ISSN 2075-1729
www.mdpi.com/journal/life

Peer-Review Record:

Phylogeny and Taxonomy of Archaea: A Comparison of the Whole-Genome-Based CVTree Approach with 16S rRNA Sequence Analysis

Guanghong Zuo, Zhao Xu and Bailin Hao

Life **2015, *5*, 949-968, doi:10.3390/life5010949**

Reviewer 1: Anonymous
Reviewer 2: Anonymous
Editor: Roger A. Garrett, Hans-Peter Klenk and Michael W. W. Adams (Guest Editor of Special Issue "Archaea: Evolution, Physiology, and Molecular Biology")

Received: 9 December 2014 / Accepted: 9 March 2015 / Published: 17 March 2015

First Round of Evaluation

Round 1: Reviewer 1 Report and Author Response

In the manuscript titled "Phylogeny and Taxonomy of Archaea: A Comparison of Whole-Genome-Based CVTree Approach with 16S rRNA Sequence Analysis", Zuo *et al.* are describing a comprehensive analysis of archaeal phylogeny with a genome-based alignment free method, and then comparing the findings to 16S rRNA based phylogenies.

The idea to use whole genome data for archaeal phylogenies is interesting, as the 16S rRNA phylogenies can be poor in resolving relationships among archaeal lineage due to GC content bias in hyperthermophilic Archaea. However, I am personally not convinced that whole genomes bring additional advantages, or if we are better off by just using a conserved set of proteins. Perhaps an additional discussion point would be the advantages of using CVtree approach with regular concatenated proteins.

Leaving that opinion aside, I think the alignment-free methodology is interesting, however I would like to see a comparison with at least one regular alignment/treeing method, based on the same genomes the authors used, and not just a visual topological comparison with other published trees.

Response: Multi-alignment of concatenated protein (or DNA) segments is a genome-scale, but not whole-genome approach. Its applicability depends on the scope of the phylogenetic study. When dealing with not-too-distantly related species it may yield more or less useful result. However, in a study covering many phyla it is very difficult, if not impossible, to collect a common set of conserved proteins.

Moreover, the concatenation method can never lead to very convincing conclusion, as give or take a few proteins may change the result. The phylogenomics people have noticed this problem, see, e.g.,

> *O. Jeffroy, H. Brinkman, F. Delsuc, H. Philippe (2008) Phylogenomics: the beginning of incongruence? Trends in Genetics, 22(4): 225–231.*

An example from the Bacteria domain is the relationship of the closely related Shigella and Escherichia coli strains. Concatenation of different number of genes led to different way of mixing-up of the two groups, but CVTree gave unambiguous separation of the strains as different species in the same genus Escherichia, see:

> *G.-H. Zuo, Z. Xu, B.L. Hao (2013) Shigella strains are not clones of Escherichia coli but sister species in the genus Escherichia. Genomics Proteomics Bioinformatics, 11: 61–65.*

In order to carry out multi-alignment of concatenated sequences, a postdoc or well-trained PhD student equipped with the corresponding software is required. In contrast, with genome sequencing becoming a common practice in many labs it costs no additional work for a bench-microbiologist to get phylogenetic and taxonomic information by using a convenient and publically available tool such as the CVTree web serve. Well, we would be glad to see comparison of CVTree phylogeny with multi-alignment of concatenated proteins if anyone finds a way to do it for so many diverse phyla, but we do not consider it as a doable job.

As far as I can see, statistical support on the branches is missing, so I have no way of assessing if this branching order is valid. Bootstrapping or jackknifing are by no means the final word on the significance of branches, however an explanation as to how the user should assess the significance of braches would be good, *i.e.*, branch length.

Response: These points were discussed in the "Material and Method" section added at the suggestion of Reviewer 2. The following was copied from the manuscript:

> *"Traditionally a newly generated phylogenetic tree is subject to statistical re-sampling tests such as bootstrap and jackknife. CVTree does not use sequence alignment. Consequently, there is no way to recognize informative or non-informative sites. Instead we take all the protein products encoded in a genome as a sampling pool for carrying out bootstrap or jackknife tests (citing our 2004 paper). Although it was very time-consuming, CVTrees did have well passed these tests (citing our 2010 paper). However, successfully passing statistical re-sampling tests only tells about the stability and self-consistency of the tree with respect to small variations of the input data. It is by far not a proof of objective correctness of the tree. Direct comparison of all branchings in a tree with an independent taxonomy at all ranks would provide such a proof, The 16S rRNA phylogeny cannot be verified by the Bergey's taxonomy, as the latter follows the former. However, agreement of branchings in CVTree with the Bergey's taxonomy would provide much stronger support to the tree as compared to statistical tests. This is the strategy we adopt for the CVTree approach."*
>
> *"There are two aspects of a phylogenetic tree: the branching order (topology) and the branch lengths. Branching order is related to classification and branch length to evolution time. Calibration of branch lengths is always associated with the assumption that mutation rate*

remains more or less a constant across all species represented in a tree, an assumption that cannot hold true in a large-scale phylogenetic study like the present one. Therefore, branching order in trees is of primary concern, whereas calibration of branch lengths makes less sense. Accordingly, all figures in this paper only show the branching scheme without indication of branch lengths and bootstrap values".

The fact that Thermofilum is placed outside Crenarchaeota in Figures 3 and 4 is a little disturbing. I haven't come across such a placement in other phylogenetic analyses of Archaea, for example in Brochier-Armanet et al 2008 Nat Rev Microb, or Rinke *et al.* 2013 Nature. I believe this needs to better explained in the manuscript, rather than just saying "this fact is noted".

Response: Yes, this is an apparent discrepancy of CVTree from 16S (and 23S) analysis for the given set of 179 archaeal genomes. However, in an on-going study of ours (not published yet) using a much larger data set this violation no longer shows up; both Korarchaeota and Crenarchaeota restore their phylum status. Taking into account the fact that both Korarchaeota and Thermofilaceae are represented by single species for the time being, their placement certainly requires further study with broader sampling of genomes.

I am also not very convinced with the placement of Nanoarchaeota. It seems like this phylum is moving around with the addition of new sequence data (for example in Rinke *et al.* Figure 2 tree, they are on an entirely different branch than Euryarchaeota). Though the authors also rightfully point out that the reduced genome size may have something to do with this placement.

Response: Highly degenerated genomes of many symbiont organisms tend to move around, in particular, to the baseline of a tree and thus distorts the overall structure of the tree. Therefore, it is better not to mix them with free-living organisms in a study. We rephrased the corresponding paragraph in the manuscript:

"The nanosized archaean symbiont Nanoarchaeum equitans has a highly reduced genome (490,885 bp). It is the only described representative of a newly proposed phylum Nanoarchaeota and it cuts into the otherwise monophyletic phylum Euryarchaeota. We note that the monophyly of Euryarchaeota was also violated by Nanoarchaeum in some 16S rRNA trees, see, e.g., Figure 4 in a 2009 microbial survey as well as (c) and (d) in our Figure 3. It has been known that tiny genomes of endosymbiont microbes often tend to move towards baseline of a tree and distort the overall picture. In fact, we have suggested skipping such tiny genomes when studying bacterial phylogeny, see, e.g., (citing our 2010 paper) and a note in the home page of the CVTree Web Server. In the present case we may at most say that Nanoarchaeota probably makes a separate phylum, but its cutting into Euryarchaeota might be a side effect due to the tiny size of the highly reduced genome".

The placement of Halobacteria (due to interfering Nanoarchaeota, I presume) is also a little disturbing. I would recommend that the authors provide a discussion of this. Especially with regards to other archaeal trees. For instance, in the tree of Armanet *et al* 2011 that the authors also refer to, the placement of Halobacteria with respect to Nanoarchaeota is very different.

Response: Yes, there was certain disturbing effect of the tiny and lonely Nanoarchaeum genome, yet the Halobacteria is a very specific clade, forming a tightly connected group and moving around as a whole, mainly due to the biased acidity of their constituent amino acids. We anticipate that the relative placement of Halobacteria with respect to other groups may stabilize when more genomes are used to construct a tree.

In summary, I find the piece interesting, but parts of the discussion are rather weak, therefore I am suggesting a major revision. Another reason for major revision is the style that the manuscript is written. I am not a native speaker, but given that I had to read sentences several times, I suspect the manuscript can benefit from an English language check.

Response: We have done a major revision of the manuscript. A new "Material and Method" section has been added. Such issues as statistical resampling tests (bootstrap and jackknife), calibration of branch length, the meaning and choice of the peptide length K, etc., were discussed in the new section. Figures 1 and 2 were combined to a new Figure 1; Figures 3 and 4 were combined to become a new Figure 2. Figure captions were made more detailed. The whole text was checked for language flaws and many places were rephrased.

Round 1: Reviewer 2 Report and Author Response

Summary

Zuo *et al.* present a comparative analysis of the taxonomic classification of the Archaea domain. About 180 archaeal genomes were used to calculate a new tree topology using CVTree. The tree was compared with several 16S rRNA trees reported in the literature, and the differences were minor. Results also provide additional support to recently proposed archaeal phyla and halobacterial orders. Authors emend classification of some strains. Authors present a new interactive tree-visualization tool which enables direct validation of taxonomic groups according to their monophyly.

Evaluation

I found major concerns along this manuscript, and suggest a substantial revision. Authors should define well their objectives, also make sure that the conclusions are novel (which ones are novel, which are supportive to previous published work, *etc.*), and avoid redundancies in the text. Besides, would be desirable that authors provide more objective criteria for high taxa circumscription based on their methodology.

General Comments

The topic is relevant for microbial taxonomy. This kind of research should be encouraged further because old taxonomic paradigms must be systematically reviewed based on new genomic data.

I would really appreciate a "Methods" part where the CVTree is explained shortly, and the tree-viewer is explained with more detail. The absence of branch lengths and bootstraps should be discussed here. Other technical aspects of the paper (e.g., sequence dataset), parameters, criteria ... all could be well organized in this part.

Response: A "Material and Method" section has been added where the CVTree algorithm, the interactive tree-viewer, statistical resampling tests (bootstrap, jackknife), calibration of branch lengths, etc., were discussed in slightly more detail.

Archaeal phylogeny has already been studied in detail, with 16S and other marker genes, and with genomic approaches too. Some of the undersigning authors had already published on this before, although with smaller input datasets. Therefore, the fact that 16S topology is quite stable and comparable with other approaches is already known.

Response: Yes, 16S rRNA phylogeny is quite stable and it almost defines the present taxonomy. We have given due credit for this. In general, CVTree does not challenge 16S rRNA analysis but complement it.

Taxonomists have traditionally circumscribed the high taxa (specially orders and classes) with great subjectivity, *i.e.*, without well accepted criteria. In terms of phylogenetic trees one premise has always been clear, a taxon must be monophyletic. This principle has been used in the present work to reconsider the status of some high taxa. However, authors do not explain objective criteria to properly interpret the rank of the clades, which impedes making a profound evaluation of the archaeal classification. Therefore, although authors have strong tools and dataset, they just achieved a small revision of the high taxa which is, indeed, quite biased by the underlying 16S guidelines.

Response: A robust phylogenetic tree comes with a fixed branching order of leaves. One looks at the leaf names and their taxonomic lineage and tries to map the latter to the branches. To this end we added the following paragraphs in the "Material and Method" section.

> *"There are two aspects of a phylogenetic tree: the branching order (topology) and the branch lengths. Branching order is related to classification and branch length to evolution time. Calibration of branch lengths is always associated with the assumption that mutation rate remains more or less a constant across all species represented in a tree, an assumption that cannot hold true in a large-scale phylogenetic study like the present one. Therefore, branching order in trees is of primary concern, whereas calibration of branch lengths makes less sense. Accordingly, all figures in this paper only show the branching scheme without indication of branch lengths and bootstrap values."*
>
> *"Branching order in a tree by itself does not bring about taxonomic ranks, e,g, class or order. The latter can be assigned only after comparison with a reference taxonomy which is not a rigid framework but a modifiable system. Though a dissimilarity measure figures in the CVTree algorithm, it is not realistic to delineate taxa by using this measure at least for the time being. Even if defined in the future, it must be lineage-dependent. For example, it cannot be expected that the same degree of dissimilarity may be used to delineate classes in all phyla. In addition, monophyly is a guiding principle in comparing branching order with taxonomy. Here monophyly must be understood in a pragmatic way restricted to the given set of input data and the reference taxonomy. If all genomes from a taxon appear exclusively in a tree branch, the branch is said to be monophyletic."*

I have noted some lack of scientific rigor according to: many wrong taxonomic names and typos, scarce figure legends, few comments about the missing branch lengths or bootstraps (!), redundancy in text and figures and fragments which are really difficult to understand. Authors should pay attention to language, explanations and text organization.

Response: We have reorganized the manuscript mainly by adding a new "Material and Method" section where discussions on branch length, statistical resampling, meaning and choice of K, etc., were given. The original Figure 1 was deleted with some related points explained in the text accompanying the original Figure 2. All figure captions have been rewritten for clarity.

Authors shouldn't forget (particularly in conclusion) that the resolution power of 16S for high ranks (genus and above) is currently well accepted. And the number of non-redundant 16S entries available is much much larger than that of archaeal genomes. 16s data offer a much comprehensive view of the archaeal diversity, including deep branches.

Response: A few more sentences were added in the "Conclusion" regarding the power and achievement of the 16S rRNA analysis.

Introduction

- L34–40: for me was difficult to follow this reasoning. Please rephrase.
- L40 "Since at present (…) are not covered (...)" is a weak reason for choosing high ranks. I recommend to shortly summarize why high ranks are so important, and why do you choose order as the lowest considered rank.

Response: We tried to rephrase the paragraph by changing, deleting, or adding a few words as follows:

> *"In this paper we study Archaea phylogeny across many phyla. This is in contrast with phylogeny of species in a narrow range of taxa, e.g., that of vertebrates (a subphylum) or human versus close relatives (a few genera). Accordingly, the phylogeny should be compared with taxonomy at large, or, as Cavalier-Smith (citing cavalier-smith 2002) put it, with "megaclassificaton" of prokaryotes. Although in taxonomy the description of a newly discovered organism necessarily starts from the lower ranks, higher rank assignments are often incomplete or lacking. At present the ranks above class are not covered by the Bacteriological Code. The number of plausible microbial phyla may reach hundreds and archaeal ones are among the less studied. According to the 16S rRNA analysis, the major archaeal classes and their subordinate orders have been more or less delineated. Therefore, in order to carry out the aforementioned cross verification we make emphasis on higher ranks such as phyla, classes, and orders. A study using 179 Archaea genomes should provide a framework for further study of lower ranks."*

Part 3.1

The figure legends require more rigorous explanation. Would be interesting to remember that the branch lengths are not taken into account, or were omitted.

Response: Branching order in a tree is directly related to taxonomy, while branch lengths have more to do with evolution. For large-scale phylogenetic study across many phyla the former is more important than the calibration of branch lengths.

The latter is based on the assumption that mutation rate is more or less constant. This assumption cannot hold when dealing with many phyla.

In Line 80—authors write some explanations to understand the tree figures. I would suggest to put this text before the first tree figure.

Response: This is done in the newly added "Material and Method" section.

Figure 1 is a bit redundant. A short comment about the inclusion of that particular sequence into the Thermoplasmatales can be added into the legend of Figure 2.

Response: The original Figure 1 was deleted and a few words added to the legend of the original Figure 2, now the new Figure 1.

However, the reclassification of Methanomassiliicoccus into Thermoplasmataceae needs more explanation. To be objective, authors should address the following question, why in the same family and not in another new family?

Response: Judging by the cluster labeled as Euryarchaeote{0+3} in Figure 2 Methanomassiliicoccus was not reclassified into Thermoplasmataceae but to an yet un-specified class.

L106–107: maybe a bit inappropriate on that position. I would suggest to add a comment on the figure legend instead.

Response: The sentence has been moved to the legend of Fig. 1 and slightly rephrased.

Part 3.2

L109–118: Summarize and move to introduction. Those sentences are of general importance for the topic and not specific to part 3.2.

Response: Done.

L119–123: if the *K* issue is relevant to understand the text then please add a proper explanation. If not, then keep it simple and avoid entering into the *K* issue (L122–123, L132, L135, L139, 183, and also remove this $K = 6$ from Figure 3.)

Response: the K issue is discussed in the newly added "Material and Method" section; so scattered mentioning of K has been deleted from the rest of text.

Add more explanations in legend of Figure3.

Response: Done in the caption of Figure 2.

L143: Sounds clearer if you avoid mixing class and phylum, for example: "The placement of phylum Korarchaeota, as a closest neighbor of family Thermofilaceae, violates the monophyly of phylum Crenarchaeota."

Response: We have rewritten the paragraph as:

> *"The new phylum Korarchaeota violates the monophyly of the phylum Crenarchaeota by drawing to itself the family Thermofilaceae. However, in an on-going study of ours (not published yet) using a much larger data set, this violation no longer shows up; both*

Korarchaeota and Crenarchaeota restore their phylum status. Taking into account the fact that both Korarchaeota and Thermofilaceae are represented by single species for the time being, their placement certainly requires further study with broader sampling of genomes."

L146: No need to explain the 6+2 if it is properly explained in the figure legend.

Figure 3 is redundant. I would recommend to avoid presenting different versions of the same tree; just the final tree is OK (use final/valid labels) and all important explanations in the text or legend. Perhaps the whole reasoning in Lines 142–180 is not so relevant for the current objective of comparing CVT, LTP, Bergeys? Or, perhaps, define this objective more clearly.

Response: From the original Figures 3 and 4 only one has been kept and the legend rewritten. In fact, the whole paragraph changed to:

"The newly proposed phylum Thaumarchaeota appears to be non-monophyletic as an outlying strain Candidatus Caldiarchaeum subterranum was assigned to this phylum according to the NCBI taxonomy. The NCBI assignment might reflect its position in some phylogenetic tree based on concatenated proteins, e.g., Figure 2 in […]. However, in the original paper reporting the discovery of this strain […] and in recent 16S rRNA studies, e.g., […], Candidatus Caldiarchaeum subterranum was proposed to make a new phylum Aigarchaeota. CVTrees support the introduction of this new phylum. A lineage modification of Candidatus Caldiarchaeum subterranum from Thaumarchaeota to Aigarchaeota would lead to a monophyletic Thaumarchaeota{7}."

L156–166. If I understood right, there was good support for Candidatus Aciduliprofundum as part of a clade called DHEV2, which is a sister clade of Thermoplasmatales. However in the present work the authors intend here to reclassify Aciduliprofundum into family Thermococcaceae of Thermococcales. This needs further explanation. Since the new affiliation is quite in disagreement with previous observations, and this is not properly justified in the results/discussion, the final statement "this modification would hold as long as no new facts challenge it" seems unacceptable. In addition, why should Aciduliprofundum be regarded as member of Thermococcaceae and not as another distinct family?

Response: The problem of taxonomic placement of Aciduliprofundum is a good example to demonstrate how one extract information from CVTrees. In the Reysenbach et al. Nature 2006 paper it was taken as the first cultivated member of the DHEV2 (deep-sea hydrothermal euryarchaeate 2) clade based on a maximum-likelihood 16S rRNA tree. Unfortunately, all other 13 members of this clade were represented by 16S rRNA sequences only and no genome data are available so far. The NCBI taxonomy gave an incomplete lineage: Archaea; Eurytchaeota; unclassified Euryarchaeota; missing taxonomic assignment at the rank class and below.

In order to make use of CVTree we must touch on the K-issue a little more. The alignment-free comparison of genomes in CVTree is implemented by counting the number of K-peptides in the protein products encoded in a genome followed by subtraction of a random background caused by neutral mutations. The peptide length K looks like a parameter, but it is actually not a parameter. Using a longer K emphasizes species-specificity, while a shorter K takes into account more common features with neighboring species. However, we never adjust K: a fixed K is used for all genomes to construct a tree, but one may construct a series of trees for K = 3, 4, 5, 6, 7, 8, … We have shown repeatedly that

K = 5 and 6 lead to best results in the sense of agreement with taxonomy, so usually only a K = 6 tree is given in publications.

Let us look at a subtree, i.e., part of a tree, containing the organisms of interest. If the branching order in all trees built for different Ks turns out to be same, it would be a strong support to the branching order. In most cases the branching order varies with K: K = 3 and 4 make sense, K = 5 and 6 yield the best, K = 7 and 8 become slightly worse, etc. For too big a K, even if the closest strains remain grouped together the whole tree may tend to become a star-tree, i.e., every small clade stands in its own and their mutual placements become less meaningful. Therefore, inspection of trees for a range of K-values provides an additional dimension to evaluate the results.

For Aciduprofumdum we have a stable pair

(<C>Thermococci{18}, <G>Aciduprofumdum{2}) at K=3, 5, 6, 7.

At K = 4 we have

(<C>Thermococci{18}, (<G>Staphylothermus{2}, <G>Aciduprofundum{2}))

In all these cases Thermoplasmata stands farther away from the above pair. However, at K = 8 and 9, when the overall tree picture has been largely distorted, Aciduliprofundum does stand closer to Thermaplasmata. Putting together all the above results we tend to consider the pair (<C>Thermococci{18}, <G>Aciduliprofundum{2}) as reflecting a more probable relation. Confined to the available data for the time being one may assign <G>Aciduliprofundum to <C>Thermococci, e.g., to denote the pair as

<C>Thermococci{20}=(<F>Thermococcaceae{18}, <G>Aciduliprofundum{2})

leaving its family unclassified or assign it to a new family. Without further phenotypic and chemotaxonomic evidence it is better not to introduce new taxon names if the present naming scheme is capable to accommodate the leaves without conflict. This was why we wrote "this modification would hold as long as no new facts challenge it". Anyway, taxonomy has always been a work in progress. One has to be prepared for modifications when new data appear. To make a long story short, we have rewritten the paragraph as:

> *"The Candidatus genus Aciduliprofundum is considered a member of the DHEV2 (deap-sea hydrothermal vent euryarchaeotic 2) phylogenetic cluster. No taxonomic information was given in the original papers [55,56]. The NCBI Taxonomy did not provide definite lineage information for this taxon at the class, order, and family ranks. According to [55] the whole DHEV2 cluster was located close to Thermoplamatales in a maximum-likelihood analysis of 16S rRNA sequences. A similar placement was seen in [54] where a Bayesian tree of the archaeal domain based on concatenation of 57 ribosomal proteins put a lonely Aciduliprofundum next to Thermoplasmata. However, in CVTrees, constructed for all K-values from 3 to 9, Aciduliprofundum juxstaposes with the class Thermococci{18}. An observation in [56] that this organism shares a rare lipid structure with a few species from Thermococcales may hint on its possible association with the latter. If we temporarily presume a lineage*
>
> *<C>Thermococci<O>Unclassified<F>Unclassified<G>Aciduliprofundum…*

one might have a monophyletic class <C>Thermococci{20}. Since none of the 13 DHEV2 members listed in [55] has a sequenced genome so far, CVTree cannot tell the placement of the DHEV2 cluster as a whole for the time being. It remains an open problem as whether DHEV2 is close to Thermoplasmata or to Thermococci, or a new class is needed to accommodate DHEV2."

L163: If I'm right the current observation actually does not support the previous work done by Brochier-Armanet.

Response: No, we did not mean it.

L167–173: The names are wrongly written (please check the original submission). Authors have to explain with more clarity, why is this clade of rank class. If that is the case, is it a single-order class? A single family order? *etc.*

Response: We should first explain how these inappropriate names appeared. We have insisted to use the directory name at the NCBI FTP site as genome name. However, in November 2013 NCBI announced that they would not release genomes of different strains of the same species as before. In a period thereafter NCBI sometimes put several genomes in a directory and we had to extract the data and to assign a name from the "Source" line of the GenBank file. This caused some confusion. For example, as of February 27, 2015, a directory name at NCBI remained "archaeon_Mx1201_uid196597" and we had to change it to:

Candidatus_Methanomethylophilus_alvus_Mx1201_uid196597

Now all "wrong names" as pointed out by the Reviewer no longer appear in figures. In the text we tried to refer to their names as complete as possible.

L174–180: I don't understand the reasoning along this paragraph. In addition "haolphiic_archaeon_DL31" is not well written, please be careful when copying names from other source. Also, a similar question about the objectiveness for detecting high taxa: why not to create new family?

Response: The genome name at NCBI FTP site is "halophilic_archaeon_DL31_uid72619". The uid number was dropped when mentioned in the text. We put it back and capitalized the first letter to "Halophilic", still an illegal genus name.

Figure 4: the reclassifications must be clearly justified in the text.

Response: It is Figure 2 in the revised manuscript. We discussed it at some length.

L191: organisms can't be validly published, but their names.

Response: No, organism cannot be published. Thanks for correcting our mistake.

Part 3.3

L221–222: Sure, but authors do not provide explanations about how do they know that a clade in a tree is a family, an order, a class, *etc.* There is a lack of criteria to reclassify the leaves.

Response: One cannot tell the rank of a node/leaf in a tree by simply looking at it. A reference taxonomy is alwys needed. We put the following in the "Material and Method" section to explain it:

> *"Branching order in a tree by itself does not bring about taxonomic ranks, e,g, class or order. The latter can be assigned only after comparison with a reference taxonomy which is not a rigid framework but a modifiable system. Though a dissimilarity measure figures in the CVTree algorithm, it is not realistic to delineate taxa by using this measure at least for the time being. Even if defined in the future, it must be lineage-dependent. For example, it cannot be expected that the same degree of dissimilarity may be used to delineate classes in all phyla. In addition, monophyly is a guiding principle in comparing branching order with taxonomy. Here monophyly must be understood in a pragmatic way restricted to the given set of input data and the reference taxonomy. If all genomes from a taxon appear exclusively in a tree branch, the branch is said to be monophyletic."*

L224: I don't understand "3063 identical nucleotide positions". Why identical?

Response: The phrase "3063 identical nucleotide positions" was copied from the caption of Figure 4 of the cited Nunoura et al. 2011 paper without much thinking. We simply deleted it.

L239–244: hard to read, please rephrase.

Response: The whole paragraph has been rewritten as:

> "*The nanosized archaean symbiont Nanoarchaeum equitans has a highly reduced genome (490,885 bp [44]). It is the only described representative of a newly proposed phylum Nanoarchaeota and it cuts into the otherwise monophyletic phylum Euryarchaeota. We note that the monophyly of Euryarchaeota was also violated by Nanoarchaeum in some 16S rRNA trees, see, e.g., Figure 4 in a 2009 paper [61] as well as (c) and (d) in our Figure 4. It has been known that tiny genomes of endosymbiont microbes often tend to move towards baseline of a tree and distort the overall picture. In fact, we have suggested skipping such tiny genomes when studying bacterial phylogeny, see, e.g., [29] and a note in the home page of the CVTree Web Server [21]. In the present case we may at most say that Nanoarchaeota probably makes a separate phylum, but its cutting into Euryarchaeota might be a side effect due to the tiny size of the highly reduced genome."*

Conclusion

I disagree that CVTree approach is independent of 16S, because authors are using the current accepted classification (which is mainly 16S-based) to validate the observed clades.

Response: As a method CVTree is independent of 16S rRNA analysis. First, it uses protein products in a genome instead of RNA segments in the genome. Second, it does not do sequence alignment. CVTree generates stable trees but cannot tell which branch corresponds to what taxon. Only after comparison with the existing classification and nomenclature one would be able to make connections with taxonomy. In this sense it does depend on 16S rRNA taxonomy. Anyway, CVTree does not challenge 16S rRNA analysis but makes it more convincing in most cases. The revealed discrepancies call for further study.

Why at higher ranks, genomic approaches are more effective? That needs more explanation. And authors should also consider the large benefits of 16S data availability, specially at high ranks (genus and above) where the 16S has good resolution.

Response: In fact, genomic approaches are more effective at species level and below due to their high resolution power. At high ranks CVTree may be more effective in the sense that it does not require additional work. Suffice it to put genomes in CVTree web server and the branches come out, then compare them with a reference taxonomy.

Other corrections

- L23: phynotypic → phenotypic — *Done*
- L27: genomic era → the genomic era — *Done*
- L39: branchings in trees are → branching order in trees is — *Done*
- L40: classes → class — *Done*
- L47: L48 and Fig5: Crearchaeota → Crenarchaeota — *Done*
- L50: Fevridicoccales → Fervidicoccales — *Done*
- L52: Thermoplasmat → Thermoplasmata — *Done*
- L90: Thermopasmata → Thermoplasmata — *Done*
- L92: Thermaplasmataceae → Thermoplasmataceae — *Done*
- L105: on → of — *Done*
- L134: monophyly or non-monophyly → monophyletic or non-monophyletic — *Done*
- L164: rbosomal → ribosomal — *Done*
- L206: Korarcgaeota → Korarchaeota — *Done*
- L208: erenow → herenow — *Even "herenow" does not seem to be an correct English word; we changed it to "so far".*
- L212: was → were — *Done*
- L225: aligned 5993 amino acids → 5993 aligned amino acids. — *Done*
- L254: 15S → 16S — *Done*

Second Round of Evaluation

Round 2: Reviewer 1 Report and Author Response

I found the revised version of this manuscript quite good, and I thank the authors for responding thoroughly to all my comments.

Response: We thank the Reviewer for the detailed comments/suggestions given in the previous report and the suggestion of doing spelling-check this time. We have gone through the final manuscript carefully once more.

Round 2: Reviewer 2 Report and Author Response

Authors have substantially improved the article, including language corrections, and have provided extensive clarifications to all initial criticisms. I recommend acceptance for publication in Life Sciences journal. I had just a few minor suggestions.

L35-38: I don't get well the sentences which start from "This is in contrast ... ". Please can you specify a bit more?

Response: We changed "is in contrast with" to "is distinct from" and added a phrase "focusing on taxonomy of higher ranks" at the end of a sentence. Now the sentences read:

> *"This is distinct from phylogeny of species in a narrow range of taxa, e.g., that of vertebrates (a subphylum) or human versus close relatives (a few genera). Accordingly, the phylogeny should be compared with taxonomy at large, or, as Cavalier-Smith \cite{cavalier-smith2002} put it,with "megaclassificaton" of prokaryotes, focusing on taxonomy of higher ranks."*

- L45: should provide → provides — *Done*
- L47–49: move to conclusions? —*Moved to the conclusion section and the first word "Though" replaced by "In addition, since"*

L81–85: I think this paragraph is interrupting a bit the text flow. I suggest deletion.

Response: The whole paragraph has been deleted. This paragraph was added to the revised manuscript because one of the Reviewers asked "Does CVTree still require input genome data to be annotated to gene features, i.e., protein or CDS?" Well, this question reminds us that for many so-called "Permanent Draft" genomes it may be worthwhile returning to our early practice of using whole genome nucleotide sequences without distinguishing coding and non-coding segments. Although it did not lead to better results as compared with using translated protein products, but it is doable on un-annotated contigs. We will try this later.

L244–245: I think the text within {} deviates the attention. I suggest delete that part, ending sentence with "monophyletic Thaumarchaeota" is also ok.

Response: We have deleted all what appeared within the curly brackets and kept only "monophyletic Thaumarchaeota" as suggested. In fact, we could not tell how these words appeared there; there was none in our draft manuscript.