Genomics 93 (2009) 525–533

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Estimation of bacterial species phylogeny through oligonucleotide frequency distances

Mahoko Takahashi [a,b], Kirill Kryukov [b], Naruya Saitou [a,b,*]

[a] Department of Genetics, School of Life Science, Graduate University for Advanced Studies, Mishima 411-8540, Japan
[b] Division of Population Genetics, National Institute of Genetics, Mishima 411-8540, Japan

ARTICLE INFO

Article history:
Received 4 October 2008
Accepted 30 January 2009
Available online 12 February 2009

Keywords:
Distance between bacteria
GC content
Genome comparison
Identification

ABSTRACT

Classification of bacteria is mainly based on sequence comparisons of certain homologous genes such as 16S rRNA. Recently there are challenges to classify bacteria using oligonucleotide frequency pattern of nonhomologous sequences. However, the evolutionary significance of oligonucleotides longer than tetra-nucleotide is not studied well. We performed phylogenetic analysis by using the Euclidean distances calculated from the di to deca-nucleotide frequencies in bacterial genomes, and compared these oligonucleotide frequency-based tree topologies with those for 16S rRNA gene and concatenated seven genes. When oligonucleotide frequency-based trees were constructed for bacterial species with similar GC content, their topologies at genus and family level were congruent with those based on homologous genes. Our results suggest that oligonucleotide frequency is useful not only for classification of bacteria, but also for estimation of their phylogenetic relationships for closely related species.

Introduction

Phylogenetic relationship of organisms is usually estimated by comparing homologous genes [e.g., [1]]. The 16S rRNA gene is most typically used for classification of bacteria [e.g., [2,3]. However, the classification method is still a matter of controversy. The accuracy of phylogenetic analysis using a single gene depends on the selected gene that may not truly reflect the whole evolutionary history of organisms in question. Moreover, the 16S rRNA classification has been useful only for taxa above the rank of species so that the 16S rRNA analysis does not have high ability for taxa lower than species. Another way is to use complete genomes to classify bacterial species. Many methods based on whole genome comparisons have been proposed; e.g., comparison of the presence and absence of orthologous or family genes, or overall gene content [4–7], presence of conserved insertions and deletions [8,9], or conservation of gene order [10–13].

As an alternative method to whole bacterial genome comparison, many studies have shown that di-nucleotide frequencies within DNA sequences exhibit species-specific signals [14–19]. Species-specific signals for oligomers up to a length of four nucleotides have also been detected [20,21]. Phylogenetic analysis using trees based on tetra-nucleotide frequencies demonstrate a level of congruence with trees based on single genes, such as 16S rRNA [22]. These studies have been revealing the effectiveness of tetra-nucleotide frequency. However, phylogenetic analysis using oligonucleotide frequencies longer than tetra-nucleotides are not studied well. In addition, most of the analyses using oligonucleotide frequency focused on the classification of bacteria, not on the estimation of phylogenetic relationships.

The previous studies raised new questions. Does longer oligo-nucleotide usage pattern have more power to classify bacterial species because closely related species share similar characteristics in their genomes? At which level in taxonomy oligonucleotide frequency-based trees can reconstruct phylogenetic relationships? We therefore conducted phylogenetic analyses by constructing trees using up to deca-nucleotide sequence (10 bp) frequencies. The aim of this study is to investigate at which level of taxonomic rank the oligonucleotide frequency is the most effective in estimating the phylogenetic relationship of bacteria.

Results

Comparison between oligonucleotide frequency-based and homologous gene-based trees

We first constructed two homologous sequence-based trees that are widely accepted; concatenated universal seven-gene trees (Fig. 1A) and 16S rRNA gene tree (Fig. S1). Because their tree topologies were very similar for bootstrap values higher than 90%, a more reliable concatenated seven-gene tree was used for further comparison. We then constructed phylogenetic trees based on mono to deca-nucleotide (1 bp to 10 bp) frequency, and the representative one based on tri-nucleotide frequency is shown in Fig. 1B. Other oligonucleotide frequency-based trees are shown in Figs. S2(A)–(I). The relationship

* Corresponding author. Division of Population Genetics, National Institute of Genetics, Mishima, 411-8540, Japan. Fax: +81 559 81 6789.
E-mail address: saitounr@lab.nig.ac.jp (N. Saitou).
URL: http://sayer.lab.nig.ac.jp/~saitou/ (N. Saitou).

0888-7543/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2009.01.009

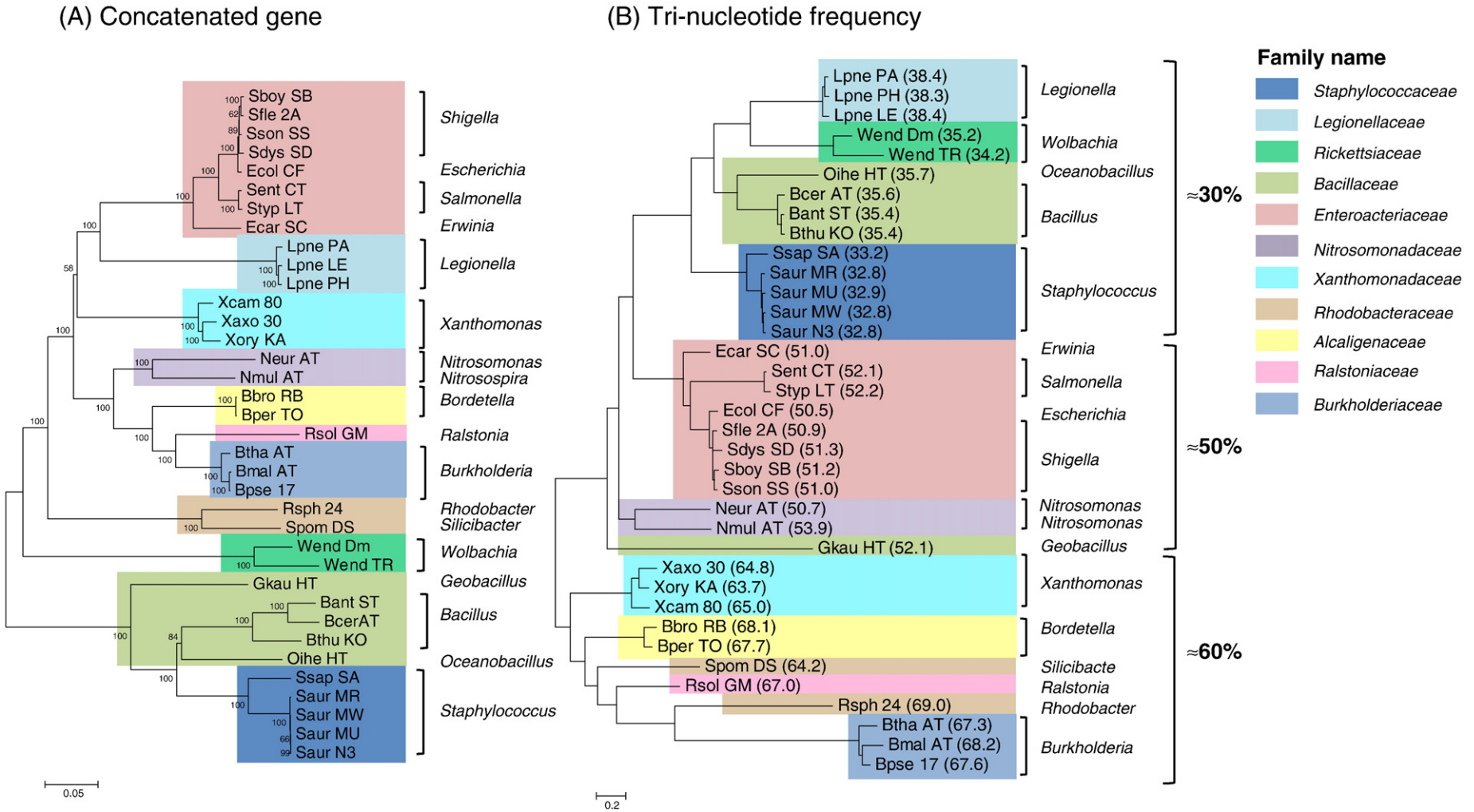

Fig. 1. Phylogenetic trees of 36 bacterial species. Bacterial species in the same colored box belong to the same family. Family names and their colors are shown in the right side of the figures in italics. Genus names are on the right side of each tree. (A) Concatenated gene-based tree of 36 bacterial species. The tree was built based on an alignment of the concatenated sequences of 7 genes (16S rRNA, 23S rRNA, gyrB, pyrH, recA, rpoA and rpoD). Bootstrap percentage values, based on 500 resamplings, are shown at the internal nodes. (B) Tri-nucleotide frequency-based tree of 36 bacterial species. Tri-nucleotide is the shortest length of oligonucleotide word to show the small topological distance to concatenated gene or 16S rRNA gene trees. GC contents of bacterial genome are shown next to each species name. Percentage levels (≈30%, ≈50% and ≈60%) of GC content are on the right most side of the tree.

of bacterial species based on mono-nucleotide frequency (Fig. S2(A)) is essentially the same as the linear representation of GC contents of each bacterial genome (Fig. S3). This is because the tree was constructed by using pairwise distances, and these reflect difference of GC contents. There are three groups in Fig. S2(A) and Fig. S3, but there exists no clear clustering pattern within each group. However, the trees created by using di to deca-nucleotide frequencies (Fig. 1B and S2(B)–(I)) could group closely related bacterial species into the same cluster. Although di to hepta-nucleotide frequency-based trees were very similar, trees based on octa- to deca-nucleotide frequencies (Figs. S2(G)–(I)) showed less ability to group closely related species into a cluster.

We compared the topologies of oligonucleotide frequency-based trees with those of the concatenated gene-based tree to examine their ability to reconstruct the phylogenetic relationship among bacterial species. We then calculated the topological distance whose value is zero when trees have identical topology, and the distance is large when trees do not have identical topology. We found that tri-nucleotide frequency-based tree had the smallest topological distance and mono-nucleotide frequency tree had the largest topological distances to homologous gene-based trees. The range of topological distances between the concatenated gene-based tree and all oligonucleotide frequency-based trees showed large values; 23–59 and 21–55, respectively (data not shown).

However, species that belong to the same family or genus were grouped into the same cluster and their phylogenetic relationships were highly congruent with those of homologous gene-based tree. In contrast, the topologies higher than family level were quite different from those of concatenated gene-based tree. These large topological distances were from the incongruence of phylogenetic relationships higher than family level in oligonucleotide-based trees. These results indicate that oligonucleotide frequency is useful for differentiating bacterial species into, at least, closely related bacterial species under family level.

Analysis of congruence among trees for similar GC content genomes

We found that the oligonucleotide frequency-based tree grouped closely related bacterial species into the same cluster. Further examination of tree topology in Fig. 1B shows that species with similar GC content were clustered together irrespective of their phylogenetic relationships. It implies that the GC content difference has great influence on bacterial classification using oligonucleotide frequency. We therefore constructed the oligonucleotide frequency-based trees using bacterial species genomes with similar GC content to reduce the influence of GC content difference. Thirty six bacterial species were separated into three GC content groups (32–38%, 50–53% and 63–69%; see Table S1), and the phylogenetic relationship of bacterial species of each species was estimated using oligonucleotide frequency (mono- to deca-nucleotides). We then calculated the topological distance (d_T) between these trees and homologous gene-based trees.

The 32–38% GC content group: Figs. 2A–C show the 16S rRNA tree, the concatenated seven-gene tree, and the tri-nucleotide frequency-based tree, respectively. The tri-nucleotide frequency tree is shown as a representative tree because its oligonucleotide length is the shortest

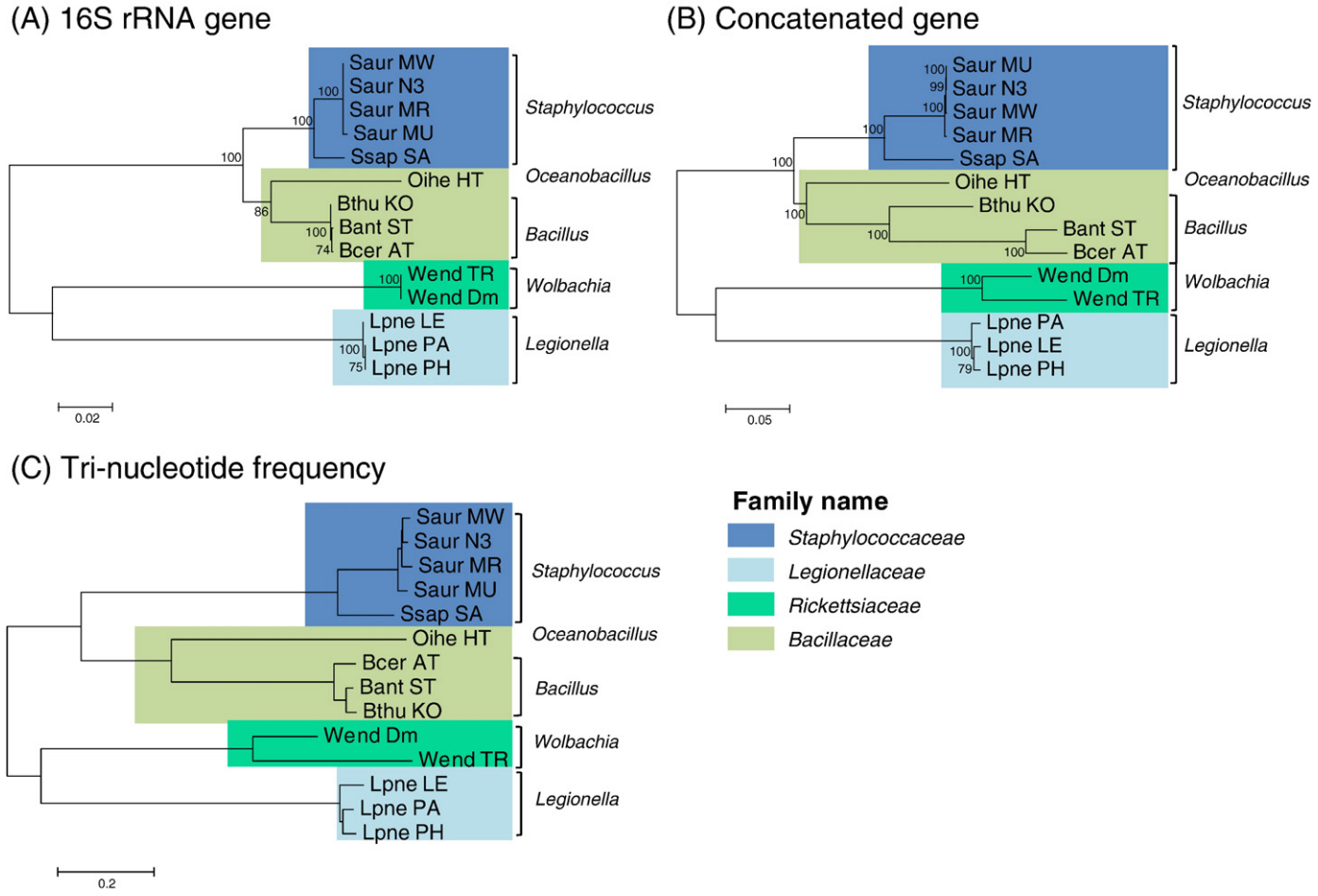

Fig. 2. Phylogenetic trees of bacterial species with GC content 32–38%. Bacterial species in the same colored box belong to the same family. Family names are shown on the right side of the figures in italics. Genus names are on the right side of each tree. (A) 16S rRNA gene-based tree. (B) Concatenated gene-based tree. Bootstrap percentage values (500 resamplings) are shown at the internal branches. (C) Tri-nucleotide frequency-based tree. Tri-nucleotide is the shortest length of oligonucleotide word to reconstruct a tree that shows the best match topology to (A) and/or (B) trees. GC contents of bacterial genome are shown next to each species name.

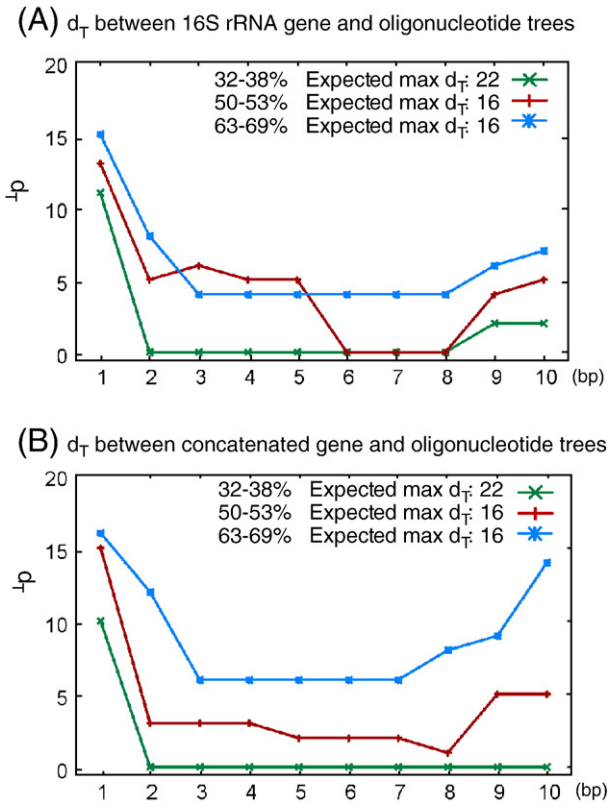

Fig. 3. Topological distance between oligonucleotide frequency and gene-based trees. The branch that have the bootstrap value equal or higher than 90% is used for calculation of topological distances (d_T) between oligonucleotide frequency-based tree and 16S rRNA tree/Concatenated gene tree. *X*-axis and *Y*-axis indicate oligonucleotide length, and d_T, respectively. Expected maximum d_T is the distance when the topologies of all internal branches are different. GC content 32–38% (green), GC content 50–53% (red), GC content 63–69% (blue). (A) d_T between 16S rRNA gene and oligonucleotide tree. (B) d_T between Concatenated gene and oligonucleotide tree.

among trees showing the best match to the topology of the homologous gene-based trees (Figs. 3A and B). Trees based on other oligonucleotide frequencies are shown in Figs. S4(A)–(I). There were distantly related species classified in different orders according to the current classification by 16S rRNA in this GC content group. In contrast, some genera contained two to five closely related species. Most oligonucleotide frequency-based trees were very congruent with homologous gene-based tree (Figs. 3A and B). Tri- to octa-nucleotide frequency-based trees had identical topology ($d_T = 0$) to the 16S rRNA tree, and tetra to hexa-oligonucleotide trees had almost the same topology ($d_T = 2$) to the concatenated gene-based tree (Fig. 2C and S4 (D)–(G)). We made further analysis to examine at which level of taxonomic rank the oligonucleotide frequency trees reconstruct phylogenetic relationships. We used both 16S rRNA and concatenated seven gene-based trees for the analysis at family level. However, we did not use the 16S rRNA tree for the analysis at genus level because the 16S rRNA tree did not show the ability to differentiate closely related species in the genus *Staphylococcus* and *Bacillus*. At family level, di- to deca-nucleotide frequency-based trees had the same topology ($d_{TF} = 0$), and at the genus level (Fig. 4), tetra- to nona-nucleotide frequency-based trees had almost the same topology ($d_{TG} = 0$-2) as that of homologous gene-based trees (Fig. 5A).

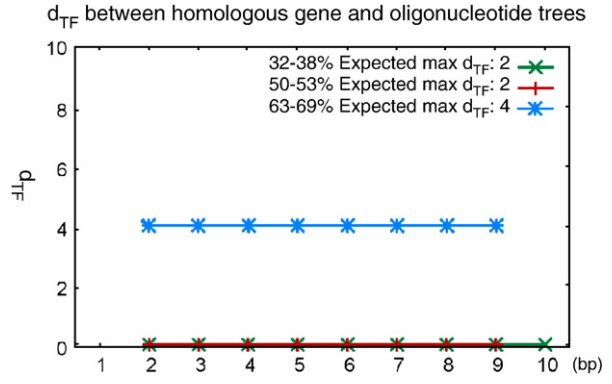

Fig. 4. Topological distance at family level (d_{TF}). Topological distances at family level (d_{TF}) between oligonucleotide frequency-based tree and 16S rRNA based/concatenated gene-based tree are the same, and only d_{TF} between Concatenated gene and oligonucleotide tree is shown. d_{TF} is not calculated when all species (3 or more species) that belong to the same genus or family do not make a cluster. Topological distances within the cluster at genus level (d_{TG}) are ignored. *X*-axis and *Y*-axis indicate oligonucleotide length, and d_{TG}, respectively. Expected maximum d_{TF} is the distance when the topologies of all internal branches are different. GC content 32–38% (green), GC content 50–53% (red), GC content 63–69% (blue).

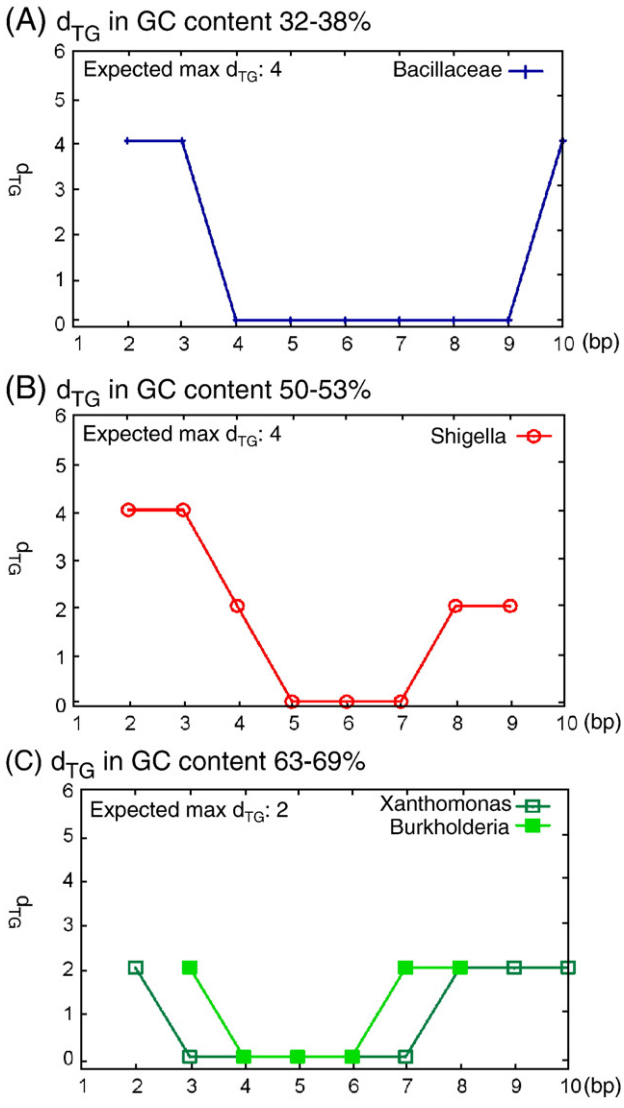

Fig. 5. Topological distance within genus level cluster (d_{TG}). Topological distances within the cluster at genus level (d_{TG}) between di- to deca-nucleotide frequency-based tree and concatenated gene-based tree are calculated. d_{TG} is not calculated when all species (3 or more species) that belong to the same genus do not make a cluster. d_{TG} is not calculated when the cluster has a bootstrap value less than 90%. When the cluster contains two or more subspecies, they are counted as one species, and d_{TG} is not calculated. *X*-axis and *Y*-axis indicate oligonucleotide length, and d_{TG}, respectively. Expected maximum d_{TG} is the distance when the topologies of all internal branches are different. (A) d_{TG} in GC content 32–38%. (B) d_{TG} in GC ontent 50–53%. (C) d_{TG} in GC content 63–69%.

The 50–53% GC content group: Figs. 6A–C show the 16S rRNA tree, concatenated seven-gene tree, and octa-nucleotide frequency-based tree, respectively. The octa-nucleotide frequency tree is shown as a representative tree because its oligonucleotide length is the shortest among trees showing the best match to the topology of the homologous gene-based trees (Figs. 3A and B). Other oligonucleotide frequency-based trees are shown in Figs. S5(A)–(I). This GC content group also contained species that belong to different orders. In family *Enterobacteriaceae*, there were four genera, two of which contain closely related species. The number of genera in this family was the largest in all families used in this study. Oligonucleotide frequency-based trees in this GC content group also showed congruence with the homologous gene-based trees. Hexa to nona-nucleotide frequency-based trees showed identical topology to the 16S rRNA tree (Figs. S5(F)–(H)). Octa and nona-nucleotide frequency-based trees were highly congruent with the homologous gene trees ($d_T = 1$, Figs. 3A and B and 6C).

The main incongruence between oligonucleotide frequency-based and homologous gene-based trees was observed in the *Enterobacteriaceae* cluster. In tri to hepta-nucleotide frequency-based trees, *E. coli*, belonging to genus *Escherichia*, is grouped with the *Shigella* cluster. Species in family *Enterobacteriaceae* share most of their genes [23–25], and this similarity might cause difficulty in the estimation of phylogenetic relationships among the species in *Enterobacteriaceae*. Indeed, bootstrap values at some branches in the *Enterobacteriaceae* cluster were not 100% even in the concatenated seven-gene tree. For these reasons, classification of many closely related species needs longer oligonucleotides, and octa and nona-nucleotide frequencies enable the best reconstruction of phylogenetic relationship.

At the family and genus level, this GC content group showed the similar result to that of GC 32-38% group. Di to nona-nucleotide frequency-based trees had the identical topology as that of homologous gene-based trees in the analysis at family level (Fig. 4). At genus level, penta to hepta-nucleotide frequency-based trees showed the identical topology to that of concatenated seven-gene tree (Fig. 5B).

The 63–69% GC content group: Figs. 7A–C show the 16S rRNA tree, the concatenated seven-gene tree, and the tetra-nucleotide frequency-based tree, respectively. The tetra-nucleotide frequency tree is shown as a representative tree because its oligonucleotide length is the shortest among trees showing the best match to the topology of the homologous gene-based trees (Figs. 3A and B). Other oligonucleotide frequency-based trees are shown in Figs. S6(A)–(I). This GC content group contains species that belong to different orders. Three genera contained closely related species. Oligonucleotide frequency-based trees in this group showed the lowest congruence with homologous gene-based trees among the three GC content groups ($d_T = 4$). Oligonucleotide frequency-based trees in this group did not reconstruct the identical phylogenetic relationships at family level (Fig. 4). However, in tetra- to octa-nucleotide frequency-based trees, closely related species at genus level were grouped as estimated by homologous gene-based trees. In addition, they reconstructed the same phylogenetic relationships within clusters at genus level in tetra to hexa-nucleotide frequency trees ($d_T = 0$, Fig. 5C).

The incongruence between oligonucleotide frequency-based trees and homologous gene-based trees were observed in the topology among genera *Bordetella*, *Ralstonia*, and *Burkholderia*. In homologous gene-based trees, these three genera formed a cluster with 100% bootstrap probability (Figs. 7A and B). In oligonucleotide frequency-

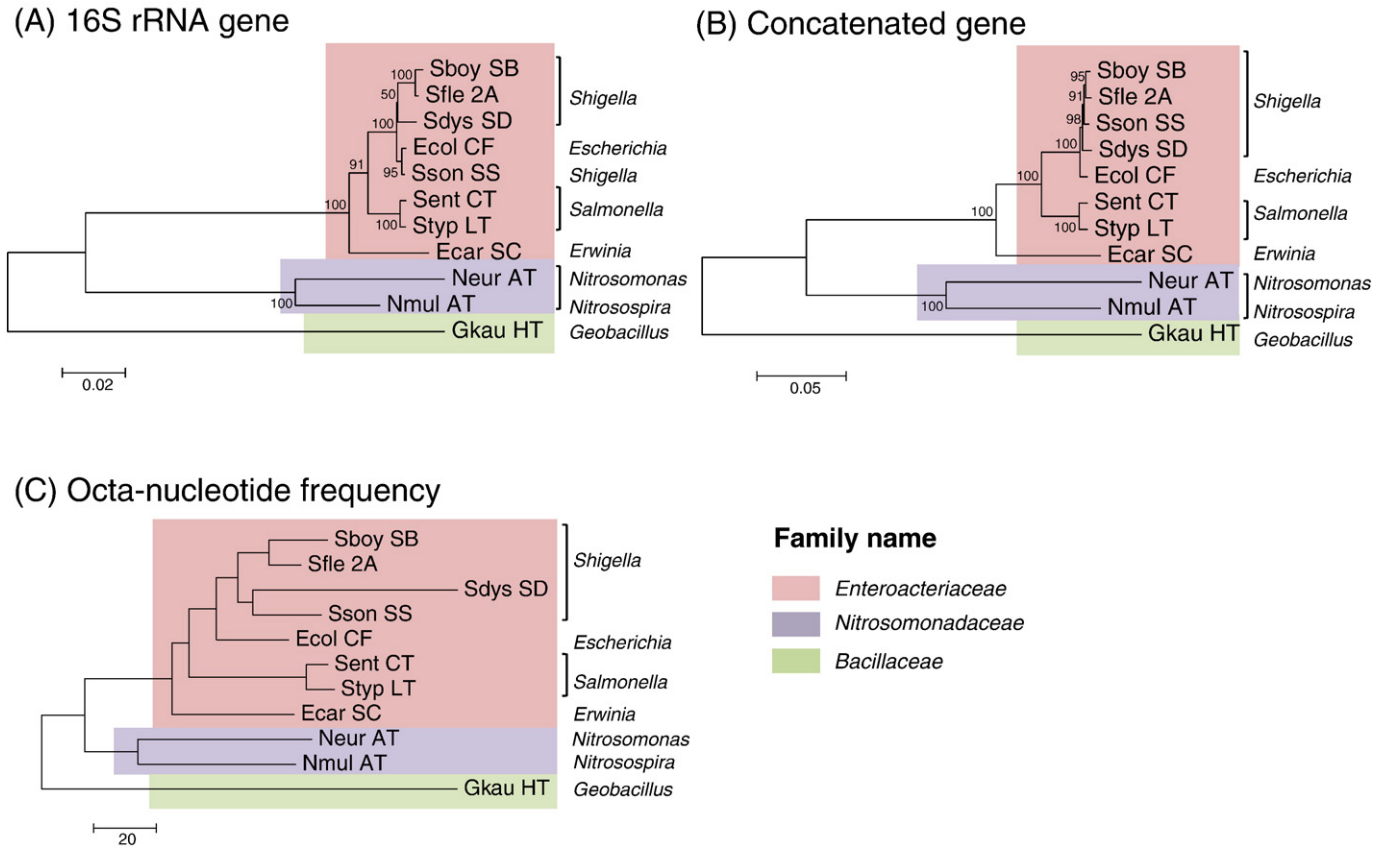

Fig. 6. Phylogenetic trees of bacterial species with GC content 50–53%. Bacterial species in the same colored box belong to the same family. Family names and their colors are shown in the right side of the figures in italics. Genus names are on the right side of each tree. (A) 16S rRNA gene-based tree. (B) Concatenated gene-based tree. Bootstrap percentage values (500 resamplings) are shown at the internal nodes. (C) Octa-nucleotide frequency-based tree. Octa-nucleotide is the shortest length of oligonucleotide word to reconstruct a tree that shows the best match topology to (A) and/or (B) trees. GC contents of bacterial genome are shown next to each species name.

based trees (e.g., Fig. 7C), however, genus *Burkholderia* showed more distant relationship to the other two genera (*Bordetella* and *Ralstonia*), and genus *Xanthomonas* clustered with these two genera instead of genus *Burkholderia*. One reason for this incongruence might be the difference of genomic structure in species belonging to genus *Burkholderia*. The genomes of *Burkholderia* species consist of two circular DNAs and have an unusually high number of simple sequence repeats. Their densities in *Burkholderia* are over 2-fold compared to that of the other bacteria with similar GC content. The *Burkholderia* genome also contains a high number of insertion sequences dispersed throughout the genome [26–28]. These characteristics of the *Burkholderia* genome may affect the composition of its oligonucleotide frequency, and result in longer distances between *Burkholderia* and other four families longer than those of homologous genes.

There is another incongruence related to *Xanthomonas* and *Ralstonia*. They are taxonomically classified into different classes, *Gammaproteobacteria* and *Betaproteobacteria*, respectively. In spite of the long evolutionary distance between them, *Ralstonia* and *Xanthomonas* formed a cluster in the oligonucleotide frequency-based tree. The genera *Xanthomonas* and *Ralstonia* are plant pathogens. Yet, other clustered species except for Spom_DS and Rsph_24 are animal pathogens. It is known that pathogenicity islands often differ in GC content from the genomic DNA [29–31]. Although we used bacterial whole genomes in this analysis, the amount of the sequence related to pathogenicity is usually much less than that of whole genome, and may not affect the overall phylogenetic relationship. However, plant or animal pathogen bacterial species genomes contain significant proportion of pathogenicity islands, and thus their phylogenetic relationship based on oligonucleotide frequency might have been twisted from their original, pathogenicity island-free genome sequences.

Effective length of oligonucleotides for classifying bacterial species

Previous studies have not examined the possibility of reconstruction of phylogenetic relationships by using oligonucleotide frequency. Our analyses of oligonucleotide trees using bacterial species with similar GC content (32–38%, 50–53% and 63–69%) demonstrated that tetra to octa-nucleotide frequency-based trees are powerful to reconstruct almost the same topologies at family level and at genus level as homologous gene-based trees. The important difference between phylogenetic relationships at family and genus levels is the range of effective length in oligonucleotides. For the phylogenetic relationship at family level, di to octa-nucleotide frequency-based trees in GC 32–38% and GC 50–53% groups showed identical topology to that of homologous gene-based trees (Fig. 4). In contrast, no oligonucleotide frequency-based trees reconstructed the identical phylogenetic relationship to that of homologous gene-based trees in GC 63–69%. However, tetra to octa-nucleotide frequency-based trees for this GC content group showed a small d_{TF} ($d_{TF}=4$). These results indicate that for all GC content groups, tetra to octa-nucleotide frequencies are appropriate to reconstruct phylogenetic relationships at family level.

Each GC group had the different range of effective oligonucleotide lengths for estimation of the phylogenetic relationship within genus level cluster. Tri to nona-nucleotide frequency-based trees in GC 32–38% group, penta to hepta-nucleotide-based trees in GC 50–53%, and tetra to hepta-nucleotide frequency-based trees in GC 63–69% showed almost the same topology as homologous gene-based trees (see Figs. 5A, B and C). This result suggests that the range of penta and hexa-nucleotide frequencies is powerful to estimate phylogenetic relationship at genus level. Taken together, oligonucleotide frequency analysis is best used for estimation of phylogenetic relationship at genus level. For the phylogenetic analysis over the

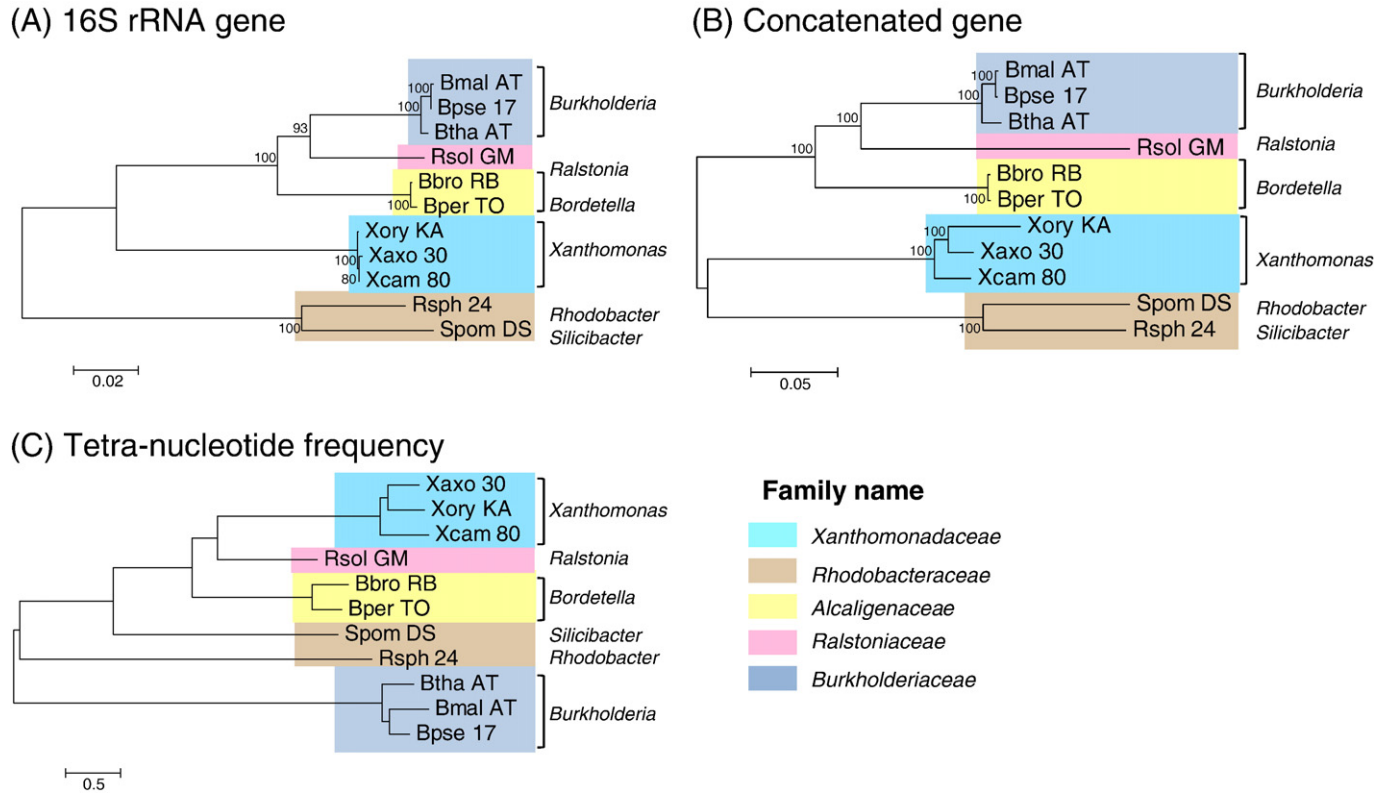

Fig. 7. Phylogenetic trees of bacterial species with GC content 63–69%. Colored boxes indicate the same family level. Family names and their colors are shown in the right side of the figures in italics. Genus names are on the right side of each tree. (A) 16S rRNA gene-based tree. (B) Concatenated gene-based tree. Bootstrap percentage values (500 resamplings) are shown at the internal nodes. (C) Tetra-nucleotide frequency-based tree. Tetra-nucleotide is the shortest length of oligonucleotide word to reconstruct a tree that shows the best match topology to (A) and (B) trees. GC contents of bacterial genome are shown next to each species name.

genus levels, these effective lengths may be different depending on the level, family, order and class in question. For example, the effective length estimation of entire tree topology in GC 50–53% group (see Figs. 3A and B) is hexa and octa-nucleotides. It is likely that when one family contains many genera, the reconstruction of the phylogenetic relationship needs longer oligonucleotide words. These results imply that at higher than genus level, the range of effective length in oligonucleotides to estimate phylogenetic relationships varies and depends on the number of species and their relatedness that are analysed.

Analysis of phylogenetic networks

We constructed phylogenetic networks to depict incompatible signals that are not shown in phylogenetic trees. In the phylogenetic network, parallel edges represent the splits estimated from the data because the split networks provide only an implicit representation of evolutionary history [32]. The presence of parallelograms in splits trees is a hallmark of recombination or horizontal gene transfer [33]. The phylogenetic networks based on concatenated seven genes for GC 63–69% group, GC 32–38% group and GC 50–53% group are shown in Figs. 8A, S7(A), and S7(B), respectively. The splits of these phylogenetic networks showed clear signals and little evidence for horizontal gene transfer or other forms of reticulate evolution. It suggests that topologies of phylogenetic trees we used in this study are plausible.

The phylogenetic networks based on oligonucleotide frequencies, however, revealed clear reticulations. Fig. 8B shows a representative phylogenetic network based on tetra-nucleotide frequency in GC 63–69%. Other phylogenetic networks in GC 32–38% group, GC 50–53% group and GC 63–69% group are shown in Figs. S7(A)–(C), respectively. In GC content 63–69% group, most networks showed some reticulations. Although split "Rsol_GM, *Xanthomonas*, *Bordetella*" vs. "Rsph_24, Spom_DS, *Burkholderia*" was very short, the corresponding branch length in the phylogenetic tree (Fig. 7C) was relatively long. The edge corresponding to split "Rsph_24, Spom_DS" vs. all other species was rather long. However, it was not displayed in the phylogenetic tree (Fig. 7C) since it is incompatible with the split grouping Rsph_24 and *Burkholderia*. This result suggests that oligonucleotide frequencies have information that is useful to classify bacterial species at family level in GC content 63–69% group.

Discussion

We showed that, at family level, tetra to octa-nucleotide frequency-based trees constructed the phylogenetic relationship most similar to 16S rRNA and concatenated gene trees. For the phylogenetic relationships at genus level, we demonstrated that trees based on penta and hexa-nucleotides estimated identical phylogenetic relationships as those of homologous gene trees. In addition, phylogenetic network analysis revealed that oligonucleotide frequency-based trees that showed less ability to estimate phylogenetic relationships still contained information to reconstruct the phylogenetic relationship. Previous studies on oligonucleotide frequency suggested that especially tetra-nucleotide frequency had species-specific characteristics and could be applied for bacterial classification [e.g., [21]]. Pride et al. [22] suggested the possibility of estimating phylogenetic relationship thorough tetra-oligonucleotide frequency. Our results are in agreement with these findings. Yet, the previous studies did not mention about the difference among the levels of phylogenetic relationship reconstructed by using frequencies from variation of oligonucleotide length in details.

Pride et al. [22] used tetra-nucleotide frequency to estimate the bacterial phylogenetic relationship. However, it is not immediately obvious whether this tetra-nucleotide frequency is the most effective. A counter example is reconstruction of phylogenetic relationship among species in family *Enterobacteriaceae*. The tetra-nucleotide frequency was not enough to reconstruct phylogenetic relationship at genus level, because many related genera such as *Shigella* were used. At family level, however, the tetra-nucleotide frequency-based tree reconstructed the phylogenetic relationship as shown in Fig. 4. Tetra-nucleotide word may be suitable for classification, but not always for estimation of phylogenetic relationships.

Interestingly, the longest oligonucleotide word frequency (e.g. deca-nucleotide) had less power to estimate the topology at family level than shorter oligonucleotides In general, closely related species share similar sequences in their genomes. There can be many longer oligonucleotide words that are shared among genomes of closely related bacteria. However, our result suggests up to nona-nucleotide is useful for bacterial classification (not always for estimation of phylogenetic relationship). Probably oligonucleotide frequency distances among species are too large to estimate the correct topology when longer oligonucleotide word frequencies are used. It means that species-specific signatures are lost because of this noise. If weighting for similar

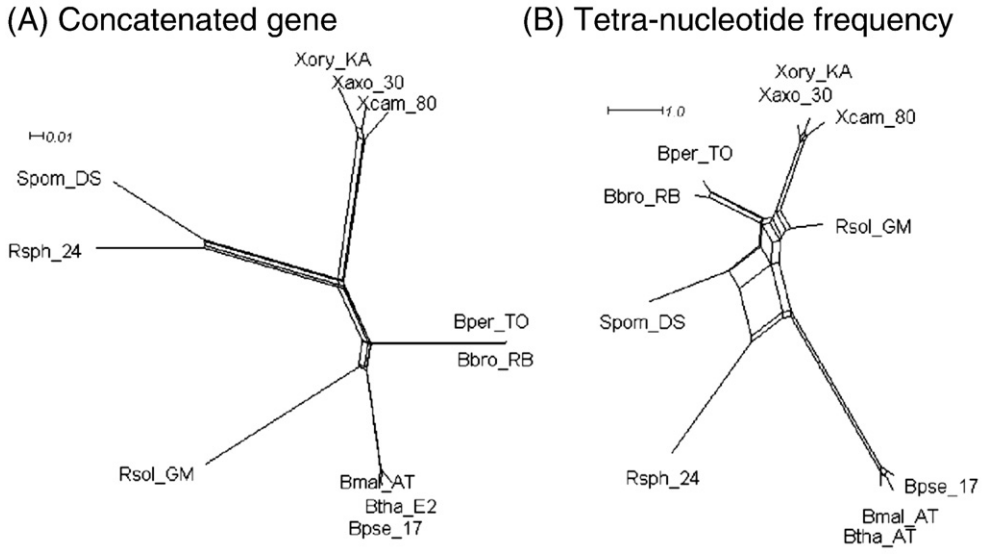

Fig. 8. Phylogenetic network of bacterial species with GC content 63–69%. (A) Phylogenetic network for Concatenated gene (B) Phylogenetic network for tetra-nucleotide frequency. Tetra-nucleotide is the shortest length of oligonucleotide word to reconstruct a tree that shows the best match topology to homologous gene tree.

deca-nucleotides is applied, the ability for bacterial classification in deca-nucleotide might be enhanced.

One of the important findings of our study is that oligonucleotide frequency-based trees reconstruct phylogenetic relationships among closely related species. In addition to the characteristics of oligonucleotide frequency, given that the 16S rRNA classification has been useful for taxa above the rank of species, oligonucleotide frequency-based trees may cover the shortcoming of the 16S rRNA classification. Furthermore, since oligonucleotide frequency shows the characteristic of the genome as a whole, and does not depend on multiple alignments of homologous sequences, it may be applicable on the genomes that have not been annotated, and there will be no need to obtain homologous gene sequences and their multiple alignments.

Whole genome sequences were used to obtain oligonucleotide frequencies in this study. When metagenome sequences are compared, however, only partial genomic sequences may be available. We thus used such partial sequences (3 kb to ≈7 kb or 2% of genomic sequences), and estimated the minimum sequence length that has enough signals to reconstruct the phylogenetic relationship. Oligonucleotide frequencies from short genome sequences had the ability to differentiate the species that belong to, at least, the same family (see Supplementary Table S2). This result thus supports the possibility of application of oligonucleotide frequencies on the metagenome analysis [34,35]. Our results, however, did not succeed to cluster closely related bacterial species using sequences shorter than 1% of the whole genome. This is incompatible with the result of Abe et al. [21] who claimed that only 1 kb sequences were enough to classify many bacterial species. They used SOM [36] that requires huge computation time, and this difference on method may affect the results.

Our results did not rule out the problem related to GC content when estimating phylogenetic relationships among distantly related species at family level or higher. Some genera with large difference in GC content were not clustered even if these belong to the same family; for example, genera *Geobacillus* and *Bacillus* belonging to family *Bacillaceae* are not clustered in Fig. 1B. In any case, if GC content is drastically different between two species, their nucleotide sequences are also different, and these can accumulate only after relatively long evolutionary time.

Materials and methods

Nucleotide sequences

We used 36 bacterial complete genomes for this analysis. Sequence data were obtained from DDBJ (http://www.ddbj.nig.ac.jp/anoftp-e.html). These 36 species were grouped into three different GC content groups, 32–38%, 50–53% and 64–69%, which contain 14, 11, and 11 species, respectively.

Oligonucleotide frequency: Program Count Motifs (http://kirill-kryukov.com/study/tools/count-motifs/; Kirill Kryukov, unpublished) was used to calculate the oligonucleotide (from mono to deca-nucleotide) frequencies. This program counts the information about all sequence fragments of up to deca-nucleotide long. It also compares the observed motif frequencies with the expected frequencies. The expected frequencies are computed by two methods — using GC content and di-nucleotide composition of the original sequence dataset. We compensated all frequencies of oligonucleotides by dividing the observed number of occurrences by the expected number for calculation of Euclidean distance.

Constructing oligonucleotide frequency-based trees: Euclidean distances (Dt) based on oligonucleotide frequency differences were calculated as

$$Dt = \sqrt{\sum_{k=1}^{N} |F1(W) - F2(W)|^2},$$

where N is the number of oligonucleotide type calculated as 4^W (W is the length of nucleotide word), $F1(W)$ and $F2(W)$ represent the frequency for each type of the W length oligonucleotides for organisms 1 and 2, respectively. Each distance was calculated from di to hexa-nucleotide frequencies. The phylogenies were constructed separately for the three groups classified in terms of GC content. The phylogeny was constructed by using the neighbor-joining method [37] and distance matrices using MEGA 4 [38].

Topological distance

Topological distance (d_T) between the homologous gene-based tree and the oligonucleotide tree as measured by the "partition metric" [39]:

$$d_T = 2[\mathrm{Min}(q_1, q_2) - P] + |q_1 - q_2|,$$

where q_1 and q_2 are the total numbers of partitions (interior branches) for trees 1 and 2 that are compared, respectively, and P is the number of partitions that are shared with the two trees. We ignored the branches that have bootstrap values less than 90%. When calculating topological distance at family level (d_{TF}), we ignored the topology within the small clusters at genus level. In contrast, when calculating topological distance within the clusters at genus level (d_{TG}), we focused on each topology within the genus level clusters and one outgroup cluster, and ignored the topology over the family level.

Homologous gene-based phylogenetic tree

Nucleotide sequences for 16S rRNA genes were downloaded from DDBJ, while the remaining six homologous genes (23S rRNA, gyrB, pyrH, recA, rpoA, and rpoD) were downloaded from NCBI (http://www.ncbi.nlm.nih.gov/) and these sequences are concatenated into one sequence in each species. These concatenated sequences were aligned by using ClustalW [40], and their neighbor-joining trees with bootstrap values were created with MEGA4.

Bacterial taxonomy

We followed lineage information given in each entry of DDBJ/EMBL/GenBank International Sequence Database.

Constructing phylogenetic network trees

The phylogenetic network was constructed by using the neighbor-net method [41] from distance matrices using SplitTree4 [32]. Software was downloaded from http://www.splitstree.org.

Acknowledgments

We are grateful to Drs. Kiyoshi Ezawa and Stefan Grünewald for their useful discussion and valuable advice. This study was supported by a grant in aid for scientific studies from Ministry of Education, Science, Sport, and Culture, Japan, to N.S.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ygeno.2009.01.009.

References

[1] M. Nei, S. Kumar, Molecular Evolution and Phylogenetics, Oxford University Press, New York, 2000.
[2] C.R. Woese, G.E. Fox, Phylogenetic structure of the prokaryotic domain: the primary kingdoms, Proc. Natl. Acad. Sci. U. S. A. 74 (1977) 5088–5090.
[3] C.R. Woese, O. Kandler, M.L. Wheelis, Towards a natural system of organisms: proposal for the domains archaea, bacteria, and eukarya, Proc. Natl. Acad. Sci. U. S. A. 87 (1990) 4576–4579.

[4] B. Snel, P. Bork, M.A. Huynen, Genome phylogeny based on gene content, Nat. Genet. 21 (1999) 108–110.
[5] F. Tekaia, A. Lazcano, B. Dujon, The genomic tree as revealed from whole proteome comparisons, Genome Res. 9 (1999) 550–557.
[6] S.T. Fitz-Gibbon, C.H. House, Whole genome-based phylogenetic analysis of free-living microorganisms, Nucleic Acids Res. 27 (1999) 4218–4222.
[7] A.K. Bansal, T.E. Meyer, Evolutionary analysis by whole-genome comparisons, J. Bacteriol. 184 (2002) 2260–2272.
[8] R.S. Gupta, Protein phylogenies and signature sequences: a reappraisal of evolutionary relationships among archaebacteria, eubacteria, and eukaryotes, Microbiol. Mol. Biol. Rev. 62 (1998) 1435–1491.
[9] R.S. Gupta, The branching order and phylogenetic placement of species from completed bacterial genomes, based on conserved indels found in various proteins, Int. Microbiol. 4 (2001) 187–202.
[10] T. Dandekar, B. Snel, M. Huynen, P. Bork, Conservation of gene order: a fingerprint of proteins that physically interact, Trends Biochem. Sci. 23 (1998) 324–328.
[11] M.A. Huynen, P. Bork, Measuring genome evolution, Proc. Natl. Acad. Sci. U. S. A. 95 (1998) 5849–5856.
[12] T. Kunisawa, Gene arrangements and phylogeny in the class Proteobacteria, J. Theor. Biol. 213 (2001) 9–19.
[13] M. Suyama, P. Bork, Evolution of prokaryotic gene order: genome rearrangements in closely related species, Trends Genet. 17 (2001) 10–13.
[14] S. Karlin, I. Ladunga, Comparisons of eukaryotic genomic sequences, Proc. Natl. Acad. Sci. U. S. A. 91 (1994) 12832–12836.
[15] S. Karlin, A.M. Campbell, J. Mrázek, Comparative DNA analysis across diverse genomes, Annu. Rev. Genet. 32 (1998) 185–225.
[16] S. Karlin, C. Burge, Dinucleotide relative abundance extremes: a genomic signature, Trends Genet. 11 (1995) 283–290.
[17] S. Karlin, Global dinucleotide signatures and analysis of genomic heterogeneity, Curr. Opin. Microbiol. 1 (1998) 598–610.
[18] H. Nakashima, K. Nishikawa, T. Ooi, Differences in dinucleotide frequencies of human, yeast, and *Escherichia coli* genes, DNA Res. 4 (1997) 185–192.
[19] H. Nakashima, M. Ota, K. Nishikawa, T. Ooi, Genes from nine genomes are separated into their organisms in the dinucleotide composition space, DNA Res. 5 (1998) 251–259.
[20] S. Karlin, J. Mrazek, A. Campbell, Compositional biases of bacterial genomes and evolutionary implications, J. Bacteriol. 179 (1997) 3899–3913.
[21] T. Abe, et al., Informatics for unveiling hidden genome signatures, Genome Res. 13 (2003) 693–702.
[22] D.T. Pride, R.J. Meinersmann, T.M. Wassenaar, M.J. Blaser, Evolutionary implications of microbial genome tetranucleotide frequency biases, Genome Res. 13 (2003) 145–155.
[23] F. Yang, et al., Genome dynamics and diversity of *Shigella* species, the etiologic agents of bacillary dysentery, Nucleic Acids Res. 33 (2005) 6445–6458.
[24] J. Wei, et al., Complete genome sequence and comparative genomics of *Shigella flexneri* serotype 2a strain 2457T, Infect. Immun. 71 (2003) 2775–2786.
[25] J. Parkhill, et al., Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica*, Nat. Genet. 35 (2003) 32–40.
[26] W.C. Nierman, et al., Daugherty, Structural flexibility in the *Burkholderia mallei* genome, Proc. Natl. Acad. Sci. U. S. A. 101 (2004) 14246–14251.
[27] M.T.G. Holden, et al., Genomic plasticity of the causative agent of melioidosis, *Burkholderia pseudomallei*, Proc. Natl. Acad. Sci. U. S. A. 101 (2004) 14240–14245.
[28] H.S. Kim, et al., Bacterial genome adaptation to niches: divergence of the potential virulence genes in three *Burkholderia* species of different survival strategies, BMC Genomics 6 (2005) 174.
[29] P. Lio, M. Vannucci, Finding pathogenicity islands and gene transfer events in genome data, Bioinformatics 16 (2000) 932–940.
[30] S. Karlin, Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes, Trends Microbiol. 9 (2001) 335–343.
[31] W. Hsiao, I. Wan, S.J. Jones, F.S. Brinkman, IslandPath: aiding detection of genomic islands in prokaryotes, Bioinformatics 19 (2003) 418–420.
[32] D.H. Huson, D. Bryant, Application of phylogenetic networks in evolutionary studies, Mol. Biol. Evol. 23 (2006) 254–267.
[33] E.C. Holmes, M. Worobey, A. Rambaut, Phylogenetic evidence for recombination in dengue virus, Mol. Biol. Evol. 16 (1999) 405–409.
[34] H. Teeling, A. Meyerdierks, M. Bauer, R. Amann, F.O. Glöckner, Application of tetranucleotide frequencies for the assignment of genomic fragments, Environ. Microbiol. 6 (2004) 938–947.
[35] T. Abe, H. Sugawara, M. Kinouchi, S. Kanaya, T. Ikemura, Novel phylogenetic studies of genomic sequence fragments derived from uncultured microbe mixtures in environmental and clinical samples, DNA Res. 12 (2005) 281–290.
[36] T. Kohonen, Self-organized formation of topologically correct feature maps, Biol. Cybern. 43 (1982) 59–69.
[37] N. Saitou, M. Nei, The neighbor-joining method: a new method for reconstructing phylogenetic trees, Mol. Biol. Evol. 4 (1987) 406–425.
[38] K. Tamura, J. Dudley, M. Nei, S. Kumar, MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0, Mol. Biol. Evol. 24 (2007) 1596–1599.
[39] A. Rzhetsky, M. Nei, A simple method for estimating and testing minimum-evolution trees, Mol. Biol. Evol. 9 (1992) 945–967.
[40] J.D. Thompson, T.J. Gibson, F. Plewniak, F. Jeanmougin, D.G. Higgins, The Clustal_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools, Nucleic Acids Res. 25 (1997) 4876–4882.
[41] D. Bryant, V. Moulton, Neighbor-net: an agglomerative method for the construction of phylogenetic networks, Mol. Biol. Evol. 21 (2004) 255–265.