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

UOIuBIH
ORSinBIH
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

Hypervariable Regions in 16S rRNA Genes for the Taxonomic Classification

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

Article Info

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

Keywords:
16S rRNA gene, Variable regions, Taxonomic, LongestCommonSubsequence

ABSTRACT: 16S ribosomal RNA (rRNA) gene sequences are reliable markers for the taxonomic classification of microbes and widely used in environmental microbiology. Production of 16S rRNA gene amplicons in large amounts, encompassing the full length of genes is not yet feasible, because of the limitations of the current sequencing techniques. They are mostly in short reads of length less than 300 base pairs. Hence, the selection of the most efficient hypervariable regions for phylogenetic analysis and taxonomic classification is a current research area. It is found that nine hypervariable regions (V1–V9), resides in bacterial 16S ribosomal RNA (rRNA) genes. Family, genus, and species-specific sequences within a given hypervariable region constitute useful targets for diagnostic assays and other scientific investigations. In this study systematic studies that compare the relative advantage of hypervariable regions grouped as V1–V2–V3, V4–V5–V6, and V7–V8–V9 for specific diagnostic goals are done. In the present research, the built in function Longest–Common–Subsequence in computer algebra package MATHEMATICA is used to create an in silico pipeline to evaluate the taxonomic classification sensitivity of the hypervariable regions compared with the corresponding full-length sequences. Conclusions: Our results suggest that V4–V5–V6 region might be an optimal sub-region for the design of universal primers with superior phylogenetic resolution for bacterial phyla.

1. INTRODUCTION

One of the most important tasks for microbiologists is to profile microbial community to explore bacteria in the environmental niches, and in various ecosystems, since bacteria contribute immensely to global energy conversion and the recycling of matter. To profile microbial community in the samples from human gut is also important since bacteria are an important component of human health.

Unfortunately most bacteria cannot be cultured or isolated under laboratory conditions, and hence our understanding of the kingdom Bacteria remains limited (Rinke et al., 2013).

Until the development of high-throughput sequencing technology, as mainstream methods in studies of bacterial communities and diversity, terminal restriction fragment length polymorphism (Liu et al., 1997), denaturing gradient gel electrophoresis analysis (Muyzer et al., 1993), fluorescent in situ hybridization (Wagner et al., 1998) and Genechips (He et al., 2012) were used.

Recently, meta-genomic methods just like the use of 16S sequencing technology (NGST), caused a remarkable expansion of our knowledge regarding uncultured bacteria.

The 16S rRNA gene sequence contains both highly conserved regions for primer design and hypervariable regions to identify phylogenetic characteristics of microorganisms. It is first used in 1985 for phylogenetic analysis (Lane et al., 1985)[9], and marker gene for profiling bacterial communities (Tringe et al., 2008) [10].

Because of the limitation of the sequencing technology, the 16S rRNA gene sequences used in most studies are partial sequences, while full-length 16S rRNA gene sequences consist of nine hypervariable regions that are separated by nine highly conserved regions (Baker, et al., 2003; Wang, et al., 2009)[11, 12].

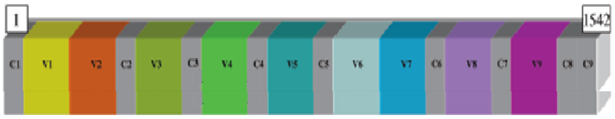

Figure 1. 16S rRNA gene sequences consist of nine hypervariable regions that are separated by nine highly conserved regions.

Positions of the hypervariable sub-regions of the 16S rRNA sequences with their start and end positions as in E. Coli is given in Table 1, as in the below (Yang et al. 2016).

Table 1 Positions of the hypervariable sub-regions of the 16S rRNA sequences, their start and end positions (E. Coli)

Hyper	s. position	e. position	Length
V1	8	96	88
V2	97	306	209
V3	307	487	180
V4	488	746	258
V5	747	885	138
V6	886	1029	143
V7	1030	1180	150
V8	1181	1372	191
V9	1373	1468	95

2. MATERIALS AND METHODS

Data contained in the high quality ribosomal RNA databases Greengenes, SILVA, and RDP are downloaded. 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 create a pipeline to annotate unknown bacteria through their 16S rRNA gene short reads, a new similarity between gene sequences is introduced. According to this similarity measure, it is seen that, average in-class similarities are statistically significantly higher than the inter-class 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

To find the level of similarity of two substrings of 16S rRNA gene sequences, assume in Figure 1., (a) is a part of a gene reported for a bacteria, and (b) is a part of a gene reported for another, or the same bacteria.

(a) GGCTAACTAGTGTAGAGGTGAAATGATTTAGAT TAGGTGGCAA….

(b) ......................GTGTAGAGGTGAAATGCGTAGAT

Figure 2. The longest common subsequence of two genes

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

GTGTAGAGGTGAAATG

We note the number of base pairs in this string, 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. We note also the number of base pairs in this string. This process is repeated till the common substrings are shorter than a threshold. 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 string.

2.2. In-class / Inter-class Similarities

For the taxonomic classes in family, genus, and species levels, the mean similarity of genes in a class and genes in different classes are computed for 198.510 genes for Greengenes, 1.488.662 genes for SILVA, and 1.350.270 for RDP. The numbers of taxonomic classes in family, genus, and species levels are shown in Table 2. in the below.

Table 2. Numbers of taxonomic classes in family, genus, and species levels.

Levels	Greengenes	SILVA	RDP
Phylum	86	80	51
Class	232	424	126
Order	366	843	391
Family	466	2117	110
Genus	1949	5317	354
Species	2389	183284	

Average in-class and inter-class similarities are computed in family, genus, and species levels. The results are shown in Table 3. in the below.

Table 3. Average in-class and inter-class similarities in family, genus, and species levels.

	Family		Genus		Species	
Databases	In	Between	In	Between	In	Between
Green	51.34	19.34	54.48	14.94	71.69	17.46
SILVA	46.37	16.87	75.26	20.26	33.82	13.23
RDP	58.80	27.64	42.94	21.44	-	-
Mean	50.17	21.28	57.56	18.88	52.76	15.35

It is seen that the difference between average in-class and inter-class similarities in family, genus, and species levels are statistically significant, and is used in another article to annotate unknown bacteria (Gursoy, and Can, 2019). It is natural to expect that these differences will also be reflected to shorter segments, hypervariable regions.

The nine hypervariable regions are grouped as V1–V2–V3, V4–V5–V6, and V7–V8–V9. Using the sub strings that approximately cover each these groups, the annotation of sampled genes are annotated in three taxonomic levels family, genus, and species.

3. RESULTS AND DISCUSSION

When all sub strings that approximately cover each of these three groups are used to annotate sampled genes in three taxonomic levels family, genus, and species a satisfactory accuracy is reached.

Table 4. Average accuracy of annotation of sampled genes by the use of short strings that cover hypervariable regions in the taxonomic level species

	Whole	VI-V2-V3	V4-V5-V6	V7-V8-V9
Greengenes	73.19	69.13	66.70	61.66
SILVA*	62.25	66.00	61.50	61.25

*The species for which at least 100 genesreported are cosidered

Table 5. Average accuracy of annotation of sampled genes by the use of short strings that cover hypervariable regions in the taxonomic level genus

	Whole	VI-V2-V3	V4-V5-V6	V7-V8-V9
Greengenes	78.95	87.36	86.38	77.88
SILVA*	90.20	85.40	87.80	93.70
RDP	76.56	65.58	47.18	22.55

*The species for which at least 100 genesreported are cosidered

Table 6. Average accuracy of annotation of sampled genes by the use of short strings that cover hypervariable regions in the taxonomic level family

	Whole	VI-V2-V3	V4-V5-V6	V7-V8-V9
Greengenes	98.98	91.20	94.64	89.48
SILVA	74.01	67.96	77.18	69.84
RDP	64.22	73.39	66.97	29.36

From tables 4, 5, and 6 it is seen that it is possible to annotate short reads around 500 base pairs using the three gene libraries with their 198.510 genes for Greengenes, 1.488.662 genes for SILVA, and 1.350.270 for RDP with an acceptable accuracy.

However if the a short reads are less than 500 base pairs, the nnotation accuracy drops down as in Table 7.

Table 7 Average accuracy of annotation of sampled genes by the use of short strings that cover a part symmetrical from the center of the hypervariable region group V4-V5-V6 in the taxonomic levels of the database Greengenes.

Levels	# levels	50	150	300	600	FULL
Phylum	86	59	76	86	87	94
Class	232	68	77	83	91	90
Order	366	63	82	87	90	89
Family	466	45	84	91	94	95
Genus	1949	35	72	81	87	89
Species	2389	26	49	52	64	71

From Table 7 it is seen that for a satisfactoryly accurate annotation using Longest Common Substring similarity measure and the three libraries Greengenes, SILVA, and RDP, short reads must not be much less than 500 base pairs.

REFERENCES

Baker GC, Smith JJ, Cowan DA. Review and re-analysis of domain-specific 16S primers. J Microbiol Methods. 2003; 55(3): 541–55.

He ZL, Van Nostrand JD, Zhou JZ. Applications of functional gene microarrays for profiling microbial communities. Curr Opin Biotech. 2012; 23(3):460–6

Lane DJ, Pace B, Olsen GJ, Stahl DA, Sogin ML, Pace NR. Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. Proc Natl Acad Sci U S A. 1985;82 (20): 6955–9.

Liu WT, Marsh TL, Cheng H, Forney LJ. Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA. Appl Environ Microbiol. 1997;63(11):4516–22.

Muyzer G, de Waal EC, Uitterlinden AG. Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. Appl Environ Microbiol. 1993;59(3):695–700.

Gursoy, O., and Can, M. (2019) Taxonomic Classification of Bacteria Using Common Substrings, Southeast Europe Journal of Soft Computing Vol.8 No.1 March, (33-39).

Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng J-F, et al. Insights into the phylogeny and coding potential of microbial dark matter. Nature. 2013;499 (7459): 431–7.

Tringe SG, Hugenholtz P. A renaissance for the pioneering 16S rRNA gene. Curr Opin Microbiol. 2008;11(5):442–6.

Wagner M, Noguera DR, Juretschko S, Rath G, Koops HP, Schleifer KH. Combining fluorescent in situ hybridization (FISH) with cultivation and mathematical modeling to study population structure and function of ammonia-oxidizing bacteria in activated sludge. Water Sci Technol. 1998;37(4–5):441–9

Wang Y, Qian P-Y. Conservative fragments in bacterial 16S rRNA genes and primer design for 16S ribosomal DNA amplicons in metagenomic studies. PLoS One. 2009; 4(10): e7401.

Yang, B., Wang, Y., 2 and Qian, P-Y. (2016) Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis, BMC Bioinformatics 17:135