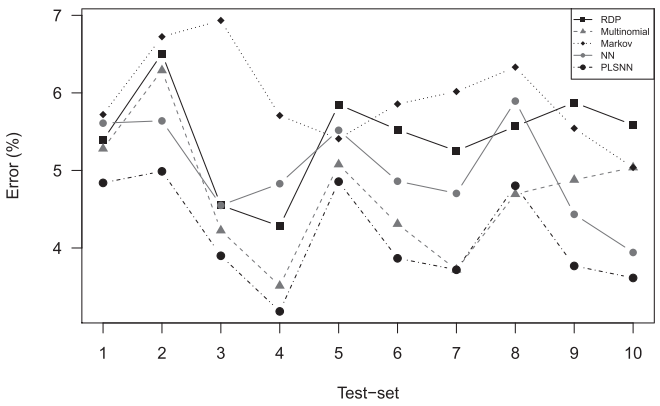


Fig. 3 Classification error for each test-set. For `Trainingset9` classification error for all the five methods are displayed for each of the 10 different test-sets from the 10-fold cross validation for full-length sequences. The `SilvaSet` gave similar results. Hence, the results are discrete values for every *K*-mer length and the connecting lines are merely to aid visual interpretation

data sets. PLSNN, on the other hand, now performed significantly poorer than RDP. The details of the results can be seen in Tables 1 and 2.

A difference between 4.2 % (full-length PLSNN `Trainingset9`) and 5.4 % (full-length RDP `Trainingset9`) error may seem small, but for building the taxonomy itself, there is no excuse for ignoring any improvement in methods. In principle the error should be zero. In a more practical use, where we want to classify a large number of sequences, a difference in 1 % means many misclassified sequences. Computation time is also an issue that should be taken into consideration. The RDP, Multinomial and Markov methods are fast and easy to both train and use for subsequent classification. All nearest-neighbour methods, including NN and PLSNN, are slower since they require distance computations for each new sequence to every sequence in the training set. The PLSNN method requires heavy computations during training, but once this has been done, new sequences are classified faster (and better) than with NN since distances are computed in a smaller sub-space.

The Markov method appears to be the clear loser in our tests. Not only does it give poorest best-case results, but we also noticed that the best word length for the Markov classifier changed from four to seven depending on the data set. The uncertainty in word length makes this method unstable and unreliable and it is discarded as a fruitful approach for 16S sequence classification.

In the PLSNN method we employ the PLS method as a preprocessing of the count data, finding linear combinations of the K-mer counts having maximum class information. If we consider word length seven there are $4^7 = 16384$ different K-mers. A full-length 16S sequence has around 1500 words of this length, which means more than 90 % of these K-mers occur zero times in any given sequence. Not all K-mers of this length can be equally important and a dimension reduction must be possible. We found that for $K > 6$ a reduction to 2000 dimensions gave the best PLS-performance. Thus, for $K = 7$ we reduce the coordinate-space from more than 16000 dimensions to 2000 before computing distances. Still, 2000 dimensions is remarkably large, but of course affected by the fact that we want to classify into a huge number of distinct genera. If the training set includes 1800 different genera, it is perhaps not surprising that we need at least this many dimensions to get a proper resolution to discriminate between them. This huge number of categories, as well as the considerable size variation between them seen in Fig. 1, makes this a rather special classification

Table 2 p-values for pairwise comparison of methods. Results from ANOVA on the effect of methods. The RDP is considered our control level and the p-values stated in the table below are the pairwise comparison for the four other methods to RDP

	Trainingset9		SilvaSet	
Method	Full-length	Fragments	Full-length	Fragments
PLSNN	< 0.001(−)	0.002(+)	< 0.001(−)	< 0.001(+)
Multinomial	0.016(−)	0.026(−)	< 0.001(−)	< 0.001(−)
NN	0.293(−)	0.895(+)	< 0.001(−)	< 0.001(+)
Markov	0.198(+)	< 0.001(+)	< 0.001(+)	< 0.001(+)

The signs in the parentheses indicate if a method gave smaller (−) or larger (+) errors than RDP

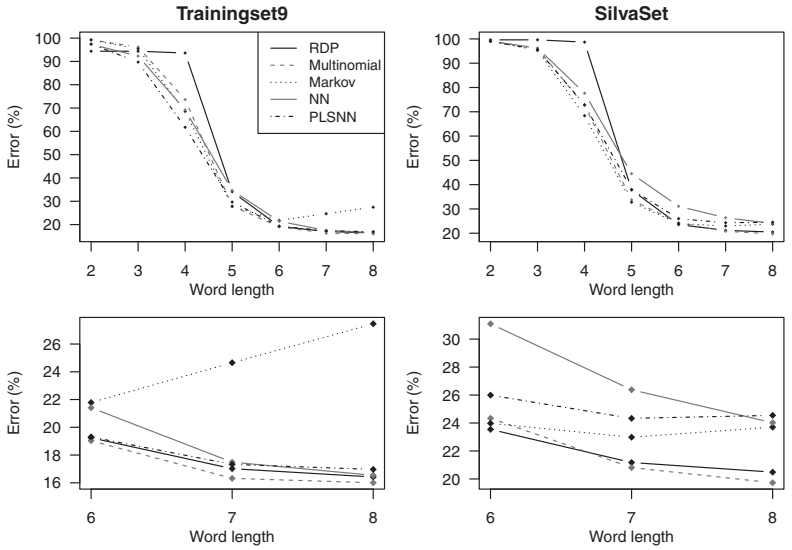


Fig. 4 Classification error for fragments. The top panels display the classification error for sequence fragments using all methods on word lengths $K2 - K8$. Sequences were split into 10 (partly overlapping) fragments of 200 bases, and all fragments were classified. The bottom panels are the same results only zoomed at the last three word lengths ($K6 - K8$). Hence, the results are discrete values for every K-mer length and the connecting lines are merely to aid visual interpretation

problem with several methodological challenges worth pursuing.

In [19] flat priors were used in their RDP-classifier. The results presented above also employ this strategy, assuming all genera are equally likely to occur in a new 16S sequence. If genera with many sequences in the training set are truly more widespread, this should be taken into account, and priors reflecting the abundance of each genus in the training set should produce better classifications. On the other hand, if a small training sample is due to an unexplored or newly discovered genus frequency-weighted priors supplies no further information to the data. We tested the RDP-classifier and the Multinomial method with both prior strategies on *Trainingset9*. The results were surprisingly similar regardless of priors. For word length eight the misclassified sequences were practically identical for the two cases, both for full-length and fragmented sequences. With this lack of differences we conclude that, unless very good arguments for the opposite can be provided, flat priors should be used. A flat prior means a single parameter (probability) is used for the entire population instead of (Ockham's razor) favours the simpler solution.

In the results we observed an error-plateau or barrier below which no K-mer based method seemed to reach. Data sets like `Trainingset9` and `SilvaSet` will always contain some proportion of questionable classifications partly since the actual relatedness between various genera is unknown, but also because the 16S gene itself is not a flawless marker. Variability between copies within the same genome as well as recombination events have been reported even for this highly conserved gene [33, 34]. If some sequences have been assigned to an incorrect genus from the beginning, classification errors seems unavoidable. Wang et al. tested their naïve Bayes classifier (RDP-classifier) on two different data sets in their work [19] from 2007. They reported the classification errors at genus level as 8.6 % and 7.9 % for the `Bergey corpus` and the `NCBI corpus`, respectively. The difference from our errors for the same method can be explained by a data set effect, presumably the data sets we have been 'improved' by eliminating some obvious mis-assignments since 2007. This emphasizes the importance of training data for classification performance [35].

All the sequences that were classified faulty by at least one of the methods were extracted and investigated further. For full-length sequences from `Trainingset9` this consisted of 725 sequences, and the errors were distributed over methods as shown in Fig. 5. First, we noticed all methods made some unique errors, from

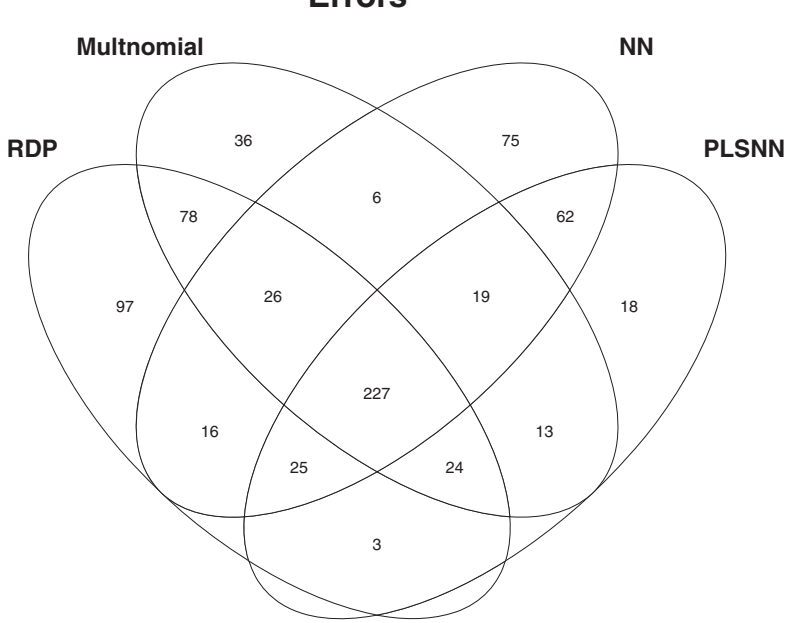


Fig. 5 The distribution of errors over methods. The Venn-diagram shows how errors distribute among the different methods. The number in each sector corresponds to the number of mis-classified sequences. The results are from `Trainingset9` and full-length sequences

18 for PLSNN to 97 for RDP. The pairwise relations showed that RDP and Multinomial shared 78 common errors, and NN and PLSNN shared 62. All other pairs had much fewer common errors (3,6,13 and 164), as expected from how the methods are designed. We also noticed that 227 sequences were classified faulty by all methods (center sector), and among these, 176 were assigned to the same genus by all five methods. These 176 sequences belong to 127 unique genera and 42 of these genera contained only two sequences in the full data set.

To investigate further the effect of genus-size, we have in Fig. 6 plotted the error percentages for sizes two to ten. As expected, small genera had elevated risk of being classified wrongly. Genera of size two means there are two sequences in total, one sequence in the training set and one in the test set. Recognizing a genus based on one previously observed sequence is of course very difficult. The genera with only one sequence present (singletons) are not shown as they always will have 100 % error. The figure shows that more errors were made for genera consisting of few sequences and this skewness in abundance poses a challenge to all statistical learning methods. One may argue that to improve classifications we need better data more than we need better methods, and that a larger data set is not necessarily a better data set. The `SilvaSet` is three times larger than `Trainingset9` but still relatively more errors were made. We agree that better data is essential, but better data and better methods are also interleaved, since no data set is completely independent of methods, and manual curation is certainly no guarantee against classification errors.

In the introduction we mentioned that previous studies show that we should consider all positions along the 16S sequences to get optimal genus-classification. Still, in Fig. 7, we see that some fragments are more informative than others. Fragment four gave a considerably better classification than the other fragments. Please note that we chopped the 16S sequences into ten partly overlapping fragments, all of length 200 bases. Thus, fragment four is located relative to each sequence length, and does not correspond exactly to a hyper-variable region, but is in most cases around region V3-V4, which is known to be the most informative part of the 16S gene. In this perspective it seems likely that there is something to gain from utilizing position-specific information. *K*-mer based methods do not take into account where in the sequence the different words are located, and there

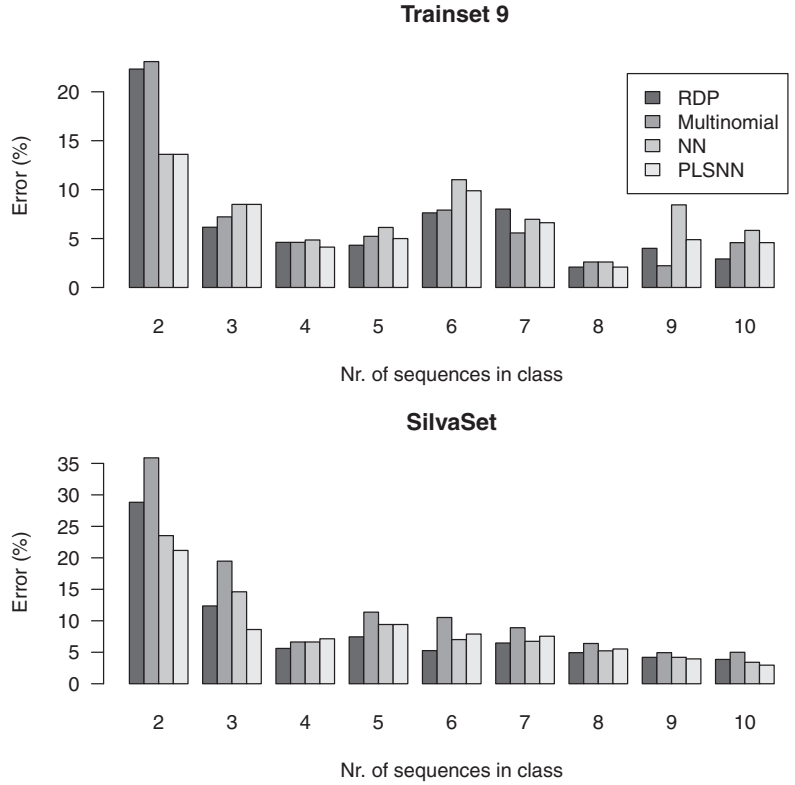


Fig. 6 Error distribution over class-sizes. The horizontal axis is the class-size (number of sequences in a genus) and the vertical axis is the error percentage averaged over all genera of the same size. The upper panel is the result from `Trainingset9` and the lower panel for the `SilvaSet`. Only class-sizes up to 10 is shown, for larger classes the error-percentages are small

may be a potential for improving the methods along this line.

Conclusion

We have compared the popular RDP method to four other K-mer based methods with respect to classification of prokaryotes based on 16S sequences. The differences in classification performance are significant, but all methods apart from the Markov method seem to stabilise on a classification error less than 6.6 % for word length bigger than seven for full-length sequences. Small extensions to the RDP method, such as counting the frequencies instead of just present/absent, seem to be an advantage, as also pointed out by [22]. On full-length 16S sequences, the Preprocessed nearest-neighbour method stands out as the best, and should be considered for high-precision jobs. With shorter 'reads' as input, the naïve Bayes based Multinomial method proves to be the method with least classification errors and therefore the method, out of the five presented methods, which is the optimal option for rapid taxonomic assignments.

The study also reveals the importance of high-quality data for improving the classifications further. All methods seem to level out at some error which is inherent in the various data sets, and it is not likely that improved methods as such will lower this barrier. We have pointed out the special features of this type of data; a large number of categories (genera) in combination with an extreme skewness in their sizes. A key to improve classification is to obtain gold standard training sets in which all efforts have been made to have as few genera as possible with only a few sequences. Increasing the number of representative sequences from one to three or four can greatly increase the classification accuracy.

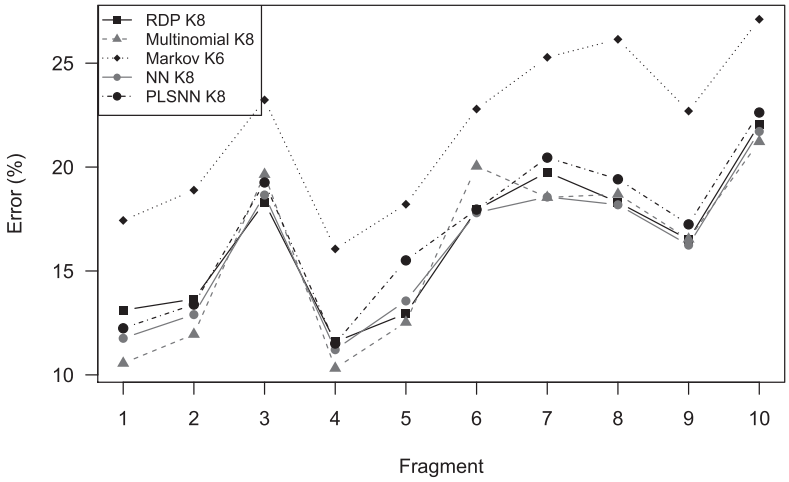


Fig. 7 Position specific error. The average error for each method on each of the 10 fragments. The fragments corresponds roughly to (partly overlapping) regions from the start to the end of each 16S sequence. These results are from `Trainingset9`, but the results from `SilvaSet` were similar. Hence, the results are discrete values for every *K*-mer length and the connecting lines are merely to aid visual interpretation

The *K*-mer methods examined here ignore the position specific information that is most likely important to discriminate certain genera. For further improvement of classification, pattern-recognition methods that takes into account position specific information through the 16S sequences may be a good place to start.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
The project was initiated by LS, KHL and HV. All authors have been involved in the development of the approach. KHL, LS and HV did the programming. HV and LS drafted the manuscript. All authors have read and approved the final version.

Acknowledgements
Hilde Vinjes scholarship has been fully financed by the Norwegian University of Life Sciences.

Author details
[1]Department of Chemistry, Biotechnology and Food Sciences, Norwegian University of Life Sciences, Oslo N-1432 Ås, Norway. [2]Nofima AS, Osloveien 1, Oslo 1430 Ås, Norway.

Received: 13 February 2015 Accepted: 6 June 2015
Published online: 01 July 2015

References
1. Özlem Taştan Bishop. 2014. Bioinformatics and Data Analysis in Microbiology. Rhodes University, South Africa: Caister Academic Press.
2. Woese CR, Stackebrand E, Macke TJ, Fox GE. A phylogenetic definition of the major eubacterial taxa. Syst Appl Microbiol. 1985;6:143–51.
3. Woese CR. Bacterial evolution. Syst Appl Microbiol. 1987;51:221–71.
4. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, Fierer N, Knight R. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc Natl Acad Sci USA. 2011;108(Suppl 1):4516–22.
5. Claesson M, Wang Q, O'Sullivan O, Greene-Diniz R, Cole J, Ross R, et al. Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions. Nucleic Acids Res. 2010;38:e200.
6. Tringe S, Hugenholtz P. A renaissance for the pioneering 16S rRNA gene. Curr Opin Microbiol. 2008;11:442–6.
7. Woese CR, Fox GE. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci U S A. 1977;74(11):5088–90.
8. Pace NR. A molecular view of microbial diversity and the biosphere. Science. 1997;276:734–40.
9. Harmsen D, Karch H. 16S rDNA for diagnosing pathogens: a living tree. ASM News. 2004;70:19–24.
10. Greengenes database. 2015. [http://greengenes.lbl.gov/cgi-bin/nph-index.cgi]. Accessed date May 18, 2015.
11. Ribosomal Database Project. 2015. [http://rdp.cme.msu.edu/]. Accessed date May 18, 2015.
12. SILVA database. 2015. [http://www.arb-silva.de/]. Accessed date May 18, 2015.
13. Ludwig W, Strunk O, Klugbauer S, Klugbauer N, Weizenegger M, Neumaier J, Bachleitner M, Schleifer KH. Bacterial phylogeny based on comparative sequence analysis. Electrophoresis. 1998;19(4):554–68.
14. Kolaczkowski B, Thornton JW. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature. 2004;431:980–4.
15. Bergeys. 2015. [http://www.bergeys.org/]. Accessed date May 28, 2015.
16. Rudi K, Zimonja M, Næs T. Alignment-independent bilinear multivariate modelling (AIBIMM) for global analyses of 16S rRNA gene phylogeny. Intl J Syst Evol Microbiol. 2006;56:1565–75.
17. Liu Z, DeSantis TZ, Andersen GL, Knight R. Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. Nucleic Acids Research. 2008;36:e120.
18. Soergel D, Dey N, Knight R, Brenner S. Selection of primers for optimal taxonomic classification of environmental 16S rRNA gene sequences. ISME J. 2012;6:1440–4.
19. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. Appl Enviromental Microbiol. 2007;73:5261–67.
20. Science Watch. 2015. [http://archive.sciencewatch.com/dr/erf/2011/11decerf/11decerfCole/]. Accessed date May 18, 2015.

21. Vinje H, Almøy T, Liland KH, Snipen L. A systematic search for discriminating sites in the 16S ribosomal RNA gene. Microb Inf Experimentation. 2014;4:2.
22. Liu K, Wong T. Naïve Bayesian Classifiers with Multinomial Models for rRNA Taxonomic Assignment. IEEE/ACM Trans Comput Biol Bioinformatics. 2013;10(5):1334–9.
23. Stone M. Cross-validatory choice and assesment of statistical predictions. J R Stat Soc Serie B-Methodological. 1974;36:111–47.
24. R. 2015. [http://www.r-project.org/]. Accessed date May 18, 2015.
25. Davidsen T, Rødland EA, Lagesen K, Seeberg E, Rognes T, Tønjum T. Biased distribution of DNA uptake sequences towards genome maintenance genes. Nucleic Acids Res. 2004;32(3):1050–8.
26. Maidak BL, Larsen N, McCaughey MJ, Overbeek R, Olsen GJ, Fogel K, Blandy J, R WC. The Ribosomal Database Project. Nucleic Acids Res. 1994;22(17):3485–7.
27. Wold S, Martens H, Wold H. The Multivariate Calibration Problem in Chemistry solved by the PLS Method. Lect Notes Math. 1983;973:286–93.
28. Nguyen DV, Rocke DM. Tumor classification by partial least squares using microarray gene expression data. Bioinformatics. 2002;18:39–50.
29. Nguyen DV, Rocke DM. Multi-class cancer classification via partial least squares with gene expression profiles. Bioinformatics. 2002;18:1216–26.
30. Aarøe J, Lindahl T, Dumeaux V, Sæbø S, Tobin D, Hagen N, Skaane P, Lönneborg A, Sharma P, Børresen-Dale A. Gene expression profiling of peripheral blood cells for early detection of breast cancer. Breast Cancer Res. 2010;12:R7. doi:10.1186/bcr2472.
31. Mehmood T, Martens H, Sæbø S, Warringer J, Snipen L. Mining for genotype-phenotype relations in Saccharomyces using partial least squares. BMC Bioinformatics. 2011;12(318):318.
32. Mehmood T, Bohlin J, Kristoffersen AB, Sæbø S, Warringer J, Snipen L. Exploration of multivariate analysis in microbial coding sequence modeling. BMC Bioinformatics. 2012;13:97. doi:10.1186/1471-2105-13-97.
33. Vetrovsky T, Baldrian P. The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. PLoS ONE. 2013;8(2):e57923. doi:10.1371/journal.pone.0057923.
34. Kitahara K, Miyazaki K. Natural and experimental evidence for horizontal gene transfer of 16S rRNA. Mobile Genet Elem. 2013;3(1):e24210.
35. Werner JJ, Koren O, Hugenholtz P, DeSantis TZ, Walters WA, Caporaso JG, et al. Impact of training sets on classification of high-throughput bacterial 16s rRNA gene surveys. ISME J. 2012;6:94–103.

Paper III

The ConTax data: Improved supervised learning of prokaryotic taxonomy

Hilde Vinje*[1], Kristian Hovde Liland[1,2], Lars Snipen[1]

[1]Department of Chemistry, Biotechnology and Food Sciences, Norwegian University of Life Sciences, P.O.Box 5003, N-1432 Ås, Norway

[2]Nofima - Norwegian Institute of Food, Fisheries and Aquaculture Research, Osloveien 1, N-1430 Ås, Norway

Email: Hilde Vinje*- hilde.vinje@nmbu.no; Kristian Hovde Liland - kristian.liland@nofima.no; Lars Snipen - lars.snipen@nmbu.no;

*Corresponding author

Abstract

A major issue in 16S-based taxonomic classification of prokaryotes is the lack of an official taxonomy. The same DNA sequence can have different taxon assignments when considering different databases. Regardless of methods used, this database discrepancy leads to a larger than necessary variance in taxonomic assignments. For both exploration of the sequence space and for training effective and robust classification methods there is an obvious need for a stable and reliable training data set. In this paper we present data sets with consensus taxonomy, the ConTax sets, for the 16S rRNA gene. It derives from three well-known and often used repositories of 16S rRNA; The Silva comprehensive ribosomal RNA database, the Greengenes 16S rRNA gene database, and the Ribosomal Database Project (RDP). A sequence is included in the ConTax set if it is marked as high quality, longer than 1200 bases, and consists strictly of `A`, `C`, `G` or `T`(`U`). The major new feature of this set is that a sequence is included if and only if it is found in at least two of the three databases, with no diverging taxon assignments down to genus level. The ConTax set comprises $664,199$ 16s rRNA sequences. From a method training perspective, the enormous number of sequences in some of the taxa provides no additional information for classification purposes. Therefore, a trimmed ConTax set is also presented. The trimming is designed to maintain the major information for each taxon, and the trimmed set consists of $38,784$ sequences. The results presented in this paper confirm that the performance of different classification methods trained and tested on the ConTax set improves substantially compared to other established data sets.

INTRODUCTION

There is no such thing as an official taxonomy of prokaryotes. This illuminating fact means there are no comprehensive gold standard data sets giving a clear picture of the microbial world as we know it. Of course, some of the branches in the bacterial tree of life are widely accepted and will most likely remain unchanged for all foreseeable future. But, a significant part of the taxonomy is still evolving, and advances in sequencing technologies have spawned a plethora of tools for exploring this 'biological dark matter' [1, 2] resulting in many new taxa [3] and an ever increasing diversity [4].

Since Woese [5] the small subunit 16S ribosomal RNA has been the central genetic marker for recognizing prokaryotes. Even if whole-genome data are now widely available, the 16S gene is still the dominating

marker for assigning bacteria to their proper position in the tree of life, at least down to the genus level. There are several methods for the classification of bacteria based on this marker, e.g. [6–8]. A common approach is still to simply BLAST any 16S sequence fragment against a database, and look for the nearest neighbours, even if this is suboptimal both in speed and precision compared to the mentioned methods. Given the countless studies where a 16S based profiling of some microbial community is a part, the number of (mis)classifications based on this marker is growing as ever before.

Regardless of the tool used, the 16S-based classification of bacteria is a typical example of a supervised learning problem. This means an observed 16S sequence (fragment) is assigned to one group (or bin or lineage) based on how similar it is to previously assigned sequences. These previously assigned sequences are usually referred to as the training data (the database), and as pointed out in [9], the quality of these data affects the classification results severely. In a recent paper [10] we examined the performance of several supervised learning methods for this problem, and one major conclusion was that improving the training data will probably have greater impact than improving the statistical methods, even if there are differences between methods as well.

There are several important aspects of such a training data set. Werner *et al.* [9] emphasize the importance of a wide ranging training set, i.e. as many branches as possible of the phylogenetic tree should be represented. This also means we need many variants of every taxon, in order to span the evolutionary variance within taxa. We can extend this criterion by demanding that taxa should also be present by approximately the same number of representatives. Virtually all supervised learning methods will perform better if the training data are approximately balanced compared to a highly skewed set. Unfortunately, the latter situation is common in frequently used databases. Some genera are over-represented with a huge number of sequence variants, while many others are barely present.

Another criterion, partly in conflict with the first, is that training data are 'correct'. This seems like the bottleneck since there is no official taxonomy. If sequences have been faulty assigned in the training data, this will of course lead to further errors when using the trained algorithm on new sequences. Given the more or less diverging taxonomy assignments in todays databases, the closest we can get to a 'correct' data set is to look for sequences where we have some consensus across several sources with respect to the classification.

The motivation for our work here has been to come up with a database of 16S sequences which is spanning as many genera as possible and at the same time only includes sequences with a very high likelihood of having a correct taxon label. This comprehensive data set should be used as a supervised learning training

set or database in order to give best possible classifications based on 16S sequence data. In this study we have collected a large number of 16S sequences from three major public repositories. Through a list of quality filtering and analyses described below we arrive at a set of full-length 16S consensus-classified sequences spanning the majority of the tree-of-prokaryotes in the List of Prokaryotic names with Standing in the Literature (LPSN, [11]). Several versions of this ConTax data set is available in the public R-package named `microcontax`.

MATERIALS AND METHODS

The ConTax set

Here we describe the major steps for how the ConTax set was assembled.

Step 1 - Data retrieval and filtering:

In this study we focused on data downloaded from the following three repositories: The Silva comprehensive ribosomal RNA database (http://www.arb-silva.de/, [12]), the Greengenes 16S rRNA gene database (GG, http://greengenes.lbl.gov/cgi-bin/nph-index.cgi, [6]) and the Ribosomal Database Project (RDP, http://rdp.cme.msu.edu/, [13]). All three are devoted specifically to ribosomal RNA, and have huge collections of high quality 16S data, including taxonomic assignments.
The data obtained from these three sources were all subject to the same quality filtering:

- Sequences with alien characters (not `A`, `C`, `G` or `T`(`U`)) were discarded.
- Sequences shorter than 1200 bases were discarded.
- Sequences without a genus assignment were discarded.

Within each of the three filtered data sets we also deleted duplicated sequences. In a few cases we found that identical sequences were given different classifications even within the same database. Such sequences were eliminated from the data set.

Step 2 - Consensus taxonomy:

From the filtered data sets we extracted sequences with consensus taxonomy. A sequence found in all three databases was approved if its genus assignment was identical in all three cases. We denote this subset the SRG (Silva, RDP, Greegenes) set. In addition, we also considered sequences found in two of the three

sources, and again required matching genus assignments in the corresponding repositories. These three subsets were denoted SR, SG and RG, respectively.

The assigned genera in the repositories contained some names not found among the standard names in the LPSN database. However, the only manual editing we did was to allow textual comparisons across the databases, like replacing the text *Armatimonas/Armatimonadetes_ gp1* with just *Armatimonas* etc.

It follows from this that there will also be a number of conflict sequences, i.e. sequences found in two or three databases, but with differing genus assignment. Such sequences can give us information about where the taxonomy is uncertain (diverging opinions), but we will not pursue these data in this paper.

Step 3 - Eliminating subsequences:

All collected sequences (union of SRG, SR, SG and RG) were grouped by their genus assignment. Within each genus all sequences were compared to all longer sequences. In all cases where a shorter sequence turned out to be a subsequence of a longer, the shorter was discarded.

All sequences left after these three steps is the full ConTax set. In the supplied R-package this data set is referred to as `contax.full`. In addition the the sequences themselves, each sequence has a unique tag, information about which repositories it originates from, and the full taxonomy (domain, phylum, class, order, family, genus).

Trimming large genera

Typical for all 16S data sets of a considerable size is that some genera have a large number of (slightly) different sequences. For most practical uses this density is not needed, and will only slow down classifications without improving precision. We made a trimmed version of the `contax.full` database as follows.

Let M be the median genus size in the data set. Let $N > M$ be the number of sequences in a large genus. We then trimmed this large genus to contain

$$N_t = M + \lceil \sqrt{N - M} \rceil$$

sequences, where the $\lceil . \rceil$ means rounding up to the nearest integer.

For each genus we made a multiple alignment using the Infernal software [14] and the SSU-model for bacteria or archaea downloaded from Rfam [15]. The alignments were trimmed at each end, eliminating positions with more than 50% gaps, and trimmed inside by eliminating position with excess of 90% gaps.

For each genus-alignment simple p-distances [16] were computed between all sequences and the *medoide* sequence was detected. The *medoide* sequence is the sequence with the smallest sum of distances to all other sequences in the genus. These sequences are available as the `medoids` subset in the R-package.
The trimming was done in two steps. I) To avoid extreme sequences we first excluded sequences where the distance to its nearest neighbour was larger than the average distance between all sequences within the respective genus. II) From the remaining sequences the medoide sequence was included as the first member of the trimmed set. All subsequent members of the trimmed set were selected as the sequence having the largest sum of distances to the previously selected members, and this was iterated until N_t sequences were included.
In Figure 1 we illustrate the trimming idea graphically. The trimmed set is named `contax.trim` in the R-package.

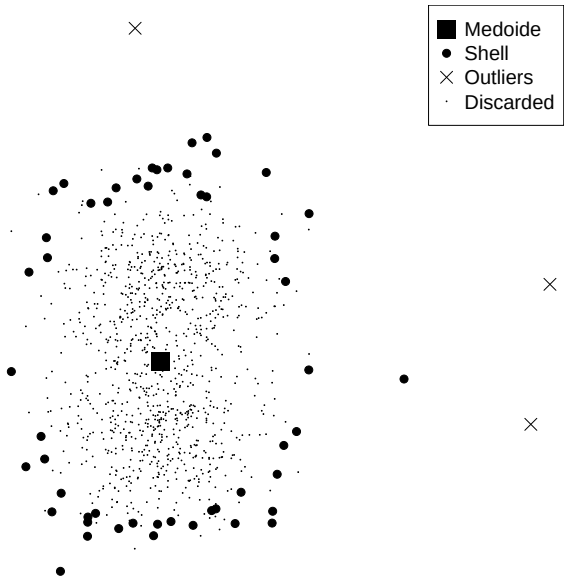


Figure 1: Trimming. An illustration of the trimming procedure. Each marker corresponds to a sequence in a large genus. In Step 1 the extreme sequences, marked as crosses, are discarded. In Step 2 the medoide sequence, marked as large filled square, is included as the first member of the trimmed set. The remaining steps select the 'shell' members marked as large filled circles. The small dots (and crosses) are the sequences discarded by the trimming.

Classification

The classification accuracy, or error level, is always partly a product of inexact data and imperfect methods. In order to compute accuracy levels, we used three distinct supervised learning methods: Tthe RDP classifier, the Multinomial classifier and the nearest-neighbour BLAST. The two first are described in detail in [7, 10, 17]. The BLAST approach means a sequence is classified by making a `blastn` search against the training data and classify to the same taxon as the best BLAST-hit (largest bit-score).

A 10-fold cross-validation procedure was used to estimate accuracy for a given data set and method, see [10] for a detailed description.

Accuracy is most commonly presented as a total result for the entire data set. It can also be computed per genus, which is also presented in the Result section. However, from a user-perspective, the accuracy is not the optimal measure of reliability. The reason is that it conditions upon the correct genus. When classifying a new sequence, we never know the correct genus, hence we do not know which accuracy is relevant for the current sequence. Instead, the Positive Predictive Values (PPV) are more informative. This is the proportion of correctly classified sequences who are all classified to the same genus [18]. Hence, a genus with a large PPV (close to 1.0) indicates that classifications to this genus are most often correct. PPV is an estimate of $Pr(\text{correct}|\text{predicted genus})$ while accuracy estimates $Pr(\text{correct}|\text{actual genus})$. In case a genus is never predicted, its PPV value is simply unknown.

Classifications were also done based on the amplicon-sequences collected *in silico* by primer pair $515F/R806$ to see how well we may classify sequences in a typical metagenome sample. These primers are suggested by the Earth Microbiome Project (http://www.earthmicrobiome.org/) to match the $V4$ region of the 16S gene. The 10-fold cross-validation procedure was again used to estimate accuracy. Hence, a method was trained on full length sequences from the training set and tested on the considerable shorter amplicon-sequences extracted from the test set sequences.

Implementation

An R data package, `microcontax`, is available for free on The Comprehensive R Archive Network [19]. The package consists of various versions of the ConTax data along with some functions for extracting and manipulating sequences and/or taxonomy information. The latter include taxonomy lookup and related functions. All data can be stored as FASTA-files and used, e.g. as a BLAST database.

We have also made available another free R package, `microclass`, containing R implementations of the classification methods used in this manuscript. This package also contains trained models based on the

ConTax data.

RESULTS

The ConTax data

From the three major data repositories on 16S rRNA, we downloaded all available sequences, and filtered them to obtain sets of unique high-quality sequences with assignments at the genus level, as described in Step 1 of the Methods section. This resulted in $1,064,763$ sequences from Silva, $912,409$ from RDP and $624,078$ from the Greengenes data repository. In Figure 2 we show how the repositories overlap with respect to sequence identity after this first step.

For each of the intersections in Figure 2 we collected all sequences with identical genus assignment, and a summary is listed in Table 1. Finally, within every genus all shorter subsequences were also discarded, resulting in the `contax.full` set (last column of Table 1).

Table 1: For each of the four intersections between the repositories is listed the number of sequences left after each of the three filtering steps described in the Methods section. SRG means the intersection Silva-RDP-Greengenes, etc. After step 1 we find the number of high-quality sequences in each intersection, after step 2 the number of sequences with consensus taxonomy only and after the last step all subsequences have been eliminated. The last column corresponds to the `contax.full` sequences.

Subset	Step 1	Step 2	Step 3
SRG	393696	366194	349617
SR	381176	312987	288899
SG	21819	7358	7107
RG	25382	19592	18576
Sum	822073	706131	664199

The `contax.full` set has $664,199$ sequences covering 1774 different genera, 316 families, 138 orders, 68 classes, 32 phyla and 2 domains.

Trimming

As expected, the `contax.full` set is extremely skewed with 156 genera having only 1 sequence while the largest genus (*Staphylococcus*) has $76,484$ sequences. The average size of a genus is 374 sequences and the median is 21. All genera with more than 21 sequences were trimmed according to the procedure described in the Method section. In total 878 genera were trimmed and the largest genus (*Staphylococcus*) was reduced from $76,484$ to 245 sequences. A total of $38,784$ sequences were left in the trimmed data set named `contax.trim` in the R-package. The trimmed version covers the exact same taxa as the full data set, only with less density for the larger genera.

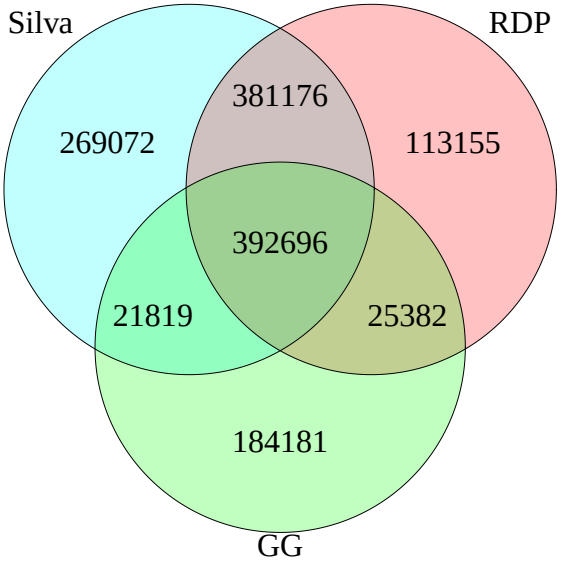


Figure 2: Venn diagram. The Venn diagram illustrates how the RDP, Silva and the Greengenes data repositories overlap with respect to sequence identity. The numbers refer to the number of unique 16S sequences in each sector.

Figure 3 is a visualization of data set skewness. The area-under-curve (auc) is a measure of skewness, with 0.5 as the minimum (all genera same size) and 1.0 the maximum. The `contax.full` data set (grey full curve) has an auc of 0.95, where the largest 10% of the genera covers more than 80% of the data set. This is reduced to 0.73 for `contax.trim`. The `Trainset9` (auc= 0.83) is un-trimmed like `contax.full`, but much smaller. The `Silvaset` (auc=0.70) has been trimmed in a manor similar to `contax.trim`, described in [10].

Classifications

The major idea of collecting the ConTax set is to have a data set where the assigned classifications are very likely to be correct. If this is the case, cross-validated classifications should produce few errors, given that the supervised learning method used is fairly good.

This was tested by running a 10-fold cross validation over the `contax.trim` data set, using three well known and much used supervised learning methods. The exact same procedure was repeated for two previously used training data sets, the `Trainingset9` from RDP, the `Silvaset` from the Silva database [10], in order to compare results.

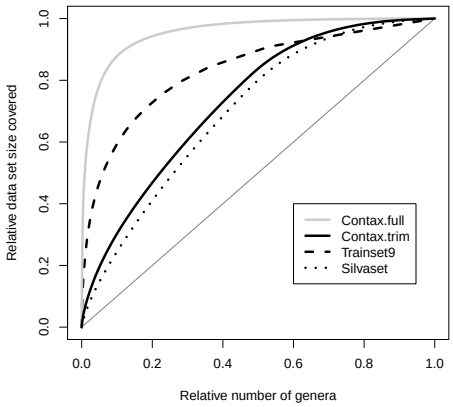


Figure 3: Data set skewness. The curves display data set skewness as follows: All genera in a data set have been sorted in descending order according to size (number of sequences) and arranged along the horizontal axis, which is then normalized to $(0, 1)$. The relative contribution of each genus to the full data set size is then accumulated along the vertical axis. A data set where all genera have the same size (no skewness) would result in a straight line from $(0, 0)$ to $(1, 1)$, indicated in the plot, with an area-under-curve of 0.5. A maximum skewness (one huge genus plus only singeltons) means an area-under-curve would be close to 1.0.

The total accuracy in classification for all three data sets and all three methods are displayed in Table 2. Singleton genera are not included in any of the results since they always will appear as mis-classified, by any method.

Table 2: The classification results (% correctly classified sequences) with the different methods and for various data sets. For both the RDP-classifier and the Multinomial a word-length of $K = 8$ was used.

METHOD	Silvaset	Trainset9	Contax.trim
RDP	93.45	94.59	99.03
Multinomial	94.21	95.54	99.26
BLAST	93.75	94.40	98.05

In addition to the total accuracy we also inspected the accuracy within each genus. We only report the results for `contax.trim` with the Multinomial method, as the results follow the same pattern as in Table 2. For the 1618 non-singleton genera the average accuracy was 98.59%. 1389 genera were perfectly classified (no errors), while two genera were completely missed: *Thermosyntropha* and *Candidimonas*. These consisted of only two sequences each. A total of 13 genera classified 50% or worse, all consisting of two sequences each. Figure 4 shows how errors accumulate over genus sizes. The first bar marks the number of errors made for genera of size 2, and then accumulated over gradually larger taxa. Of a total of 287

mis-classified sequences (horizontal line), 75% of them are seen among genera of size 39 or less (vertical line).

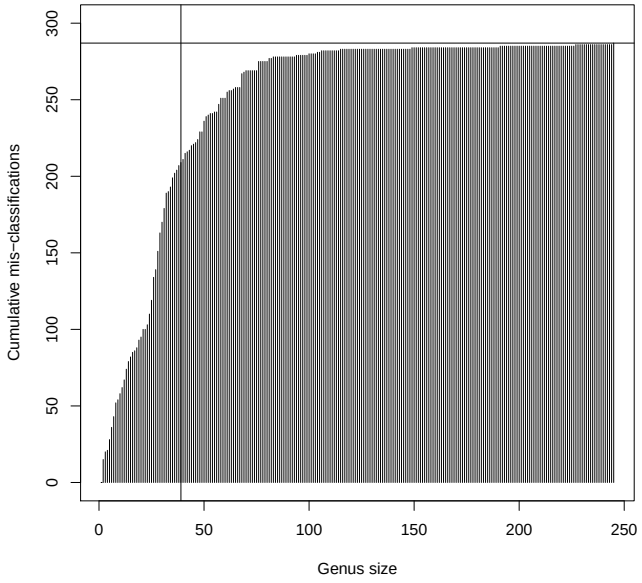


Figure 4: Errors and genus size. The figure shows how the number of errors accumulates over genus size. These results are based on the Multinomial method and the `contax.trim` set. The horizontal axis lists genus sizes in ascending order. The smallest size is 2 sequences (singeltons discarded), and these genera produce 15 errors. The vertical line marks that 75% of the 287 errors are made in genera of size 39 or less.

The amplicon-sequences extracted *in silico* by the primer pair $515F/R806$ were also tested using `contax.trim` in order to get a picture of how good classifications we may expect from a metagenome sample. First, with this primer pair 7117 (out of $38,784$) sequences from `contax.trim` had no perfect match. This means close to 20% of the sequences will produce no or perhaps only a few amplicons. A total of 67 genera (out of 1774) were completely lost, producing no perfect amplicons. These genera were in general small, except *Propionibacterium* with 205 sequences where none produced a primer match. The mean amplicon length was 292 bases. Among the extracted amplicons the total classification accuracy was 93.19% for the RDP classifier and 93.83% for the Multinomial classifier.

The Positive Predictive Values (PPV) were also computed for each genus. Of the 1658 unique genera assigned a sequence, 1348, had a perfect PPV of 100%. Additionally, 198 had a PPV of more than 90%

while 70 genera ended in the interval $40 - 90\%$. The remaining had either a PPV of 0 (42 genera) or were never predicted (116 genera), leaving the PPV unknown. Again we observed that a small or non-existing PPV coincides with genera of very few sequences (1 or 2).

DISCUSSION

The ConTax set

Developing and improving methods for rapid and accurate recognition of sequence-patterns is at the very heart of modern biology. Taxonomic classification, or binning, based on the 16S rRNA gene is in focus for several good reasons, but the profiling of microbial communities is the most obvious. This means implementation of some type of supervised learning method. A fundamental requirement in all method development is a stable training data set, and the most important feature of this data set is its reliability. In order to recognize taxon A we must be certain that our methods have been trained on sequences who are in fact from taxon A. In our effort to develop and improve methods, we concluded in [10] that the lack of such data is currently the most severe bottleneck.

In Table 2 we present some results from our pursuit to establish a high-quality training data set of 16S sequences. From the three repositories we downloaded several million sequences, but after the quality filtering we arrived at a total of around 2.6 million full-length 16S gene variants. Of these, $822,073$ unique sequences were found in either two or all three repositories, distributed as shown in Table 1 (Step 1). It is obvious that the Silva and RDP databases were the most similar with respect to content. After Step 2 only sequences with identical genus assignment remain.

In the intersection between all databases (SRG) there is around 93% agreement on genus assignment. In total, the lack of consensus is around 15%, i.e. for approximately 1 out of 7 sequences the repositories disagree on the genus assignment! This is a horrifying result, indicating that BLASTing a metagenome sample against any randomly chosen database will produce a huge number of highly uncertain taxon assignments. This 'error level' is way larger than what is reported for most supervised learning method used for classification. Having an accuracy of, say 95%, in a classifier is fine, but of little help if it is trained on data where 15% of the sequences have incorrect taxon labels.

Despite our strict requirements we were able to collect as many as $664,199$ sequences in the `contax.full` data set, covering 1774 genera. Exploring this data set is in itself interesting, since it may tell us something about the 16S landscape as a whole, e.g. how well defined a genus really is, how well they are separated, etc. However, from the viewpoint of developing supervised learning methods, this data set is too large, in

the sense that many genera are too densely represented, and for this reason we produced a trimmed version.

Trimming

An optimal data set for training and testing methods should not be larger than necessary. The `contax.full` data set is extremely skewed, as expected. The largest genera contains many, very similar, sequences and only a smallish subset of these are really needed to recognize the genus. Our proposed trimming procedure is based on the idea that a genus only needs the 'core' and the 'shell' sequences, i.e. one in the centre and a number of representatives along the borders. We also discard extreme sequences in order to prevent the trimmed set to be infested by possibly incorrectly labelled sequences. By this trimming we reduce the full data set `contax.full` by around 20-fold to `contax.trim` with $38,784$ sequences. The skewness, visualized in Figure 3 and quantified by the auc-measure, is reduced substantially. Some information may of course be lost in this trimming, but the gain in time efficiency and memory usage benefit future method development. We regard this trimming as an instrumental part of designing a data set for improving taxonomic classification methods.

Classifications

From the results in Table 2 we see substantial improvement in classification accuracy for all three methods when cross-validated on `contax.trim` compared to the other two data sets. Multinomial, which turns out as the overall best classification method, only mis-classifies 287 sequences out of $38,625$ in this data set. We interpret the differences between the data sets as due to more correct genus-labels in `contax.trim` sequences. A sequence that, due to its content, clearly belongs to taxon A, but has been labelled as taxon B, will tend to produce an error when cross-validation is employed in combination with any classification method. Since the sequences in the ConTax data sets all have some degree of consensus in their taxon-labels, we presume that sequences on the border between two or more genera are under-represented. Thus, all accuracies achieved here are most likely optimistic with respect to real uncertainties when classifying new sequences, e.g. from a metagenome sample. However, as a data set for training supervised learning methods, the `contax.trim` is the best we have seen so far.

It should also be noted that extremely skewed data sets may produce artificially high total accuracy. Having an enormous number of sequences from one huge genus means virtually all of these will be correctly classified, and the total per-sequence accuracy becomes large due to this genus alone, even if almost none of the other taxa are recognized. For this reason, reported accuracies should always include per-taxon

results as well.

From the per-genus results presented we see that the errors are in most cases made for small taxa. Figure 4 is an illustration of this. Grouping taxa by their size, and accumulating errors accordingly, clearly demonstrates that the errors level off after reaching a certain size. There may be several reasons for this. First, a small taxon means very little information for training a classifier, and any method will tend to make more errors under such circumstances. Next, a small taxon is also more uncertain by nature. If we have few sequences in a genus, it is either because this has been sequenced very rarely or because all other sequences in this neighborhood has no consensus classification. The latter indicates that the taxon itself has a vague definition. Singleton taxa (1 sequence only) is the most extreme case. The ConTax data sets contains 156 singletons. This should be compared to 911 and 58 in `Trainset9` and `Silvaset`, respectively. By inspection, we found that none of the singletons from `Trainset9` and `Silvaset` are found in the ConTax sets. This reflects the uncertain nature of such genera. In order to really improve on the existing taxonomy the effort must be made to achieve safe ground with respect to these rarely seen taxa.

We also classified amplicon-sequences matching the primer pair $515F/R806$. This primer pair addresses both bacteria and archaea 16S sequences in the variable region $V4$ and is recommended especially for environmental samples [20, 21]. The per-sequence accuracy dropped from 99.26% for full-length sequences to 93.83% for the much shorter amplicons, using the Multinomial method. A drop was expected, since we have seen in a previous study that discriminating sites are scattered along the entire 16S gene, and any shorter sub-sequence (even $V4$) is bound to produce sub-optimal classifications [22]. However, the accuracy is still good compared to the bias introduced by the primers themselves. If we sampled an environment with a population identical to `contax.trim` with these primers, we would never detect close to 20% of the sequences, and 67 taxa would remain completely unseen.

The PPV for each genus is a measure of confidence once a new classification is made. It is an estimated probability of correct classification given the predicted genus. Using the Multinomial method on the `contax.trim` data set, we found that more than 75% of the 1774 genera have perfect PPV, and if we include those with PPVs above 90% we observed that more than 87% of the genera are predicted with very high confidence. Again, the small taxa are where we find low confidence.

CONCLUSION

In this paper we present the ConTax data sets specifically designed for training supervised learning methods to make taxonomic classification based on the 16S marker. Everything is publicly available for

free in the R data package `microcontax`. In our previous efforts to improve classification methods based on the 16S gene we observed an error-plateau for all tested methods [10]. One of the main conclusions was that training data quality must be improved in order to obtain better classification results. The ConTax data sets presented here is our effort to meet this requirement. These data are designed to span as many genera as possible while including only sequences with high-confidence consensus taxonomy assignments. We tested three commonly used classification methods on the trimmed ConTax set as well as two other publicly available training data sets. All methods showed a substantial improvement in cross-validated accuracy when trained on the ConTax set in comparison to the others. We registered that most of the mis-classifications are made for small genera, and any further expansion of the ConTax training data should be focussed on adding more sequences to the smallest taxa.

FUNDING

This work was supported by the Norwegian University of Life Sciences.

ACKNOWLEDGEMENTS

Hilde Vinjes scholarship has been fully financed by the Norwegian University of Life Sciences.

Conflict of interest statement.

None declared.

References

1. Siegl A, Kamke J, Hochmuth T, Piel J, Richter M, Liang C, Dandekar T, Hentschel U: **Single-cell genomics reveals the lifestyle of Poribacteria, a candidate phylum symbiotically associated with marine sponges**. *The ISME Journal* 2011, **5**:61–70.
2. McLean JA, Lombardo M, Badger JH, Edlund A, Novotny M, Yee-Greenbaum J, Vyahhi N, Hall AP, Yang Y, Dupont CL, Ziegler MG, Chitsaz H, Allen AE, Yooseph S, Tesler G, Pevzner PA, Friedman RM, Nealson KH, Venter JC, Lasken RS: **Candidate phylum TM6 genome recovered from a hospital sink biofilm provides genomic insights into this uncultivated phylum**. *Proceedings of the National Academy of Science* 2013, **www.pnas.org/cgi/doi/10.1073/pnas.1219809110**.
3. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy A, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers Y, Smith HO: **Environmental Genome ShotguncSequencing of the Sargasso Sea**. *Science* 2004, **304**.
4. Nikolaki S, Tsiamis G: **Microbial Diversity in the Era of Omic Technologies**. *BioMed Research International* 2013, **http://dx.doi.org/10.1155/2013/958719**.
5. Woese CR: **Bacterial evolution**. *Syst Appl Microbiol* 1987, **51**:221–271.
6. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL: **Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB**. *Appl Environ Microbiol* 2006, **72**:5069–5072, [http://greengenes.lbl.gov/cgi-bin/nph-index.cgi].
7. Wang Q, Garrity GM, Tiedje JM, Cole JR: **Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy**. *Applied and enviromental Microbiology* 2007, **73**:5261–5267.
8. Lanzén A, Jørgensen SL, Huson DH, Gorfer M, Grindhaug SH, Jonassen I, Øvre L, Urich T: **CREST – Classification Resources for Environmental Sequence Tags**. *PLoS ONE* 2012, **7**(11).
9. Werner JJ, Koren O, Hugenholtz P, DeSantis TZ, Walters WA, Caporaso JG, Angenent LT, T KR, , Ley RE: **Impact of training sets on classification of high-throughput bacterial 16s rRNA gene surveys**. *The ISME Journal* 2012, **6**:94–103.
10. Vinje H, Liland KH, Almøy T, Snipen L: **Comparing K-mer based methods for improved classification of 16S sequences**. *BMC Bioinformatics* 2015, **16**:205.
11. Parte A: **LPSN—list of prokaryotic names with standing in nomenclature**. *Nucleic Acid Research* 2013, **doi:10.1093/nar/gkt1111**:1–4.
12. Pruesse E, Quast C, Knittel K, Fuchs B, Ludwig W, Peplies J, Glöckner FO: **SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB**. *Nucleic Acids Research* 2007, **35**:7188–7196, [http://www.arb-silva.de/].
13. Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ, Kulam-Syed-Mohideen AS, McGarrell DM, Marsh T, Garrity GM, Tiedje JM: **The Ribosomal Database Project: improved alignments and new tools for rRNA analysis**. *Nucleic Acids Research* 2008, **37**:D141–D145, [http://rdp.cme.msu.edu/].
14. Nawrocki EP, Eddy SR: **Infernal 1.1: 100-fold faster RNA homology searches**. *Bioinformatics* 2013, **doi:10.1093/bioinformatics/btt509**:1–3.
15. Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy EW S Rand Floden, Gardner PP, Jones TA, Tate J, Finn RD: **Rfam 12.0: updates to the RNA families database**. *Nucleic Acids Research* 2014, **43**(D1), [http://rfam.xfam.org/].
16. Nei M, Kumar S: *Molecular Evolution and Phylogenetics*. Oxford university press 2000.
17. Liu K, Wong T: **Naïve Bayesian Classifiers with Multinomial Models for rRNA Taxonomic Assignment**. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2013, **10(5)**:1334–9.
18. Altman D, Bland J: **Diagnostic tests 2: Predictive values**. *BMJ: British Medical Journal* 1994, **309(6947)**:102.
19. R Core Team: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria 2015, [http://www.R-project.org/].

20. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, Fierer N, Knight R: **Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample**. *Proceedings of the National Academy of Sciences USA* 2011, **108(Suppl 1)**:4516–4522.

21. Kuczynski J, Lauber CL, Walters WA, Wegener Parfrey L, Clemente JC, Gevers D, Knight R: **Experimental and analytical tools for studying the human microbiome**. *Nature Reviews Genetics* 2012, **13**(47-58).

22. Vinje H, Almøy T, Liland KH, Snipen L: **A systematic search for discriminating sites in the 16S ribosomal RNA gene**. *Microbial Informatics and Experimentation* 2014, **4:2**.

Paper IV

`microclass`: An R-package for 16S taxonomy classification

Kristian Hovde Liland[1,2], Hilde Vinje[1], Lars Snipen*[1]

[1]Department of Chemistry, Biotechnology and Food Sciences, Norwegian University of Life Sciences, P.O.Box 5003, N-1432 Ås, Norway

[2]Nofima - Norwegian Institute of Food, Fisheries and Aquaculture Research, Osloveien 1, N-1430 Ås, Norway

Email: Kristian Hovde Liland - kristian.liland@nofima.no; Hilde Vinje - hilde.vinje@nmbu.no; Lars Snipen*- lars.snipen@nmbu.no;

*Corresponding author

Abstract

Background: Taxonomic classification based on the 16S rRNA gene sequence is important for the profiling of microbial communities. In addition to giving the best possible accuracy, it is also important to quantify uncertainties in the classifications.

Results: We present an R package with tools for making such classifications, where the heavy computations are implemented in C++ but operated through the standard R interface. The user may train classifiers based on specialized data sets, but we also supply a ready-to-use function trained on a comprehensive training data set designed specifically for this purpose. This tool also includes some novel ways to quantify uncertainties in the classifications.

Conclusions: Based on input sequences of varying length and quality, we demonstrate how the output from the classifications can be used to obtain high quality taxonomic assignments from 16S sequences within the R computing environment. The package is publicly available at the Comprehensive R Archive Network.

Keywords: R, taxonomy, 16S.

Background

The profiling of microbial communities by the sequencing of the 16S rRNA gene has become a standard approach in metagenomics [1]. This means that collected DNA is subject to a targeted sequencing to extract a chosen region of the 16S gene from all organisms in the sample. The actual content of the sample can then be described by performing a large scale taxonomic classification of these sequences, i.e. assign them to the proper taxonomic bin, also referred to as binning [2]. Since 16S-based microbial profiling has become such a widely adopted approach, it is also important that the bioinformatics tools involved are optimized to the highest standard. Today's most widely used tool for this job is the RDP-classifier [3]. It is beyond doubt a good tool for this job, but at the same time it is not perfect, and in a systematic testing of this and other approaches we found there were always other methods that performed better [4].
Alternative tools are necessary for scientific evolution, and here we present a software to be used within the popular R computing environment [5].

There are some issues that must be considered when it comes to making tools for the binning of 16S sequences. First, the pattern recognition algorithm itself must be capable of recognizing the, sometimes small, differences in DNA that separates taxa. It must also handle the huge amount of bins or categories we are facing here, thousands rather than 2-3 which is often the case in textbook literature. Precision also very much depends on the quality of the training data [6]. Since there is no official taxonomy of prokaryotes, there are no real gold standard data sets available. As we concluded in [4], this is probably the most severe bottleneck. In [7] we recently published designed data sets based on a consensus taxonomy assignment among several data repositories, which is probably the closest we get to a gold standard today. Speed is another issue. With today's sequencing technology and low prices, a data set may easily contain millions of reads. Some procedures for OTU (Operational Taxonomic Unit) picking will start out by classifying reads to pre-defined taxa [8]. Thus, the number of sequences to classify may be huge. Other procedures cluster the reads before taxonomy assignments, defining OTU's as 'spherical' groups in a space of evolutionary distances approximated by alignment percentage identity, and then classify only the cluster centroids [9]. If we are interested in recognizing specific taxonomic profiles, e.g. in forensic applications [10], the classification of all reads into pre-defined bins is clearly what we seek.
Uncertainty is the third issue. In any collection of reads there will be a number of sequences that cannot be given a high-confidence classification. There are several reasons for this. First, the taxonomy itself is not always well defined, and sometimes even high-quality sequences fall on the border between existing taxa, making the classification uncertain. Second, due to sequencing errors and chimeras some reads may be difficult to recognize, and third, some microbial communities will contain new taxa not previously seen. In the presented R-package we have implemented some algorithms that have proved efficient and/or are much used for 16S taxonomic classification. Efforts have been made to make them both fast and memory-efficient. All methods can be trained on new data, but we have also supplied the package with a ready-to-use tool that is already trained and optimized for the `contax.trim` data set from [7]. This tool also quantifies uncertainties in a new way. The `microclass` R-package, as well as its symbiotic data package `microcontax`, are freely available at the Comprehensive R Archive Network (CRAN, [11]).

Implementation

The multinomial method

Based on our previous testing of classification methods in [4] we found that the best overall results were obtained by the algorithm denoted the multinomial method [12]. Thus, we have focused the attention on

this method in this package. The function `multinomTrain` is used to train a model of this type on any data set containing FASTA-formatted sequences along with the correct taxon assignments for each sequence. The function `multinomClassify` is then used to classify new sequences based on a trained model.

Both training of a multinomial model and classification of new sequences involves counting a large number of K-mers (overlapping words of length K) in the sequences. The overhead when doing such operations is large, and efficient vectorization is difficult to achieve. A direct implementation would also require the computation of a matrix product of size $(N \times 4^K) \cdot (4^K \times M)$, where N and M are the number of sequences to classify and the number of taxa in the training data, respectively. This is a time consuming task for large N, M and K. Therefore, these computations have been implemented in C++ through the `Rcpp` [13] interface in R, and some short-cuts are made.

The nucleotide sequences are first converted to integer vectors my mapping A, C, G and T (or U) to 0, 1, 2, and 3, while all other letters are mapped to -4^{15}. The latter is done to easily discard K-mers including alien symbols when counting. For training of a multinomial model, all K-mers of each taxon are counted. The counting itself is done by sliding a window along the integer vector of each sequence and computing a position as the inner product between $[4^{K-1}, 4^{K-2}, ..., 4, 1]$ and the integers in the window. For each of the inner products, this position in the taxon's counting vector is increased by 1. The result is a matrix, $\boldsymbol{X}$, of size $(M \times 4^K)$ that holds the counts for all K-mers in all taxa. Finally, each position in the matrix is re-scaled to $\log_2\left(\frac{x_{ij}-P/4^K}{\sum x_{i\cdot}-P}\right)$, where P is the number of pseudo-counts added. This is stored in an $(M \times 4^K)$ matrix named $\boldsymbol{Q}$ to represent multinomial log-probabilities with pseudo-counts.

When classifying new sequences using the multinomial method, we avoid the mentioned matrix product by combining the K-mer counting with summing of multinomial log-probabilities. For each counted K-mer, the corresponding column in the $\boldsymbol{Q}$ matrix is added to the result, thus never explicitly creating the K-mer count matrix or performing the product with $\boldsymbol{Q}$. As such we reduce from $(4^K \cdot M)$ operations to $((n-K) \cdot M)$ for a new sequence of length n. For full 16S sequences (with $n \approx 1500$ bases) this is vastly more efficient from around $K = 6$ and up and is easily parallelized.

The `taxMachine`

Users often want a ready-to-use tool to classify (many) 16S sequences without having to perform all the training. Based on the work behind [4, 7] we have arrived at an optimized tool for classifying 16S sequences, called `taxMachine` in this package. The `taxMachine` is based on using the multinomial method with a word length of $K = 8$ and a pseudo count of 100. It has been trained on full-length 16S sequences to

recognize full or partial (reads) sequences at the genus level, using the designed and optimized `contax.trim` data set for training, see [7] for details. The `taxMachine` includes computations of classification uncertainties that requires a detailed explanation.

Classification uncertainty

Uncertainty in a taxonomic classification can be split into two types. The first type is when a sequence happens to be very close to the decision boundary between two or more taxa. We can be fairly certain it belongs to one of these taxa, but it lacks the final discriminative power to safely assign it to one of them. The second type of uncertainty occurs when something completely new is seen. This is not uncommon in metagenome samples, and should be flagged separately since it may indicate sequencing errors, chimeras or some novel type of organism.

The d-score

The first type of uncertainty is measured by what we name the d-score. Consider sequence i in a set of sequences that we want to classify. In the `taxMachine` the predicted genus of sequence i is found by computing the posterior log-probability for every genus, and classifying to the genus with maximum value. If we sort all posterior log-probabilities for sequence i in descending order, $p_{i,1}$ denotes this maximum, while $p_{i,2}$ is the second largest, etc. These log-probabilities all depend on the sequence length, since a longer sequence will in general contain more unique K-mers, and the posterior log-probability will be a sum with more (negative) terms. This is illustrated in the left panel of Figure 1. Here we have sampled fragments of random length (> 100 bases) from all sequences in the `contax.trim` data set, and then classified them, collecting the $p_{i,1}$ for sequence $i = 1, ..., 38\,781$. The $p_{i,1}$ values are clearly biased by sequence length, and their variance is also increasing for longer sequences.

We first normalize the posterior log-probability with respect to sequence length. We fitted linear regression models describing how both the mean and the standard deviation of the data in the left panel of Figure 1 varies by sequence length l (from $l = 1$ to 2500). Thus, if sequence i has length l it gets the normalized posterior log-probability

$$\tilde{p}_{i,1} = \frac{p_{i,1} - \hat{p}_l}{\hat{s}_l} \tag{1}$$

where $\hat{p}_l$ and $\hat{s}_l$ are the predicted mean and standard deviation at sequence length l, using the fitted regression models. Note that $p_{i,2}$ (and any other posterior log-probability) can be normalized in the same way, using the same fitted regression model.

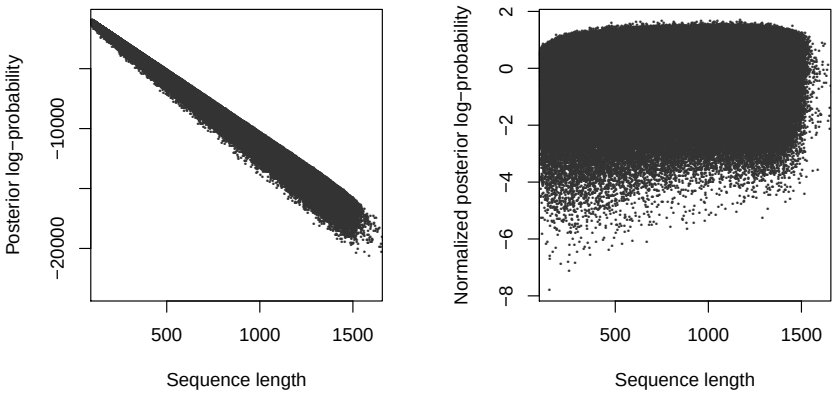


Figure 1: Posterior log-probability normalization. The left panel shows posterior log-probabilities for 38 781 sequences. The sequences are random sub-sequences of the `contax.trim` data set, spanning all lengths from 100 bases to more than 1500. Every sequence has been classified using the multinomial model trained on the full-length data, and each dot marks the maximum posterior log-probability for one sequence. There is clearly a linear trend in the values, with larger variance for longer sequences. In the right panel the same values are plotted after the normalization procedure described in the text.

The d-score of sequence i is simply the difference between the largest and the second largest normalized posterior log-probability:

$$d_i = \tilde{p}_{i,1} - \tilde{p}_{i,2} \tag{2}$$

If we are near a decision boundary we expect $d_i \approx 0$ since the second best genus is almost as good as the best. On the other hand, if $d_i >> 0$ it means the predicted genus is much more likely than any other, and we have a high confidence classification.

The r-score

The second type of uncertainty is high if we see something very different from what we have in the training data set. Consider sequence i belonging to genus g with corresponding normalized maximum posterior log-probability $\tilde{p}_{i,1}$ from (1). From all sequences belonging to genus g we computed the sample mean and sample standard deviation of the $\tilde{p}_{i,1}$'s, denoted $\bar{p}_g$ and s_g respectively. The r-score for sequence i is the standardized residual

$$r_i = \frac{\tilde{p}_{i,1} - \bar{p}_g}{\bar{s}_g} \tag{3}$$

where $\bar{s}_g$ is smoothed version of s_g as explained below. Thus, the r-score is a standardized measure of how different a sequence is from its predicted genus centre.

Different genera have different sequence diversity, which is reflected in different values of the sample standard deviation s_g. However, many genera have too few sequences to provide a reliable estimate of this standard deviation, some even have only 1 sequence making s_g impossible to compute. Thus, the $\bar{s}_g$ in (3) is based on a simple smoothing. First, all sample standard deviations where grouped by genus-size. In Figure 2 we show how smaller genera (few sequences) tend to have smaller sample standard deviations. We used the `loess` method [14] to estimate the size-specific sample standard deviation, shown as black squares in Figure 2. We denote this s_n where n is the genus-size. If genus g has size n we get the genus-specific standard deviation estimate as

$$\bar{s}_g = \sqrt{\frac{(n-1)s_g^2 + s_n^2}{n}} \tag{4}$$

When a new sequence is classified, we do not know its true genus. The predicted genus is then used as a plug-in in (3), i.e. we use $\bar{p}_g$ and $\bar{s}_g$ where g is the predicted genus. If the resulting r_i has a large negative value, it means the computed $\tilde{p}_{i,1}$ is much smaller than the average $\bar{p}_g$ for genus g, and sequence i is unlikely to belong to this genus even if this is where it maximizes the posterior probability.

Exactly how negative is the r-score for an un-recognized sequence? To guide this decision we computed the r-scores for all sequences in the `contax.full` data set [7], and from this we computed the empirical cumulative distribution function. For any given r_i value this gives us the probability of having an r-score this small, or smaller, given that the sequence was from the training data. A very small probability means the sequence is very unusual compared to the training data.

Other methods

The package also contains some alternatives to the multinomial method, mostly for comparisons. The RDP-classifier [3] is a popular tool used in many metagenome applications. The version implemented here is a stripped version without the bootstrapping effort to quantify uncertainties in the classifications. It has been implemented in C++ and accelerated similarly to the multinomial method, see above for details.

A classification using BLAST is also included, since this approach has been common. It is both slower and less precise then the other methods. It requires the BLAST+ software to be installed on the system.

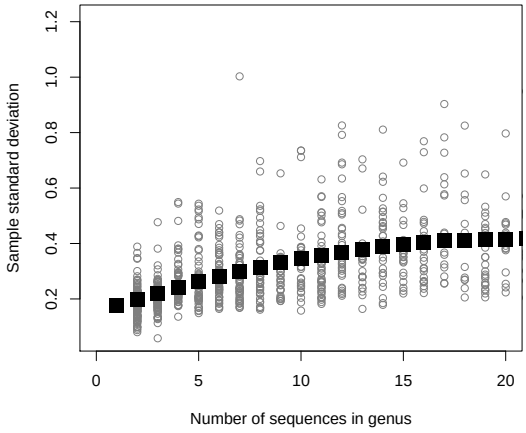


Figure 2: Smoothing genus standard deviation. The sample standard deviations for every genus (grey rings) are plotted against genus size (number of sequences). The black squares are the mean values for each genus size, after loess-smoothing as described in the text.

Results and Discussion

The `microclass` package provides optimized tools for taxonomic classification of 16S sequence data in the R computing environment. Some well established and proven methods are available to all users of R, with the possibility to train all methods on new and specialized data sets. However, a ready-made classification tool, `taxMachine`, is also supplied as an R-function. This has been optimized in several ways to produce the most accurate classifications at the genus level, without consuming too much memory. Specifically, it employs K-mers of length 8, where an increase to $K = 9$ or $K = 10$ comes at high cost in computation time and memory consumption compared to the small gain in accuracy for genus classifications. Pseudo counts have been set to 100 in the `taxMachine` as a robust compromise regardless of sequence length (see Supplementary Figure 1).

The classification of 16S is the most fundamental approach to profiling a microbial community, and due to the explosion in metagenomic research activities, tools for recognizing taxa from 16S sequences (reads) should be tuned to their optimal performance. The `taxMachine` R-function builds on a parallelized sparse-matrix implementation of the multinomial method that makes it efficient both with respect to speed and memory usage. It has been trained on the `contax.trim` data set, containing 38 871 full-length high-quality sequences covering 1774 genera, where all sequences have a consensus taxonomy, making it the

closest we get to a well-balanced gold standard training set.

As described in the Implementation section the `taxMachine` provides information about classification uncertainty, based on the posterior probabilities of the multinomial model. The very first step needed in these computations is to remove the bias from sequence length in the log-probabilities, as suggested in equation (1). The right panel of Figure 1 shows how the normalized posterior log-probabilities have no apparent trends over sequence length, as opposed to the raw-values in the left panel. This normalization makes it possible to compute uncertainty/reliability scores to sequences regardless of their exact lengths. The proposed d-score for a sequence is the difference in score between the most likely and the second most likely taxon. A d-score close to 0 means the sequence is close to a decision border, being almost equally similar to both taxa, and more likely to be mis-classified. To visualize this, we classified fragments of all sequences in the `contax.trim` data set using the `taxMachine`. We considered fragments of typical read-lengths; 120-150 and 270-300 bases, which is typical for Illumina HiSeq and MiSeq raw data, and 450-500 bases, which is typical for Roche 454 and merged (paired-end) Illumina MiSeq data. From each of the original 38 871 sequences we sampled 10 such fragments at random locations along each sequence. Comparing the predicted genus to the assigned genus, the error percentages were 1% for 450-500 bases reads, 3% for 270-300 bases and 11% for 120-150 bases, respectively, when the sequences from which the reads were generated were included in the model training (see Supplementary Table 1 for cross-validated success rates). The d-score should ideally be small for the mis-classified sequences, and large for the others. In Figure 3 we show a ROC analysis where all sequences are ranked by their d-score. Based on the large AUC statistics ($0.92 - 0.93$) we conclude that a small d-score is an effective criterion for identifying mis-classified sequences. In Figure 4 we show how the d-score distributes for the mis-classified sequences. Clearly, the majority has a d-score below 1.0 and the shorter the reads the more the d-scores are concentrated near 0. The probability of mis-classification will in general never exceed that of correct classification even for d-score almost at 0, but at 0 there is a $50 - 50$ chance of making a mistake. Various applications will require different strictness, but a classification with d-score above 1.0 can in general be considered safe. Based on the results in Figure 4 we found that among all classifications with a $d > 1$ there were 1.1%, 0.7% and 0.3% errors for input sequences of lengths 120-150, 270-300 and 450-500 bases, respectively.

We face a different type of uncertainty when we collect sequences very different from what we have seen in the training data set. In the Implementation section we describe the r-score to detect this. A negative r-score means the sequence has a lower probability than average for the assigned taxon. But how much

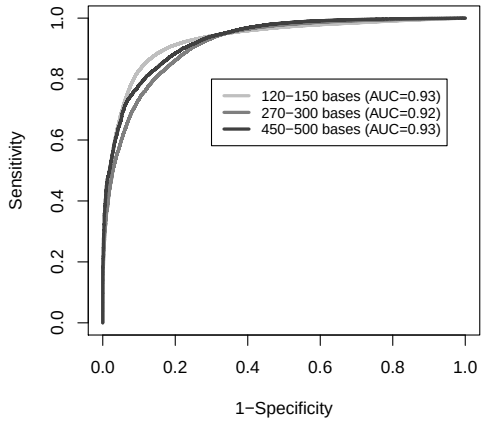


Figure 3: ROC analysis of *d*-scores. Based on the classification of read-length fragments, each sequence was either correctly or incorrectly classified. Each sequence also has a *d*-score. Ranking by *d*-score produced a separation of incorrect and correct classifications as indicated by the ROC-curves and the corresponding AUC statistics. Each curve is based on results for 387 810 sequences.

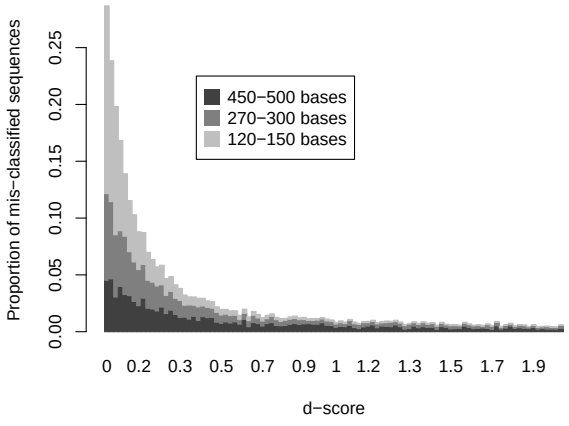


Figure 4: Histogram of *d*-scores. The histograms show the proportions of *d*-scores for the mis-classified sequences only, in the range from 0 to 2.0. This is from the same results as Figure 3, with three different read-lengths.

lower than the average is critical? To investigate this we used the same results as mentioned above, classifying sub-sequences of typical read-lengths, but in addition we also included full-length sequences. We then computed the r-score for all correctly classified sequences. Figure 5 shows the r-score densities for the various cases. It is the heavy left tail of the densities that is of interest. First, we notice there is some difference between densities for sequences of different lengths. Next, we see that even for correctly classified sequences, a very negative r-score occurs in a few cases. An r-score below -4 to -5 is rare for correctly classified sequences, and indicates an unusual sequence. The `taxMachine` also provides a probabilistic measure related to the r-score. Based on the `contax.full` data set (664 199 sequences) we computed densities similar to those in Figure 5, and from these the empirical cumulative distribution functions. The probability $Pr(r < r_i|\text{training data})$ is found from this distribution, for any given r_i. This probability reflects how unusual a sequence is compared to the training data, and if this is very small, its classification is not reliable.

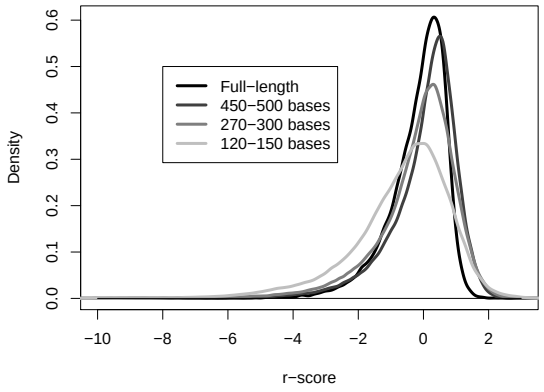


Figure 5: Densities of r-scores. Based only on correctly classified sequences, the densities show how the r-scores distribute. The densities were estimated by a non-parametric kernel smoother in R. Only negative r-scores are of interest, since a (very) negative value indicates a (very) unusual sequence.

In Figure 6 we demonstrate how the r-score histograms change when faced with sequences from unknown taxa. Here we have only focused on sub-sequences of lengths $450 - 500$, but the results were similar for other sequence lengths as well. We used a taxon-wise cross-validation, i.e. in each iteration we leave out all sequences from a taxon, train the model on the rest, and classify the sequences of the left-out taxon. This

means all classified sequences are from an unknown taxon, not part of the training data. The upper left panel shows, for comparison, how the distribution looks like without this cross-validation (mean r-score -0.1). In the upper right panel each genus has been left out, i.e. the training data contains no sequences from the genus of the classified sequence. The r-scores in general become more negative even if some are still quite large, even positive (mean r-score -13.5). This is not surprising, since many genera are quite similar, and a sequence from the neighboring genus may not look very unusual. In the lower panels we have cross-validated over order and phylum (mean r-scores -17.0 and -19.5), making the classified sequences gradually more distant from those of the training data. The lower left tail of the histograms are thinner, and a substantial number of sequences got very negative r-scores well outside the range of the plots. The proportion of sequences in the green-yellow region (large r-scores) is gradually smaller.

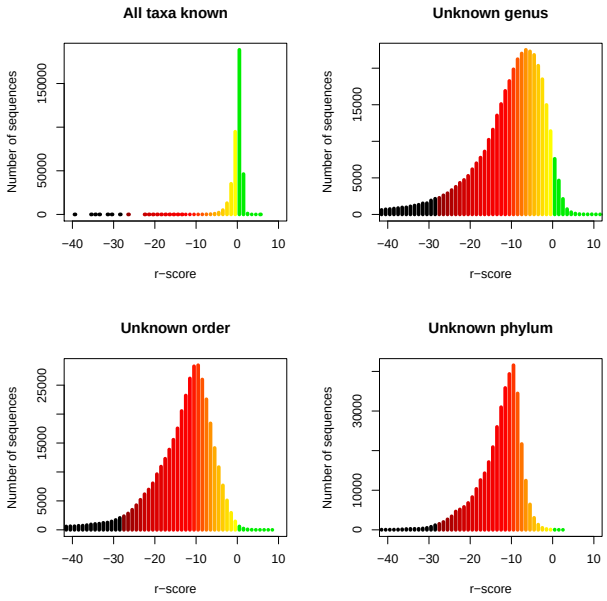


Figure 6: Effect of unknown taxa on r-scores. The four histograms show distribution of r-scores. The colors are: Green for all positive r-scores and black for scores more negative than ever observed in the `contax.full` data set. The transition from yellow to red indicates gradually smaller probabilities (from around 10^{-1} at yellow to 10^{-8} at dark red) of observing the corresponding r-score in the training set. Red colors are probabilities below 10^{-5}. The upper left panel are r-scores where all classified taxa are present in the training data, i.e. no unknown taxa. In the upper right panel each genus is unknown, i.e. when classifying a sequence from genus A, there are no sequences from this genus in the training data. In the lower panels the same procedure has been repeated but the training data lack sequences from the same order and phylum, respectively.

Figure 7 illustrates, in a similar way, the effect of sequencing errors. The sequences from the upper-left panel of Figure 6 have been corrupted with random sequencing errors at two levels, and then classified. A 1% error level will distort the r-scores, but still the majority of sequences are recognized to an acceptable level, with r-scores above -6. In total, more than 98% of the sequences are correctly classified. At 5% sequencing error the majority of the reads have r-scores well into the red and even black region, indicating unrecognised sequences. Still, more than 90% of them are correctly classified, mostly those with the larger r-scores. Chimera sequences will also result in sequences that are different in K-mer composition from its source sequences. In Supplementary Figure 2 an example of such a mixture, including d- and r-scores is shown.

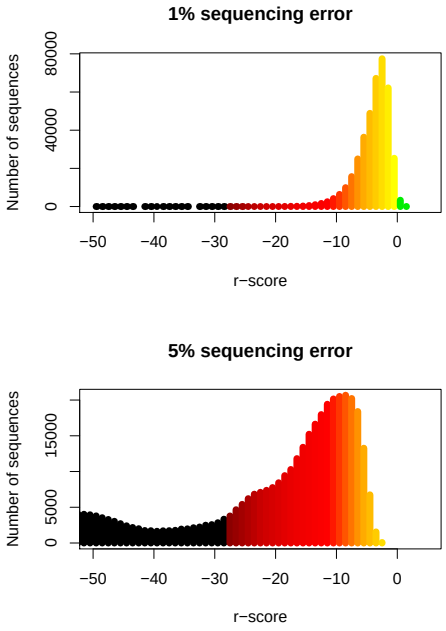


Figure 7: Effect of sequencing error. Histograms of r-scores similar to those in Figure 6, but for two levels of sequencing errors. Reads of lengths $450 - 500$ bases were corrupted at random before classification.

The r-score, and/or its corresponding probability, may be used to discard sequences that appear unusual. As always, the strictness of this procedure will depend on the application. For most applications we would not discard reads unless they are in the lower 1% or 0.1% quantile, at least (probabilities smaller than $10^{-2} - 10^{-3}$). Instead of fixing some threshold, and discarding reads, one may also use these probabilities

as weights, and give reads with small r-scores less weight. When tabulating read-counts into a taxonomic profile, this seems like a natural procedure. Conservative estimates of the expected success rates in classifying new reads and full sequences can be found in Supplementary Table 1.

Conclusions

The package `microclass` offers tools for taxonomic classification based on 16S rRNA sequence data to the R community. There are function for training classifiers on your own, specialised data sets, and then use these classifiers to classify new sequences. The `taxMachine` function has synthesized the designed training data from the `microcontax` data package with the methods of this package, and is our suggested tool for general classifications. It also implements some novel ways to express uncertainties in the classifications, indicating if the input sequences are difficult to recognize.

List of abbreviations

RDP - Ribosomal Database Project

OTU - Operational Taxonomic Unit

CRAN - Comprehensive R Archive Network

BLAST - Basic Local Alignment Search Tool

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

The R package is available for free from The Comprehensive R Archive Network [11]. It is most easily obtained by starting R and running
`install.packages("microclass",repos="http://cran.r-project.org/")` in the console window. All data used in this paper are also publicly available, in the R-package `microcontax`, at [11].

Competing interests

The non-profit corporation Nofima - Norwegian Institute of Food, Fisheries and Aquaculture Research has no competing interests related to this publication or the presented software. The authors declare that they have no competing interests.

Funding

This project has been financed by the Norwegian University of Life Sciences.

Authors' contributions

All authors have contributed significantly to all programming, documentation and preparation of this manuscript.

Acknowledgements

Not applicable.

References

1. Özlem Taştan Bishop (Ed): *Bioinformatics and Data Analysis in Microbiology.* Rhodes University Bioinformatics, Department of Biochemistry, Microbiology and Biotechnology, Rhodes University, South Africa: Caister Academic Press 2014.
2. Schloss PD, Westcott SL: **Assessing and Improving Methods Used in Operational Taxonomic Unit-Based Approaches for 16S rRNA Gene Sequence Analysis**. *Appl. Environ. Microbiol.* 2011, **77**(10).
3. Wang Q, Garrity GM, Tiedje JM, Cole JR: **Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy**. *Applied and enviromental Microbiology* 2007, **73**:5261–5267.
4. Vinje H, Liland KH, Almøy T, Snipen L: **Comparing K-mer based methods for improved classification of 16S sequences**. *BMC Bioinformatics* 2015, **16**:205.
5. R Core Team: **R: A Language and Environment for Statistical Computing**. R Foundation for Statistical Computing, Vienna, Austria 2015, [http://www.R-project.org/].
6. Werner JJ, Koren O, Hugenholtz P, DeSantis TZ, Walters WA, Caporaso JG, Angenent LT, T KR, , Ley RE: **Impact of training sets on classification of high-throughput bacterial 16s rRNA gene surveys**. *The ISME Journal* 2012, **6**:94–103.
7. Vinje H, Liland KH, Snipen L: **A consensus taxonomy for improved classification of prokaryotes** 2016. [Submitted manuscript].
8. Caporaso J, Kuczynski J, Stombaugh J, Bittinger K, Bushman F, Costello E, Fiere N, Pena A, Goodrich J, Gordon J, Huttley S GA andKelley, Knights D, Koenig J, Lozupone C, McDonald D, Muegge B, Pirrung M, Reeder J, Sevinsky J, Turnbaugh P, Walters W, Widmann J, Yatsunenko T, Zaneveld J, Knigh R: **QIIME allows analysis of high-throughput community sequencing data**. *Nature Methods* 2010.
9. Edgar R: **UPARSE: highly accurate OTU sequences from microbial amplicon reads**. *Nature Methods* 2013, **10**:996–998.

10. Leake S, Pagni M, Falquet L, Taroni F, Greub G: **The salivary microbiome for differentiating individuals: proof of principle**. *Microbes and Infection* 2016, **18**:399–405.

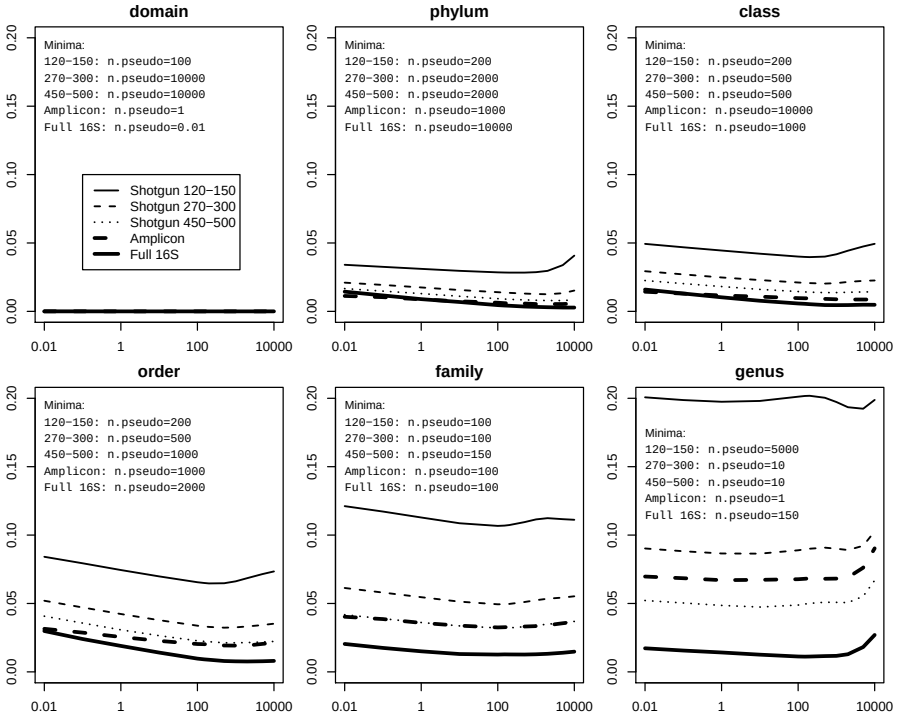
11. **Comprehensive R Archive Network**[https://cran.r-project.org/].

12. Liu K, Wong T: **Naïve Bayesian Classifiers with Multinomial Models for rRNA Taxonomic Assignment**. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2013, **10(5)**:1334–9.

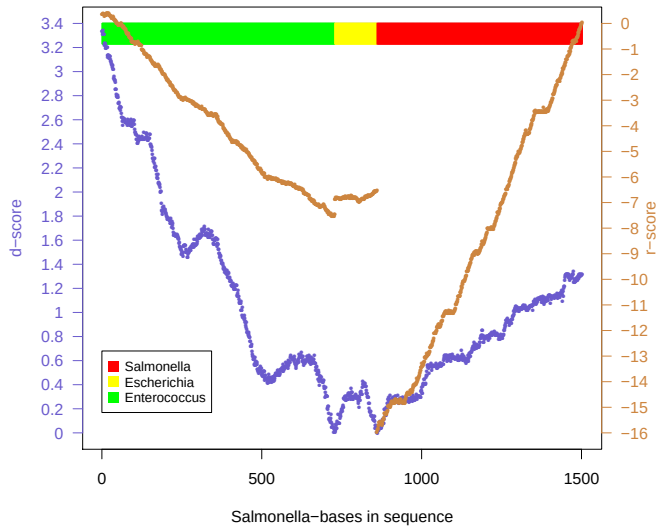
13. Eddelbuettel D, Francois R: **Rcpp: Seamless R and C++ Integration**. *Journal of Statistical Software* 2011, **40**(8):1–18.

14. Cleveland WS, Grosse E, Shyu WM: **Local regression models**. In *Statistical Models in S*. Edited by Chambers JM, Hastie TJ, Wadsworth & BrooksCole 1992:Chapter 8.

Additional Files



Supplementary Figure 1: Effect of pseudo-counts. The fraction of mis-classified sequences after 10-fold cross-validation, using various number of pseudo-counts in the training of the multinomial models (horizontal axes). This is based on the `contax.trim` data set, and models have been trained at all levels of the taxonomy, from domain to genus (panels). The effects of different choices of pseudo-counts are modest, and at the genus-level the use of 100 pseudo-counts is a reasonable compromise for all types of input sequence lengths. The amplicon sequences are obtained by using the primer pair $515F$ (`GTGYCAGCMGCCGCGGTAA`) and $806rB$ (`GTGYCAGCMGCCGCGGTAA`) to extract subsequences, in general matching the V3-V4 region of the 16S gene.



Supplementary Figure 2: Chimera example. We constructed a chimera sequence by mixing *Salmonella* and *Enterococcus*. Both sequences have 1503 bases, and the chimera starts as *Salmonella* and ends as *Enterococcus*. The horizontal axis shows the number of *Salmonella* bases, i.e. if the n first bases are *Salmonella* then the $1503 - n$ last bases are *Enterococcus*. The blue axis/dots shows how the d-score changes as we gradually mix the two sequences, and the tan axis/dots similar for the r-score. The red/yellow/green band at the top shows the classification at each chimera level. On the left side, when only a minority of the sequence is *Salmonella*, it is recognized as *Enterococcus* (green region). In the middle, it is misclassified as *Escherichia* (yellow region), which is a fairly close relative of *Salmonella*, but as the *Salmonella*-part gets majority it is recognized as *Salmonella* (red region). Notice the low d-score values in the middle section, indicating uncertain classifications. The r-scores also drop in the middle region. The 'jumps' in r-score are due to the dependency of the classified genus. The posterior log-probabilities do not change abruptly, but the r-score is related to what we expect for the assigned genus, and the latter causes the switches.

Supplemtntary Table 1: Performance of the multinomial classifier. Number of misclassified for every 1000 sequences classified in the `contax.trim` data set using 10-fold cross-validation, removal of singletons, $K = 8$ and $n.pseudo = 100$. In parentheses are the effects of changing from $K = 8$ to $K = 10$, i.e. the reduction in mis-classified sequences. Increasing K leads to a substantial increase in memory usage and computing time, as the problem grows by $O(4^K)$, and at the genus-level the gain of increasing beyond $K = 8$ is too small.

	120-150bp	270-300bp	450-500bp	515f+806rB	Full 16S
domain	0 (-0)	0 (-0)	0 (-0)	0 (-0)	0 (-0)
phylum	29(-11)	14 (-4)	9 (-3)	6 (-4)	5 (-3)
class	40 (-8)	21 (-4)	14 (-3)	10 (-5)	6 (-3)
order	65 (-9)	34 (-6)	23 (-6)	20 (-7)	10 (-6)
family	107 (-3)	49 (-5)	32 (-5)	32 (-4)	12 (-5)
genus	197(+1)	85 (-0)	45 (-1)	64 (-2)	7 (-0)