

Prokaryotic Branch of the Tree of Life:

A Composition Vector Approach

Bailin HAO^{1,2,3*} Lei GAO^{2,4}

(¹ *T-Life Research Center, Fudan University, Shanghai 200433, China;* ² *Institute of Theoretical Physics, Chinese Academy of Sciences, Beijing 100080, China;* ³ *Santa Fe Institute, Santa Fe, NM 87501, USA;* ⁴ *Present: Department of Botany and Plant Sciences, University of California, Riverside, CA 92521, USA.*

* Corresponding author: hao@mail.itp.ac.cn)

Abstract The Composition Vector Tree (CVTree) is a parameter-free and alignment-free method to infer prokaryotic phylogeny from their complete genomes. It is distinct from the 16S rRNA analysis in both the input data and the methodology. The prokaryotic phylogenetic trees constructed by using the CVTree method agree well with the Bergey's taxonomy in all major groupings and fine branchings. Thus, combined use of the CVTree approach and the 16S rRNA analysis may provide an objective and reliable reconstruction of the prokaryotic branch of the Tree of Life.

Key words prokaryotic phylogeny, taxonomy, composition vector, CVTree, Bergey's Manual

Prokaryotes are the most abundant organisms on Earth. They have been thriving for more than 3.7 billion years. They shaped most of the ecological and even geochemical environments for all organisms living on Earth. Yet our understanding of prokaryotes, particularly, their taxonomy and phylogeny, has been quite limited. No wonder merely a few years ago Carl Woese called microbiology "the science without a past" (Woese, 2000). Nevertheless, the use of 16S rRNA sequences to infer prokaryotic phylogeny, suggested by Woese and collaborators (Woese & Fox, 1977), has brought about a wealth of new knowledge. This success has reached so far that the modern prokaryotic taxonomy as reflected in the new edition of Bergey's Manual of Systematic Bacteriology (Bergey's Manual Trust, 2001-2009) is now largely based on 16S rRNA analysis. This situation, however, broaches a question of principle: the Bergey's taxonomy needs verification independent of the 16S rRNA analysis, in order to serve a function of demarcating the natural boundaries among prokaryotic species in an objective and convincing way. The fact that our newly proposed CVTree approach, using entirely different input data and methodology, supports most of the 16S rRNA results, may put the prokaryotic branch of the Tree of Life on a secure footing.

1. CVTree Approach to Prokaryotic Phylogeny

The CVTree approach was announced in 2002 (Hao et al., 2003) and has been described in Qi et al. (2004a) and Hao & Qi (2004). A Web Server has been installed for public access (Qi et al., 2004b). In brief, the input to CVTree is a collection of all translated amino acid sequences from the genome of an organism. We use the NCBI curated RefSeq (Pruitt et al., 2007) sequences in order to provide a common basis of comparison. Then the number of K-peptides is counted by using a sliding window, shifting one letter at a time along all protein sequences. These counts are kept in a fixed lexicographic order of amino acid letters to form a vector with 20^K components. A key procedure leading to the final composition vector is the subtraction of a background caused mainly by neutral mutations in order to highlight the shaping role of natural selection. This is done by using a (K-2)-th order Markovian prediction based on the number of (K-2)- and (K-1)-peptides from the same genome. A distance matrix is calculated from these composition vectors and the standard Neighbor-Joining program from the PHYLIP package (Felsenstein, 1980 – 2008) is used to generate the CVTrees. Instead of further elaboration of the method we emphasize its distinction from other more traditional methods:

1. It is an alignment-free method as each organism is represented by a composition vector with 20^K components determined by the number of distinct K-peptides in the collection of all translated protein sequences. Sequence alignment is replaced by K-peptide counting which is not challenged by the huge difference in genome size and gene number of prokaryotes.
2. It does not require the selection of RNA or protein-coding gene(s) as all translated protein products in a genome are used. Associated with this is the immunity of CVTree to lateral gene transfer (LTG). The analysis of Carl Woese on the role of LTG in cell evolution (Woese, 2002) may help to justify this point.
3. While the evaluation of traditional phylogenetic trees relies more or less on compatibility and stability arguments and various statistical tests such as bootstrapping or Jack-knifing have been invoked in this spirit, the CVTree results are verified by direct comparison with systematic bacteriology (Gao et al., 2007). The CVTrees constructed for from 69 to 432 organisms over the past 5 years bear a stable topology in major branchings from phyla down to species and strains. As compared to some traditional phylogenetic tree construction methods, the CVTree approach enjoys a nice feature of “the more genomes the better agreement” with taxonomy.
4. Moreover, the CVTree provides a parameter-free method that takes the collection of all proteins of the organisms under study as input and generates a distance matrix as output. The peptide length K, though appearing like a parameter, really controls the resolution power of the method. In fact, the CVTree method has

shown rather high resolution to elucidate the evolutionary relationship of different strains of one and the same species.

5. The high resolution power of the CVTrees provides a means to elucidate evolutionary relationships among different strains of one and the same species when the 16S rRNA analysis may not be strong enough to resolve too closely related strains.
6. While the 16S rRNA analysis cannot be applied to the phylogeny of viruses as the latter do not possess a ribosome, the CVTree method has been successfully used to construct phylogeny of coronaviruses including human SARS virus (Gao et al., 2003) and double-strand DNA viruses (Gao & Qi, 2007). It has been applied to chloroplasts as well (Chu et al., 2004).

In Fig. 1 we show the highest rank CVTree at $K=5$, adopted from the Supplementary Material of Gao et al. (2007). A highly degenerated genome of *Candidatus Carsonella ruddii* with genome size less than 160 kbp and 182 genes, much smaller than any known free-living bacteria, is excluded. Among the 431 organisms 424 are grouped under the correct phylum; the 7 outliers are not far from where they might be placed. Detailed organisms CVTrees for $K=3$ to 6 may be found in the Supplementary Material of Gao et al. (2007).

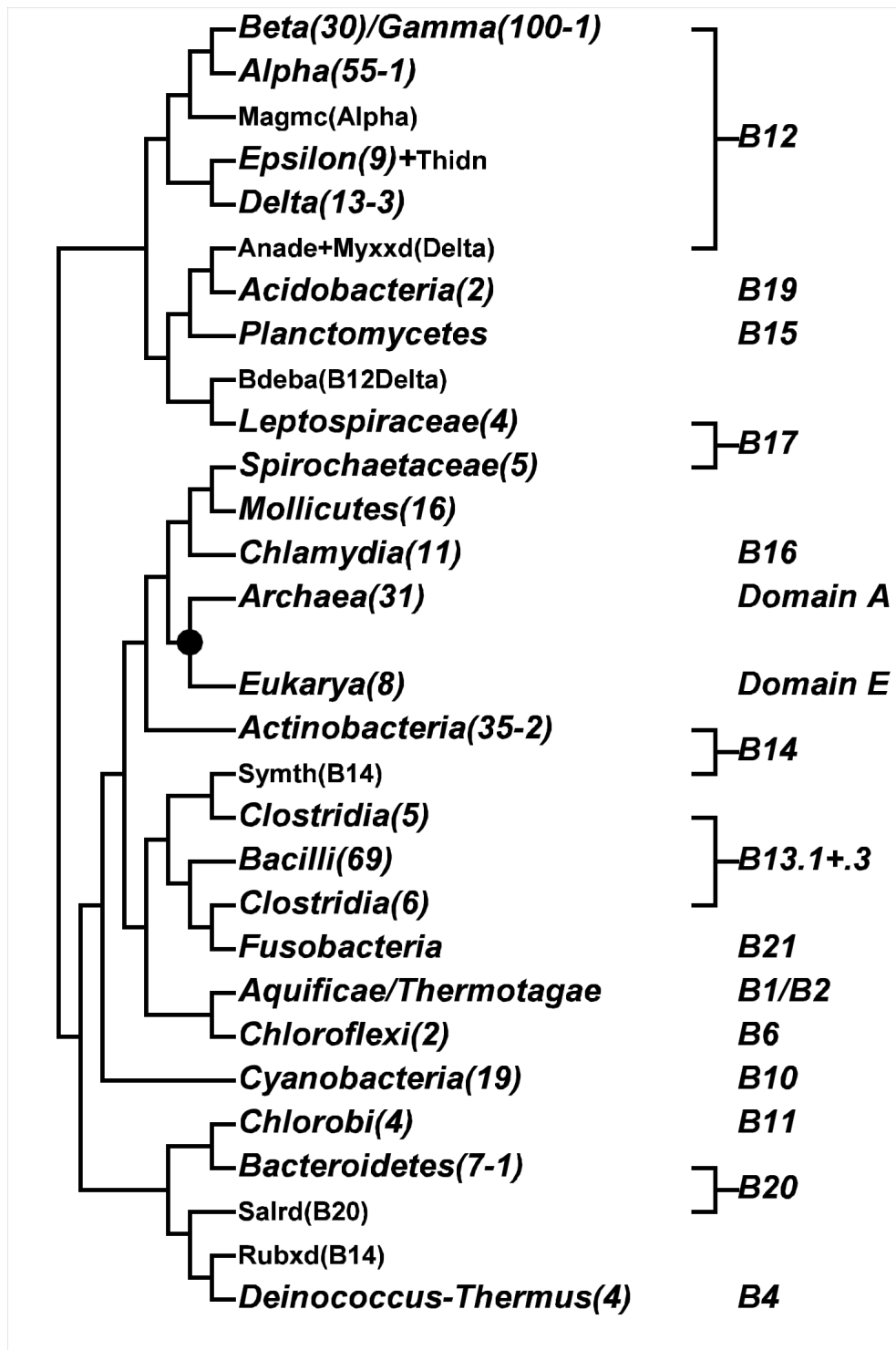


Fig. 1. The highest rank CVTree at K=5. A taxon name represents a monophyletic cluster with the number of organisms given in parentheses. For example, Gamma(100-1) is the cluster of Gammaproteobacteria with 99 organisms, the “outlier” Thidn actually finds its correct placement in the Epsilon group.. Given on the right are the phylum numbers in Bergey’s Outline Rel. 5 (Garrity et al., 2004). The black dot denotes the trifurcation point of the main domains of life. Note that this is an unrooted tree and the branches are not to scale.

2. Comparison of CVTree Phylogeny with Systematic Bacteriology

Recently, we have performed an exhaustive comparison of CVTrees based on 31 Archaea and 401 Bacteria genomes available on 31 December 2006 with biologists' systematics (Gao et al., 2007). According to the Bergey's taxonomy (Garrity et al., 2004) these genomes represent 18 phyla, 35 classes, 79 orders, 120 families, 190 genera and 327 species (We analyzed but do not mention strains here as there is no taxonomic standard at the strain level.) Among this hierarchy there are 145 taxa that contain two or more lower taxa, e.g., 62 genera that contain more than 2 species. These 145 cases were subject to comparison with the CVTrees. It turned out that in 103 cases the phylogeny was consistent with taxonomy and in 42 (29%) cases, some differences were observed. The Gao et al. (2007) and its Supplementary Material described these discrepancies case by case. It is a significant fact that most of these 42 cases have been known to biologists.

Since the submission of Gao et al. (2007) there has appeared a new release of *Taxonomic Outline of Bacteria and Archaea* (abbreviated as TOBA 7.7 below, see Garrity et al., 2007) and more than 200 new prokaryotic genomes have been sequenced. Comparison of CVTrees built on more genomes with newer or alternative taxonomic schemes have removed some more of the 42 discrepant cases. We list a few examples.

1. In the Archaea branch of CVTrees at all $K=3$ to 6 the class *Thermoplasmata* appears in Phylum *Crenarchaeota*. This is a cross-phylum discrepancy in comparison with the Bergey's Manual where the order is listed under *Euryarchaeota*. However, this placement in CVTrees agrees with the scheme given in the book *Five Kingdoms* (Margulis & Schwartz, 1998).
2. In the genus tree representing 31 Archaea species, given in Gao et al. (2007), the species *Aeropyrum pernix* from the order *Desulfurococcales* prevents the order *Thermoproteales* from forming a monophyletic branch; the species *Archaeoglobus fulgidus* from the class *Archaeoglobi* prevents the class *Methanomicrobia* from forming a monophyletic group. However, in our newly produced CVTree for 47 Archaea and 569 Bacteria (unpublished) there are 6 more species in the former group and 3 more species in the latter group. All the above-mentioned orders/classes form monophyletic groups in their own and the whole Archaea branch of the CVTree has reached full agreement with the TOBA 7.7 taxonomy. This is one of the examples of "the more the better" mentioned above.
3. The placement of *Oceanobacillus* was a cross-phylum disagreement with older releases of Bergey's Outline (Garrity et al., 2002), where it was listed under *Proteobacteria*. In all CVTrees it joins other species of Class *Bacilli* of Phylum *Firmicutes*. It was moved to *Firmicutes* in more recent releases of the Outline (Garrity et al., 2003). Being already consistent with Bergey's Outline from 2004

on this case was not counted in the 42 differences, but kept as a historical record.

4. In Outline Rel.5 (Garrity et al., 2004) the genus *Thiomicrospira* contains two species, *T. crunogena* and *T. denitrificans*. However, there was a footnote on page 87: “The identity of *T. denitrificans* is questionable as it belongs within the *Epsilonproteobacteria*.” In our CVTrees *T. denitrificans* appears within *Epsilon* group at all K. In TOBA 7.7 (Garrity et al., 2007) it was renamed *Sulfurimonas denitrificans* and put in *Epsilon* group of *Proteobacteria*.
5. In all CVTrees from K=3 to 6 the four organisms *Synechococcus* sp. WH8102, sp. CC9605, sp. 9902, and sp. CC9311 form a stable monophyletic branch which does not join other *Synechococcus* species but falls into the *Prochlorococcus* cluster. In our recent paper (Gao et al., 2007) we suggested that these organisms should be ascribed to *Prochlorococcus*. In TOBA 7.7 (Garrity et al., 2007) the only listed strain of *Synechococcus*, sp. WH8102, indeed appears under *Prochlorococcus*.
6. In CVTrees at K=5 and 6 the species *Pelodictyon luteolum* falls among the three species from Genus *Chlorobium*, preventing the latter from forming a monophyletic group. It was suggested in Gao et al. (2007) to move *P. luteolum* into Genus *Chlorobium*. Indeed, it is seen in TOBA 7.7 (Garrity et al., 2007) as *Chlorobium puteolum*.

What described above shows the predictive power of the CVTree approach. In fact, our CVTrees indicate or hint on some more taxonomic revisions. The efficiency of the CVTree Web Server is being significantly improved to cope with the situation when 5000 to 6000 prokaryotic genomes will become available in a few years according to “Sequencing the Bergey’s” Project (2007).

3. Discussion

The CVTree approach is not meant to replace the 16S rRNA analysis. Being an independent method, it supports the latter in an overwhelming majority of cases and provides valuable suggestions on taxonomic revisions. When 16S rRNA analysis does not possess enough resolution, for example, in the case of multiple strains of a species, the CVTree method supplies additional information.

The use of complete genomes is both a merit and a demerit of the CVTree approach. It is a merit as no choice of genes is made and even lateral gene transfer is taken into account to some extent. It is a demerit because the number of available complete prokaryotic genomes is always limited. However, with the progress of “Sequencing the Bergey’s” Project this limitation will soon become less severe. With wide taxonomic coverage of selected sequences it may contribute to the establishment of a whole-genome backbone for the prokaryotic branch of the Tree of Life. We mention, in addition, that CVTree method has been tested for protein families and

could yield meaningful results (Wei et al., 2004).

There is good hope that the CVTree method may be applied to such eukaryotes as fungi. As a new and successful approach the foundation of the CVTree method is still being scrutinized, see, e.g., Shi et al. (2007).

Acknowledgements. The authors thank Dr. Ying-Long Qiu for carefully reading the manuscript and making valuable suggestions. This research was partially supported by the National Basic Research Program of China (973 Program) Grant No. 2007CB814800.

References

- Bergey's Manual Trust. 2001 – 2009. *Bergey's Manual of Systematic Bacteriology*, 2nd Ed. Vol. 1-5, New York: Springer-Verlag.
- Chu K H, Qi J, Yu Z-G, Ahn V. 2004. Origin and phylogeny of chloroplasts revealed by a simple correlation analysis of complete genomes. *Molecular Biology and Evolution* 28: 70 – 76.
- Felsenstein J. 1980 – 2008. PHYLIP (Phylogeny Inference Package) version 3.5c. Available from evolution.genetics.washington.edu/phylip.html [accessed 23 January 2008].
- Gao L, Qi J. 2007. Whole genome molecular phylogeny of large dsDNA viruses using composition vector method. *BMC Evolutionary Biology* 7: 41.
- Gao L, Qi J, Sun J-D, Hao B-L. 2007. Prokaryote phylogeny meets taxonomy: An exhaustive comparison of composition vector trees with systematic bacteriology. *Science in China Ser. C Life Sciences* 50: 587 – 599.
- Gao L, Qi J, Wei H-B, Sun Y-G, Hao B-L. 2003. Molecular phylogeny of Coronaviruses including human SARS-CoV. *Chinese Science Bulletin* 48: 1170 – 1174.
- Garrity G M, Johnson K L, Bell J, Searles D B. 2002. Taxonomic Outline of Prokaryotes. *Bergey's Manual of Systematic Bacteriology*. 2nd Ed. Rel. 3.0. Springer-Verlag. DOI: 10.1007/bergeysoutline200210 [online]. Available from 141.150.157.80/bergeysoutline/main.htm [accessed 23 January 2008].
- Garrity G M, Bell J A, Lilburn T G. 2003. Taxonomic Outline of Prokaryotes. *Bergey's Manual of Systematic Bacteriology*. 2nd Ed. Rel. 4.0. Springer-Verlag. DOI: 10.1007/bergeysoutline200310 [online]. Available from 141.150.157.80/bergeysoutline/main.htm [accessed 23 January 2008].
- Garrity G M, Bell J A, Lilburn T G. 2004. Taxonomic Outline of Prokaryotes. *Bergey's Manual of Systematic Bacteriology*. 2nd Ed. Rel. 5.0. Springer-Verlag. DOI: 10.1007/bergeysoutline200405 [online]. Available from 141.150.157.80/bergeysoutline/main.htm.
- Garrity G M, Lilburn T G, Cole J R, Harrison S H, Enzeby J, Tindall B J. 2007. Taxonomic Outline of Bacteria and Archaea (TOBA). Rel 7.7, 6 March 2007. Michigan State University [on-line]. Available from www.taxonomicoutline.org

- [accessed 23 January 2008]
- Hao B-L, Qi J. 2004. Prokaryote phylogeny without sequence alignment: from avoidance signature to composition distance. *Journal of Bioinformatics and Computational Biology* 2: 1 – 19.
- Hao B-L, Qi J, Wang B. 2003. Prokaryote phylogeny based on complete genomes without sequence alignment, *Modern Physics Letters B*17: 91 – 94.
- Margulis L, Schwartz K V. 1998. *Five Kingdoms. An Illustrated Guide to the Phyla of Life on Earth.* 3rd Ed. San Francisco: W H Freeman.
- Pruitt K D, Tatusova T, Maglott D R. 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nuclear Acids Research* 35, Database Issue: D61 – D65.
- Qi J, Wang B, Hao B-L. 2004a. Whole genome prokaryote phylogeny without sequences alignment: a K-string composition approach. *Journal of Molecular Evolution* 58: 1 – 11.
- Qi J, Luo H, Hao B-L. 2004b. CVTree: a phylogenetic tree reconstruction tool based on whole genomes. *Nuclear Acids Research* 32. Web Server Issue: W45 – W47.
- Sequencing the Bergey's Project (2007). Available from www.sequencingbergeys.org [accessed 15 September 2007]
- Shi X-L, Xie H-M, Zhang S-Y, Hao B-L. 2007. Decomposition and reconstruction of protein sequences: the problem of uniqueness and factorizable language. *Journal of Korean Physical Society* 50: 118 – 124.
- Wei H-B, Qi J, Hao B-L. 2004. Prokaryote phylogeny based on ribosomal proteins and aminoacyl tRNA synthetases by using the composition distance approach. *Science in China Ser. C Life Sciences (中国科学C辑.生命科学)* 47: 313 – 321.
- Woese C R. 2000. Prokaryote systematics: the evolution of a science. In Balows A, Trupper H G, Dworkin M, Harder W, Schleifer K H, eds. *The Prokaryotes*. Vol. 3. New York: Springer-Verlag.
- Woese C R. 2002. On the evolution of cells. *Proceedings of National Academy of Science USA* 99: 8742 – 8747.
- Woese C R, Fox G E. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of National Academy of Science USA* 74: 5088 – 5090.