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

On the Accuracy of the 16S-rRNA Gene Conserved Regions

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:
Longest common subsequence, Biodiversity, Conserved regions 16S, Primer design

ABSTRACT: The study of microbial communities through sequencing the 16S rRNA gene by the use of high throughput sequencing technology has emerged as a significant improvement for the discipline. However, the short size of these sequences is a limiting factor for the taxonomic classification of bacteria and archaea. These short reads are amplified from DNA, using primers. Although several researchers claim that they succeeded to create the best universal primers, the reality is that no primer has been demonstrated to be truly universal. This suggests that conserved regions of the 16S rRNA gene is not conserved enough. The aim of this study is to evaluate the conservation degree of the conserved regions separating the hypervariable regions of the 16S rRNA genes. Data contained in Greengenes, SILVA, and RDP databases are used for the study. Primers reported as matches of each conserved region were assembled to form fifteen contigs by Martinez-Porchas et al. (2017). Under the information of the degenerate bases in primes these contıgs are multiplied to cover all possibilities of degenerate bases. In Greengenes database there are 198.510 non redundant 16S rRNA genes are reported. This number is 1.488.662 for, SILVA, and 1.350.270 for RDP. To analyze the level of conservation of a contig, one gene is selected from one database, then using the longest common subsequences, for each of these 15 contigs, the longest common subsequences are found between a contig, and a gene. Then the length of longest common subsequence is divided by the length of the contig to get the percentage of conservation of this contig in that gene. This is done for each contig, in the entire databases. Averages revealed that the segments of contigs are not as conserved as expected, 72% in Greengenes, 71% in SILVA, and 57% in RDP. It is concluded that conserved regions of the 16S rRNA genes exhibit considerable variation that has to be considered when using these conserved regions as bases for primer production.

1. INTRODUCTION

After the emergence of high throughput sequencing technology, the study of microbial communities through sequencing the 16S rRNA gene by Sanger method has been abandoned by most of the scientists. High throughput robust study of microbial communities, allowing laboratories to obtain millions of high quality sequences. Although this is a significant achievement for the study of bacterial diversity, the short size of the resulting sequences is a limiting factor for the taxonomic classification of bacteria.

A new technology based on single molecule real-time sequencing (SMRT), has been tested as possible solution for this problem with promising results by Martinez-Porchas et al. (2017). However, the perbase sequencing cost is still significantly higher than that of current high throughput sequencing platforms, and therefore not practical for most of the laboratories. For this reason, the use of 16S rRNA short reads to study bacterial diversity will remain as the main strategy during the coming years. It is also seen that, there are particular taxonomic groups of bacteria that would require a specific fraction of the 16S rRNA gene.

This fact has an impact on primer design, several primers targeting diverse hypervariable regions of the 16S rRNA gene have been used and reported as guarantee of wide coverage and good amplification (Klindworth, et.al. 2012; Sambo et al., 2018); however, none of these primers is truly universal and the coverage usually depends upon intrinsic factors such as sequence, size, position, degenerations, combinations during the primer design, chemical reagents used, amplification conditions and other PCR biases, and extrinsic factors such as the kinds of samples, such as bacterial composition and environment, and PCR inhibitors hauled in the sampling process (Janda, and Abbott, 2007). Moreover, regions of the 16S rRNA gene differ in taxonomic informativeness; thus, some regions seem to be more useful for taxonomic classification from a general perspective and others for particular taxa.

Most of the studies regarding proposal and effectiveness of different primers are usually based on the study of biological samples. These studies have been useful to extend the panorama regarding the research of bacterial communities; however, the intrinsic and extrinsic factors influencing the performance of these primers do not allow concluding if they have the best possible coverage. These kinds of results allow to conclude if one pair of primers is better than others, but do not provide conclusive information regarding coverage; for example, it is possible that only a fraction of the species thriving in any environmental sample are being covered by any combination of primers, whereas the 16S rRNA fraction of others may require different amplification conditions, or do not match with the primer sequence because some fragments of the conserved regions are probably not as conserved as expected, and so on. In this case, the information provided by environmental samples regarding the coverage of any pair of primers would be incomplete.

Furthermore, many of the primers used are frequently not validated through in silico tests, while others are proved only with a couple of thousand sequences previously selected. Additionally, the use of degenerated primers has been proposed for the amplification of DNA coding for homologous genes, covering a larger number of genes from unspecific prokaryotes (Frank, et. al. 2008).

Degenerated primers were initially designed manually, inserting degenerations after multiple alignments; however sophisticated software programs are used today (Boye et al., 1999). Thus, it is necessary to understand variations in conserved regions of the 16S rRNA gene and to carry out tests with all possible sequences, including degenerations, and combinations; which is impractical and unfeasible.

However, this can be done virtually considering all of the information contained in robust databases, not only the sequences obtained from biological samples. Moreover, the analysis of these conserved regions could provide useful information to evaluate how much conserved are these regions. Therefore, the aim of this study is to evaluate the conservation degree of the so-called conserved regions flanking the hypervariable regions of the 16S rRNA gene.

2. MATERIALS AND METHODS

In Martinez-Porchas et al. (2017), all of the primers reported for each region were aligned to generate a continuous primer contig sequence, if primers formed separated contigs by a gap, each segment was considered as sub-contig (a, b or c), as in Table 1.

Table 1 Primer contigs generated by assembling all of the primers reported for each conserved region of the 16S rRNA gene. Location is based on E. coli sequence (Martinez-Porchas et al. 2017).

N	Sequence	Location
1	AGAGTTTGATYMTGGCTCAG	8_27
2	ASYGGCGNACGGGTGAGTAA	100_119
3	ACTGAGAYACGGYCCARACTCCTA CGGRNGGCNGCAGTRRGGAA	320_363
4	GGCTAACTHCGTGNCVGCNGCYGC GGTAANAC	504_535
5a	GTGTAGMGGTGAAATKCGTAGAT	682_704
5b	CAAACRGGATTAGAWACCCNNGTA GTCCACGC	778_809
6a	AAANTYAAANRAATWGRCGGGGR CCCGCACAAG	906_938
6b	ATGTGGTTTAATTCGA	948_963
6c	CAACGCGARGAACCTTACC	966_984
7a	AGGTGNTGCATGGYYGYCGTCAGC TCGTGYCGTGAG	1045_1080
7b	TGTTGGGTTAAGTCCCRYAACGAG CGCAACCCT	1082_1114
8a	GGAAGGYGGGGAYGACG	1176_1192
8b	GGGCKACACACGYGCTAC	1219_1236
9	GCCTTGYACWCWCCGCCCGTC	1386_1406
10	GGGTGAAGTCRTAACAAGGTANCC	1486_1509

In Table 2 it is seen that these contigs contain degenerate bases R, Y, M, K, S, W, H, D, B, V, and N. The possible amino acids corresponding these degenerate bases is given in Table 2. For example K represents T or G, therefore sequence containing this degeneration was multiplied by two possibilities. This was also considered for all kinds of degenerations detected in all sequences; for instance, Y, M, S, R, W have two possibilities each, V, H, B, D three possibilities, and *N* has four as seen in Table 2.

Table 2 Degenerations detected in contigs, Y, M, S, R, W has two possibilities each, V, H, B, D have three possibilities, and *N* has four.

Key to symbols	
R	A or G
Y	C or T
M	A or C
K	G or T
S	G or C
W	A or T
H	A or T or C
D	G or A or T
B	G or T or C
V	G or A or C
N	A or T or G or C

After multiplication of contigs to cover all possibilities the number of contigs are increased as seen in Table 3.

Table 3. Possibilities the number of contigs

Name	Length	ND	D bases	Alt
1	20	2	YM	4
2	20	3	SYN	16
3	44	8	YRN	16
4	32	6	HVYN	72
5a	23	2	MK	4
5b	32	4	RWN	16
6a	33	7	YWN	32
6b	16	0	0	1
6c	19	1	R	2
7a	36	5	YN	8
7b	33	2	RY	4
8a	17	2	Y	2
8b	18	2	KY	4
9	21	3	YW	4
10	24	2	RN	8
Total	388	49		193

In Table 3, ND is the number of degenerated bases in the related contig, D bases is the degenerated bases in this contig, and Alt is the number of alternative contigs created for this contig.

2.1. Longest Common Subsequence Search

The conservation degree of the regions dividing all of the nine-hypervariable regions, V1-V9 of the 16S rRNA gene was estimated through the analysis of data contained in the high quality ribosomal RNA databases Greengenes, SILVA, and RDP. This number of non-redundant bacterial sequences with around 1,200 bases length is 198.510 for Greengenes. This number is 1.488.662 for SILVA, and 1.350.270 for RDP.

To analyze the level of conservation of a contig, one gene is selected from one database, then using the longest common subsequences, for each alternative of this contigs the longest common subsequence is found between this alternative of the contig, and a gene. Then the length of longest common subsequence of the alternative contigs is divided by the length of the contig to get the percentage of conservation of this contig in that gene. This is done for each contig, in the entire databases.

Assume in Figure 1., (a) is a piece of a gene reported for a bacteria, and (b) is the (AG) alternatives degenerate bases (MK) of contig 5a: GTGTAGMGGTGAAATKCGTAGAT.

(a) GGCTAACTA**GTGTAGAGGTGAAATG**ATT **TAGAT**TAGGTGGCAA….

(b) **GTGTAGAGGTGAAATG**CG**TAGAT**

Figure 1. The longest common subsequence of (a) a gene and (b) a contig

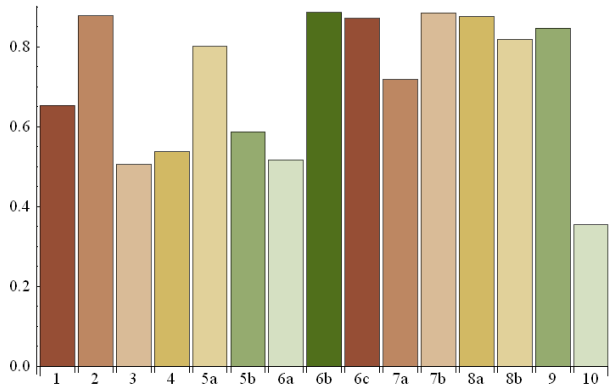
The longest common subsequence of (a) and (b) is

GTGTAGAGGTGAAATG

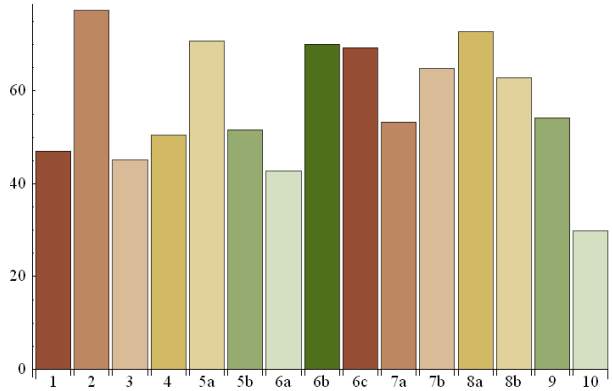
We note this LCSS, we repeat this process for all four alternatives of contig 5a, and the length of the longest LCSS of the four is divided by the length of the contig 5a to ger the ratio of conservation of contig 5a in this gene. The portion **TAGAT** of 5a is also common, but it is not counted. This is done for all genes in the database. The average of these percentages are the percentage of the conservation of the contig 5a. This is done for all fifteen contigs and for all genes in the data bases Greengenes, RDP, and SILVA.

3. RESULTS AND DISCUSSION

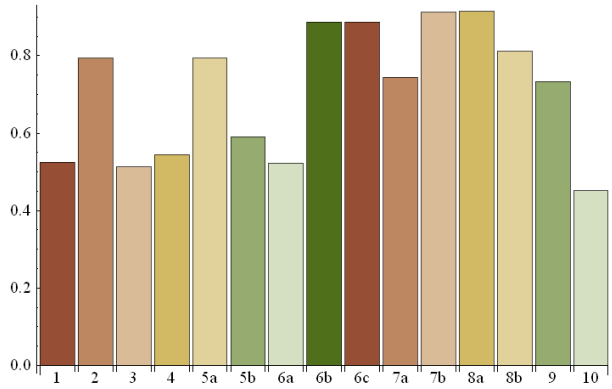
When all contigs are checked against all genesin the database, we obtained percentages of the conservation of the fifteen contigs in and all genes in the data bases (a) Greengenes, (b) RDP, and (c) SILVA as in Figure 2.



(a)



(b)



(c)

Figure 2. Percentages of the conservation of the fifteen contigs in and all genes in the data bases (a) Greengenes, (b) RDP, and (c) SILVA.

These results are also shown in Table 4.

Table 4. Percentages of the conservation of the fifteen contigs in and all genes in the data bases Greengenes, SILVA, and RDP

Name	L (E Coli)	Greeng%	SILVA%	RDP%
1	8-27	53.78	52.46	47.07
2	100-119	80.06	79.41	77.33
3	320-363	58.71	51.34	45.08
4	504-535	51.75	54.47	50.50
5a	682-704	88.59	79.41	70.82
5b	778-809	87.10	59.05	51.62
5c	906-938	71.78	52.24	42.74
6a	948-963	88.56	88.68	70.10
6b	966-984	87.52	88.71	69.35
7a	1045-1080	81.86	74.43	53.31
7b	1082-1114	84.71	91.21	64.85
8a	1176-1192	35.43	91.42	72.73
8b	1219-1236	53.78	81.13	62.91
9	1386-1406	80.06	73.31	54.22
10	1486-1509	58.71	45.18	29.78
Mean		72.49	70.83	57.49

Thus, considerable coverage variability was observed in conserved regions located at the extremes of the 16S rRNA gene 1 and 10 where the average conservation is not more than 60%. These results could call into question the suitability of some of the primers that have been used for long time. In spite of these variabilities, LCSS analysis revealed that there are yet particular segments within each region with acceptable conservation degree to be considered for the study of prokaryotic diversity. In this regard, regions 3, 5a, 6b-9 and 6a for which high conservation percentages are observed in all three data bases can be useful to design primers that are more suitable to profiling and comparing microbial communities; however, additional considerations have to be taken into account to design primers (Wang& Qian, 2009).

REFERENCES

Boye, K., Høgdall, E., and Borre, . (1999) Identification of bacteria using two degenerate 16S rDNA sequencing primers, Microbiological ResearchVolume 154, Issue 1, Pages 23-26

Frank, J.A, Reich, C.I., Sharma, S., Weisbaum, J.S., Wilson, B.A., and Olsen, G.J. (2008) Critical Evaluation of Two Primers Commonly Used for Amplification of Bacterial 16S rRNA Genes, Applied And Environmental Microbiology, p. 2461–2470 Vol. 74, No. 8

Janda, J.M., and Abbott, S.L. (2007) 16S rRNA Gene Sequencing for Bacterial Identification in the Diagnostic Laboratory: Pluses, Perils, and Pitfalls, J Clin Microbiol. 2007 Sep; 45(9): 2761–2764.

Klindworth, A., Pruesse, E., Schweer, T., Peplies, J. Quast, C., Horn, M., and Glöckner, F.O. (2012) Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies, Nucleic Acids Res. 2013 Jan; 41(1)

Martinez-Porchas M., Villalpando-Canchola E., Ortiz Suarez LE, Vargas-Albores F. (2017) How conserved are the conserved 16S-rRNA regions? PeerJ.;5:e3036.

Sambo, F., Finotello, F., Lavezzo, E., Baruzzo, G., Masi, G., Peta, E., Falda, M., Toppo, S., Barzon, L., and Di Camillo, B. (2018) Optimizing PCR primers targeting the bacterial 16S ribosomal RNA gene, BMC Bioinformatics, 19:343