Gene 429 (2009) 59–64

Contents lists available at ScienceDirect

Gene

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

Phylogenetic construction of 17 bacterial phyla by new method and carefully selected orthologs

Tokumasa Horiike [a], Daisuke Miyata [b], Kazuo Hamada [c], Satoshi Saruhashi [d], Takao Shinozawa [d], Sudhir Kumar [e], Ranajit Chakraborty [f], Tomoyoshi Komiyama [g], Yoshio Tateno [h],*

[a] *Department of Integrated Genetics, National Institute of Genetics, Mishima, Shizuoka, 411-8540, Japan*
[b] *Department of Commerce and Economics, Chiba University of Commerce, Ichikawa, Chiba 272-8512, Japan*
[c] *Radiance Ware Co., IOC Honjo-Waseda, Waseda Research Park, Honjo Saitama, 367-0035, Japan*
[d] *Department of Integrative Bioscience and Biomedical Engineering, Graduate School of Science and Engineering, Waseda University, Waseda Research Park, IOC Honjo-Waseda, Honjo, Saitama, 367-0035, Japan*
[e] *Center for Evolutionary Functional Genomics, The Biodesign Institute, Arizona State University, Tempe, Arizona 85287-5301, USA*
[f] *Center for Genome Information, Department of Environmental Health, University of Cincinnati, Cincinnati, Ohio 45267-0056, USA*
[g] *Department of Clinical Pharmacology, Tokai University School of Medicine, Kanagawa, 259-1193, Japan*
[h] *Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Mishima, Shizuoka, 411-8540, Japan*

ARTICLE INFO

Article history:
Received 30 August 2008
Received in revised form 3 October 2008
Accepted 6 October 2008
Available online 21 October 2008

Received by T. Gojobori

Keywords:
Bacterial phylogeny
Concatenated tree
Out-paralog
Ortholog database
Thermophilic eubacteria
Tree evaluation

ABSTRACT

Here, we constructed a phylogenetic tree of 17 bacterial phyla covering eubacteria and archaea by using a new method and 102 carefully selected orthologs from their genomes. One of the serious disturbing factors in phylogeny construction is the existence of out-paralogs that cannot easily be found out and discarded. In our method, out-paralogs are detected and removed by constructing a phylogenetic tree of the genes in question and examining the clustered genes in the tree. We also developed a method for comparing two tree topologies or shapes, ComTree. Applying ComTree to the constructed tree we computed the relative number of orthologs that support a node of the tree. This number is called the Positive Ortholog Ratio (POR), which is conceptually and methodologically different from the frequently used bootstrap value. Our study concretely shows drawbacks of the bootstrap test. Our result of bacterial phylogeny analysis is consistent with previous ones showing that hyperthermophilic bacteria such as Thermotogae and Aquificae diverged earlier than the others in the eubacterial phylogeny studied. It is noted that our results are consistent whether thermophilic archaea or mesophilic archaea is employed for determining the root of the tree. The earliest divergence of hyperthermophilic eubacteria is supported by genes involved in fundamental metabolic processes such as glycolysis, nucleotide and amino acid syntheses.

1. Introduction

The construction of the correct phylogenetic tree remains a key issue in evolution. Generally speaking, a phylogenetic tree, once correctly constructed, would be used as a contour map in biology. However, many lingering problems exist with the construction of the correct tree. One of them is that trees constructed by using single genes are often inconsistent to one another. This inconsistency is frequently observed in the construction of bacterial phylogeny (Brown and Doolittle, 1997). In theory this problem is resolved or alleviated by incorporating as many orthologous genes, proteins, domains or genome fragments as possible (Tateno et al., 1982), and several methods have been developed along this line. They include those based on statistical properties of genomes (Qi et al., 2004; Grishin et al., 2000), gene (or domain) contents (Snel et al., 1999; Tekaia et al., 1999; Wolf et al., 2001, 2002; House and Fitz-Gibbon, 2002; Korbel et al., 2002; Horiike et al., 2004; Dutilh et al., 2004; Fukami-Kobayashi et al., 2007), gene orders (Wolf et al., 2001; Korbel et al., 2002), and concatenated orthologs (Wolf et al., 2001; Brown et al., 2001; Brochier et al., 2002; Daubin et al., 2001; Henz et al., 2004; Gadagkar and Kumar, 2005; Ciccarelli et al., 2006).

Unfortunately, none of those methods is perfect in that they tend to yield inaccurate relationships particularly for distantly related species due to disturbing factors such as horizontal gene transfer, loss of out-paralog, and/or unusual base compositions (Fitch, 2000; Delsuc et al., 2005; Snel et al., 2005). Nevertheless, two of them are worth mentioning, because they have been used more frequently than the others. One is to construct a consensus tree (supertree) that is made up with consistent parts of individual trees each of which is constructed

Abbreviations: COG, Clusters of Orthologous Groups of proteins; HTG, Horizontally Transferred Gene; MBGD, MicroBial Genome Database; NJ, Neighbor-Joining method of tree construction; OTU, Operational Taxonomic Unit; POR, Positive Ortholog Ratio; UPGMA, Unweighted Pair-Group Method with Arithmetic mean.

* Corresponding author. Tel.: +81 55 981 6857; fax: +81 55 981 6858.
E-mail address: ytateno@genes.nig.ac.jp (Y. Tateno).

0378-1119/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2008.10.006

from a different data source (Bininda-Emonds et al., 2002; Daubin et al., 2002). The other is the alignment concatenated tree, which is obtained by using the concatenated multiple alignment of amino acid or nucleotide sequences (Brown et al., 2001). It is reminded that the two types still suffer from the loss of reliability due at least to loss of out-paralog and horizontally transferred genes (HTGs).

Therefore, we first examined and refined the extant ortholog databases of bacterial genomes to exclude as many HTGs and out-paralogs as possible. We then constructed a concatenated tree and a supertree of bacterial phyla by using the refined database. Furthermore, we developed a method for evaluating the nodes of a constructed tree, Positive Ortholog Ratio (POR), and applied it to our concatenated tree. It is noted that our method is conceptually and methodologically different from the bootstrap test (Felsenstein, 1985). In our bacterial phylogeny construction and evaluation we particularly focused on the phylogenetic position of thermophilic eubacteria that is directly related with the problem of the earliest eubacterial cluster or the earliest species that appeared on earth.

2. Materials and methods

2.1. Preparation of the ortholog dataset

We examined and selected the microbial genome database (MBGD, http://mbgd.genome.ad.jp/) (Uchiyama, 2003) as our primary data resource, because it was found to contain fewer paralogs than the clusters of orthologous groups of proteins (COG, http://www.ncbi.nlm.nih.gov/COG/) (Tatusov et al., 1997). The orthologous proteins (orthologs) in MBGD are obtained by constructing a phylogenetic tree of the possible orthologs in question. However, there is one problem in MBGD that the tree is constructed by UPGMA (Michener and Sokal, 1957), which is known to be less reliable than other methods such as the neighbor-joining (NJ) method (Saitou and Nei, 1987; Saitou and Imanishi, 1989). The use of UPGMA would thus wrongly sort out out-paralogs that seriously disturb the construction of the correct tree topology. Therefore, we instead applied NJ method to the source data of MBGD for sorting out out-paralogs and obtain more reliable orthologs than the original ones. The procedure of identifying and excluding out-paralogs in the present study is as follows.

Let us suppose that the unrooted tree in Fig. 1 was constructed by NJ method for eight possible orthologs from eubacteria (e) and archaea (a). Let us then assume that the tree is divided at the arrow or root into two clusters, one containing (e1, e2, e3, e4, a1, a2) and the other containing (e5, a3). In this case (e5, a3) is excluded, and (e1, e2, e3, e4, a1, a2) is kept in the refined database. If one of the two clusters contains only eubacteria that share at least one species in the other, we compare the number of species of eubacteria in both clusters and choose the one with a larger number and the archaea discarding the eubacterial cluster with a smaller number of species. We discard the case in which the eubacteria in the two clusters do not share a species, because we cannot place the root in the tree. The same is true for the case where the one cluster contains only archaea. Of course, there are many cases in which a constructed tree contains only one cluster, suggesting that no out-paralogs are included in the tree.

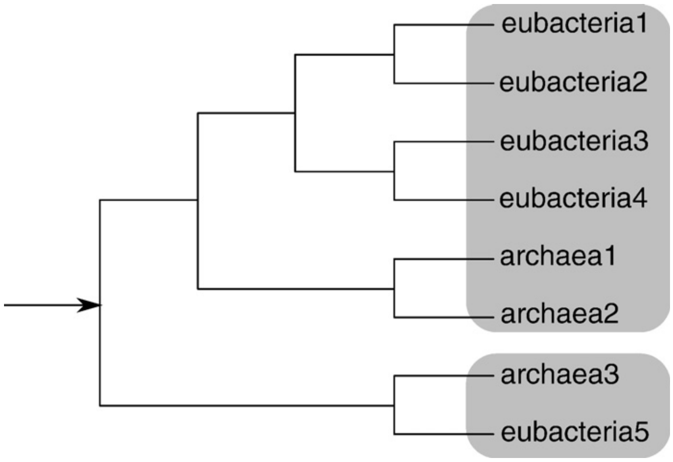

Fig. 1. Phylogenetic tree for finding out-paralogs. Eight species in the tree are divided at the arrow into two clusters. In this case the smaller cluster is excluded from our study.

2.2. Phylogeny construction

Let us assume that we have N ortholog groups with m_i ($i = 1, 2, 3, \ldots, N$) species each, where m_i is the number of species for the i-th group, and max $m_i = M$. The orthologs of every N cluster are aligned by MAFFT (Katoh et al., 2002) and edited by Gblocks (Castresana, 2000). The latter includes a procedure to retain gap(s) in the aligned sequences when half or more of M species have amino acid(s) at the corresponding residue(s), or discard them from all species when less than half have amino acid(s). The aligned amino acid sequences are used for constructing a phylogenetic tree as follows.

An initial tree is constructed for one of the N groups by NJ method in BioNJ (Saitou and Nei, 1987, Gascuel, 1997), and it is used in the maximum likelihood method in PhyML (Felsenstein, 1981, Guindon and Gascuel, 2003) as the initial tree to produce the final tree. (The parameters used for PhyML are: “JTT” for substitution model, “estimated” for proportion of invariable sites, “estimated” for gamma distribution parameters, “4” for the number of substitution categories, “yes” to optimize topology, “yes” to optimize branch length and rate parameter, and “BIONJ” for starting tree(s).) This is repeated N times for N ortholog groups and N individual trees are constructed. We next concatenate the aligned sequences of N ortholog groups for M species. If a species does not have all N orthologs, the missing orthologs are filled with gaps. As a result, we obtain M concatenated sequences that are then used for constructing a concatenated tree by the same procedure as above.

2.3. Evaluation of the constructed phylogenetic tree

The concatenated tree is expected to be more accurate than each of the N individual trees, as the former is more resistant to the existence of false orthologs and the variation of the number of amino acid substitutions than the latter. The question then is how much accuracy the former has. To evaluate the accuracy of the concatenated tree, we compare the concatenated tree to every N individual tree by using the method we newly developed. The new method, ComTree, allows us to compare the topology of a pair of trees even for a pair having different numbers of OTUs and no roots.

ComTree works as follows. In Fig. 2, one concatenated and three individual trees are given. The question here is how to evaluate the internal node shown in the closed circle in the concatenated tree by comparing it to the three individual trees. There are 3 branches extending from the node which reach out to three sets of OTUs, (A, B, C), (D, E, F) and (G, H), respectively. Now, every individual tree is examined if it has a node with three branches extending each to one of the OTUs in the three OTU sets of the concatenated tree. Let us denote the number of such individual trees as Np. If such a tree further satisfies the condition that the three OTU sets are not intermingled with those in the concatenated tree, the tree is qualified to support the node in the concatenated tree. Let us also denote the number of the qualified individual trees as Nq. Then, we define the Positive Ortholog Ratio (POR) as,

$$\mathrm{POR} = \mathrm{Nq}/\mathrm{Np}. \qquad (1)$$

POR indicates the relative number of orthologs that support a node of the concatenated tree. Note that there are N–Np individual trees

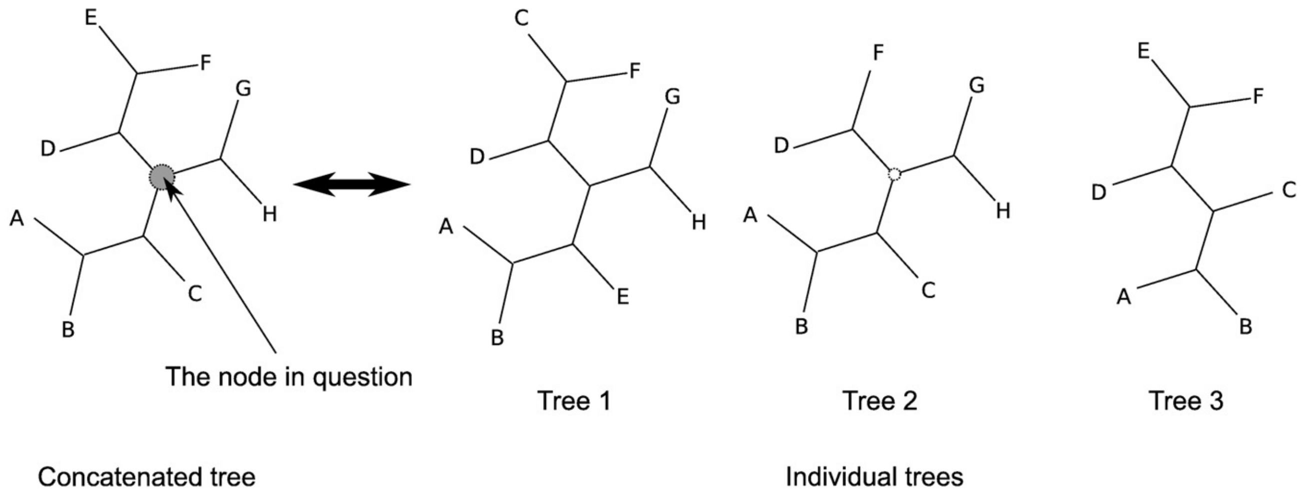

Fig. 2. Definition of POR. A concatenated tree and three individual trees are given. The node in circle in the concatenated tree has three branches, branch 1 extending to OTUs, (A, B, C), branch 2 extending to OTUs (D, E, F) and branch 3 extending to OTUs (G, F). The node in open circle in tree 2 has three branches each of which respectively extends to at least one OTU in each of the three OTU sets of the concatenated tree. The three OTU sets in tree 2 are not intermingled with those in the concatenated tree, while those in tree 1 are. Tree 3 has neither OTU H nor OTU G.

that do not contribute to POR. It is noted that ComTree is independent from the number of OTUs. In the present case, each of trees 1and 2 has an OTU in one of the three OTU sets of the concatenated tree. Tree 2 further satisfies the condition mentioned above, while tree 1 does not. Tree 3 does not have G or H in one OTU set of the concatenated tree, and does not contribute to POR. Therefore, Np = 2, Nq = 1 and N–Np = 1. The value of POR in the present case is thus 0.5, which means that 50% of the orthologs used for the POR computation support the node in question.

3. Results and discussion

3.1. Construction of molecular phylogeny of 17 bacterial phyla

According to MBGD, there are 17 (= *M*) bacterial phyla that are supposed to cover most, if not all, known bacterial species. The 17 phyla are composed of 3 archaea and 14 eubacteria. To gain a large-scale view of bacterial phylogeny, we chose one species that was representative in each of the 17 phyla. Table 1 lists the selected species in the 17 phyla. By the procedures mentioned in the materials and methods we first obtained 227 (= *N*) possible ortholog clusters each of which had eight species or more including at least one archaea species. After the removal of out-paralogs by the procedure mentioned also in the materials and methods, *N* was reduced to 102. The presence or absence of an ortholog in each species and its functional relevance are given in Supplementary Data S1. For the 102 clusters we first obtained a set of 102 individual trees. We then concatenated the multiple alignments of the 102 groups for each of the 17 species. The total length of each concatenated sequence was 33,250 residues including gaps. By using the concatenated sequences we finally constructed a tree with POR values as shown in Fig. 3a.

As the root of the concatenated tree in Fig. 3a is located in the lineage between archaea and eubacteria, the eubacterial cluster of the tree demonstrates that Thermotogae (tma) diverged first, Aquificae (aae) diverged second, then one cluster including gram positive bacteria and another including proteobacteria diverged one after another. While the POR values for the clusters of the archaea are high (0.75, 0.91), those for the eubacterial clusters are not except perhaps for the thermophilic clusters (Fig. 3a). In particular, no individual tree supports the cluster of Chlorobium (cte) and Bacteroidetes (bfr), because no individual tree belonged to the Nq group in this case.

To examine the validity of the concatenated tree further, we also constructed the supertree by applying the Most Similar Supertree (dfit) method with the default settings (the Clann version 3, Creevey and McInerney, 2005) to the 102 individual trees. In this version dfit method starts with the choice of the guide or initial supertree. The initial supertree is then compared to each of the 102 individual trees and a similarity value between them is computed. It is noted that the comparison in dfit method is different from ComTree in that while the former deals with the whole tree topology, the latter focuses on the individual nodes of the trees in comparison. The total of the 102 similarity values is the score of the initial tree. Next, the initial tree is modified to be the second supertree by the nearest neighbor interchange method, and the same procedure is repeated for the second tree to produce the second score. The whole procedure is repeated to produce the next score until the score no longer increases. The tree with the maximum score is the final supertree that is given in Fig. 3b for the present case.

The comparison between the concatenated tree and supertree reveals three points. First, the archaea clusters are consistent between the both trees. The clusters have a large POR value on each node in the concatenated tree. Secondly, Aquificae, Thermotogae and Chloroflexi are distantly related with other eubacteria in both trees. Finally, the other clusters with the low POR values are not consistent with those in the supertree, suggesting that those clusters should be studied further using more orthologs.

Table 1
List of species chosen from 17 bacterial phyla

	Abbreviation	Phylum	Species
1	cdi	Actinobacteria	*Corynebacterium diphtheriae* NCTC13129
2	aae	Aquificae	*Aquifex aeolicus* VF5
3	bfr	Bacteroidetes	*Bacteroides fragilis* YCH46
4	cmu	Chlamydiae	*Chlamydia muridarum* Nigg MoPn
5	cte	Chlorobium	*Chlorobium tepidum* TLS
6	det	Chloroflexi	*Dehalococcoides ethenogenes* 195
7	syc	Cyanobacteria	*Synechococcus elongatus* PCC 6301
8	dra	Deinococcus	*Deinococcus radiodurans* R1
9	bsu	Firmicutes	*Bacillus subtilis* 168
10	fnu	Fusobacteria	*Fusobacterium nucleatum* ATCC 25586
11	rba	Planctomycetes	*Rhodopirellula baltica* SH 1
12	eco	Proteobacteria	*Escherichia coli* K12 MG1655
13	lic	Spirochaetes	*Leptospira interrogans serovar* Lai str. 56601
14	tma	Thermotogae	*Thermotoga maritima* MSB8
15	ape	Crenarchaeota	*Aeropyrum pernix* K1
16	afu	Euryarchaeota	*Archaeoglobus fulgidus* DSM 4304
17	neq	Nanoarchaeota	*Nanoarchaeum equitans* Kin4-M

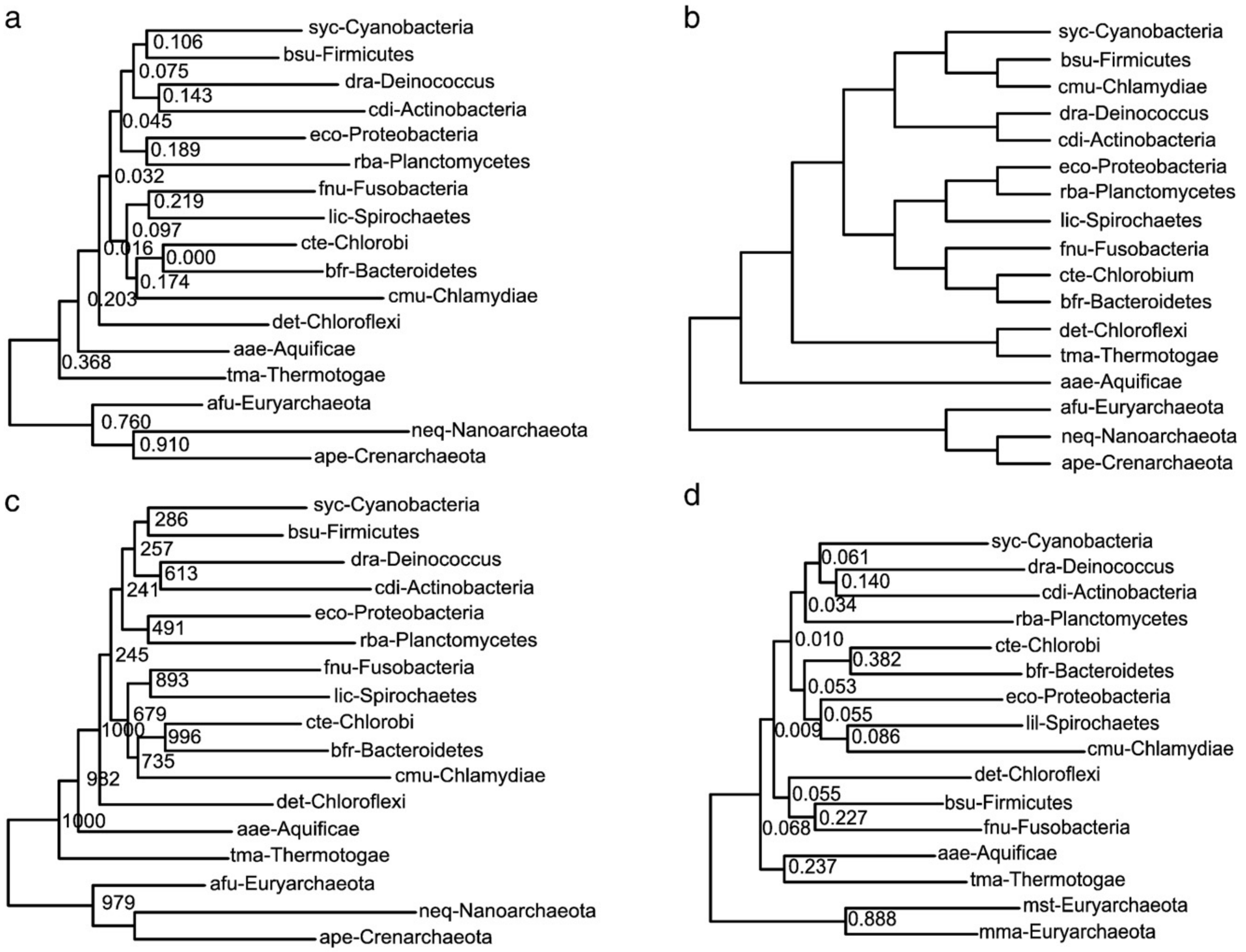

Fig. 3. Concatenated trees and a supertree for the 17 bacterial phyla. (a) A concatenated tree with thermophilic archaea as the out-group. The number given on each node is a POR value. (b) A supertree with thermophilic archaea as the out-group. (c) A concatenated tree with thermophilic archaea as the out-group. The number given on each branch is a bootstrap value from 1000 samplings. (d) A concatenated tree with mesophilic archaea as the out-group and POR values.

3.2. Difference between the bootstrap and POR evaluations

When one constructs a phylogenetic tree, one usually obtains the bootstrap values on the branches of the tree as a regular practice. Therefore, it is worthwhile to compare the bootstrap with the POR values in our constructed trees. Before the comparison let us point out a notable difference between the two methods. While node X in Fig. 4 separates a given OTU set (A, B, C, D, E, F) into three groups, (A, B), (C, D) and (E, F) in POR, the corresponding branch Y in the bootstrap test splits the same OTU set into two groups, (A, B, C, D) and (E, F). The bootstrap test ignores the further classification of (A, B, C, D), suggesting that POR examines the tree topology more accurately than the bootstrap test.

A bootstrap value for each branch of the concatenated tree was computed by conducting 1000 trials, as shown in Fig. 3c. The comparison of the trees in Figs. 3a and c reveals the fundamental discrepancy between the POR and bootstrap values. There are two branches with a bootstrap value of 100% in the tree in Fig. 3c. Therefore, one would think that the divergences or separations by these branches are correct or reliable. However, the tree in Fig. 3a shows that the POR values on the corresponding nodes are 0.016 and 0.368. The value 0.016 indicates that almost no ortholog supports the node in question. For other clusters in the trees the discrepancy between the bootstrap and POR values is also evident. The cluster of Chlorobium and Bacteroidetes with POR=0 mentioned above is supported by the bootstrap value of 99.6%. These results warn that the bootstrap value is not a reliable indicator for evaluating a constructed tree. The previous study has also shown that the bootstrap test for concatenated sequence tree is not always reliable (Gadagkar and Kumar, 2005).

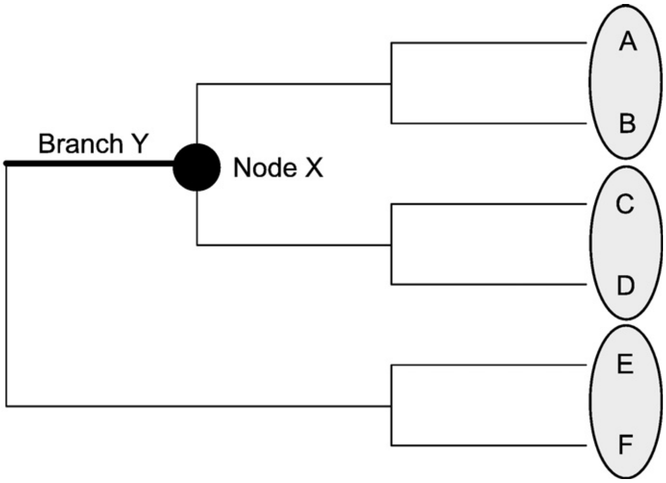

Fig. 4. Tree showing a node for POR and a branch for bootstrap test. Node X is the subject in POR, while branch Y is the subject in the bootstrap test.

As is well known the bootstrap test is based only on one dataset of orthologs that is the source of repeated sampling. Therefore, the sampled sets are not mutually independent, which means that the bootstrap value has no statistical meaning. One cannot seriously argue about what 100% of the bootstrap value really means, not mention to a value smaller than that. In POR different individual trees are constructed for different orthologs. More gravely, a sampled DNA or an amino acid sequence in the bootstrap test is totally artificial and biologically meaningless. One cannot construct a biologically meaningful tree by using biologically meaningless data. We admit that the bootstrap test is a resort to testing the accuracy of a constructed tree in the age of gene. It is not really appropriate in the age of genome and ultra-high speed sequencing.

3.3. Phylogenetic position of thermophilic eubacteria

One of the key issues in bacterial phylogeny perhaps is the phylogenetic position of thermophilic eubacteria. There are two conflicting views on it. One asserts that they are the earliest species in eubacteria (Woese, 1987) and the other opposes to that (Korbel et al., 2002). Our concatenated tree in Fig. 3a is in agreement with the former view showing that thermophilic eubacteria such as Thermotogae (tma) and Aquificae (aae) diverged first in the eubacterial cluster. Actually, the earliest divergence of Thermotogae (tma) and Aquificae (aae) is supported by 27 (=Nq) different orthologs given in Supplementary Data S2. One may think, however, that the number of the supporting orthologs is small compared with the total number of 102 (=*N*). This is because 36 orthologs belong to the Np group, and the remaining 66 orthologs belong to the *N*–Np group, which does not contribute to POR (see Materials and Methods). The supporting orthologs are those involved in basic metabolic processes such as glycolysis, nucleic acid synthesis and amino acid synthesis. They originated in ancient evolutionary time, perhaps in the era of the pre-prokaryote, which spans from the origin of life to the divergence of the first prokaryote (Kumada et al., 1993). The 17 bacterial phyla did not yet diverge in this era.

On the contrary to our study, certain researchers (Woese, 1987; Yarza et al., 2008) used only *16SrRNA* gene for their studies of bacterial phylogeny. Therefore, one might still think that *16SrRNA* gene alone is good enough to study bacterial phylogeny. We disagree. As we repeatedly mentioned, one gene is not enough to construct the correct tree. In fact, there are cases in which *16SrRNA* gene did not lead to the correct tree (Ragan, 1988; Saruhashi et al., 2007). In addition, the bootstrap test those authors used is not appropriate to examine the correctness of a constructed tree, as mentioned above.

There is the possibility that similar amino acid contents in proteins of eubacterial and archaeal theromophiles would forcefully bring the former to the outmost cluster of the eubacteria (Cambillau and Claverie, 2000; Fukuchi and Nishikawa, 2001). To examine this possibility we also used mesophilic archaea instead of thermophilic archaea and constructed another concatenated tree given in Fig. 3d. As the two trees in Figs. 3a and d together show, the similar amino acids do not seem to affect the construction of bacterial phylogeny. The two trees also suggest that these thermophilic clusters are unaffected by the choice of archaea species with different genome sizes. In addition, they do not support the possibility of a large extent of horizontal gene transfer between thermophilic archaea and thermophilic eubacteria (Aravind et al., 1998), though there might have been to some extent.

Our result is in agreement with previous studies (Grishin et al., 2000; Snel et al., 1999; Wolf et al., 2001, 2002; House and Fitz-Gibbon, 2002; Korbel et al., 2002; Brown et al., 2001; Henz et al., 2004), as long as the divergence of thermophilic eubacteria is concerned. However, other researchers have shown that other species than thermophilic eubacteria diverged first in the eubacterial cluster (Qi et al., 2004; Korbel et al., 2002; Dutilh et al., 2004; Fukami-Kobayashi et al., 2007; Ciccarelli et al., 2006). Therefore, the argument about the phylogenetic position of thermophilic eubacteria is not resolved yet, and more studies are required to elucidate the earliest cluster of the eubacteria. To resolve the argument completely we may also have to estimate the divergence time of the earliest eubacterial species, whatever it will be. There is a report which states that photosynthetic organisms had evolved and were living in a stratified ocean supersaturated in dissolved silica 3.4 billion yeas ago (Tice and Lowe, 2004). With the success of the estimation we will then be able to elucidate the earliest known species that appeared on earth.

Acknowledgments

TH was supported by grant from the Ministry of Education, Science, Sports, Culture and Technology of Japan.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.gene.2008.10.006.

References

Aravind, L., Tatusov, R.L., Wolf, Y.I., Walker, D.R., Koonin, E.V., 1998. Evidence for massive gene exchange between archaeal and bacterial hyperthermophiles. Trends Genet. 14, 442–444.

Bininda-Emonds, O.R., Gittleman, J.L., Steel, M.A., 2002. The (super)tree of life: procedures, problems, and prospects. Annu. Rev. Ecol. Syst. 33, 265–289.

Brochier, C., Bapteste, E., Moreira, D., Philippe, H., 2002. Eubacterial phylogeny based on translational apparatus proteins. Trends Genet. 18, 1–5.

Brown, J.R., Doolittle, W.F., 1997. Archaea and the prokaryote-to-eukaryote transition. Microbiol. Mol. Biol. Rev. 61, 456–502.

Brown, J.R., Douady, C.J., Italia, M.J., Marshall, W.E., Stanhope, M.J., 2001. Universal trees based on large combined protein sequence data sets. Nat. Genet. 28, 281–285.

Cambillau, C., Claverie, J.M., 2000. Structural and genomic correlates of hyper-thermostability. Biol. Chem. 275, 32383–32386.

Castresana, J., 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. 17, 540–552.

Ciccarelli, F.D., et al., 2006. Toward automatic reconstruction of a highly resolved tree of life. Science 311, 1283–1287.

Creevey, C.J., McInerney, J.O., 2005. Clann: investigating phylogenetic information through supertree analyses. Bioinformatics 21, 390–392.

Daubin, V., Gouy, M., Perriere, G., 2001. Bacterial molecular phylogeny using supertree approach. Genome Inform. 12, 155–164.

Daubin, V., Gouy, M., Perriere, G., 2002. A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. Genome Res. 12, 1080–1090.

Delsuc, F., Brinkmann, H., Philippe, H., 2005. Phylogenomics and the reconstruction of the tree of life. Nat. Rev. Genet. 6, 361–375.

Dutilh, B.E., Huynen, M.A., Bruno, W.J., Snel, B., 2004. The consistent phylogenetic signal in genome trees revealed by reducing the impact of noise. J. Mol. Evol. 58, 527–539.

Felsenstein, J., 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17, 368–376.

Felsenstein, J., 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39, 783–791.

Fitch, W.M., 2000. Homology a personal view on some of the problems. Trends Genet. 16, 227–231.

Fukami-Kobayashi, K., Minezaki, Y., Tateno, Y., Nishikawa, K., 2007. A tree of life based on protein domain organizations. Mol. Biol. Evol. 24, 1181–1189.

Fukuchi, S., Nishikawa, K., 2001. Protein surface amino acid compositions distinctively differ between thermophilic and mesophilic bacteria. J. Mol. Biol. 309, 835–843.

Gadagkar, S.R., Kumar, S., 2005. Maximum likelihood outperforms maximum parsimony even when evolutionary rates are heterotachous. Mol. Biol. Evol. 22, 1241–2139.

Gascuel, O., 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol. Biol. Evol. 14, 685–695.

Grishin, N.V., Wolf, Y.I., Koonin, E.V., 2000. From complete genomes to measures of substitution rate variability within and between proteins. Genome Res. 10, 991–1000.

Guindon, S., Gascuel, O., 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52, 696–704.

Henz, S.R., Huson, D.H., Auch, A.F., Nieselt-Struwe, K., Schuster, S.C., 2004. Whole-genome prokaryotic phylogeny. Bioinformatics 21, 2329–2335.

Horiike, T., Hamada, K., Miyata, D., Shinozawa, T, 2004. The origin of eukaryotes is suggested as the symbiosis of pyrococcus into gamma-proteobacteria by phylogenetic tree based on gene content. J. Mol. Evol. 59, 606–619.

House, C.H., Fitz-Gibbon, S.T., 2002. Using homolog groups to create a whole-genomic tree of free-living organisms: an update. J. Mol. Evol. 54, 539–547.

Katoh, K., Misawa, K., Kuma, K., Miyata, T., 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30, 3059–3066.
Korbel, J.O., Snel, B., Huynen, M.A., Bork, P., 2002. SHOT: a web server for the construction of genome phylogenies. Trends Genet. 18, 158–162.
Kumada, Y., et al., 1993. Evolution of the glutamine synthetase gene, one of the oldest existing and functioning genes. Proc. Natl. Acad. Sci. U. S. A. 90, 3009–3013.
Michener, C.D., Sokal, R.R., 1957. A quantitative approach to a problem of classification. Evolution 11, 490–499.
Qi, J., Wang, B., Hao, B.I., 2004. Whole proteome prokaryote phylogeny without sequence alignment: a K-string composition approach. J. Mol. Evol. 58, 1–11.
Ragan, M.A., 1988. Ribosomal RNA and the major lines of evolution: a perspective. Biosystems 21, 177–187.
Saitou, N., Imanishi, T., 1989. Relative efficiencies of the Fitch–Margoliash, maximum-parsimony, maximum-likelihood, minimum-evolution, and neighbor-joining methods of phylogenetic tree construction in obtaining the correct tree. Mol. Biol. Evol. 6, 514–525.
Saitou, N., Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4, 406–425.
Saruhashi, S., Hamada, K., Horiike, T., Shinozawa, T., 2007. Determination of whole prokaryotic phylogeny by the development of a random extraction method. Gene 392, 157–163.
Snel, B., Bork, P., Huynen, M.A., 1999. Genome phylogeny based on gene content. Nat. Genet. 21, 108–110.
Snel, B., Huynen, M.A., Dutilh, B.E., 2005. Genome trees and the nature of genome evolution. Annu. Rev. Microbiol. 59, 191–209.
Tateno, Y., Nei, M., Tajima, F., 1982. Accuracy of estimated phylogenetic trees from molecular data. I. Distantly related species. J. Mol. Evol. 18, 387–404.
Tatusov, R.L., Koonin, E.V., Lipman, 1997. A genomic perspective on protein families. Science 278, 631–637.
Tekaia, F., Lazcano, A., Dujon, B., 1999. The genomic tree as revealed from whole proteome comparisons. Genome Res. 9, 550–557.
Tice, M.M., Lowe, D.R., 2004. Photosynthetic microbial mats in the 3,416-Myr-old ocean. Nature 431, 549–552.
Uchiyama, I., 2003. MBGD: microbial genome database for comparative analysis. Nucleic Acids Res. 31, 58–62.
Woese, C.R., 1987. Bacterial evolution. Microbiol. Rev. 51, 221–271.
Wolf, Y.I., Rogozin, I.B., Grishin, N.V., Koonin, E.V., 2002. Genome trees and the tree of life. Trends Genet. 18, 472–479.
Wolf, Y.I., Rogozin, I.B., Grishin, N.V., Tatusov, R.L., Koonin, E.V., 2001. Genome trees constructed using five different approaches suggest new major bacterial clades. BMC Evol. Biol. 1, 8.
Yarza, P. et al., 2008. The All-Species Living Tree project: a 16S rRNA-based phylogenetic tree of all sequenced type strains. Syst. Appl. Microbiol. 31, 241–250.