Region-specific Similarity Cutoff for Bacterial Species Circumscription in 16S rRNA-based Studies

Mincheol Kim[1]*, and Hyun-Seok Oh[2]

[1]Korea Polar Research Institute, Incheon; [2]School of Biological Sciences, Seoul National University, Seoul, South Korea

Summary

16S ribosomal RNA is the most widely used molecular marker in microbial community studies. Generally, 16S rRNA gene sequences are clustered into operational taxonomic units (OTUs) based on sequence similarity and the OTU defined at 97% similarity cutoff is commonly used as a 'proxy' for bacterial species. The problem is that the cutoff originally set for the full length 16S rRNA has also been used in subregions of 16S rRNA without proper validation. Here, we tested first the suitability of using sequence similarity itself in three subregions of 16S rRNA and then determined the optimal similarity cutoff for species demarcation in each region on the basis of genomic evidence. The effect of alignment and distance calculation methods on similarity cutoff values was additionally investigated. F-measure analysis revealed that V1-V3 region showed the highest taxonomic accuracy, a similar extent to that of full length, and lowered in other two regions, V3-V5 and V6-V9 regions. Optimal similarity cutoff corresponding to bacterial species boundary differed by regions. Similar degree of cutoff was obtained for V3-V5 (99.0%) and V6-V9 regions (99.1%), whereas a much lower level of cutoff was estimated for V1-V3 region (98.0%). Alignment methods have little influence on the optimal cutoff values, while the cutoff values differed more by distance calculation methods. Taken together, our results show that region-optimized similarity cutoff should be used when clustering subregions of 16S rRNA into OTUs and among three subregions V1-V3 is the most appropriate target for similarity-based OTU clustering.

Questions

1. What is the species-level similarity cutoff when partial 16S rRNA gene sequences are used in bacterial community studies?

2. Is 'sequence similarity' itself a robust parameter in partial 16S rRNA gene as was the case for the full length gene?

3. How does the similarity cutoff change with differing alignment and similarity calculation methods?

Materials & Methods

Bacterial genome data collection
- 15,844 prokaryote genomes from NCBI (Jan 2014)
- Excluding single-cell genomes and assembled genomes from metagenome data
- Genome quality filtering
  (e.g. # of contigs < 2000, N50, abnormal G+C content, etc.)
- 14,314 high-quality genomes were finally obtained!!

Average Nucleotide Identity
- BLAST-based reciprocal ANI calculation (Goris et al., 2007)

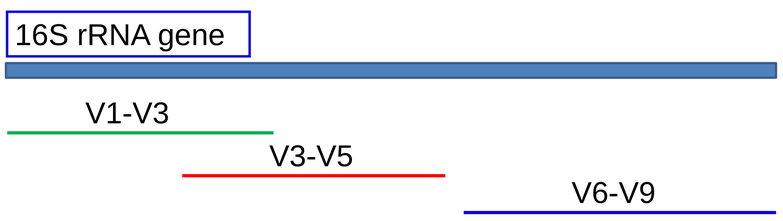
16S rRNA gene extraction
-16S rRNA extraction from genomes (HMMER3)

Sequence alignment and similarity calculation
- Pairwise sequence alignment (Myers and Miller, 1988)
- Multiple sequence alignment using mothur
- Similarity calculation with three different gap penalties

```
ACGTGGTGAC
⋮⋮  ⋮⋮⋮⋮ ⋮
AC--GAAG-C
```

Without gaps	2/7 = 0.286
Any gap as a mismatch	4/9 = 0.444
Every position in gaps	5/10 = 0.500

Statistical analyses
-Precision, recall, F measure, and twofold cross validation



Results

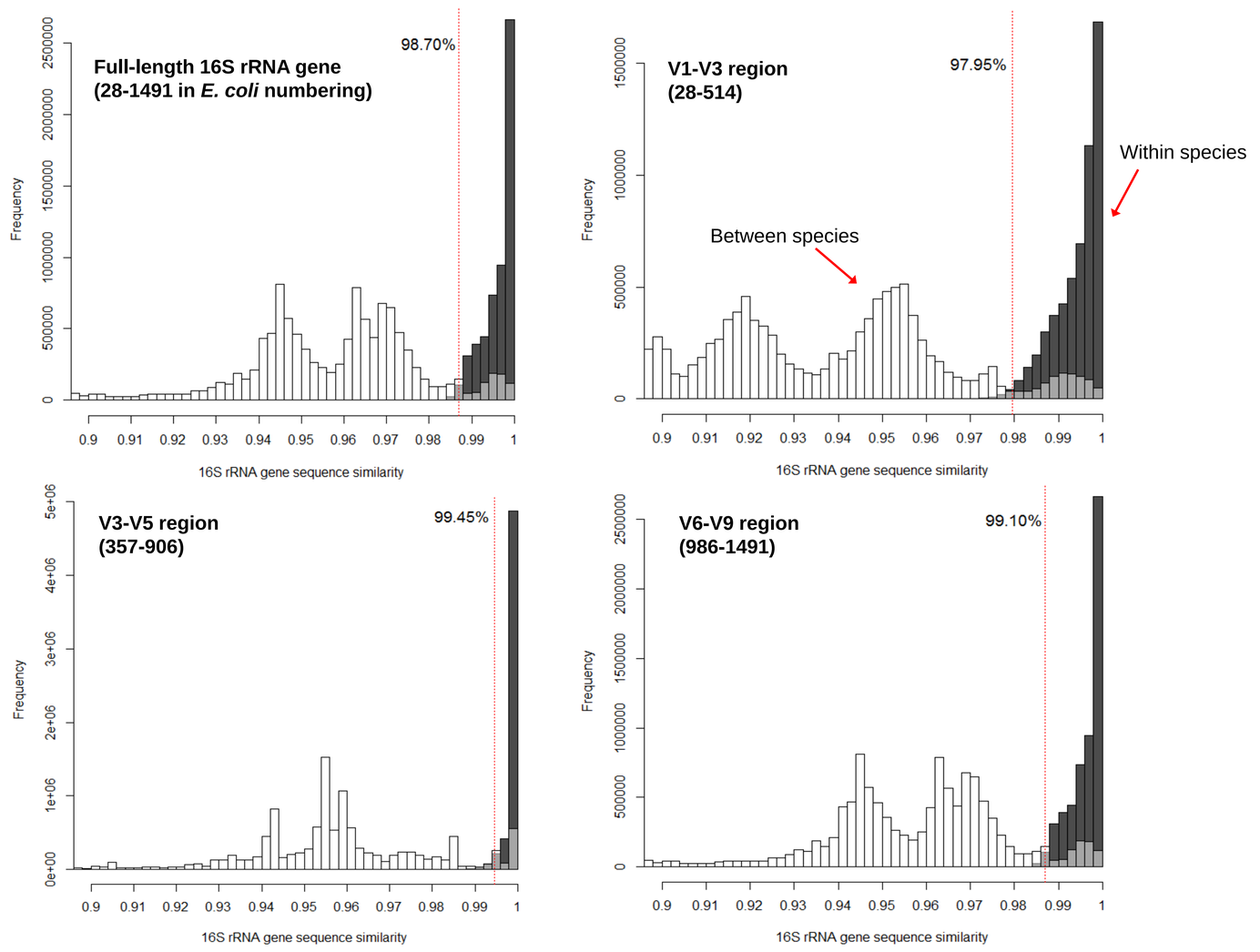


Fig. 1. Region-specific similarity cutoff for bacterial species demarcation. 95% ANI value corresponding to 70% DNA-DNA hybridization (DDH) was used for circumscribing bacterial species.

V1-V3 shows the best accuracy !!

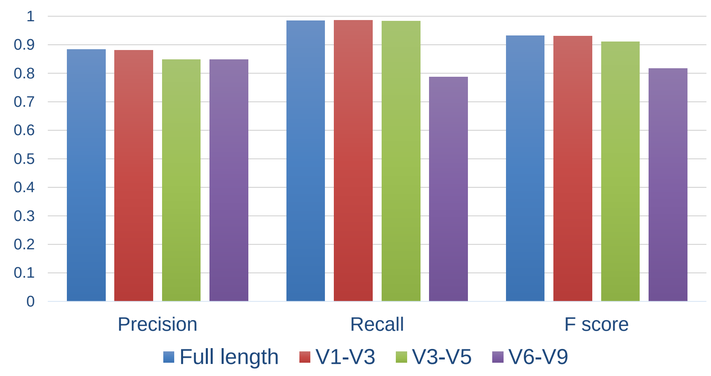


Fig. 2. Taxonomic accuracy of region-specific cutoff values.

Table 1. Effect of alignment and similarity calculation methods on similarity cutoff values

	Gap penalty	Alignment methods	
		Pairwise Alignment	Multiple Alignment
Full length	No gap	98.7%	98.7%
	One gap	98.7%	**98.6%**
	Each gap	98.7%	98.6%
V1-V3	No gap	98.0%	98.0%
	One gap	98.0%	**97.8%**
	Each gap	98.0%	97.8%
V3-V5	No gap	99.5%	99.5%
	One gap	99.1%	**99.1%**
	Each gap	99.1%	99.1%
V6-V9	No gap	99.1%	99.1%
	One gap	98.9%	**99.1%**
	Each gap	99.0%	99.1%

Previous 97% similarity cutoff "highly" underestimates the true bacterial diversity !!

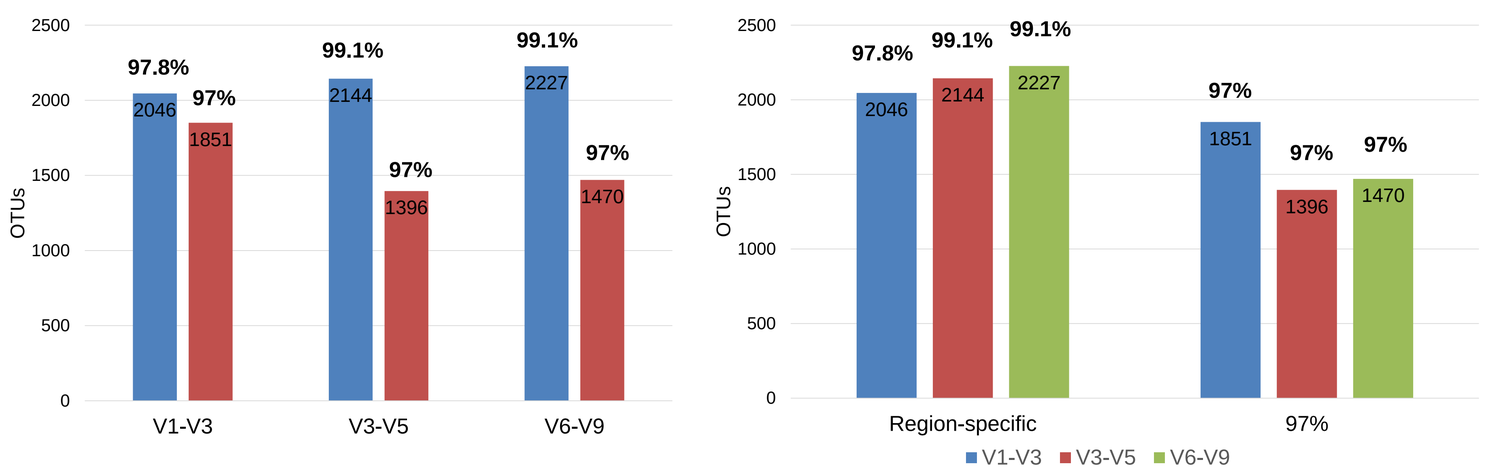


Fig. 3. Differing OTU richness with different similarity cutoff values. All available 16S rRNA gene sequences were extracted from 14,314 high-quality genomes with 2,509 bacterial species. The number of OTUs were calculated by applying both region-specific cutoff values and a single cutoff value (97%).

Conclusions

- Different similarity cutoff values should be used when estimating bacterial OTUs in partial 16S rRNA-based studies
- V1-V3 region showed the best accuracy in sequence similarity-based OTU clustering

e-mail: mincheol@kopri.re.kr