Prokaryote phylogeny meets taxonomy: An exhaustive comparison of composition vector trees with systematic bacteriology

GAO Lei1,2, QI Ji1**, SUN JianDong2,3 & HAO BaiLin1,2,4†

1 Institute of Theoretical Physics, Chinese Academy of Sciences, Beijing 100080, China;
2 T-Life Research Center, Fudan University, Shanghai 200433, China;
3 Beijing Genomics Institute, Chinese Academy of Sciences, Beijing 101300, China;
4 Santa Fe Institute, Santa Fe, NM87501, USA

We perform an exhaustive, taxon by taxon, comparison of the branchings in the composition vector trees (CVTrees) inferred from 432 prokaryotic genomes available on 31 December 2006, with the bacteriologists’ taxonomy—primarily the latest online Outline of the Bergey’s Manual of Systematic Bacteriology. The CVTree phylogeny agrees very well with the Bergey’s taxonomy in majority of fine branchings and overall structures. At the same time most of the differences between the trees and the Manual have been known to biologists to some extent and may hint at taxonomic revisions. Instead of demonstrating the overwhelming agreement this paper puts emphasis on the biological implications of the differences.

composition vector method, CVTree, Bergey’s taxonomy, prokaryote phylogeny, prokaryote taxonomy, 16S rRNA

Prokaryotic taxonomy has been a long-standing problem. Ever since the time of Linnaeus, classification of organisms has been based on morphological features and emphasis has been put more on grouping similar species than on pursuing their evolutionary relationship. It was the proposal by Zuckerkandl and Pauling[1] to use conserved protein sequences as evolutionary “clocks” that put phylogeny and taxonomy on molecular basis. This approach has been more or less successful for various eukaryotic taxa. A few years ago the National Science Foundation of USA launched the AToL (Assembling the Tree of Life) project (see the project’s homepage: atol.sdsc.edu) that aimed at reconstructing the evolutionary origins of all living organisms. Yet ironically, in a recent Science paper[2] addressed on building the Tree of Life, the branch for the most numerous organisms on Earth, namely the prokaryotic branch, was lacking.

The retardation of prokaryotic phylogeny is not incidental. There are too few morphological features available for traditional phylogenetic reconstructions. No appropriate “clocks” were discovered for prokaryote molecular phylogeny until Carl Woese and collaborators proposed to use the small subunit ribosomal RNA (16S rRNA in prokaryotes) as a molecular clock[3]. The 16S rRNA trees have been so successful that the new edition of the Bergey’s Manual of Systematic Bacteriology[4] (hereafter abbreviated as the Bergey’s Manual) follows “a phylogenetic framework based on analysis of the nucleotide sequence of the small ribosomal subunit RNA, rather than a phenotypic structure” (Garrity’s Preface to ref. [4]).

However, phylogeny and taxonomy are not synonyms. A correct taxonomy should agree with phylogeny in major and finer branchings. And, a faithful phylogeny should reflect the evolutionary history of species. Being
1 Material and method

1.1 Complete genomes

We fetched all the 432 complete prokaryotic genomes available from the National Center for Biotechnological Information (NCBI) ftp-site (ftp://ftp.ncbi.nih.gov/Genomes/Bacteria) on 31 December 2006. These are the sequences with accession numbers prefixed by NC. The complete collection of translated amino acid sequences of an organism, i.e., the .faa files, is used. These sequences have the merit of being curated by the NCBI staff and thus may provide a common background for comparison. Eight eukaryotic genomes are included as outgroup. All organism names, their abbreviations and accession numbers are listed in the Supplementary Material[12]. We did not include plasmid sequences and other extrachromosomal elements.

1.2 The CVTree method

Since the CVTree method has been described in our previous publications[9–11] we only give a brief account here.

In the CVTree method each organism is represented by a Composition Vector whose components correspond to the numbers of various (overlapping) K-peptides (for a fixed K) in all the translated amino acid sequences from an organism’s genome. The numbers are modified by subtracting a statistical background to highlight the role of selective evolution. The subtraction procedure is based on a (K–2)-th order Markov prediction and therefore the minimal K starts from 3. Using the distance matrix thus obtained, a neighbor-joining tree is produced by a standard program in the Phylip package (http://evolution.genetics.washington.edu/phylip.html).

We have reconstructed all CVTrees from K = 3 to 6. These trees are given in the Supplementary Material[12]. The peptide length K, though looking like a parameter, is a measure of resolution of the method. The prokaryote CVTrees constructed over the years from 72 genomes at first to 432 genomes in the present work have shown that K = 5–6 yields the best result in the sense of consistency with the biologists’ taxonomy. This agrees with biologist’s view that “six amino acids are sufficient” for identification of a protein[13]. The justification of the CVTree approach is still under way, see, e.g., ref. [14].

1.3 Prokaryote taxonomic reference

There is no official standard for prokaryotic taxonomy.
However, the classification scheme in Bergey’s Manual of Systematic Bacteriology\textsuperscript{14} is now widely accepted by microbiologists as the best approximation to an official classification\textsuperscript{15}. The new edition of Bergey’s Manual is based on 16S rRNA phylogeny as well as on classical morphological and physiological observations\textsuperscript{15}. The Bergey’s taxonomy is somewhat more conservative in that it relies on cultured type strains whereas only a tiny fraction of prokaryote species can be grown in culture.

In this paper and its Supplementary Material\textsuperscript{12} a lineage in the Bergey’s Manual or its online Outline\textsuperscript{16} is abbreviated as B13.3.2.6.2 or B_{13}C_{3}O_{2}F_{6}G_{2} for Phylum BXIII (Firmicutes), Class III (Bacilli), Order II (Lactobacillales), Family VI (Streptococcaceae), Genus II (Lactococcus). We call this a “Bergey’s code”\textsuperscript{19}. It must be noted, however, that the Bergey’s code is merely a convenient shorthand which may change with each new edition of the Bergey’s Manual. The current code corresponds to Rel. 5.0 of the online Outline\textsuperscript{16}.

The National Center for Biotechnological Information (NCBI) provides a Taxonomic Browser (http://www.ncbi.nlm.nih.gov/Taxonomy). Though containing a disclaimer not being “an authoritative source for nomenclature or classification”, the NCBI taxonomy is more dynamic and up-to-date. We refer to NCBI taxonomy especially when it differs from Bergey’s and speaks in favor of the CVTree phylogeny. Sometimes we refer to the taxonomic list at European Bioinformatics Institute (EBI) (http://www.ebi.ac.uk/genomes/bacteria.html) for additional information. Occasionally we also refer to some other taxonomy, e.g., that reflected in the book Five Kingdoms\textsuperscript{17}.

## 2 Comparison of CVTrees with taxonomy

We have performed an exhaustive, taxon by taxon, comparison of CVTree phylogeny with the latest Bergey’s taxonomy from strains and species up to classes and phyla. A detailed analysis is given for the Archaea branch in the next subsection. Similar details for Bacteria are provided in the Supplementary Material\textsuperscript{12} and only a summary is given in the subsequent sections.

In making comparison of the branches of a phylogenetic tree with taxonomy, one should clearly bear in mind that taxonomic ranks such as phylum, class, order, family, genus and species are invented by human being. Only the two extremes, e.g., the clustering of species in a genus and the grouping of all lower taxa into a higher taxon such as phylum or class, make more sense (Charles Darwin mentioned repeatedly the species, genera and families in his Origin of Species, but rarely referred to higher taxa). Accordingly, two guiding principles are followed in our analysis. At the strain and species level we examine whether the members get or stay together with K increasing; we call this “convergence”.

At the high end we check whether the subordinate members form a monophyletic cluster under the highest taxon, taking mutual positions of the members as a secondary factor.

### 2.1 Three domains of life

The discovery of three domains of life on the Earth by Carl Woese and collaborators\textsuperscript{18} is a significant progress in understanding the living world. It serves as a touchstone whether in a phylogeny the three domains of life are unambiguously resolved. In Table 1 we show the clustering of the 440 genomes in CVTrees into the three domains. The convergence with K increasing shows off clearly.

<table>
<thead>
<tr>
<th>K=3</th>
<th>K=4</th>
<th>K=5</th>
<th>K=6</th>
</tr>
</thead>
<tbody>
<tr>
<td>7E</td>
<td>8E</td>
<td>8E</td>
<td>8E</td>
</tr>
<tr>
<td>1E in B</td>
<td>25A in B</td>
<td>25A in B</td>
<td>31A</td>
</tr>
<tr>
<td>1A(Arcfu) in B</td>
<td>1A(Methj) in B</td>
<td>1A(Naneq) in B</td>
<td>1A(Naneq) in B</td>
</tr>
<tr>
<td>4A in B</td>
<td>4A in B</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 1** Clustering of the 440 genomes into 3 domains

**Abbreviations:** A, Archaea; B, Bacteria; E, Eukarya.

However, there is a minor proviso concerning the bacterial endosymbiont Carsonella ruddii\textsuperscript{19}, see subsection 2.3.7 on the placement of higher taxa.

### 2.2 Analysis of the Archaea branch

We have performed a comprehensive analysis of the Archaea branch of CVTree made of 31 species to show the way how the comparison was carried out. The 31 Archaea genomes are listed in the Supplementary Material\textsuperscript{12}. Their taxonomic distribution is given in Table 2.

<table>
<thead>
<tr>
<th>Phylum</th>
<th>Class</th>
<th>Order</th>
<th>Family</th>
<th>Genus</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>A2</td>
<td>8</td>
<td>9</td>
<td>12</td>
<td>18</td>
<td>23</td>
</tr>
<tr>
<td>A3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>10</td>
<td>13</td>
<td>17</td>
<td>23</td>
<td>31</td>
</tr>
</tbody>
</table>

**Table 2** Taxonomic distribution of the 31 Archaea
Among the 23 genera 5 contain more than one species. Among the 17 families 4 contain more than one genus. Among the 10 classes 3 contain more than one family. Among the 3 phyla only one contains more than one class. These numbers are based on the Bergey’s taxonomy. We have only referred to the Bergey’s taxonomy so far. Now comes comparison with the phylogenetic trees. By inspecting all the $K = 3$ to 6 CVTrees we see that at the species level:

(1) There are two genera containing two species: *Pyrobaculum* and *Thermoplasma*. The species in the corresponding genus always stay together for $K = 3$ to 6. These genera are denoted as *Pyrobaculum*(II) and *Thermoplasma*(II), respectively in the trees. We use Roman numerals to denote the number of species in a genus;

(2) There are three genera containing three species: *Sulfolobus*, *Methanosarcina*, and *Pyrococcus*. The species in the corresponding genus always stay together for $K = 3$ to 6. These genera are labeled as *Sulfolobus*(III), *Methanosarcina*(III), and *Pyrococcus*(III), respectively in the trees.

When there are three or more lower taxa in a taxon their mutual relationships are also worth scrutinizing. The *Pyrococcus* genus appears as (Pyrfu, (Pyrab, Pyrho)) for all $K$. The *Sulfolobus* appears as (Sulso, (Sulac, Sulto)) at $K = 3, 5, 6$, but as (Sulac, (Sulso, Sulto)) at $K = 4$. The *Methanosarcina* genus appears as (Metbf, (Metac, Metma)) at $K = 4, 5, 6$ with (Metma, (Metac, Metbf)) at $K = 3$.

In general, we see a “convergence” with $K$ increasing: $K = 4$ is better than $K = 3$ although with sporadic fluctuations, $K = 6$ is identical or better than $K = 5$. By “better” we mean it is closer to the standard taxonomy. Throughout this paper this convergence will show off repeatedly at different taxonomic levels.

Using the abbreviations introduced above, the Ar-

---

**Figure 1**  The genus trees for the Archaea branch taken from the 440-genome CVTrees at $K = 3$ to 6.
chaea branch of the CVTrees at \( K = 3 \) to 6 is given in Figure 1. These are genus trees as their 23 leaves correspond to the 23 genera, see Table 2.

The three families that contain two genera are:

1. Methanobacteriaceae containing Methanosphaera and Methanobacterium;
2. Methanosarcinaceae containing Methanosarcina(III) and Methanococcoides;
3. Thermococcaceae containing Thermococcus and Pyrococcus(III).

They all converge from \( K = 3 \) to 6.

There is one family \( A_2C_5F_1 \) (Halobacteriaceae) representing the whole class \( C_4 \) for the time being that contains four genera: Haloarcula (Halma), Natronomonas (Natpd), Halobacterium (Halsa), and Haloquadra tum (Halwd). They converge at all \( K \) with slight variation at \( K = 4 \). Even when this family disunes from the main Archaea cluster at \( K = 3 \) and 4 they stay together, denoted as “4A in B” in Table 1. The species Halwd has not been listed in the Bergey’s Manual yet, but its belonging to this family is evident in all CVTrees.

Therefore, the CVTrees at \( K = 5 \) and 6 agree with the Bergey’s taxonomy at all the species, genus and family levels. In order to compare the ordering of higher taxa we redraw the \( K = 5, 6 \) trees in Figure 2 using the Bergey’s codes for higher taxa. The relative positions of leaves are the same as that in Figure 1. It is readily seen from Figures 1 and 2 that there are only three discrepancies with the Bergey’s taxonomy:

1. A seemingly cross-phylum discrepancy: the class \( A_2C_5 \) (Thermoplasmata) from the phylum \( A_2 \) (Euryarchaeota) is stably clustered into the phylum \( A_1 \) (Crenarchaeota) in all trees from \( K = 3 \) to 6. However, this placement agrees with the classification scheme of some biologists, given, e.g., in Five Kingdoms\textsuperscript{[17]}. Therefore, it does not make a real problem in the CVTrees.

![Figure 2](image-url)  
**Figure 2**  
Higher taxa trees at \( K = 5, 6 \) for the Archaea using Bergey’s codes.
(2) The placement of Aerpe prevents $A_1C_1O_1$ to form a monophyletic cluster.

(3) The placement of Arcfu prevents $A_2C_7$ to form a monophyletic cluster.

The last two cases hint at possible taxonomic revision of Aerpe and Arcfu. For example, assigning Aerpe to $O_1F_7$ and Arcfu to $C_3$ would resolve the problem.

An overall inspection of the trees from $K = 3$ to 6 shows the good convergence with $K$. At $K = 3$ and 4 the 31 Archaea do not form a monophyletic cluster. They form a monophyletic cluster at $K = 5$ and 6. However, the newly discovered phylum Nanoarchaeota, represented by the only species Nanoq, which has not been listed in the Bergey’s taxonomy, does not appear as a separate phylum at $K = 5$; it does so at $K = 6$. The class $A_3C_4$ (Halobacteria), being a monophyletic clade at all $K$, reflected as (4A in B) in Table 1, joins the phylum Euryarchaeota finally at $K = 6$, thus the $K = 6$ tree supports the designation of Nanoq to a new phylum.

In total, the detailed analysis of the 31-genome Archaea branch of CVTrees with the Bergey’s Manual only reveals problematic taxonomic assignments of Aerpe and Arcfu.

### Table 3  Taxonomic distribution of the 401 bacterial genomes

<table>
<thead>
<tr>
<th>Phylum</th>
<th>Class</th>
<th>Order</th>
<th>Family</th>
<th>Genera</th>
<th>Species</th>
<th>Strain</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1 (Aquificae)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>B2 (Thermotogae)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>B4 (Deinococcus-Thermus)</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>B6 (Chloroflexi)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>B10 (Cyanobacteria)</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>15</td>
<td>19</td>
</tr>
<tr>
<td>B11 (Chlorobi)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>B12 (Proteobacteria)</td>
<td>5</td>
<td>33</td>
<td>53</td>
<td>99</td>
<td>157</td>
<td>208</td>
</tr>
<tr>
<td>B13 (Firmicutes)</td>
<td>3</td>
<td>7</td>
<td>14</td>
<td>22</td>
<td>58</td>
<td>96</td>
</tr>
<tr>
<td>B14 (Actinobacteria)</td>
<td>3</td>
<td>9</td>
<td>16</td>
<td>16</td>
<td>31</td>
<td>35</td>
</tr>
<tr>
<td>B15 (Planctomycetes)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>B16 (Chlamydia)</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>7</td>
<td>11</td>
</tr>
<tr>
<td>B17 (Spirochaetes)</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>7</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>B19 (Acidobacteria)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>B20 (Bacteroidetes)</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>B21 (Fusobacteria)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>15</td>
<td>25</td>
<td>66</td>
<td>103</td>
<td>167</td>
<td>296</td>
</tr>
</tbody>
</table>

### Table 4  Number of taxa contained in an higher taxon (domain Bacteria)

<table>
<thead>
<tr>
<th>i =</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>13</th>
<th>$\Sigma_{i=1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strains/Species</td>
<td>242</td>
<td>31</td>
<td>14</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>54</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Species/Genus</td>
<td>110</td>
<td>30</td>
<td>9</td>
<td>8</td>
<td>5</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>34</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genera/Family</td>
<td>69</td>
<td>23</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Families/Order</td>
<td>41</td>
<td>16</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Orders/Class</td>
<td>15</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Classes/Phylum</td>
<td>11</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>488</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>184</td>
<td></td>
</tr>
</tbody>
</table>
denoting the \(i\)-th number in the \(j\)-th row of Table 4 by \(n^j_i\), the sums \(a^j = \sum_{i \geq 1} n^j_i\) and \(b^j = \sum_{i \geq 1} i \times n^j_i\) satisfy the relation \(a^{j-1} = b^j\) for \(j = 1\) to 6 with \(a^0\) being the number of all bacterial strains, i.e., \(a^0 = 401\) in this paper. In fact, these \(a^j\)-s appear in the last row of Table 3. For details see the Supplementary Material[12].

The total number of taxa that contains two or more lower taxa is given in the last column of Table 4. Put in words, there are 54 species that contain two or more strains, 57 genera that contain two or more species, 34 families that contain two or more genera, etc. In total, there are 184 such taxa. We emphasize that all the numbers in Tables 3 and 4 are produced by fitting the 401 genomes into the Bergey’s taxonomy without reference to any phylogeny.

Now we compare the CVTrees from \(K = 3\) to 6 with the taxonomy and check how these 184 cases show themselves as branching points. The comparison is exhaustive in the sense that all the cases are analyzed without ignoring any exceptions. It turns out that among the 184 cases 138 completely agree with the branchings in the trees and 46 taxa reveal some differences, see Table 5 for statistics at various taxonomic levels. At low taxonomic ranks such as species, genera and families the agreement is overwhelming. It is natural that at higher ranks (orders, classes and phyla) there are relatively more differences; even taxonomists usually disagree on placement of higher taxa. As seen from the detailed description in the Supplementary Material[12] these 46 cases include some really minor ones. It is a remarkable fact that most of the 46 differences between CVTree and Bergey’s taxonomy have been known to biologists to some extent or correspond to disagreement between different taxonomic schools.

2.3.1 Trivial cases. Various phyla are represented unevenly in the dataset. Phyla B1, B2, B15 and B21 are represented only by one species. Phyla B6, B11, B16, B19 and B20 show simple agreement with CVTrees at all lower taxonomic ranks as long as a limited number of genomes are available. All these phyla enter our discussion only when it comes to deal with mutual position of phyla in a tree. We leave the description of these trivial cases to the Supplementary Material[12]. In what follows we present a brief summary of comparison for five phyla: the *Cyanobacteria* (B10), the *Proteobacteria* (B12), the *Firmicutes* (B13), the *Actinobacteria* (B14), and the *Spirochaetes* (B17).

### Table 5 Summary of comparing taxonomy with CVTrees

<table>
<thead>
<tr>
<th>Case</th>
<th>Number of taxa contained</th>
<th>(i&gt;1) cases compared with CVTrees</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strains/Species</td>
<td>242</td>
<td>54 47 7</td>
</tr>
<tr>
<td>Species/Genus</td>
<td>110</td>
<td>57 46 11</td>
</tr>
<tr>
<td>Genera/Family</td>
<td>69</td>
<td>34 26 8</td>
</tr>
<tr>
<td>Families/Order</td>
<td>41</td>
<td>25 14 11</td>
</tr>
<tr>
<td>Orders/Class</td>
<td>15</td>
<td>10 5 5</td>
</tr>
<tr>
<td>Classes/Phylum</td>
<td>11</td>
<td>4 0 4</td>
</tr>
<tr>
<td>Total</td>
<td>488</td>
<td>184 138 46</td>
</tr>
</tbody>
</table>

2.3.2 Phylum B10 (*Cyanobacteria*). The 19 organisms and their convergence in the CVTrees at strain and species levels are listed in Table 6. The CVTrees for \(K = 3\) — 6 are given in Figure 3. We see that at \(K = 4, 5\) and 6, the 19 organisms do form a monophyletic cluster, a fact indicating the correctness of putting them in one and the same phylum.

### Table 6 The 19 organisms in Phylum B10

<table>
<thead>
<tr>
<th>Bergey’s taxonomy</th>
<th>CVTrees</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class</td>
<td>Subsection</td>
</tr>
<tr>
<td>C 1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>6</td>
</tr>
</tbody>
</table>

The problem in classifying the *Cyanobacteria* may be seen from the difference in Bergey’s taxonomy and NCBI taxonomy. In both Bergey’s taxonomy and NCBI taxonomy there is only one Class *Cyanobacteria*. However, they differ at the next taxonomic rank. In Bergey’s taxonomy there are 5 unnamed Subsections. In the NCBI taxonomy there are 7 named Orders among which *Chroococcales* corresponds to Subsection I, *Oscillatoriales* corresponds to Subsection III and *Nostocales* corresponds to Subsection IV. In addition, there are new orders such as *Gloeobacterales* and *Prochlororales*. The CVTrees may help to revise the taxonomy.

Many of the problems come from the *Prochlorococcus* species. These smallest known photosynthetic bacte-
ria were discovered in the late 1980s. They are abundant in oceans and play a substantial role in global carbon cycle. In the Bergey’s taxonomy *P. marinus* belongs to Class *Cyanobacteria* Subsection I Family 1.1 Form Genus XI without special names at the intermediate ranks. In the NCBI taxonomy *P. marinus* belongs to class *Cyanobacteria* order *Prochlorales* family *Prochlorococcaceae*, i.e., a whole new lineage has been introduced. As of 31 December 2006 complete genomes from 5 ecotypes of *P. marinus* were published (Table 7).

Although the names of these 5 ecotypes look like different strains in one same species, treating them as different species does not lead to any problem in the CVTrees. The stable grouping (Promp, Prom9) in all CVTrees from $K = 3$ to $6$ agrees with these two highlighted adapted ecotypes being evolutionarily closer (see Table 7). We denote them as Prom9(II). The grouping (Proma, Promt) in CVTrees for $K = 4$ to $6$ with an exception at $K = 3$ also agrees with the two ecotypes eSS120 and eNATL2A being closer. We denote these two as Promt(II). These shorthands are used in Figure 3.

As regards the difference between the CVTrees and the Bergey’s taxonomy, we mention the following:

1. The 4 strains of *Synechococcus* sp. (Synpx(4)) do not get into the designated genus *Synechococcus*. Instead Synpx(4) joins the *Prochlorococcus* species Promm. In fact, these strains live in the same habitat as Promm and are considered “potential competitor” of the latter. The grouping (Promm, Synpx(4)) persists from $K = 3$ to $6$. It is reasonable to recognize Synpx(4) strains as members of *Prochlorales*.

2. Although *Thermosynechococcus elongatus* has not been listed in the Bergey’s Outline its lineage may be determined up to the genus as $B_{10}C_{1}O_{1}F_{1}G_{1}$ from the CVTrees. Apparently it is closer to *Synechococcus elongatus*.

3. The Bergey’s designation of the genera *Gloeobacter* and *Prochlorococcus* to the same family as *Synechococcus* is not justified in the CVTrees. The NCBI taxonomy putting *Prochlorococcus* in the order *Prochlorales* and *Gloeobacter* in the order *Gloeobacteriales* might be more reasonable than the Bergey’s.

### Table 7 The five ecotypes in the genus *Prochlorococcus*

<table>
<thead>
<tr>
<th>Ecotype</th>
<th>Name</th>
<th>Abbr.</th>
<th>Remark</th>
</tr>
</thead>
<tbody>
<tr>
<td>eMIT9312</td>
<td><em>P. marinus</em> str. MIT 9312</td>
<td>Prom9</td>
<td>near surface, high-light adapted</td>
</tr>
<tr>
<td>eMED4</td>
<td><em>P. marinus</em> MED4</td>
<td>Promp</td>
<td>as above</td>
</tr>
<tr>
<td>eSS120</td>
<td><em>P. marinus</em> subsp. marinus marinus str. CCMP1375</td>
<td>Proma</td>
<td>deep water, low-light adapted</td>
</tr>
<tr>
<td>eNATLA2</td>
<td><em>P. marinus</em> NATL2A</td>
<td>Promt</td>
<td>as above</td>
</tr>
<tr>
<td>eMIT9313</td>
<td><em>P. marinus</em> str. 9313</td>
<td>Promm</td>
<td>as above</td>
</tr>
</tbody>
</table>

**Figure 3** Convergence of all the 19 organisms in Phylum *Cyanobacteria*. For abbreviations of species in *Prochlorococcus* see Table 7.

### 2.3.3 Phylum B12 (*Proteobacteria*). The Phylum B12 is represented by 208 organisms in the dataset. The Ber-
Bergey’s taxonomy divides this phylum into 5 classes/groups. We discuss these groups one after another.

The Alpha group is represented by 55 genomes. The CVTrees converge well at the strain, species, genus and family levels, so we only look at the orders. According to the Bergey’s Outline the 55 organisms come from 6 orders and there is a newly sequenced species Magnetococcus MC-1 (Magmc) which has not been listed in the Bergey’s. The CVTrees in terms of orders are given in Figure 4. This is yet another good example of convergence of the CVTrees. The 6 orders form a monophyletic cluster at $K = 4, 5$ and 6. These trees hint at the possibility of assigning a new order to Magmc within the Alpha group.

**Figure 4** Convergence of orders in the Alphaproteobacteria group with $K$ increasing. Note that at $K = 4$–6 the six orders defined in the Bergey’s Manual form a monophyletic cluster. The newly sequenced Magnetococcus MC-1 is abbreviated as Magmc.

The 30 organisms in the Beta group converge well at the strain, species, genus and family levels. The 6 orders...
in the group form a monophyletic cluster for \( K = 5 \) and 6, with only one order standing out at \( K = 4 \) and a more scattered placement at \( K = 3 \). See Figure 5 for the order convergence with \( K \).

There are 101 organisms in the dataset being assigned to the Gamma group. All but two organisms converge well to a monophyletic cluster. Leaving the detailed analysis to the Supplementary Material\(^{[12]}\), we concentrate on a feature common to the 16S rRNA trees and the CVTrees. Woese and coworkers observed in their study of 16S rRNA trees, “The Beta subdivision is actually a highly divergent branch within Gamma, and these two together show a sister relationship to Alpha”\(^{[22]}\). This happens in all CVTrees as well. However, we now have more to say on this point. How the Gamma group is separated by the Beta group is clearly seen from the split of the order Enterobacteriales, represented by the family Enterobacteriaceae. This is one of the most studied bacterial families. In all CVTrees for \( K = 3 - 6 \) this family is divided by the Beta group into two subgroups with sharp contrast in their genome size. The upper subgroup consists of 28 organisms from 7 genera. They always form a monophyletic cluster with minor variations of internal relations, see Table 8 for the minimal genome size in each genus.

Table 8 The smallest genome size in the upper group of Enterobacteria

<table>
<thead>
<tr>
<th>Species</th>
<th>Abbr.</th>
<th>Genome size (bp)</th>
<th>Gene number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Escherichia coli</td>
<td>EcoliK</td>
<td>4 639 675</td>
<td>4 237</td>
</tr>
<tr>
<td>Erwinia carotovora</td>
<td>Erwt</td>
<td>5 064 019</td>
<td>4 472</td>
</tr>
<tr>
<td>Photobacter luminescens</td>
<td>Phll</td>
<td>5 688 987</td>
<td>4 683</td>
</tr>
<tr>
<td>Salmonella enterica</td>
<td>Salpa</td>
<td>4 585 229</td>
<td>4 093</td>
</tr>
<tr>
<td>Shigella dysenteriae</td>
<td>Shids</td>
<td>4 369 232</td>
<td>4 274</td>
</tr>
<tr>
<td>Sodalis glossinidius</td>
<td>Sodgl</td>
<td>4 171 146</td>
<td>2 432</td>
</tr>
<tr>
<td>Yersinia pestis</td>
<td>Yerp</td>
<td>4 534 590</td>
<td>3 981</td>
</tr>
</tbody>
</table>

All the 8 organisms in the lower group of Enterobacteriaceae are endocellular symbionts of insects. Their genomes have undergone a reductive evolution and now have very small size, see Table 9 for the maximal genome size in each genus. Therefore, it is a common problem of CVTrees and the 16S rRNA trees that they could not distinguish early evolved genomes from those resulted by reductive evolution simply from their being located on lower part of a branch.

Among the 13 organisms of the Delta Group there is an unquestionable monophyletic core made of 10 leaves. The only representative Bdeba of the order Bdellovibrionales and 2 species Anade and Myxxd from the order Myxococcales are outliers at all \( K = 3 \) to 6. For Bdeba we cannot say anything until more related genomes become available. For the latter case we note that the taxonomic position of the Myxococcales has been a long-standing problem. Some years ago there was even suspicion that these species might not belong to bacteria at all, see ref. \([23]\). This situation is denoted as Delta (13-3) in Figure 6.

Table 9 The largest genome size in the lower group of Enterobacteria

<table>
<thead>
<tr>
<th>Species</th>
<th>Abbr.</th>
<th>Genome size (bp)</th>
<th>Gene number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baumannia cicadellinicola</td>
<td>Bauch</td>
<td>686 194</td>
<td>595</td>
</tr>
<tr>
<td>Clostridium kelloggii</td>
<td>Blobp</td>
<td>791 654</td>
<td>610</td>
</tr>
<tr>
<td>Buchnera aphidicola</td>
<td>Bucap</td>
<td>641 454</td>
<td>546</td>
</tr>
<tr>
<td>Wigglesworthia brevipalpis</td>
<td>Wigbr</td>
<td>697 724</td>
<td>611</td>
</tr>
</tbody>
</table>

For the Epsilon group we note only that T. denitrificans from the Gamma group firmly joins this group at all \( K = 3 \) to 6, supporting the observation of Bergey’s Outline that we cite in subsection 2.3.7. 2.3.4 Phylum B13 (Firmicutes). The phylum B13 is represented by 96 organisms in our dataset. The Firmicutes are a very diverse group. “Hugenholtz recognizes at least four other phyla within the Firmicutes.” (Footnote 3 on p. 2 of the Bergey’s Outline\(^{[16]}\)). However, two of its three classes, namely Bacilli and Mollicutes, form close but separate monophyletic groups in the CVTrees at \( K = 5 \) and 6. The class Clostridia splits into two parts at \( K = 5 \) but combines together and joins the Bacilli at \( K = 6 \). Several species in B13 have many strains sequenced. For example, there are 9 strains in Staphylococcus aureus, 11 strains in Streptococcus pyogenes. Due to the high resolution power of the CVTree method these species provide a nice chance to explore the evolutionary order of strains within a species. In fact, the internal relationship of strains converges well, see the Supplementary Material\(^{[12]}\) for details. In view of the strain convergence within species in the majority of cases in all phyla including B13, a few non-convergent cases call for special attention. For example, the 3 strains of Bacillus cereus as well as the 4 strains in Chlamydomphila pneumoniae (from B16) change their mutual position with varying \( K \). Whether these are caused by rapid variability of strains requires further investigation.

We also note that a new genus Oceanobacillus was
introduced into Bergey’s Manual in Ver. 3 (July 2002) of the Outline under Gammaproteobacteria in B12. However, in all CVTrees it groups with other members of Bacillaceae (B13) from the outset. The Bergey’s Outline has moved it to B13 since Ver. 4 (October 2003). Therefore, this case is no longer considered a difference between the CVTrees and the Bergey’s taxonomy.

2.3.5 Phylum B14 (Actinobacteria). In the first edition of the Bergey’s Manual of Systematic Bacteriology (1986) the Actinobacteria were designated an order within the phylum of Firmicutes. In the forthcoming second edition of the Bergey’s Manual they are promoted to a new phylum. There are 35 organisms assigned to phylum B14 in the dataset and all but two converge well in the CVTrees at the strain, species, genus and family levels, justifying the correctness of the establishment of a new phylum for the Actinobacteria[^24]. For the two outliers see the Supplementary Material[^12].

We note an example of the necessity of keeping taxonomy different from phylogeny in a certain context. The species Mycobacterium bovis always gets into the group made of the two strains of M. tuberculosis (the pathogen for human tuberculosis) as if it is a new strain of the latter. The genomic sequence of M. bovis is more than 99.95% identical to that of M. tuberculosis with about 1% reduction in genome size, yet the tropism of their infectious pattern is noticeable[^25]. The clinic difference of the two species may be caused by gene expression, a factor not taken into account in all sequence-based phylogeny for the time being. A similar
situation happens in the mixing-up of the *Escherichia Coli* and *Shigella* strains in phylogenetic trees, see the Supplementary Material for details. Therefore, in spite of phylogenetic closeness of the species there are good reasons to keep them different in taxonomy.

2.3.6 Phylum B17 (*Spirochaetes*). For the phylum *Spirochaetes* we note a prominent fact in all CVTrees. The two families *Spirochaetaceae* and *Leptospiraceae* never get together for all $K = 3 - 6$. In fact, as it was indicated in the first edition of the Bergey’s Manual (1984), “*Treponema* and *Leptospira* are assigned to the same order due to their common spirochete-like morphology”. They may well belong to two separate phyla.

2.3.7 Placement of Higher Taxa. Placement of higher taxa has always been under debate among taxonomists. It is even more so with respect to prokaryotes. Even the notion of prokaryote species has been challenged many times and as recently as in 2006. In view of this situation, the clustering of the overwhelming majority of the 432 organisms into a few monophyletic branches in CVTrees, that clearly corresponds to the biologist’s taxonomy, is an encouraging fact. The best convergent CVTree in terms of the highest taxonomic ranks at $K = 6$ is shown in Figure 6, that at $K = 5$ given in the Supplementary Material.

An inspection of Figure 6 shows that among the 432 organisms there are only 8 outliers:

(1) Two outliers from *Gammaproteobacteria* are denoted as Gamma(101-2). The bacterial endosymbiont *Carsonella ruddii* (Carru) has a highly reduced genome of 160 kbp with 182 protein-coding genes, much less than the smallest known free-living bacteria (see, e.g., ref. [28] and follow-up discussions in the literature). It should be discarded in a phylogenetic study of free-living prokaryotes. However, we keep it in this work to show that it does not make much trouble to the overall structure of the trees except for its own wrong position. Another outlier *Thiomicrospira denitrificans* (Thidn) simply should not be counted at all. Actually it gets stably into the Epsilon group in all CVTrees from $K = 3$ to 6. This has been noted in the Bergey’s Outline: Footer note 229 on page 87 says “The identity of *T. denitrificans* is questionable as it belongs within the Epsilonproteobacteria”.

(2) The two outliers from Actinobacteria(35-2), Rubxd and Synth, did not get very far from the greater cluster of which the main body of 33 Actinobacteria appear to be a branch.

(3) The 3 outliers in Delta(13-3) have been discussed before in Subsection 2.3.3 on *Proteobacteria*.

(4) The only outlier Salrd from Bacteroidetes(7-1) did remain in a greater cluster.

One should admit it is an excellent agreement between phylogeny and taxonomy that there are only 7 exceptions in the placement of 432 organisms. In addition, one may indicate a few more features in the grouping of higher taxa:

(1) The class *Mollicutes* from *Firmicutes* should clearly make a separate phylum. The remaining two classes join together at $K = 6$ to make another possible new phylum.

(2) The *Leptospiraceae* splits from the phylum *Spirochaetes* (B17) to form a possible new phylum.

(3) The 3 phyla B1, B2 and B6 form a greater monophyletic cluster at $K = 5$ and 6. More genomes are required to test the generality of this observation.

3 Discussion and conclusions

It is a remarkable fact that the CVTrees agree so well with the Bergey’s taxonomy which is based more and more on the 16S rRNA analysis. The CVTree approach and the 16S rRNA analysis use, so to speak, “orthogonal” data from the genome and utilize different methodology to infer phylogenetic information. Yet they support each other in an overwhelming majority of branchings and clusterings of taxa, thus providing a reliable framework to demarcate the natural boundaries among prokaryote species.

Very few discrepancies have been known between the CVTrees and the 16S rRNA trees. One example came from *Methanopyrus kandleri* which did not join other known methanogenes according to the rRNA analysis, but it got into the methanogens in CVTrees.

If we recollect that only a few years ago whole-genome phylogeny “does not resolve the major branchings of the Bacteria”, the high resolution power of CVTree method from strains up to classes and phyla is really an achievement. However, the use of complete genomes is both a merit and a demerit of the CVTree approach. It is a merit because no choice of sequences and genes is made, thus greatly reducing the subjectivity and bias of the inferred phylogeny. It is a demerit since the availability of complete genomes will always limit the scope of study. At present time more than 6250
than 6250 prokaryote names have been included in the new edition of Bergey’s Manual. In the *Proteobacteria* phylum alone 72 families 425 genera and 1875 species are listed in the Manual. Among them 123 species from 81 genera and 53 families are represented by complete genomes as of 31 December 2006. A few thousand new taxa are expected to be added to the second edition of the Bergey’s Manual. By the completion of the new edition, the possible taxonomic revisions suggested by the CVTrees may be checked on a wider scale. Therefore, a study similar to what is reported in this paper may well provide a core of the phylogenetic tree and provide further test on the predictive power of the CVTree method.

So far we have relied on qualitative results of the CVtree approach, mainly, on the tree topology. The composition vectors, however, contain much more information. How to make use of additional information and further justify the CVTree method is on our research agenda.

The Supplementary Material

The *Supplementary Material* \(^{[12]}\) contains: lists of all genomes used in this study, their abbreviations, NCBI accession numbers, and Bergey’s code; the original CVTrees for $K = 3$ to 6 as text files; and an exhaustive, taxon by taxon, comparison of CVTrees with the biologist’s taxonomy.

*The authors acknowledge the support from Fudan University. BLH thanks a discussion with Dr. Tim Lilburn at ATCC with participation of Prof. George Garrity by telephone connection as well as discussion with Prof. Whitman at UGA. Correspondence on Actinobacteria with Prof. Chenglin Jiang is also gratefully acknowledged.*

---