Classification Methods for 16S rRNA Based Functional Annotation

Rafal Kulakowski, Adi Lausen, Etienne Low-Decarie and Berthold Lausen

Abstract Microbial communities play an essential role in Earth's ecosystems. The goal of this study was to investigate whether the functional potential of microorganisms forming these diverse communities can be directly identified using a *16S rRNA marker gene* with supervised learning methods. The recently developed FAPROTAX database has been used along with the SILVA database to produce a training set where 16S rRNA sequences are linked to a number of metabolic functions. Since gene sequences cannot be explicitly used as feature vectors by most classification algorithms, the present research aimed to investigate possible feature engineering approaches for 16S rRNA. Techniques based on *Multiple Sequence Alignment* (MSA) and *N-grams* are proposed and tested. The results showed that the feature representation based on the N-grams outperformed MSA, especially when implemented with large and diverse

Rafal Kulakowski · Adi Lausen · Berthold Lausen
University of Essex, Wivenhoe Park, Colchester CO4 3SQ
✉ rkulaka@essex.ac.uk
✉ a.lausen@essex.ac.uk
✉ blausen@essex.ac.uk

Etienne Low-Decarie
Data Centre of Expertise and Astronauts, Life Sciences and Space Medicine Canadian Space Agency, Government of Canada, 6767 Route de l'Aéroport, Saint-Hubert, QC, Canada, J3Y 8Y9
✉ etienne.decarie@gmail.com

Archives of Data Science, Series A
(Online First)
KIT Scientific Publishing
Vol. 4, No. 1, 2018

DOI: 10.5445/KSP/1000085951/17
ISSN 2363-9881

functional groups. This suggests that a clustering-like alignment procedure results in a biased feature representation of the marker gene. Since classifiers trained using *Random Forest* and *Support Vector Machines* techniques were able to accurately detect a range of functional groups it is concluded that the 16S rRNA gene provides substantial information for the direct identification of functional capabilities.

1 Introduction

Microorganisms, such as bacteria and archaea, can be found almost everywhere on earth. Their importance in global climate regulation, as well as human health (Widder et al, 2016) is becoming increasingly apparent due to the recent advancements in sequencing technologies. The term "*Next Generation Sequencing*" (NGS) technologies refers to a wide range of sequencing platforms such as "*Illumina*" and "*SOLiD*", which are at the forefront of the effort to improve our understanding of the microbial world. These platforms enable a wide range of sequencing methods to be implemented (for a comprehensive review, see Goodwin et al (2016); Quail et al (2012)).

DNA Approaches to Microbial Community Characterization

The two sequencing strategies in metagenomics typically applied to obtain relevant data for microbial community research are *Shotgun Sequencing* and *Targeted Sequencing*. While the former involves attempting to sample all genes in all organisms found in a given environment, the latter targets a specific easily identifiable marker gene that most, or all, microorganisms possess. 16S ribosomal DNA represents such a marker gene and is most frequently used for nucleic acid-based detection and identification of bacteria and archea (Quast et al, 2013). A number of conserved regions of 16S allow it to be easily detected, whilst its nine hyper-variable regions can be used to distinguish between microorganisms on the basis of genetic differences.

In an ideal scenario having the comprehensive genomic information about a microbial community is preferable to sequencing a single marker gene. In practice, however, shotgun sequencing often recovers only short fragments of the genome and thus provides highly ambiguous output. Moreover, in situations where a highly complex microbial community is surveyed, the attained information often

represents a small fraction of the total diversity within the targeted community (Delmont et al (2012); Tyson et al (2012)), and the amount of additional sequencing required for comprehensive covering often remains unknown.

The main task for which 16S rRNA sequences is currently used is known as *Taxonomic Annotation*. The taxonomic profiles of microbial communities provide insight about the species composition in the surveyed environment. However, the phylogenetic information is often less relevant for understanding the underlying ecological processes than the information about the functional capabilities of inhabiting microbes within complex communities (Louca et al, 2016). Therefore, a number of methods have been recently developed to retrieve this type of information from environmental samples. Most, however, require comprehensive genomic data, which may be difficult to acquire, because of technological limitations (Ekblom and Wolf, 2014).

16S rRNA Based Functional Annotation

Functional community profiling tools capable of using 16S rRNA data include FAPROTAX (Louca et al, 2016) and PICRUSt (Langille et al, 2013). Both these packages work with taxonomically annotated data presented in *Operational Taxonomic Unit* (OTU; Sokal and Sneath, 1963) table format. An OTU table can be used as an input with the FAPROTAX package, a database-driven solution to 16S rRNA functional annotation, which uses the latest literature to map OTUs to biochemical functions. Assuming that no error was committed during taxonomic annotation, using this package guarantees correct assignment of functions for taxa present in the database. It does not, however, allow predicting biochemical functions for microorganisms not found within the FAPROTAX database.

The PICRUSt package (Langille et al, 2013) can provide predicted functional profiles using available databases of genomes by inferring the gene content of those taxa not present in the databases. This process begins by using taxonomic information to identify the microbe's position in a phylogenetic tree and finding its closest relatives with full genome content in the database. The genome content of these microorganisms is then used to estimate the gene content of their shared ancestors, which are later used for the prediction of the queried microbe (Langille et al, 2013). Once the relevant functional genes have been inferred, the functional capabilities can be estimated. This approach is highly reliant on the database of reference genomes and is more likely to provide false predictions, if no genomes of microbes with similar homology to that

of the unknown examples are found within the database. Since there is no simple linear link between homology and functional capabilities (see Figure 3), distantly related microbes can often perform similar metabolic functions and, thus, PICRUSt's predictions are expected to suffer from homology related bias. In fact, as shown by Sun et al (2019) PICRUSt's accuracy degrades sharply when implemented with non-human samples, which have less representation in most genome databases, and thus it cannot be considered a reliable tool when inferring functional composition of non-human animal, soil or marine environments.

The present work aimed to examine whether instead of involving the step of taxonomic annotation with 16S rRNA in the classification process (which relies on the importance of phylogenetic information), the biochemical functions can be inferred directly from the marker genes. By implementing supervised learning methods, we tested whether 16S rRNA sequences could hold appropriate information for predicting a range of functional capabilities of microorganisms. Furthermore, we asessed possible approaches to represent these marker genes as features. Finally, we aimed to provide pre-processing pipelines in the R programming language to facilitate further development in methodology for 16S rRNA based classification of biochemical functions.

2 Dataset

To assess the potential of supervised learning methods for the 16S rRNA based function classification of microbial communities a dataset with a relevant number of microorganisms with known functional capabilities and their 16S rRNA sequences was required. To construct such a dataset, we used the SILVA database (Quast et al, 2013), which stores over four million 16S rRNA quality checked sequences with reliably identified and manually curated taxonomies. All sequences have been downloaded from the SILVA database (release 119) website in a single FASTA file format and processed in R to create an OTU table. In this format the taxonomic information was used in FAPROTAX to map biochemical functions to microbes based on the latest literature (Louca et al, 2016).

Over two million 16S rRNA sequences found in the SILVA database have been assigned at least one of the 90 biochemical function by FAPROTAX. The distribution of sequence lengths has three picks at approximately 500, 800 and 1300 bp's. This means that many sequences used in our classification problem

had at least some sequence fragments missing. Thus, the symbols found at a given position in two different sequences were likely to come from two different real positions in the original 16S rRNA genes.

3 Methods

Classification is a supervised learning task in which a d-dimensional feature vector $x = (x^{(1)}, ..., x^{(d)})^T \in X \subseteq R^d$ is used to predict a class label $y \in Y$. Decision models, known as classifiers $C : R^d \rightarrow Y$, are usually trained on a subset of available examples with known labels and their performance is evaluated based on the accuracy of predictions for previously unseen examples. Functional annotation of microbial communities based on 16S rRNA can be tackled as a set of k binary sequence classification problems, each describing whether a microorganism represented by a 16S rRNA sequence possesses a given functional capability. Therefore, we aimed to train k binary classifiers to map a sequence s to class label y, thus $C : s \rightarrow y \in \{1, 0\}$. Let $k \in \{1, \ldots, m\}$. If the class label $y = 1$ is assigned to a function s by the k-th classifier, then it is predicted that a microorganism can perform the k-th function, otherwise, if $y = 0$, it cannot.

The 16S rRNA sequences can be represented in mathematical notation as: $s = (s^{(1)}, ..., s^{(l)}) \in S^l$ with an alphabet of symbols $S \in \{A, C, G, U\}$, which does not match the predefined format of the feature vectors required by the classification algorithms. Whilst in practice some techniques, including Naïve Bayes and Tree classifiers can work with non-numeric variables, most would require all features to be numeric. In our training data we included sequences of varying lengths that are likely to come from different regions of the 16S rRNA gene. This means that one cannot assume that a symbol $s_i^{(j)}$, which is a j-th element of the sequence vector for i-th observation, corresponds to the same feature as $s_{i+1}^{(j)}$, an element in the same position in the sequence of another observation. For this reason, a "*Feature Representation*" (FR) method is required, so that FR: $s \rightarrow x$, where $x = (x^{(1)}, ..., x^{(d)}) \in X$ with the same number of dimensions d for each observation i.

The process of converting a raw data into usable feature vectors is referred to as "*Feature Engineering*" and it has a crucial influence on the performance of the classifiers. The expected accuracy of the resulting classifiers is the most impor-

tant criterion by which a given feature representation should be evaluated. The dimensionality of the resulting feature space should also be reduced if $d >> N$, where N is the number of observations, and should include only those features which have a significant impact on the decision of the classifier. Moreover, if interpretability is an important factor, one may choose to further decrease the number of features and avoid a complex transformation of the feature vectors.

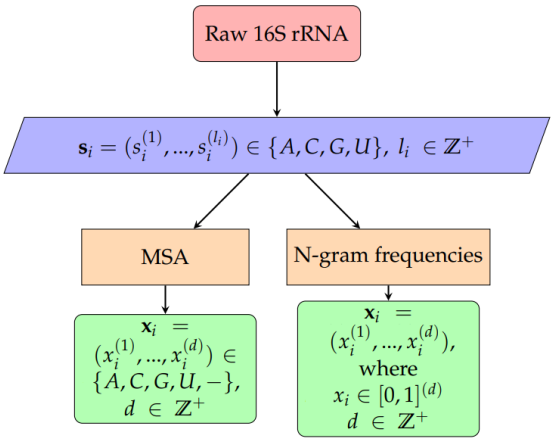

Figure 1: Feature Engineering approaches. The left-hand panel shows how categorical feature vectors are derived via the MSA approach. The numerical feature vectors derived via the N-gram approach are shown in the right panel. d = number of dimensions of MSA feature space; δ = number of dimensions of N-gram feature space.

In this paper, we propose two approaches to feature engineering of 16S rRNA sequences. The first approach, based on *Multiple Sequence Alignment* produces a categorical feature vector allowing for a number of simple classification methods to be used. The second one, based on *N-grams* (also known as *k-mers*), results in a preferable, numerical feature space, facilitating the use of the state-of-the-art classification methods. Both feature engineering approaches are summarised in Figure 1.

Multiple Sequence Alignment (MSA)

For the first approach to represent 16S rRNA as a feature vector, we tested whether aligning sequences in a way that standardises their structure may

result in a useful feature space for our classification problem. Alignment procedures are frequently used when analysing RNA or DNA data to identify regions of similarity between sequences. Significant similarity may indicate a functional, structural or evolutionary relationship between the sequences (Dong and Pei, 2007). During an alignment, gaps are inserted to signal missing information or sequence deletions and, regions of similar structures are shifted to the same columns. The objective of this procedure is usually related to minimisation of the number of columns with different characters and the number of times a gap is inserted. MSA is defined as an alignment of three or more sequences (Carrillo and Lipman, 1988). Whilst aligning two sequences is a trivial task, aligning a large number of sequences can be computationally expensive and thus involves a search for a possibly suboptimal solution using a heuristic method (for a comprehensive review of different approaches to sequence alignment, see Notredame (2002) and Thompson et al (2011)).

$$\mathbf{s} := \begin{cases} s_1 = \left(s_1^{(1)}, \ldots, s_1^{(l_1)}\right) \\ s_2 = \left(s_2^{(1)}, \ldots, s_2^{(l_2)}\right) \\ \vdots \\ s_N = \left(s_N^{(1)}, \ldots, s_N^{(l_N)}\right) \end{cases} \xrightarrow{\textbf{MSA}} \mathbf{x} := \begin{cases} x_1 = \left(x_1^{(1)}, \ldots, x_1^{(d)}\right) \\ x_2 = \left(x_2^{(1)}, \ldots, x_2^{(d)}\right) \\ \vdots \\ x_N = \left(x_N^{(1)}, \ldots, x_N^{(d)}\right) \end{cases}$$

Figure 2: Feature Engineering using the Multiple Sequence Alignment. The orginal set of sequences with different lengths l_i is represented by $\mathbf{s}$ and the aligned table by $\mathbf{x}$. N denotes the number of observations available and d the dimensionality of the resulting feature space.

The resulting aligned table (see Figure 2) can potentially be used as for classifier training due to the standardised structure and length of feature vectors. In R, packages capable of aligning multiple sequences of DNA or RNA include MUSCLE (Edgar, 2004) and MSA (Bodenhofer et al, 2015). Aligning sequences from the SILVA database often produces tables with over 2000 columns, resulting in high-dimensional categorical feature vectors. Such feature spaces can only be used by classifiers capable of handling categorical variables as the aligned tables contain only characters and the gap symbols "-". The candidate methods would include Naïve Bayes, Nearest Neighbours and Tree classifiers. Alternatively,

other classification algorithms such as Support Vector Machines and Deep Neural Networks would require either *One-Hot-Encoding* (i.e., constructing feature vectors with dummy variables) or *Feature Hashing* (Moody, 1989) to be implemented in order to attain a numerical feature representation. The former approach would result in a significant increase in dimensionality, as well as sparsity of the resulting feature space, which is likely to have a negative effect on the performance of the classifiers. Although, *Feature Hashing* techniques do not suffer from the same limitations, an effective encoding method needs to be found for the application to be successful (Weinberger et al (2009) & Attenberg et al (2009)). Such manipulations of a feature representation could improve the performance of classifiers trained on the resulting feature space. However, the current research did not test these approaches.

N-grams

The conventional classification methods require the feature vectors to be comprised of numerical variables. For our second approach to feature engineering for 16S rRNA, we will investigate whether converting the sequences into a desirable, numerical format by using *n-grams* will result in classifiers with good expected accuracy rates. In the field of text analytics, the phrase n-grams refers to the counts or frequencies of words or characters of length n found in a given text sequence. A n-gram based feature space is typically constructed by combining these counts into a single vector, which can be used as a feature representation of the original data. For RNA and DNA sequences n-grams are also reffered to as *k-mers* and typically represent the counts of short sequence segments of length n. They are extensively applied in metagenomics, including some taxonomic annotation pipelines (Wang et al (2007) & Lu et al (2017)), MSA algorithms (Edgar, 2004) and are essential for clustering (Kaisers et al, 2018) and alignment free sequence analysis (Huang (2016) & Rahman et al (2018)).

The N-gram counts can be computed in R using the biogram package (Burdukiewicz et al, 2017). In our classification problem, *1-gram* counts represent the number of times each symbol $\{A, C, G, U\}$ appeared in a single sequence $s_j = (s_j^{(1)}, ..., s_j^{(l)})$. Consequently, the *2-gram* counts report the number of times a pattern of two symbols $\{AA, AC, ..., UG, UU\}$ is found. The dimensionality of this type of feature vectors grows exponentially with the length of sequence segments that are counted. Having four symbols in the alphabet, the dimensionality $\delta = 4^n$ if only *n-grams* of length n are used. If

n-grams with $n = \{k, ..., K\}$ with $k, K \in \mathbb{Z}_+$ and $k \leq K$ are used, the number of dimensions is $\delta = \sum_{n=k}^{K} 4^n$. The feature vectors with n-grams of length 1 or 2 are unlikely to provide enough detail about the patterns within the 16S rRNA sequences to discriminate between microorganisms with different functional potential effectively. However, as we choose to use counts of longer sequence segments the number of features is going to cause the classifier training to be more computationally expensive.

Dimensionality Reduction

A number of dimensionality reduction methods have been developed to improve the robustness and interpretability of classifiers built on high-dimensional datasets. These are typically classified as either *Feature Selection* or *Feature Extraction* (see van der Maaten et al (2009), Xu et al (2017), for details) .

The feature selection methods for dimensionality reduction attempt to identify a subset of the original features which are the most informative for a given machine learning task. Such methods include: *Filtering*, *wrapper*, and *embedded* strategies.

Filtering of features is processed using chosen criteria such as a correlation, information gain or gain ratio. It benefits from being relatively computationally inexpensive compared to other strategies. It is especially suited for high-dimensional feature spaces and may also be used as a first step of a longer dimensionality reduction procedure.

The *wrapper* methods (Kohavi and John, 1997) include search algorithms which explore the possible combinations of features to find a set which allows good classifiers to be built. The search space for such optimisation problems is usually too large and the evaluation of new solutions computationally too expensive for an exhaustive search. Thus heuristic methods, such as a *Genetic Algorithm* (Goldberg, 1989) or *Particle Swarm Optimization* (Kennedy and Eberhart, 1995), are implemented to find a local optimum.

The *embedded* methods perform feature selection as part of model construction. The popular LASSO method (Tibshirani, 1996) is an example of such an approach. It reduces the feature space by constructing a linear model with a L1 penalty. This shrinks the coefficients of many features to zero and the algorithm uses the remaining features for final model building. LASSO has been originaly built for regression, but it can also be used in classification tasks. A binary classification task can be approached as a regression problem with

$y \in (0, 1)$ and the non-zero coefficients of LASSO may be used as feature space by a classification algorithm.

The feature extraction methods represent a completely different approach to dimensionality reduction. These methods transform the original high-dimensional data into a feature space with fewer dimensions. Depending on the chosen method, this can be done in a supervised or unsupervised manner. *Principle Component Analysis* (PCA) (Pearson, 1901) is currently considered a standard method for constructing an unsupervised, linear mapping of high-dimensional data onto a lower dimensional space. The resulting, low-dimensional representations, made of linearly uncorrelated variables known as *Principle Components*, are used extensively for data mining and visualisation, and can be used as an alternative (low-dimensional) feature representation for classification tasks.

The majority of the aforementioned dimensionality reduction methods are not capable of dealing with the type of feature space solely based on the MSA approach, and thus dimensionality reduction has only been tested with the N-gram based approach.

4 Experimental Setup

To test the representations of a 16S rRNA gene for detecting metabolic capabilities two sets of experiments were conducted. We trained a number of binary classifiers to identify whether a given 16S rRNA represents a microorganism which does or does not perform a given function. Five different functional groups have been chosen for these experiments: *Ureolysis*, *methanotrophy*, *fermentation*, *phototrophy* and *chemoheterotrophy*, each with different levels of prevalence and varying phylogenetic diversity (for definitions of these functional groups see Table 1). The proposed approaches to feature representation of 16S rRNA, based on N-grams and MSA, were also tested and compared.

Table 1: Definitions of functional groups.

Function	Definition	Prevalence
Ureolysis	The breakdown of urea into ammonia and carbon dioxide.	2.0 %
Methanotrophy	Describes prokaryotes that metabolize methane as their only source of carbon and energy. They can be either bacteria or archaea and require single-carbon compounds to survive.	0.4 %
Fermentation	A process that releases energy from a sugar or other organic molecule, does not require oxygen or an electron transport system, and uses an organic molecule as the final electron acceptor.	36.2 %
Phototrophy	Using light energy to synthesize sugars and other organic molecules from carbondioxide.	3.6 %
Chemoheterotrophy	Describes organisms which get their energy from the oxidation of inorganic minerals or consume other organisms to produce carbon.	71.3 %

Note: The prevalence rates refer to the frequency of a functional group in the processed dataset, i.e., with sequences from SILVA database and functional labels annotated by FAPROTAX (Louca et al, 2016).

The phylogenetic diversity of each function is illustrated in Figure 3. As it can be observed, the metabolic capabilities are not mutually exclusive, as some functions overlap and can be considered a more specific form of another. The chemoheterotrophy function can be performed by microorganisms present in almost every part of the phylogenetic tree. Furthermore, the microorganisms possessing this trait often can perform other functions, including fermentation, methanotrophy and ureolysis. This demonstrates how widely spread the trait is; in the SILVA dataset (release 119) over 1.5 million OTUs (see Table 2) were assigned the chemoheterotrophy function. Fermentation is another widely spread biochemical function. It is, however, more conserved then chemoheterotrophy, as most of the observations with this trait form clusters at the right-hand side of the tree, with some small clusters at the top of the tree. Ureolysis and Methanotrophy functions are mostly found within few well-defined clusters distributed across the left and top-right parts of the phylogenetic tree respectively, which demonstrates that these traits are phylogenetically diverse but at the same

time locally conserved. According to our sample, the phototrophy is by far the most conserved of the five functions, as most of the corresponding observations are found in three medium to large clusters at the left and bottom branches of the tree with some outliers located at the top. Moreover, the species with this trait do not appear to belong to other popular functional groups.

Mean accuracy rates of classifiers built on 50 independent samples are reported. For each trial a random sample of 400 observations was taken from the dataset constructed using the SILVA and FAPROTAX databases. A class balance was enforced for both training and testing of the classifiers by undersampling the majority class. A 10-fold cross-validation method was used to estimate expected accuracy rates for each sample and the mean accuracy rates across all 50 samples.

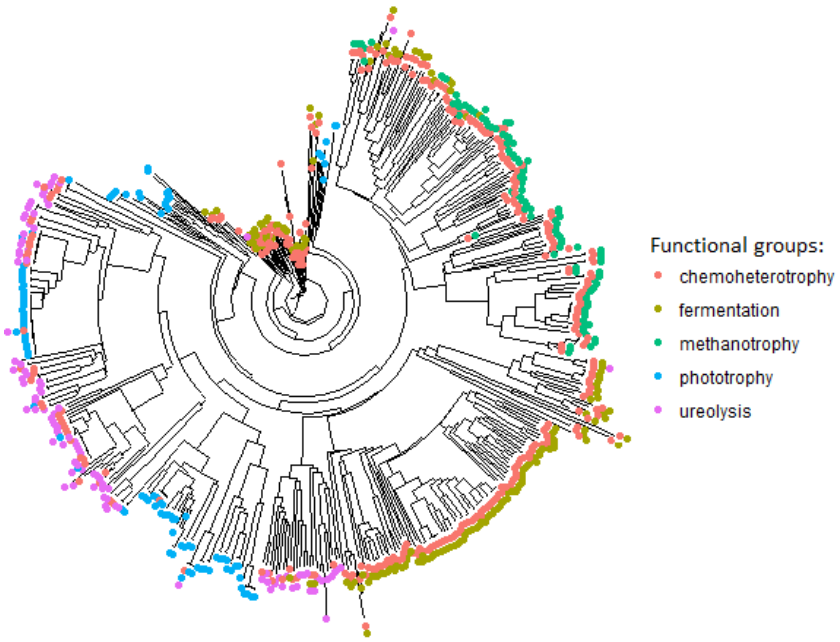

Figure 3: Phylogenetic Tree. Includes 500 microorganisms which belong to at least one of 5 functional groups: chemoheterotrophy, fermentation, methanotrophy, phototrophy, ureolysis. Each function is represented by at least 100 members. The distance matrix used to draw the tree is derived from an aligned table, thus sequence similarity is measured by how often two sequences have matching symbols in each column of the aligned table. The sequence alignment was implemented with *msa* package in R.

For both proposed approaches, Naïve Bayes and Random Forest classifiers have been applied using *e1071* (Meyer et al, 2017) and randomForest (Liaw and Wiener, 2015) R packages. For the N-gram approach, SVMs were trained trained with the e1071 package. The radial kernel has been chosen for all SVM classifiers and 100 trees have been built for each random forest classifier. In the series of experiments reported here we have used $n = \{1, 2, 3, 4, 5\}$ to construct N-gram based feature vectors.

For dimensionality reduction, we implemented LASSO and Principle Component Analysis (PCA) techniques using the *glmnet* package in R (Simon et al, 2011). With the implementation of LASSO, all features which where assigned coefficient values of zero have been removed from the feature vectors. Those techniques have reduced the number of features from 1364 to between 50 and 90 (depending on the sample and the functional group classified). The glmnet package automatically searches for a good value of the λ parameter (the default value of 0.01 for minimum λ is used to start the search). For the PCA method we have used the 50 most relevant components as features. No dimensionality reduction has been implemented with the MSA based feature representation, however, non-unique columns and those with too many gaps (> 75 %) were removed from the aligned tables.

5 Results

Multiple Sequence Alignment

The first set of experiments focused on addressing the question whether accurate classifiers can be trained with Naïve Bayes and Random Forest methods, using the categorical feature vectors from the aligned table attained by implementing MSA. The results are summarised in Table 2.

Overall, Random Forest classifiers had higher accuracy rates (between 83.4 % and 90.1 %) than the Naïve Bayes classifier (66.1 % – 75.7 %). The lowest accuracy rates for the Random Forest classifier were obtained for the largest and the most phylogenetically diverse functional group (chemoheterotrophy), while the less numerous and diverse functional labels, such as methanotrophy and photrophy were idenfied with higher accuracies.

Table 2: MSA-based classification results.

		Mean Accuracy	
Function Name	***No. of Observations***	***Naïve Bayes***	***Random Forest***
Ureolysis	47,809	0.707	0.869
Methanotrophy	9,473	0.757	0.892
Fermentation	853,671	0.662	0.876
Phototrophy	84,293	0.661	0.901
Chemoheterotrophy	1,683,041	0.695	0.834
Overall Mean Accuracy		**0.696**	**0.874**

Note: Mean accuracy rates of the classifiers built on 50 independent random samples of 400 microorganisms. A 50-50 class balance has been kept for both training and testing. The number of observations reported in the second column represents the number of microorganisms in SILVA dataset (release 119) which where assigned a given function by the FAPROTAX package.

N-grams

The second set of experiments examined the use of N-gram based feature spaces for biochemical function classification with 16S rRNA sequencing data. Overall, all three classification techniques had mean accuracy rates over 80 % (see Table 3). The SVM outperformed the other two classification methods, for all functions except ureolysis, where Random Forest provided the best results.

Table 3: Ngram-based classification results.

		Mean Accuracy	
Function Name	***Support Vector Machines***	***Naïve Bayes***	***Random Forest***
Ureolysis	0.860	0.839	0.878
Methanotrophy	0.954	0.901	0.911
Fermentation	0.946	0.792	0.890
Phototrophy	0.916	0.781	0.904
Chemoheterotrophy	0.858	0.815	0.852
Overall Mean Accuracy	**0.907**	**0.827**	**0.887**

Note: Mean accuracy rates of the classifiers built on 50 independent random samples of 400 microorganisms. A 50-50 class balance has been kept for both training and testing.

The effects of implementing LASSO and PCA techniques for dimensionality reduction on classification accuracy rates is summarised in Table 4. The application of LASSO resulted in a similiar performance across all three classifiers. Although this method provided the best accuracy rates for the Naïve Bayes classifier, SVM and Random Forest performed better with the PCA based feature space. The classifiers built on an original N-gram based feature space where no dimensionality reduction method was implemented (1364 features) outperformed on average both dimensionality reduction approaches.

Table 4: N-gram based Classification with dimensionality reduction methods.

	Mean Accuracy		
Dimensionality Reduction	***Support Vector Machines***	***Naïve Bayes***	***Random Forest***
LASSO	0.885	0.838	0.872
PCA	0.892	0.736	0.887
None	0.907	0.827	0.887

Note: Aggregate mean accuracy rates of the classifiers built for all five functional groups based on 50 independent random samples. A 50-50 class balance have been kept for both training and testing.

N-grams vs. MSA

Comparing the results from Tables 3 and 4 one can observe that the classifiers trained on the N-gram based feature representation performed on average better than those following the MSA approach. The performance of the Naïve Bayes classifier has been affected by the change of feature representation as the mean accuracy rate increased by 0.131 when using N-grams instead of aligned tables. No substantial increase (0.013) in the accuracy of the Random Forest classifier indicates that it is well-suited to deal with high-dimensional, categorical feature spaces produced by MSA.

Table 5 displays the results of the best performing N-gram based classifiers (SVM) without dimensionality reduction across the five functions in comparison to the best performing MSA based classifier (Random Forest). As it can be observed, for the N-gram approach, each function was detected with accuracy rates between 86 % - 95 % by the SVM classifier. Furthermore, this approach outperformed the Random Forest classifier with MSA, 4 of out 5 times.

Table 5: Comparison between MSA and Ngram approaches.

		Mean Accuracy	
Function Name	***No. of Observations***	***N-gram SVM***	***MSA RF***
Ureolysis	47,809	0.860	0.869
Methanotrophy	9,473	0.954	0.892
Fermentation	853,671	0.946	0.876
Phototrophy	84,293	0.916	0.901
Chemoheterotrophy	1,683,041	0.858	0.834
Overall Mean Accuracy		**0.907**	**0.874**

Note: The best performing classification techniques have been used for a final comparison. The mean accuracy rates of the classifiers built on 50 independent random samples of 400 microorganisms. A 50-50 class balance has been kept for both training and testing. The number of observations reported in the second column represents the number of microorganisms in SILVA dataset (release 119) which where assigned given function by the FAPROTAX package.

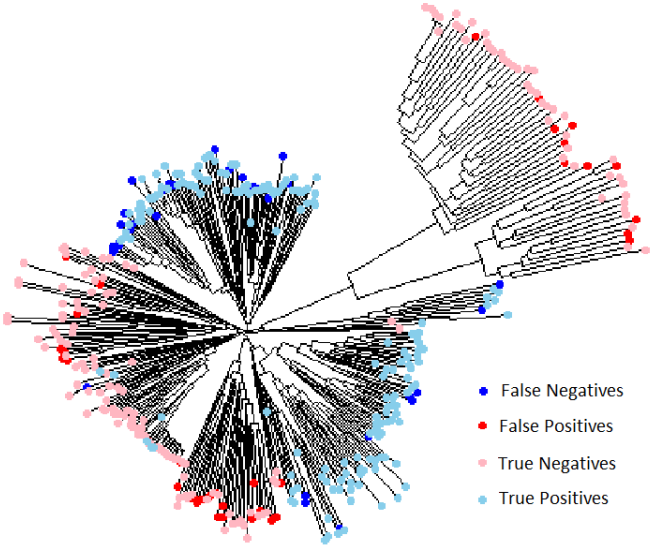

Figure 4: The results of MSA-based Random Forest fermentation mapped to the phylogenetic tree. The distance matrix used to draw the tree is derived from an aligned table, thus sequence similarity is measured by how often two sequences have matching symbols in each column of the aligned table. All 400 observations used to construct the tree where not included in the training set.

Figures 4 and 5 illustrate the distribution of misclassification errors in a phylogenetic tree. One can observe that both feature representation approaches resulted in classifiers capable of non-linear mapping between phylogeny and function. The parts of the trees which include examples from both the positive ($y = 1$) and negative ($y = 0$) classes (this includes the bottom and the bottom-left part of both trees) are a significant source of errors. Furthermore, both classifiers committed errors even when other examples from the same, well-defined clusters were correctly classified (see top right part of Figure 4). The N-gram based approach is notably making less errors of this type. The frequent errors inside those well-defined clusters suggest that not enough information has been provided to the classification algorithm to detect functional clusters. It is, thus, likely, that the N-gram based approach requires less data points to create informative features as it disregards positional information, and produces vectors with substantial number of informative features when compared to an MSA based feature representation.

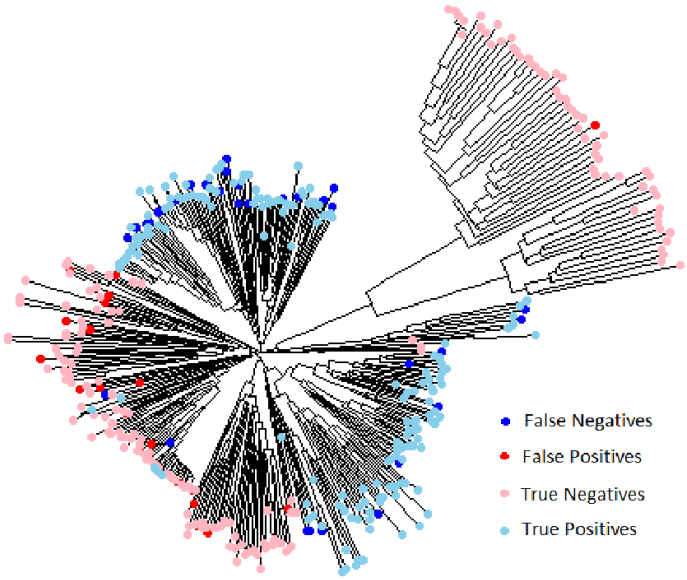

Figure 5: The results of N-gram-based SVM fermentation classifier mapped to the phylogenetic tree. The distance matrix used to draw the tree is derived from an aligned table, thus sequence similarity is measured by how often two sequences have matching symbols in each column of the aligned table. All 400 observations used to construct the tree where not included in the training set.

6 Discussion & Conclusion

In this paper we have proposed the application of machine learning methods for functional annotation based on the 16S rRNA gene. After constructing a valid training dataset, we have tested two feature engineering approaches, one based on the Multiple Sequence Alignment and another on N-grams. Our results demonstrated that 16S rRNA contains relevant information for identifying functional capabilities as both approaches could be used to train classifiers with satisfactory accuracy rates.

Although one could only speculate about the predictive functional potential of the 16S rRNA marker gene, as shown by our results the gene can be used with relative success for mapping phylogeny to function. In other words, non-linear methods of classification such as Random Forest and radial Support Vector Machines, can be trained to assign functions in a manner which does not appear to be biased towards local (i.e., closely related) phylogenetic clusters. In addition, our results suggest that the N-gram based feature space is more suitable than MSA for this classification task. The frequencies of the shorter sequence segments have proven to be an effective form of feature representation for 16S rRNA, despite not taking into the account the information about the positions in the original sequence. Considering that MSA is a clustering-like procedure, applying it for the purpose of feature representation means that the structure of the resulting feature vectors has been determined by the process which aims to maximise sequence similarity. Thus, when this method is applied, the feature representation of a new unseen sequence will be dependent on the stored set of sequences used for alignment and, therefore, it is biased. This might explain the weaker performance of the classifiers built with this approach. Another reason that might explain the weaker results obtained from MSA is the higher dimensionality of the feature space and the lower numbers of highly informative features.

The implemented dimensionality reduction methods, applied with the N-gram approach, did not have a notable effect on the performance of the classifiers and, in most cases, resulted in slightly weaker models. As shown by our results the LASSO method did improve the accuracy of the simplest classification algorithm (Naïve Bayes), whereas Random Forest and Support Vector Machines did not benefit from the reduction in the number of features. The potential reason for the random forest classifier remaining stable when used with high-dimensional datasets is that it implements a form of implicit feature selection

when constructing decision trees. The SVM classifier, on the other hand, is known to behave similarly to regularized algorithms, as it implements a more robust loss function (hinge loss) when finding decision boundaries. Moreover, the eigenvalues of the kernel matrix typically decay quickly as the dimensions become less informative and, thus, in many high-dimensional situations SVMs effectively classify according to few most relevant dimensions. Nevertheless, other dimensionality reduction techniques, such as improved variants of LASSO (Bach (2008) & Zare et al (2013)) and non-linear PCA (Scholz et al, 2005), or feature selection methods developed for genomic data (Mahmoud et al, 2014) may improve the performance of classifiers.

Better accuracy rates may also be achieved by implementing alternative classification techniques, including *Deep Learning* or ensemble methods, such as *Optimal Tree Ensembles* (Khan et al, 2019), *Random Projection* (Cannings and Samworth, 2017) or ensembles of subset of kNN classifiers (Gul et al, 2018). Moreover, to obtain more robust results, future studies would benefit from conducting similar experiments with a larger number of observations. This would not only improve the expected accuracy but will also lead to a better understanding regarding the predictive power of these methods.

Conclusion

Our results demonstrate that 16S rRNA contains relevant information for direct identification of functional capabilites. In addition, our findings provide evidence that the N-gram approaches produce preferable feature spaces when compared to MSA. Taken altogether, the results of this study can serve as a point of departure for the development of future machine learning pipelines for 16S rRNA based functional annotation.

Acknowledgements We acknowledge support from grant number ES/L011859/1, from The Business and Local Government Data Research Centre, funded by the Economic and Social Research Council to provide researchers and analysts with secure data services.

References

Attenberg J, Weinberger K, Dasgupta A, Smola A, Zinkevich M (2009) Collaborative Email-Spam Filtering with the Hashing Trick. 6th Conference on E-Mail and Anti-Spam (CEAS2009), California (USA). URL: https://pdfs.semanticscholar.org/b18e/b1900ba773dceed4cd0719234c4d0ecc3065.pdf.

Bach FR (2008) Bolasso: Model Consistent Lasso Estimation Through the Bootstrap. In: Proceedings of the 25th International Conference on Machine Learning (ICML), Association of Computing Machinery (ACM), New York (USA), pp. 33–40. DOI: 10.1145/1390156.1390161.

Bodenhofer U, Bonastesta E, Horejs-Kainrath C, Hochreiter S (2015) MSA: An R Package for Multiple Sequence Alignment. Bioinformatics 31(24):3997–3999. DOI: 10.1093/bioinformatics/btv494.

Burdukiewicz M, Sobczyk P, Lauber C (2017) biogram: N-Gram Analysis of Biological Sequences. URL: https://CRAN.R-project.org/package=biogram. R package version 1.4.

Cannings TI, Samworth RJ (2017) Random-Projection Ensemble Classification. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 79(4):959–1035. DOI: 10.1111/rssb.12228.

Carrillo H, Lipman DJ (1988) The Multiple Sequence Alignment Problem in Biology. Journal of Applied Mathematics 48(5):1073–1082. DOI: 10.1137/0148063.

Delmont TO, Simonet P, Vogel T (2012) Describing Microbial Communities and Performing Global Comparisons in the Omic Era. The ISME Journal 6:1625–1628. DOI: 10.1038/ismej.2012.55.

Dong G, Pei J (2007) Sequence Data Mining, Advances in Database Systems Book Series (ADBS), Vol. 33. Springer US, Boston (USA). DOI: 10.1007/978-0-387-69937-0.

Edgar RC (2004) MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput. Nucleic Acids Research 32(5):1792–1797. DOI: 10.1093/nar/gkh340.

Ekblom R, Wolf JBW (2014) A Field Guide to Whole-Genome Sequencing, Assembly and Annotation. Evolutionary Applications 7(9):1026–1042. DOI: 10.1111/eva.12178.

Goldberg DE (1989) Genetic Algorithms in Search, Optimization and Machine Learning, 1st edn. Addison-Wesley Longman Publishing Co., Inc., Boston (USA). ISBN: 02-0115-767-5.

Goodwin S, McPherson JD, McCombie WR (2016) Coming of Age: Ten Years of Next-Generation Sequencing Technologies. Nature Reviews Genetics 17(6):333–351. DOI: 10.1038/nrg.2016.49.

Gul A, Perperoglou A, Khan Z, Mahmoud O, Miftahuddin M, Adler W, Lausen B (2018) Ensemble of a Subset of kNN Classifiers. Advances in Data Analysis and Classification 12(4):827–840, Springer, Berlin, Heidelberg (Germany). DOI: 10.1007/s11634-015-0227-5.

Huang HH (2016) An Ensemble Distance Measure of k-mer and Natural Vector for the Phylogenetic Analysis of Multiple-Segmented Viruses. Journal of Theoretical Biology 398:136–144. DOI: 10.1016/j.jtbi.2016.03.004.

Kaisers W, Schwender H, Schaal H (2018) Hierarchical Clustering of DNA k-mer Counts in RNAseq Fastq Files Identifies Sample Heterogeneities. International Journal of Molecular Sciences (IJMS) 19(11):3687–3701. DOI: 10.3390/ijms19113687.

Kennedy J, Eberhart R (1995) Particle Swarm Optimization. Proceedings of the International Conference on Neural Networks (ICNN'95) 4:1942–1948, Institute of Electrical and Electronics Engineers (IEEE), Perth (Australia). DOI: 10.1109/ICNN.1995.488968.

Khan Z, Gul A, Perperoglou A, Miftahuddin M, Mahmoud O, Adler W, Lausen B (2019) Ensemble of Optimal Trees, Random Forest and Random Projection Ensemble Classification. Advances in Data Analysis and Classification, pp. 1–20, Springer, Berlin, Heidelberg (Germany). DOI: 10.1007/s11634-019-00364-9.

Kohavi R, John GH (1997) Wrappers for Feature Subset Selection. Artificial Intelligence 97(1-2):273–324. DOI: 10.1016/S0004-3702(97)00043-X.

Langille MGI, Zaneveld J, Caporaso JG, McDonald D, Knights D, et al (2013) Predictive Functional Profiling of Microbial Communities Using 16S rRNA Marker Gene Sequences. Nature Biotechnology 31:814–821.

Liaw A, Wiener M (2015) R: Breiman and Cutler's Random Forests for Classification and Regression. URL: `https://cran.r-project.org/web/packages/randomForest/randomForest.pdf`. R package version 4.12.

Louca S, Parfrey LP, Doebeli M (2016) Decoupling Function and Taxonomy in the Global Ocean Microbiome. Science 353(6305):1272–1277. DOI: 10.1126/science.aaf4507.

Lu J, Breitwieser FP, Thielen P, Salzberg SL (2017) Bracken: Estimating Species Abundance in Metagenomics Data. PeerJ Computer Science 3:e104. DOI: 10.7717/peerj-cs.104.

van der Maaten L, Postma E, van den Herik J (2009) Dimensionality Reduction: A Comparative Review. Tech. Rep., Tilburg University, Tilburg Centre for Creative Computing, Tilburg (The Netherlands). URL: `https://lvdmaaten.github.io/publications/papers/TR_Dimensionality_Reduction_Review_2009.pdf`.

Mahmoud O, Harrison A, Perperoglou A, Gul A, Khan Z, Metodiev M, Lausen B (2014) A Feature Selection Method for Classification Within Functional Genomics Experiments Based on the Proportional Overlapping Score. BMC Bioinformatics 15:274–294, Springer Nature. DOI: 10.1186/1471-2105-15-274.

Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F, Chang C, Lin C (2017) R: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), Vienna University of Technology. URL: `https://cran.r-project.org/web/packages/e1071/e1071.pdf`. R package version 1.8.

Moody J (1989) Fast Learning in Multi-Resolution Hierarchies. Vol. 1. Morgan-Kaufmann, pp. 29–39. URL: `https://papers.nips.cc/paper/175-fast-learning-in-multi-resolution-hierarchies`.

Notredame C (2002) Recent Progresses in Multiple Sequence Alignment: A Survey. Pharmacogenomics 3(1):131–144. DOI: 10.1517/14622416.3.1.131.

Pearson K (1901) On Lines and Planes of Closest Fit to Systems of Points in Space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 2(11):559–572. DOI: 10.1080/14786440109462720.

Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y (2012) A Tale of Three Next Generation Sequencing Platforms: Comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics 13(1):341–354. DOI: 10.1186/1471-2164-13-341.

Quast C, Pruesse E, Yilmaz P, Gerken J, Shweer T, Yarza P, Peplies J, Glöckner FO (2013) The SILVA Ribosomal RNA Gene Database Project: Improved Data Processing and Web-Based Tools. Nucleic Acids Research 41(D1):590–596. DOI: 10.1093/nar/gks1219.

Rahman A, Hallgrimsdottir I, Eisen M, Pachter L (2018) Association Mapping from Sequencing Reads Using *k*-mers. eLife e32920, Flint J (ed). DOI: 10.7554/eLife.32920.

Scholz M, Kaplan F, Guy CL, Kopka J, Selbig J (2005) Non-Linear PCA: A Missing Data Approach. Bioinformatics 21(20):3887–3895. DOI: 10.1093/bioinformatics/bti634.

Simon N, Friedman J, Hastie T, Tibshirani R (2011) Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. Journal of Statistical Software 39(5):1–13. DOI: 10.18637/jss.v039.i05.

Sokal R, Sneath P (1963) Principles of Numerical Taxonomy. W. H. Freeman and Co., San Francisco (USA), London (UK).

Sun S, Jones RB, Fodor AA (2019) Inference Based PICRUSt Accuracy Varies Across Sample Types and Functional Categories. bioRxiv 655746, Cold Spring Harbor Laboratory (CSH). DOI: 10.1101/655746.

Thompson JD, Linard B, Lecompte O, Poch O (2011) A Comprehensive Benchmark Study of Multiple Sequence Alignment Methods: Current Challenges and Future Perspectives. PLOS ONE 6(3):e18093. DOI: 10.1371/journal.pone.0018093.

Tibshirani R (1996) Regression Shrinkage and Selection via the lasso. Journal of Royal Statistical Society, Series B (Methodological) 58(1):267–288. DOI: 10.1111/j.2517-6161.1996.tb02080.x

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF (2012) Community Structure and Metabolism Through Reconstruction of Microbial Genomes from the Environment. The ISME Journal 6:1625–1628. DOI: 10.1038/ismej.2012.55.

Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. Applied and Environmental Microbiology 73(16):5261–5267. DOI: 10.1128/AEM.00062-07.

Weinberger K, Dasgupta A, Langford J, Smola A, Attenberg J (2009) Feature Hashing for Large Scale Multitask Learning. Association of Computing Machinery (ACM), New York (USA). DOI: 10.1145/1553374.1553516.

Widder S, Allen RJ, Pfeiffer T, Curtis TP, Wiuf C, et al (2016) Challenges in Microbial Ecology: Building Predictive Understanding of Community Function and Dynamics. The ISME Journal 10:2557–2568.

Xu X, Liang T, Zhu J, Zheng D, Sun T (2017) Review of Classical Dimensionality Reduction and Sample Selection Methods for Large-Scale Data Processing. Neurocomputing 328:5–15. DOI: 10.1016/j.neucom.2018.02.100.

Zare H, Haffari G, Gupta A, Brinkman R (2013) Scoring Relevancy of Features Based on Combinatorial Analysis of Lasso with Application to Lymphoma Diagnosis. BMC Genomics 14:S14–Suppl1). DOI: 10.1186/1471-2164-14-S1-S14.