VOL.8 NO.1 March 2019 - ISSN 2233 – 1859

Operations Research Society in Bosnia and Herzegovina

Southeast Europe Journal of Soft Computing
Available online: http://scjournal.ius.edu.ba

IUS Soft Computing Research Group

Taxonomic Classification of Bacteria Using Common Substrings

M. Can
O. Gürsoy

Faculty of Engineering and Natural Sciences,
International University of Sarajevo International University of Sarajevo,
Hrasnicka Cesta 15, Ilidža 71210 Sarajevo,
Bosnia and Herzegovina
mcan@ius.edu.ba
ogursoy@ius.edu.ba

Article Info

Article history:
Article received on 10 January 2019
Received in revised form 1 February 2019

Keywords:
16S rRNA gene, LongestCommonSubsequence, Taxonomic classification

ABSTRACT: For the taxonomic classification of microbes, 16S ribosomal RNA (rRNA) gene sequences are widely used in environmental microbiology as reliable markers. Although the massive sequencing of 16S rRNA gene amplicons encompassing the full length of genes is not easy, because of the limitations of the current sequencing techniques, in databases Greengenes, RDP, and SILVA millions of rRNA gene sequences are uploaded. In this research, first a new similarity measure LCSS, for full length genes is defined. Then it is found that sequences reported for the same bacteria species demonstrate around 53% average sequence similarity in Greengenes and SILVA databases, while average similarity among genes reported for different bacteria species is around 15% only. This is 63%, and 20% respectively at genus level for the three data bases Greengenes, RDP, and SILVA. Hence, species, and genus-specific sequences constitute useful targets for diagnostic assays and other scientific investigations. In the present research, the built in function LongestCommonSubsequence is used repeatedly in computer algebra package MATHEMATICA to create an in silico pipeline for taxonomic classification uploaded new full-length sequences. Conclusions: Our results suggest that LongestCommonSubsequence similarity can be used for taxonomic classification of unknown bacteria through their full 16S ribosomal RNA (rRNA) gene sequences.

1. INTRODUCTION

Bacteria contribute immensely to global energy conversion and the recycling of matter in almost all environments explored. The flora of the human gut has been extensively explored for potential associations with the appearance of many human diseases (Duvallet et. al. 2017; Forbes, et. al. 2016; Turnbaugh et. al. 2006) [1-3] The collection of microbes and their genes that exist within and on the skin of the human body, are known as the microbiome. Humans and microbes have established a symbiotic association over time, and changes in this collaboration is linked to several diseases . The rich microbial diversity of environments such

\

important ecological inferences (Yilmaz, et al. 2016: Fierer, 2017) Determining which microbial taxa are at the surface and at the ocean beds, how they survive are the driving questions in marine microbial ecology. [4-6].

As a result of these research activities, there are now a substantial number of microbial community datasets deposited in sequence archives, as an example, the European Nucleotide Archive currently holds over 600,000 environmental samples (Mitchell et al., 2017)[7], and the rate of deposition is climbing. To draw biological

information from this large amount of data requires accurate and reliable in slico tools and methods.

Taxonomic Assignments

Inference of community composition through taxonomic classification has been one of the crucial steps in almost all microbiome-based analyses. For more than three decades now, the common approach for taxonomic assignment of microbial species has been the classification of ribosomal RNA (rRNA) sequences. Pace, et. al, (1986)[8] developed two approaches. In one approach, suitable for mixed populations of limited complexity, less than about ten different organisms, they isolated 5S rRNA, sorted out the various species-specific molecules by high-resolution gel electrophoresis. Individual 5S rRNA types then are sequenced and, with reference to existing files of 5S rRNA sequences, the phylogenetic affinities of organisms contributing the analyzed 5S rRNAs are defined.

In the second approach for mixed populations of unlimited complexity, 16S rRNA genes are "shotgun-cloned" using DNA purified from natural samples. It does not matter that the original DNA was from a mixed population of organisms; the rRNA gene clones are selected on an individual basis, as isolated recombinant bacteriophage. The different types of cloned rRNA genes then are sorted out in the laboratory and submitted to limited sequence analysis using a technique which provides immediate access to regions of the 16S rRNA gene particularly useful for phylogenetic evaluations. Again, by comparison of sequences with existing reference collections of complete and partial rRNA sequences, the phylogenetic affinities of the organisms in the original population are established.

Currently, the most widely used tools for this purpose are the mothur (Schloss, JG, et. al., 2009) [9] and Quantitative Insights Into Microbial Ecology (QIIME) software packages (Caporaso,D., et., al., 2010) [10]. These correspond to large toolsets that are able to process, classify, and perform downstream analyses on individual genetic markers like the 16S rRNA gene, conserved across the prokaryotic domains.

For taxonomic classification, each tool compares a set of queried sequences against a defined reference database, such as Greengenes (McDonald, S., et. al,, 2016) [11], NCBI (Federhen , 2012) [12], RDP (Cole, JR., et. al., 2013) [13], or SILVA (Yilmaz, P., et. al., 2014) [14], assigning the most likely taxonomic lineages. Ultimately, the success of these analyses is not only dependent on the breadth and diversity of annotated sequences available in public repositories, but also on the accuracy of the classification algorithms used by each of the tools. By default, QIIME makes use of the UCLUST clustering method (Edgar, RC., 2010) [15] to assign biological sequences to a reference database, while mothur reimplements the naive Bayesian RDP classifier, developed by Wang et al. (Wang Q., et al. , 2007)[16].

Two other tools, MAPseq (Matias Rodrigues, JR., et. al., 2017)[17] and QIIME 2 (Nicholas, A., 2018)[18], have recently been released, QIIME 2 has officially replaced QIIME as of January 2018. QIIME 2 also makes use of a naive Bayes classifier (Bokulich NA, 2018) [19], and MAPseq is a k-mer search approach that outputs confidence estimates at different taxonomic ranks.

Microbiome studies frequently strive to associate microbial diversity signatures with a phenotype of interest. However, focusing solely on high-level taxonomic ranks, such as classes and orders, can severely underestimate the degree of variation observed between sample groups. To circumvent this, highly discriminative approaches are needed to be able to pinpoint the most significant taxa warranting further validation.

For assessing the performance of MAPseq, mothur, QIIME, and QIIME 2 with different reference databases, (Almeida et. al. 2018) limited their analyses to classification at the lineage level instead of operational taxonomic units, as it allows a more consistent and easier interpretation of the results. Species assignment of every queried sequence would be the desired outcome, but as was previously shown (Golob JL, 2017)[21], the limited resolution of the 16S rRNA locus precludes an accurate classification at this level. Furthermore, there is significant inconsistency in species nomenclature across all reference databases, for example, RDP does not report taxon names below genus. In their work, they calculated the degree of recall and precision at the genus and family ranks, as in our opinion they provide the best compromise between classification accuracy and resolution.

Table 1. The level of recall across all software tools, data-bases, and sub-regions for taxons family and genus (Almeida, A., et. al 2018).

			Family	Genus
Software	Database	Sub-r	Recall	Recall
MAPseq	Greengenes	V3-V4	88.3	58.9
MAPseq	NCBI	V3-V4	81.7	51.7
MAPseq	SILVA	V3-V4	67.2	46.5
mothur	RDP	V3-V4	85.4	50.5
mothur	SILVA	V3-V4	82.9	40.8
QIIME2	Greengenes	V3-V4	93.2	69.2
QIIME2	SILVA	V3-V4	93.6	69.0
QIIME	Greengenes	V4	59.4	45.1
QIIME	SILVA	V4	66.4	57.5

2. MATERIALS AND METHODS

In this research work, a new similarity between gene sequences is introduced. According to this similarity measure, it is seen that, average inclass similarities are

statistically significantly higher than the interclass averages. It is concluded that this similarity measure can be used to annotate unknown bacteria at all taxon levels.

2.1. Longest Common Subsequence Search

The average inclass similarities and interclass averages are compared through the analysis of data contained in the high quality ribosomal RNA databases Greengenes, SILVA, and RDP. The number of non-redundant bacterial 16S ribosomal RNA (rRNA) gene sequences with around 1,200 base pairs is 198.510 for Greengenes. This number is 1.488.662 for SILVA, and 1.350.270 for RDP.

To find the level of similarity of two gene sequences, assume in Figure 1., (a) is a gene reported for a bacteria, and (b) is a gene reported for another, or the same bacteria.

(a) GGCTAACTAGTGTAGAGGTGAAATGATTTAGAT TAGGTGGCAA….

(b) .....GTGTAGAGGTGAAATGCGTAGAT

Figure 1. The longest common subsequence of two genes

The longest common subsequence of (a) and (b) is

GTGTAGAGGTGAAATG

Then we remove this common subsequence from both sequences. Then look for next longest common substring. If there is no longer one this time the string

TAGAT

may be the second longest common subsequence. It is seen that ten iterations of this process is optimal.

Then we add the lengths of these common substrings and normalize by dividing this sum, to the length of the shorter gene.

2.2. In class and Interclass Similarities

The average inclass similarities and interclass averages are computed for family, genus and species taxon levels in the three databases Greengenes, SILVA, and RDP. The results are show n Tables 2-3.

Table 1. similarities in class/ Inter Class for family level

Databases	In Class	Inter Class
Greengenes	51.34	19.34
SILVA	46.37	16.87
RDP	58.80	27.64
Mean	52.17	21.28

Table 2. similarities in class/ Inter Class for genus level

Databases	In Class	Inter Class
Greengenes	71.69	17.46
SILVA	75.84	22.43
RDP	42.94	21.44
Mean	63.49	20.44

Table 3. similarities in class/inter class for species level

Databases	In Class	Inter Class
Greengenes	71.69	17.46
SILVA	33.82	13.23
Mean	52.76	15.35

It is seen that there is a significat difference between in class/inter class similarities for all levels. This fact gave us an ide that the longest common subsequence similarity can be used for annotation of unknown bacteria.

3. RESULTS AND DISCUSSION

When randomly sampled genes are annotated according to their in class and interclass genes similarities for the three databases at taxon levels family, genus and species (if relevant) the annotation accuracies in tables 4-6 are found.

Table 4. LCSS accuracies for Greengenes

Levels	# levels	Accuracy
Phylum	86	91.86
Class	232	95.96
Order	366	92.35
Family	466	93.60
Genus	1949	84.55
Species	2389	86.73

Table 5. LCSS accuracies for RDP

Levels	# levels	Accuracy
Phylum	51	90.00
Class	126	84.13
Subclass	226	80.09
Order	391	61.08
Suborder	2041	80.90
Family	110	63.30
Genus	354	68.84

Table 6. LCSS accuracies for SILVA

Levels	# levels	Accuracy
Phylum	80	92.5
Class	424	78.57
Order	843	73.47
Family	2117	74.01
Genus	5317*	90.20
Species	183284	

It is seen that the Longest Common Subsequence in class, interclass similarities can successfully used in the three taxon levels.

REFERENCES

Alexandre Almeida 1,2,*, Alex L. Mitchell 1, Aleksandra Tarkowska 1 and Robert D. Finn (2018) Benchmarking taxonomic assignments based on 16S rRNA gene profiling of the microbiota from commonly sampled environments, GigaScience, 7, , 1–10

Bokulich, NA, Kaehler, BD, Rideout, JR, Dillon,M., Bolyen,E., Knight, R., Huttley, GA., and Caporaso JG. (2018) Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin, Microbiome. 2018; 6: 90.

Caporaso JG, Kuczynski J, Stombaugh J et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods 2010;7:335–6.

Cole JR, Wang Q, Fish JA et al. Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Res 2014;42:D633–42.

Duvallet C, Gibbons SM, Gurry T et al. Meta-analysis of gutmicrobiome studies identifies disease-specific and sharedresponses. Nat Commun 2017;8:1784.

Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 2010;26:2460–1.

Federhen S. The NCBI Taxonomy database. Nucleic Acids Res 2012;40:D136–43.

Fierer N. Embracing the unknown: disentangling thecomplexities of the soil microbiome. Nat Rev Microbiol 2017;15:579–90.

Forbes JD, Van Domselaar G, Bernstein CN. The gut microbiotain immune-mediated inflammatory diseases. Front Microbiol 2016;7:1081.

Golob JL, Margolis E, Hoffman NG et al. Evaluating the accuracy of amplicon-based microbiome computational pipelines on simulated human gut microbial communities. BMC Bioinformatics 2017;18:283.

McDonald D, Price MN, Goodrich J et al. An improved Greengenes taxonomywith explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J 2012;6:610–8.

Matias Rodrigues JF, Schmidt TSB, Tackmann J et al.MAPseq: highly efficient k-mer search with confidence estimates, for rRNA sequence analysis. Bioinformatics 2017;33:3808–10.

Mitchell AL, Scheremetjew M, Denise H et al. EBI Metagenomics in 2017: enriching the analysis of microbial communities, fromsequence reads to assemblies. Nucleic Acids Res 2017;46:D726–35.

Pace NR, Stahl DA, Lane DJ et al. The analysis of natural microbial populations by ribosomal RNA sequences. Adv Microb Ecol 1986, 9:1–55.

Schloss PD,Westcott SL, Ryabin T et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol 2009;75:7537–41.

Sczyrba A, Hofmann P, Belmann P et al. Critical assessment of metagenome interpretation—a benchmark of metagen omics software. Nat Methods 2017;14:1063–71.

Turnbaugh PJ, Ley RE, Mahowald MA et al. An obesityassociatedgut microbiome with increased capacity for energyharvest. Nature 2006;444:1027–31.

Wang Q, Garrity GM, Tiedje JM et al. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterialtaxonomy. Appl Environ Microbiol 2007;73:5261–7.

Yilmaz P, Yarza P, Rapp JZ et al. Expanding the worldof marine bacterial and archaeal clades. Front Microbiol 2016;6:1524.

Yilmaz P, Parfrey LW, Yarza P et al. The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. Nucleic Acids Res 2014;42:D643–8.