

Towards Understanding Bacterial Systematics Using Topological Similarity between Phylogenetic Trees Based on 16S ribosomal RNA Genes and Conserved Proteins

Yong Wang* and Jiao-Mei Huang

Department of Life Science, Institute of Deep Sea Science and Engineering,
Chinese Academy of Sciences, San Ya, Hai Nan, PR China

Research Article

Received date: 13/06/2017

Accepted date: 25/07/2017

Published date: 31/07/2017

*For Correspondence

Yong Wang, Department of Life Science,
Institute of Deep Sea Science and
Engineering, Chinese Academy of Sciences,
San Ya, Hai Nan, PR China,
Tel: (86) 898-88381062,
Fax: (86) 898-88222506.

E-mail: wangy@sidsse.ac.cn

Keywords: Phylogenetic relationship,
Geodesic distance, Bacterial phylum,
Ribosomal RNA gene, Conserved proteins

ABSTRACT

Single-copy conserved genes and internal regions of 16S ribosomal RNA (rRNA) are potentially molecular markers for bacterial systematics. In this study, 33 highly conserved genes, the three rRNAs and internal regions of 16S rRNA were used to construct phylogenetic trees for 25 bacterial phyla. Topological dissimilarity revealed that the trees for an internal region between 515bp and 785bp in 16S rRNA are highly similar to that for full-length 16S rRNA. Pairwise comparison of the trees for conserved genes and rRNAs also showed that 16S and 23S rRNA genes were grouped with ribosomal proteins (r-proteins) L2, S3, S7, L14, S10 and S12. The 16S rRNA internal region and the r-proteins were critical in translation process, indicating a tight connection between bacterial systematics and translation process. To demonstrate efficiency of taxonomic sorting by the markers, monophyletic clades in the trees were examined for co-existence of the taxa from the same phyla. Translation initiation factor 2, 16S rRNA and 23S rRNA could assign almost all the taxa correctly into the monophyletic clades. Together, our results suggest that rRNAs and some r-proteins co-evolved due to involvement in mRNA manipulation and formation of decoding center, and may thus serve as molecular markers for bacterial systematics.

INTRODUCTION

The current widely-accepted bacterial systematics had been established based on similarity of 16S ribosomal RNA (rRNA) genes and other highly conserved genes^[1,2]. A bacterial isolate may be approximately classified into different taxonomic levels by comparing with the sequences in public database. Phylogenomic analysis using concatenated conserved genes is an alternative to locate a bacterium in the tree of life^[3,4]. Using enrichment cultivation and single-cell genomics, novel bacterial phyla were discovered and representative genomes were decoded. With a growing number of new bacterial phyla in recent years, approaches to quick identification and positioning of new taxonomic bacterial groups in rare environments are highly demanded. To study microbial composition of a high biodiversity is also a difficult task, due to lack of comprehensive evaluations of the existing markers.

Since 1970s, small subunit rRNA (ssu-rRNA) genes had been used for taxonomic identification of organisms and laid a foundation for systematic studies on the tree of life^[5,6]. Archaea as a new domain independent from Bacteria and Eukaryotes was recognized by phylogeny of 16S rRNA genes. A 16S rRNA gene sequence contains both highly conserved regions for primer design and hyper-variable regions for taxonomic positioning of a microorganism. A long-standing question is which 16S rRNA region should be targeted for PCR amplicons, particularly when a complex microbial community in an environmental sample was subjected to a study. Although the 16S rRNA gene is a reliable phylogenetic marker for systematics of bacterial phyla, it is difficult to sequence full-length 16S rRNA genes with high-throughput sequencers. Though per base costs were lowered drastically and sequencing depth was improved nowadays^[7,8], the average sequencing read length was not much increased^[9]. Considering the short read length of current sequencing platforms, selection of proper

primers is still critical for studies of bacterial community in various environments. Recent studies utilizing high-throughput technology also revealed that using suboptimal primer pairs would not unevenly amplify rRNA genes from certain species. The biases in amplification caused either under- or over-estimate of some species in a microbial community^[10-12]. Several studies have attempted to uncover the optimal primer pairs for bacterial community study^[13-15]. However, their results were under modifications as more bacterial and archaeal phyla were revealed^[16].

A more important issue is that a high copy number of 16S rRNA gene in a certain group of microbes may raise over-estimate of their composition in the community¹. Depending on ecological strategy and genome size, bacteria differ remarkably in copy number of rRNA operons^[17,18]. The wide range of copy number from one to up to twelve copies¹⁸ may result in wrong estimates of biodiversity and microbial composition in a sample. An alternative option is to use single-copy genes in bacterial genomes as molecular markers. There are at least 38 universally conserved genes in prokaryotes^[4,19]. Their ubiquitous presence in bacterial genomes may allow for precise positioning of a bacterial isolate into the systematic phylogeny of all the organisms. However, whether these conserved genes can be consistent with 16S rRNA as molecular markers in phylogenetic topology of bacterial phyla remains to be answered. It is possible that the tree of life constructed using some universally conserved genes differ profoundly from that based on the ssu-rRNA genes.

In this study, we worked on rRNA genes and 33 universally conserved genes, aiming to answer 1) which internal regions in 16S rRNA genes are approximately equivalent to the full-length gene in the classification efficiency; 2) whether 16S, 23S and 5S differ in topology pattern of phylogenetic trees constructed using different bacterial phyla; 3) how different are the phylogenetic trees reconstructed using rRNA genes and universally conserved proteins in sorting of bacterial phyla. We used geodesic distance to estimate the dissimilarity of the phylogenetic trees for rRNA genes and conserved proteins. Topological dissimilarity between the phylogenetic trees may be estimated by geodesic algorithm that can project the node and branch structure of a tree into a multi-dimensional model^[20]. Geodesic distance has also been applied to quantify discrepancy between the phylogenetic trees^[21]. Recently, a study took advantage of this method to quantitatively compare phylogenetic trees reconstructed using 38 conserved bacterial genes^[4]. Pairwise geodesic distance exhibited the topology of the tree for translation initiation factor 2 (IF2) is highly similar to the concatenated marker sequences^[4]. This result suggests that IF2 can roughly reflect the genetic relationships that were built up using all the conserved genes. With the help of geodesic algorithm, it is possible to quantify the sensitivity of rRNA genes and single-copy conserved genes for taxonomic classification. With the quantitative evaluation, the correlative relationships between different markers in classification efficiency can be elucidated. Finally, phylogenetic congruency of bacterial phyla in monophyletic clades was examined to compare the sorting efficiency of all the molecular markers.

RESULTS AND DISCUSSION

Dissimilarity of phylogenetic trees based on full-length 16S rRNA genes and variable Regions

A total of 56 bacterial representative's species from 25 bacterial phyla were used for comparison of phylogenetic relationships. Bayesian phylogenetic trees were firstly constructed for 16S rRNA genes and internal regions. Partition of the 16S Rrna genes resulted in 10 regions of R1-R10, which were started with a conserved region and ended up with a variable region (**Figure 1**). The size of the regions ranged from 102 bp to 237 bp in reference to their positions in *Escherichia coli* MG1655 (**Table 1**). Dissimilarity of the Bayesian phylogenetic inferences was estimated with geodesic distance algorithm. Agglomerative hierarchical clustering (AHC) result showed that the phylogenetic tree based on the full-length 16S rRNA gene was the nearest neighbor to that based on the region of R4. This indicates that the topology of the trees for the full-length and R4 was more similar to each other than to the other regions. R5 was the second closest region to the full-length 16S rRNA gene. Therefore, the 16S rRNA region between the positions of 515 bp and 786 bp may serve as the best target region of PCR amplicons for taxonomic classification, because the taxa were positioned in similar phylogenetic trees as the full-length 16S rRNA genes did. As suggested by a previous study^[22], R4 ranks the first in sensitivity as a reliable marker for bacterial taxonomic classification. Our study, for the first time, estimated dissimilarity in topology of phylogenetic relationships, and pairwise relationships of the different 16S rRNA regions. In contrast, the regions of R1, R2, R10 and R7 were distantly related to the internal regions, indicating that the peripheral regions were less efficient for the classification of the bacterial phyla than the regions such as R4, R5, R6 and R8. This result is against the previous recommendation of R1-R3 as the target of PCR amplicons for detection of microbial community^[13]. The heterogeneity of phylogenetic trees based on different regions is also not consistent with a recent analysis, which shows almost equal recovery efficiency for the internal regions^[23].

Table 1. Internal regions of 16S ribosomal RNA sequence. 518, The start and end positions were referred to the corresponding sites in the 16S rRNA sequence from *Escherichia coli* MG1655 (U00006).

Region	Start (bp)	End (bp)	Length (bp)
R1	1	103	103
R2	104	340	237
R3	341	514	174
R4	515	684	170
R5	685	785	101
R6	786	916	131
R7	917	1052	136
R8	1053	1225	173
R9	1226	1389	164
R10	1390	1491	102

It remains a question why the 16S rRNA internal regions differ in the classification sensitivity. The fundamental reason perhaps lies in their different roles in the translation process of mRNA molecules. The internal regions interact with different parts of translation machine and other rRNA parts, resulting in the variances in conservative degree and classification efficiency. In 3D structure of 16S rRNA, it is easy to connect the regions of R4-R5 with the translation center of ribosome complex. Within R4-R5, the important ‘690 hairpin’ (680 bp-710 bp) 24 and decoding center (G530 and neighboring sites) [24-26] had been pinpointed. The ‘690 hairpin’ is highly conserved in all three domains due to its essential role in translation process [27,28]. This hairpin was reported to mediate binding of tRNA at P-site, r-protein S11 and translation initiation factor 3 and to interact with 790 loop of 16S rRNA and domain IV of 23S rRNA [29]. However, we could not evaluate contribution of these positions in the classification of bacterial phyla as molecular markers in the present study. Also it is difficult to locate the regions (no matter conserved or variable) that are more critical in determining the taxonomic position of a given 16S rRNA sequence. Previous studies have reported that the regions of R1-R3 and R10 are located at the bottom and top of a ribosome, respectively [25,26]. In our result, R7 exhibits less phylogenetic coherence compared with its adjacent regions, and the peripheral regions R1, R2 and R10. This suggests that the hypervariable region in R7 probably serves as a stabilize of the secondary structure of 16S rRNA and has not yet been reported to have functional importance. This reminds us of the debate over the taxonomic classification and molecular marker selection. There was an association between evolutionary rate and gene dispensability [30-32]. According to this theory, genes with high dispensability usually evolve slowly but might be applied for assignment of a species to higher taxonomic levels. In return, less important regions can allow for more changes and may maintain conservation only at lower taxonomic levels. Similarly in the current study, the functionally important R4-R5 might evolve at a lower rate and tend to be more stable than the other internal regions, and thus these regions could help to achieve a more stable phylogenetic topology among the diversified bacterial phyla. The structural regions may be less conserved and, therefore, contain more polymorphisms that may occur only at lower taxonomic levels. These regions cannot be used as markers for taxonomic classification of a novel lineage at phylum level in a community. This is a possible reason to the divergence between the internal regions in classification sensitivity.

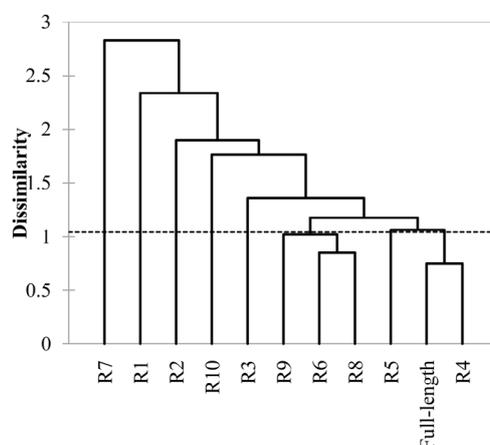


Figure 1. Dissimilarity of phylogenetic trees constructed using full-length and internal regions of 16S rRNA gene.

In our study, the functioning sites are usually quite short, compared to the whole region. Whether the several conserved sites in an rRNA gene may determine the topology of a phylogenetic tree consisting of 25 different phyla is questionable. On the other hand, the distance in topology of phylogenetic trees indicates that the weight of informative sites differed between the internal regions in hierarchical taxonomic classification. Some of the internal regions were probably more

useful as molecular markers for families and genera, rather than phyla. In this study, the length of the internal regions was different. R5 and R10 were relatively short and almost in the same length. However, R5 contains more valuable sites for taxonomic classification of the species at phylum level than R10. On the other hand, this comparison means that the difference in length between the internal regions would affect their sensitivity in classification.

Topological similarity of the phylogenetic trees based on rRNA genes and conserved genes

Further study was conducted using 33 conserved genes from the 56 bacterial species and 25 phyla. In the AHC result, four clusters were below the primary merging dissimilarity level at 2.1 (Figure 2). The first cluster consists of two subunit genes for phenyl-tRNA synthetase; the neighboring cluster is composed of translation initiation factor 2 and three r-proteins.

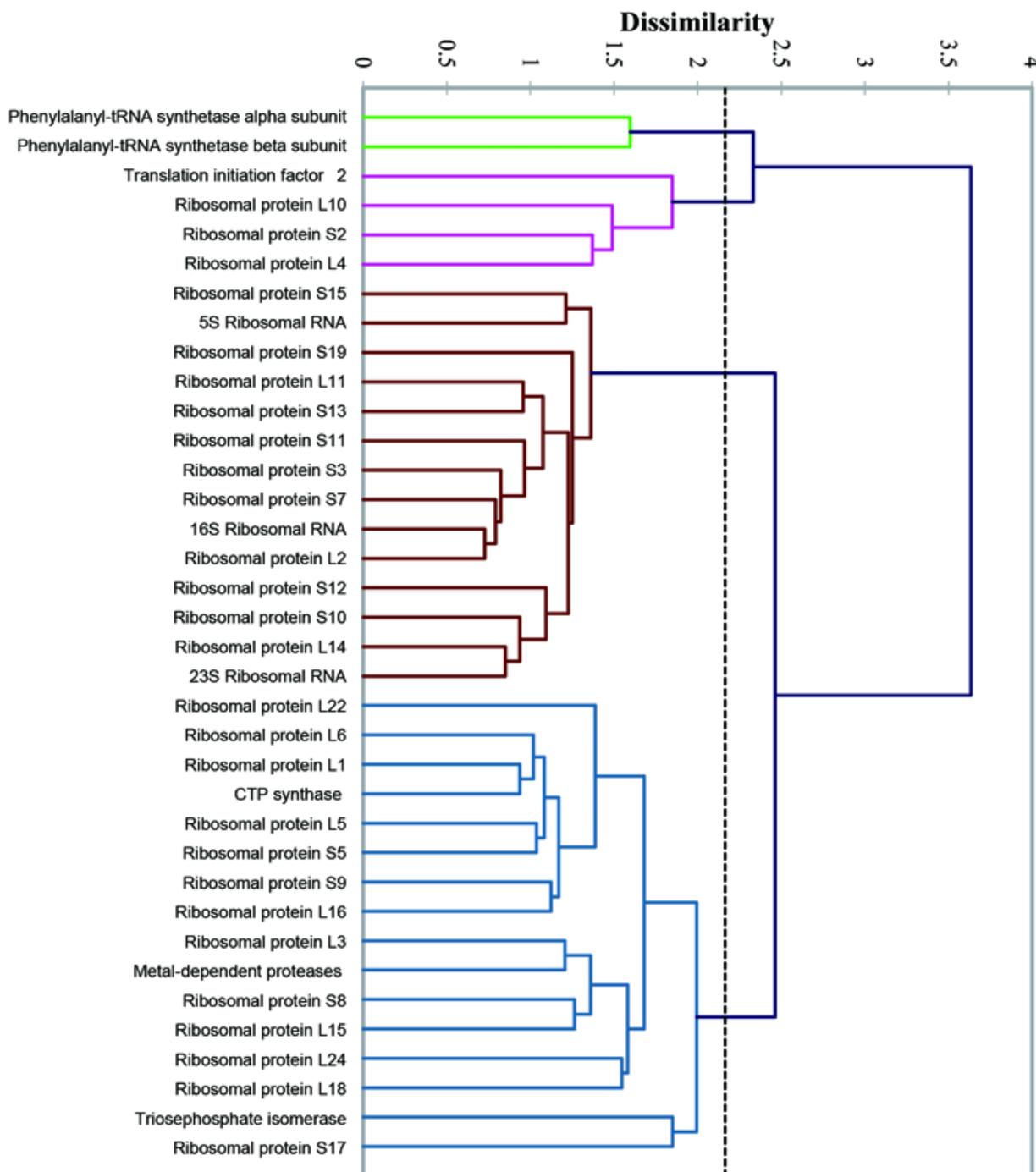


Figure 2. Dissimilarity and AHC of phylogenetic trees for bacterial phyla. The phylogenetic trees for 53 bacterial species using rRNA genes (5S, 16S and 23S) and 33 conserved genes were calculated in topological dissimilarity, which was demonstrated in AHC clustering.

The remaining proteins and rRNA genes were grouped into two big clusters. The three rRNA genes were located in one cluster, whereas their nearest neighbors are different. The trees built up using 5S rRNA genes exhibited the smallest

dissimilarity to those for r-proteins S15 and S19, while 16S rRNA genes shows affinity to r-proteins L2, S7 and S3. The topological dissimilarity also displayed a closer relationship between 23S rRNA genes and r-proteins L14, S10 and S12. Except for r-proteins L2 and L14, all these r-proteins were affiliated with 30S ribosomal subunit. In another big cluster, 13 r-proteins were neighbors of CTP synthase, metal dependent protease (COG0533) and triosephosphate isomerase.

All the ribosomal proteins that were grouped with the rRNA genes were highly related to translation function (**Table 1**). For example, r-protein L2 is a primary rRNA binding protein critical for association of 30S and 50S subunits and for tRNA binding and peptide bond formation [33]. The trees for r-proteins S3 and S7 were also highly similar to that of 16S rRNA gene (**Figure 2**). The two r-proteins bind directly to the head and lower part of the 30S subunit, respectively. The former is involved in unwinding helical structure in mRNA and is able to position an mRNA into the translation machine [34]. R-protein S7 is located just above the cleft and the decoding site, and is one of two protein components responsible for assembly initiation of bacterial small subunit ribosome [35]. S7 is also a major protein component that had been shown to cross-link with tRNA molecules bound at the A and P sites [36,37]. As one of the principal regulatory elements, S7 can also control r-protein synthesis by the translational feedback mechanism [38]. Moreover, it can stabilize decoding center of the 16S rRNA [26]. These r-proteins are heavily involved in translation process. Their similarity in topology to 16S rRNA gene suggests more contacts between these r-proteins and 16S rRNA. Experimental evidence is required to verify the closer functional association between the r-proteins and the 16S rRNA.

The phylogenetic tree for 23S rRNA showed a short distance to those of L14, S10 and S12 r-proteins. L14 is located at the interface of the small and large subunits, along with L2 [39]. Binding of translation silencing factor RsfA on L14 will result in termination of translation [40]. Ribosomal structure in 3.3Å resolution showed that S12 interacts with 23S rRNA and serves as a critical part of decoding center by modulating tRNA selection and in response to streptomycin 40. S10 is an antitermination apparatus in 70S ribosome [41]. It is regulated by r-protein L4 [42], a factor that initiates assembly of the large subunit [43]. It is interesting that 23S rRNA is grouped with S10 rather than L4 in topological comparison of the phylogenetic trees (**Figure 2**). This implies that some parts in S10 co-evolved with 23S rRNA sites that perhaps form decoding center. Further evidence is needed to support this hypothesis. Although L4 is critical in assembly of the large subunit, our result indicates that it is not congruent with 23S rRNA gene in evolution.

5S rRNA transfers information and coordinates different functional centers in ribosome [44]. Structural picture of 50S ribosomal subunit suggests that it binds r-proteins L5, L18 and L25 [45]. Our topological distance of the phylogenetic trees showed that 5S rRNA was not in the same cluster as L5 and L18, but was closer to 16S and 23S rRNAs (**Figure 2**). Again, this result indicates structural contacts did not a prerequisite for phylogenetic congruency. Perhaps, functional importance of 5S more than a role of coordinator determined its grouping with the rRNA genes and other essentially functional r-proteins.

Although some of the r-proteins are also critical in decoding process, they are not as conserved as the 33 genes in this study. An example is r-protein S1 that also mediates the initiation of translation by unwinding the secondary structure of mRNA and positioning it in decoding channel [46]. However, r-protein S1 was not among the 33 highly conserved gene lists. Internal variants and short conserved regions might affect identification of these genes, which may explain absence of r-protein S1 in identified conserved genes.

Congruency of phyla in phylogenetic trees

Accuracy of taxonomic sorting by the rRNA genes and conserved proteins might be visualized by the statistics of congruent species from the same phyla in the phylogenetic trees. We counted occurrence of the clades in which the species from the same phyla were grouped, and plotted in a black-white map for all the phylogenetic markers (**Figure 3**). The full-length 16S and 23S rRNA genes and r-protein S10 could almost cluster all the species from the same phyla into one clade. The species from one phylum could not be grouped. For 16S rRNA gene and r-protein S10, the species from Firmicutes were split into different clades; Gammaproteobacteria were not distributed in one clade in the phylogenetic tree of 23S rRNA genes. The performance of other markers was even worse in terms of correct grouping of the species from the same phyla.

Surprisingly, translation initiation factor 2 could also cluster the species into their individual phyla, and only the species from Actinobacteria and Firmicutes did not stay in the same clades (**Figure 3**). On the other hand, the phylogenetic relationships of the species exhibited a high dissimilarity between translation initiation factor 2 (IF2) and 16S rRNA genes. The two clusters where they were located were beyond the minimum emerging threshold line (**Figure 2**). With this respect, the phylogenetic relationships displayed by the two markers differed at a level beyond phylum. This suggests that the arrangement of the bacterial phyla in the trees based on IF2 and 16S rRNA was different.

For all the molecular markers, the taxa from Firmicutes, Actinobacteria and Spirochaetes were most difficult to be clustered into monophyletic clades (**Figure 3**). Our result demonstrated that the species from class of Clostridia could not be grouped with the other two from Firmicutes by most of the markers. This indicates the phylogenetic depth of Clostridia

could be reflected only by full-length 23S rRNA and four molecular markers. For Spirochaetes, a recent work revealed its high genetic distance among different classes [23]. A large number of genetic variations in the species from Spirochaetes probably blurred the informative sites useful for correct taxonomic assignment of different classes in Spirochaetes. In sum, our result indicates that partial rRNAs and most of r-proteins did not contain enough informative content that allows to fully distinguishing these taxa at phylum level.

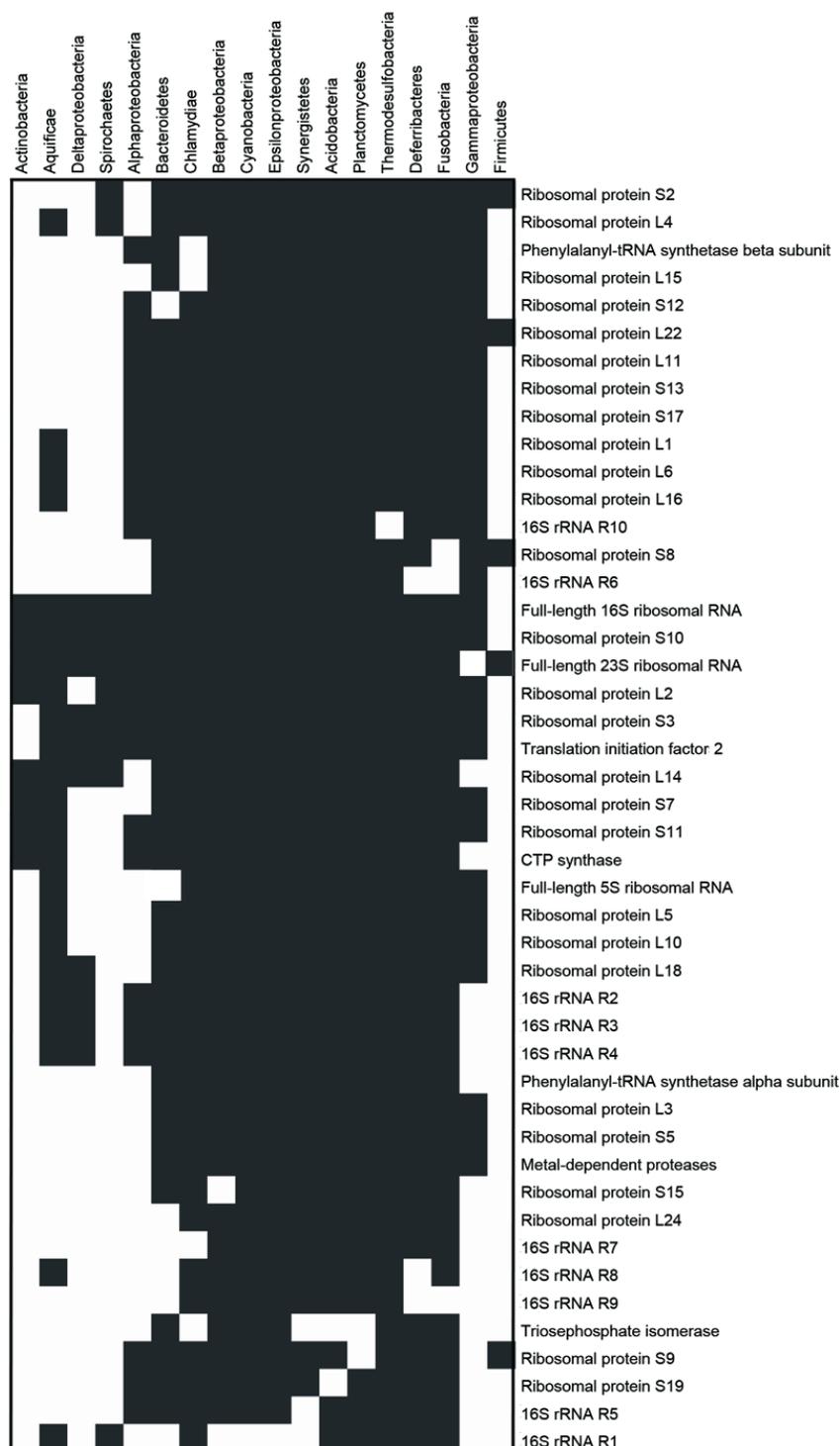


Figure 3. Congruence of bacterial phyla among phylogenetic trees. The taxa from the same phyla were examined in the phylogenetic trees reconstructed using rRNAs, 16S rRNA internal regions (R1-R10) and 33 conserved genes. If the monophyletic cluster existed in a phylogenetic tree for a phylum, a black square was depicted.

Some new phyla such as Deferribacteres and Planctomycetes did not have enough completely sequenced genomes from lower classification levels. As such, it is perhaps easier to sort the taxa in new phyla into monophyletic clades than those with representatives from different classes. Moreover, random selection of the taxa and alignment accuracy could

be the sources of the problem in taxonomic sorting of the species.

CONCLUDING REMARKS

In this study, we examined conserved genes and rRNAs in their sensitivity and efficiency in splitting bacterial species into corresponding phyla. Several r-proteins and R4-R5 internal regions in 16S rRNA may be desirable molecular markers in future work. Not all markers lead to a consistent phylogenetic topology with 16S rRNA. This reminds us existence of multiple nomenclature systems in Bacteria domain. To be cautious, we should use the current 16S rRNA-based relationships between phyla. The markers suggested in this study require further evaluation in studies of environmental communities and metagenomes, as more new phyla and unculturable bacteria were discovered.

Methods

Collection of full-length 16S rRNA genes and corresponding conserved genes: Protein sequences of all Clusters of Orthologous Groups (COGs) ^[47] were downloaded from the NCBI database (www.ncbi.nlm.nih.gov/COG/). Among the list of COGs, protein IDs of 37 highly conserved COGs (essential bacterial genes) (**Table 1**) for all the bacteria were pooled.

To obtain the bacteria with full-length 5S, 16S and 23S rRNA genes and a full set of highly conserved genes, a total of 2785 completely sequenced bacterial genomes were downloaded from the NCBI in Genbank format. In reference to the protein IDs of the COGs, 505 bacterial species contained at least 33 conserved genes. These bacterial species were sorted into corresponding phyla.

For each of the bacterial phyla, three representative genomes were selected randomly and the species from different orders were preferred. All rRNA genes (5S, 16S and 23S) were then extracted from the genomes using in-house java script. In case of presence of multiple copies, only one was retained for analysis. The list of the species was shown in **Table 1**.

The COGs and rRNA genes were aligned with Muscle ^[48], followed by manual adjustments. The 16S rRNA genes in aligned format were split between conserved regions. The start and end sites of the partition positions were referred to **Table 1**.

AHC of phylogenetic trees using geodesic distance

The aligned sequences of the conserved genes, full-length 16S rRNAs and their segregated regions were used to reconstruct Bayesian phylogenetic relationships. The best substitution model for the aligned sequences was recommended in the output of J Model Test 2.1 ^[49]. Using a model of GTR+Gamma+Invariant, a Bayesian phylogenetic inference was calculated with ten million MCMC chains using BEAST 1.8.1 ^[50]. With a Burnin setting of 2000, a consensus tree was then produced, and posterior probabilities were calculated.

The pairwise geodesic distances between the Bayesian trees were computed by using GTP geodesic algorithm ^[21]. A distance matrix was constructed as input for AHC using complete linkage model.

ACKNOWLEDGEMENTS

This study was supported by National Science Foundation of China No. 31460001 and No. 41476104. This work was also supported by the Strategic Priority Research Program of Chinese Academy of Sciences No. XDB06010201 and awards from the Sanya Institute of Deep Sea Science and Engineering (SIDSSE-201206 and SIDSSE-201305) and Hainan International Collaborative Grant (KJHZ2015-22).

REFERENCES

1. Case RJ, et al. Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. *Appl Environ Microbiol.* 2007;73:278-288.
2. Santos SR and Ochman H. Identification and phylogenetic sorting of bacterial lineages with universally conserved genes and proteins. *Environ Microbiol.* 2004;6:754-759.
3. Dodsworth JA, et al. Single-cell and metagenomic analyses indicate a fermentative and saccharolytic lifestyle for members of the OP9 lineage. *Nat Commun.* 2013;4:1854.
4. Rinke C, et al. Insights into the phylogeny and coding potential of microbial dark matter. *Nature.* 2013;499:431-437.
5. Woese CR and Fox GE. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci U S A.* 1977;74:5088-5090.
6. Fox GE, et al. The phylogeny of prokaryotes. *Science* 1980;209:457-463.

7. Bennett S. Solexa Ltd. *Pharmacogenomics* 2004;5:433-438.
8. Soares AR, et al. Next-generation sequencing of miRNAs with Roche 454 GS-FLX technology: steps for a successful application. *Methods Mol Biol* 2012;822:189-204.
9. Loman NJ, et al. Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotechnol* 2012;30:434-439.
10. Baker GC, et al. Review and re-analysis of domain-specific 16S primers. *J Microbiol Methods*. 2003;55:541-555.
11. Tringe SG and Hugenholtz P. A renaissance for the pioneering 16S rRNA gene. *Curr Opin Microbiol*. 2008;11:442-446.
12. Hamady M and Knight R. Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. *Genome Res*.2009;19:1141-1152.
13. Kim M, et al. Evaluation of different partial 16S rRNA gene sequence regions for phylogenetic analysis of microbiomes. *J Microbiol Methods*. 2011;84:81-87.
14. Klindworth A, et al. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res*. 2013;41:e1.
15. Yu Z, et al. Evaluations of different hypervariable regions of archaeal 16S rRNA genes in profiling of methanogens by Archaea-specific PCR and denaturing gradient gel electrophoresis. *Appl Environ Microbiol*. 2008;74:889-893.
16. Evans PN, et al. Methane metabolism in the archaeal phylum Bathyarchaeota revealed by genome-centric metagenomics. *Science*. 2015;350:434-438.
17. Klappenbach JA, et al. rRNA operon copy number reflects ecological strategies of bacteria. *Appl Environ Microbiol*. 2000;66:1328-1333.
18. Fogel GB, et al. Prokaryotic genome size and SSU rDNA copy number: estimation of microbial relative abundance from a mixed population. *Microb Ecol*.1999;38:93-113.
19. Wu DY, et al. Systematic identification of gene families for use as "markers" for phylogenetic and phylogeny-driven 19 ecological studies of 423 f Bacteria and Archaea and their major subgroups. *PLoS ONE*. 2013;8:0077033.
20. Billera LJ, et al. Geometry of the space of phylogenetic trees. *Adv Appl Math*. 2001;27:733-767.
21. Owen M. Computing Geodesic Distances in Tree Space. *SIAM J on Discrete Math*. 2011;25:1506-1529.
22. Wang Q, et al. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol*. 2007;73:5261-5267.
23. Yarza P, et al. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat Rev Microbiol*. 2014;12:635-645.
24. Morosyuk SV, et al. Structure and function of the conserved 690 hairpin in Escherichia coli 16 S ribosomal RNA. II. NMR solution structure. *J Mol Biol*. 2001;307:197-211.
25. Schuwirth BS, et al. Structures of the bacterial ribosome at 3.5 Å resolution. *Science*. 2005;310:827-834.
26. Schluenzen F, et al. Structure of functionally activated small ribosomal subunit at 3.3 Å resolution. *Cell*. 2000;102:615-623.
27. Van de Peer Y, et al. Database on the structure of small subunit ribosomal RNA. *Nucleic Acids Res*. 1999;27:179-183.
28. Morosyuk SV, et al. Cunningham PR. Structure and function of the conserved 690 hairpin in Escherichia coli 16 S ribosomal RNA: analysis of the stem nucleotides. *J Mol Biol*. 2000;300:113-126.
29. Morosyuk SV, et al. Structure and function of the conserved 690 hairpin in Escherichia coli 16 S ribosomal RNA. III. Functional analysis of the 690 loop. *J Mol Biol*. 2001;307:213-228.
30. Pál C, et al. Genomic function: Rate of evolution and gene dispensability. *Nature*. 2003;421:496-497.
31. Hirsh AE and Fraser HB. Protein dispensability and rate of evolution. *Nature*. 2001;411:1046-1049.
32. Koonin EV. Systemic determinants of gene evolution and function. *Mol Syst Biol* 2005;1:2005.
33. Rippa V, et al. The ribosomal protein L2 interacts with the RNA polymerase alpha subunit and acts as a transcription modulator in Escherichia coli. *J Bacteriol*. 2010;192:1882-1889.
34. Takyar S, et al. mRNA helicase activity of the ribosome. *Cell*. 2005;120:49-58.
35. Hosaka H, et al. Ribosomal protein S7: a new RNA-binding motif with structural similarities to a DNA architectural factor. *Structure*. 1997;5:1199-1208.

36. Sylvers LA, et al. Photochemical cross-linking of the anticodon loop of yeast tRNA(Phe) to 30S-subunit protein S7 at the ribosomal A and P sites. *Biochimie*. 1992;74:381-389.
37. Döring T, et al. The decoding region of 16S RNA; a cross-linking study of the ribosomal A, P and E sites using tRNA derivatized at position 32 in the anticodon loop. *EMBO J*. 1994;13:2677-2685.
38. Dean D, et al. Identification of ribosomal protein S7 as a repressor of translation within the str operon of *E. coli*. *Cell*. 1981;24:413-419.
39. Gao H, et al. Study of the structural dynamics of the *E. coli* 70S ribosome using real-space refinement. *Cell*. 2003;113:789-801.
40. Vila-Sanjurjo A, et al. Modulation of 16S rRNA function by ribosomal protein S12. *Biochim Biophys Acta*. 2007;1769:462-471.
41. Das A, et al. Evidence that ribosomal protein S10 itself is a cellular component necessary for transcription antitermination by phage? N protein. *Proc Natl Acad Sci*. 1985;4070-4074.
42. Worbs M, et al. Crystal structure of ribosomal protein L4 shows RNA-binding sites for ribosome incorporation and feedback control of the S10 operon. *EMBO J*. 2000;19:807-818.
43. Nierhaus KH. The assembly of prokaryotic ribosomes. *Biochimie*. 1991;73:739-755.
44. Dinman JD. 5S rRNA: Structure and Function from Head to Toe. *Int J Biomed Sci*. 2005;1:2-7.
45. Ban N, et al. The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science*. 2000;289:905-920.
46. Duval M, et al. *Escherichia coli* ribosomal protein S1 unfolds structured mRNAs onto the ribosome for active translation initiation. *PLoS Biol*. 2013;11.
47. Tatusov RL, et al. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res*. 2000;28:33-36.
48. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 2004;32:1792-1797.
49. Darriba D, et al. jModelTest 2: more models, new heuristics and parallel computing. *Nat Methods*. 2012;9:772.
50. Drummond AJ and Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol*. 2007;7:214.