Prokaryote Phylogeny Meets Taxonomy: An Exhaustive Comparison of Composition Vector Trees with Systematic Bacteriology

Lei Gao
Ji Qi
Jiandong Sun
Bailin Hao

SFI WORKING PAPER: 2007-06-010

SANTA FE INSTITUTE

Prokaryote Phylogeny Meets Taxonomy: An Exhaustive Comparison of Composition Vector Trees with Systematic Bacteriology

GAO Lei[1, 2], QI Ji[1, *], SUN Jiandong[2, 3] & HAO Bailin[1, 2, 4]

(Submitted to *Science in China*, Series C, Life Sciences,)
Received April 2, 2007；Revised May 7 2007

1. Institute of Theoretical Physics, Academia Sinica, Beijing 100080, China
2. T-Life Research Center, Fudan University, Shanghai 200433, China
3. Beijing Genomics Institute, Academia Sinica, Beijing 101300, China
4. Santa Fe Institute, Santa Fe, NM87501, USA

*Present address: Center for Comparative Genomics and Bioinformatics, Penn State University, 310 Wartik Building, University Park, PA16802, USA

Correspondence should be addressed to HAO Bailin(E-mail: hao@itp.ac.cn)

Abstract

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 on taxonomic revisions. In stead of demonstrating the overwhelming agreement this paper puts emphasis on the biological implications of the differences.

I Introduction

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 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[2] that aims at reconstructing the evolutionary origins of all living organisms. Yet ironically, in a recent Science paper[3] addressed to building the Tree of Life the branch for the most numerous on Earth organisms, 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[4] . The 16S rRNA trees have been so successful that the new edition of the *Bergey's Manual of Systematic Bacteriology*[5] (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 [5]).

However, phylogeny and taxonomy are not synonyms. A correct taxonomy should agree with phylogeny in major and finer branchings. On the other hand, a faithful phylogeny should reflect the evolutionary history of species. Being originally a comprehend taxonomic summary of the hard work of many generations of bacteriologists but now largely based on 16S rRNA phylogeny, the Bergey's Manual needs independent phylogenetic support and verification. In addition, one can always ask to what extent the evolution of a single gene such as 16S rRNA may reflect the evolution of the species. Indeed, even the ribosomal operons in a bacterium may be replaced by that from other species[6], how could single- or few-sequence-based phylogeny be immune from the influence of lateral gene transfer? Whether there is a phylogenetic signal in prokaryotic proteins was questioned a few years ago: "The tree … consists of almost entirely of noise at the level of bacterial phylum divisions, indicating that, even with large amounts of data, it may not be possible to reconstruct the prokaryote phylogeny using standard sequence-based methods"[7].

On the other hand, the inpouring of more and more complete prokaryotic genomes since 1995 has led to an upsurge of whole-genome phylogenetic studies, for a recent review see [8]. However, most of these methods are eventually sequence-alignment based and as such depend on many parameters — the BLAST procedure used in many studies being an example. Even some "automatic reconstruction"[9] requires manual identification of genes at some stage. Furthermore, since the actual phylogenetic tree, if any, was buried in the long evolutionary history, the quality judgement of phylogenetic reconstructions has to rely on self-consistent arguments and on statistical tests such as bootstrapping or Jack-knife resampling. Hence, molecular phylogeny of prokaryotes has become a minor industry in its own and few trees have been compared directly with biologists' tree of life in great details.

In view of what said, a few years ago we developed an alignment-free and parameter-free method[10, 11] to infer prokaryote phylogeny from whole genome data. We have built a public web server named CVTree[12]. Throughout this paper we use CVTree both as a shorthand for the method and for the tree thus obtained. More importantly, we adopt a new philosophy towards the verification of trees. Treating the CVTree method that takes genome data as input as a theoretical construction, we compare its output directly with “experimental facts” collected in the Bergey’s Manual. This paper is the summary of such a detailed and exhaustive comparison.

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[13] 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 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*[14]. 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[10, 11, 12] 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 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[15]. We have reconstructed all CVTrees from $K = 3$ to 6. These trees are given in the *Supplementary Material*. 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 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[16]. The justification of the CVTree approach is still under way, see, e.g., [17].

1.3. Prokaryote Taxonomic Reference

There is no official standard for prokaryotic taxonomy. However, the classification scheme in *Bergey’s Manual of Systematic Bacteriology*[5] is now widely accepted by microbiologists as the best approximation to an official classification[18]. The new edition of Bergey’s Manual is based on 16S rRNA phylogeny as well as on classical morphological and physiological observations[18]. 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*[14] a lineage in the Bergey’s Manual or its online *Outline*[19] is abbreviated as B13.3.2.6.2 or $B_{13}C_3O_2F_6G_2$ for Phylum BXIII (*Firmicutes*), Class III

(*Bacilli*), Order II (*Lactobacillales*), FamilyVI (*Streptococcaceae*), Genus II (*Lactococcus*). We call this a "Bergey's code"[10]. 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*[19] .

The National Center for Biotechnological Information (NCBI) provides a Taxonomic Browser[20]. 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)[21] for additional information. Occasionally we also refer to some other taxonomy, e.g., that reflected in the book *Five Kingdoms*[22] .

2. Comparison of CVTrees with Taxonomy

We perform 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*[14] and only a summary is given in the subsequent sections.

In making comparison of the branchings 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., clustering of species in a genus and grouping of all lower taxa into a highest taxon such as phylum or class, make more sense[1]. Accordingly, two guiding principles are followed in our analysis. At the strains 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[23] was 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 Table1 we show the clustering of the 440 genomes in CVTrees into the three domains. The convergence with K increasing shows off clearly.

Table 1: Clustering of the 440 genomes into 3 domains. Abbreviations: A=*Archaea*, B=*Bacteria*, E=*Eukarya*.

K=3	K=4	K=5	K=6
7E	8E	8E	8E
1E in B			
25A in B	25A in B	31A	31A
1A(Arcfu) in B	1A(Methj) in B		
1A(Naneq) in B	1A(Naneq) in B	400B	400B
4A in B	4A in B		

[1] Charles Darwin mentioned repeatedly species, genera and families in his *Origin of Species*, but rarely referred to higher taxa.

However, there is a minor proviso concerning the bacterial endosymbiont *Carsonella ruddii*[24] , see Subsection 3.3.7 on the placement of higher taxa.

2.2. Analysis of the Archaea Branch

We perform a comprehend 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*[14] . Their taxonomic distribution is given in Table 2.

Table 2: Taxonomic distribution of the 31 Archaea.

Phylum	Class	Order	Family	Genus	Species
A1	1	3	4	4	7
A2	8	9	12	18	23
A3	1	1	1	1	1
Total	10	13	17	23	31

Among the 23 genera 5 contain more than one species. Among the 17 families 4 contain more than one genus. Among the 13 orders 3 contain more than one family. Among the 10 classes 2 contain more than one order. 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 contain 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 contain 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* appear 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.

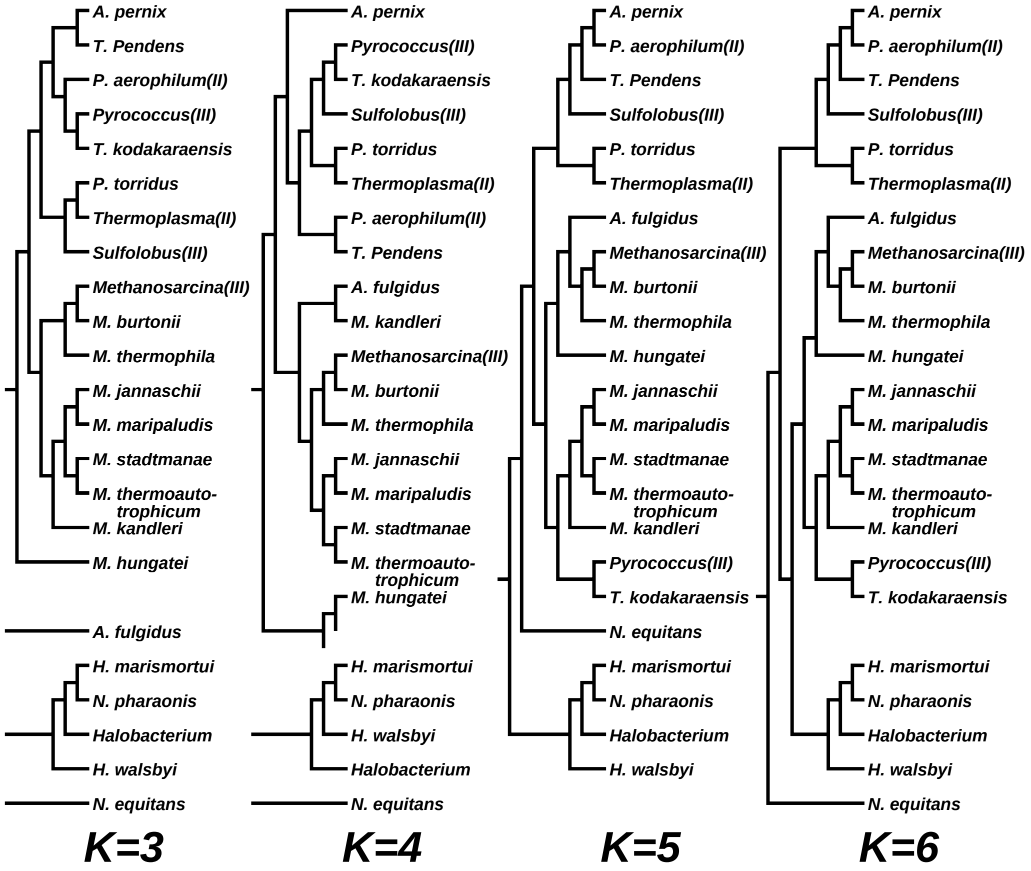

Figure 1: The genus trees for the Archaea branch taken from the 440-genome CVTrees at K = 3 to 6.

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

Using the abbreviations introduced above the Archaea branch of the CVTrees at K = 3 to 6 is given in Fig.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 $A_2C_4O_1F_1$ (*Halobacteriaceae*) representing the whole class C_4 for the time being that contains four genera: *Haloarcula* (Halma), *Natronomonas* (Natpd), *Halobacterium* (Halsa), and *Haloquadratum* (Halwd). They converge at all K with slight variation at K = 4. Even when this family disunites 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 Fig. 2 using the Bergey's codes for higher taxa. The relative positions of leaves are the same as that in Fig. 1. It is readily seen from Figs. 1 and 2 that there are only three discrepancies with the Bergey's taxonomy:

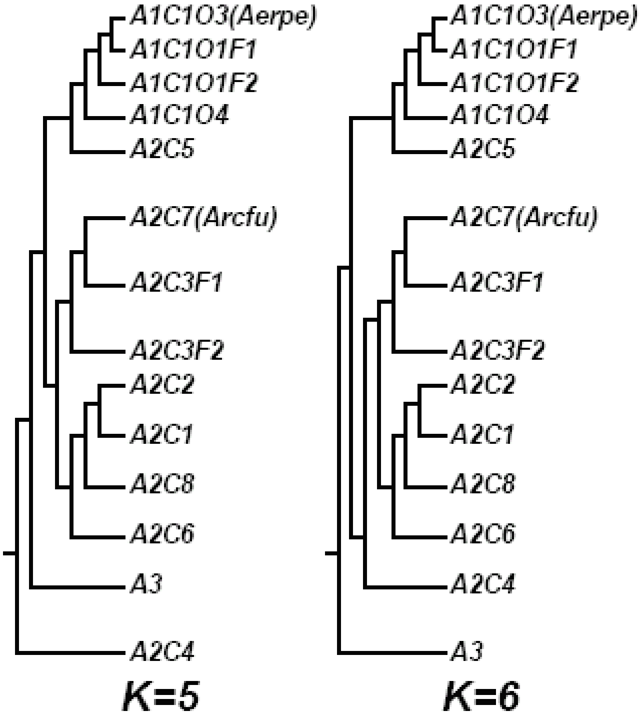

Figure 2: Higher taxa trees at *K* = 5, 6 for the Archaea using Bergey's codes.

1. A seemingly cross-phylum discrepancy: the class A_2C_5 (*Thermoplasmata*) from the phylum A2 (*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*[22] . Therefore, it does not make a real problem in the CVTrees.
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 on possible taxonomic revision of Aerpe and Arcfu. For example, assigning Aerpe to $O_1F_?$ and Arcfu to C_3 would resolve the problem.

An overall inspection of the trees from *K* = 3 to 6 shows the well 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 Naneq, 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_2C_4 (*Halobacteria*), being a monophyletic clad 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 Naneq 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.

2.3. Analysis of the Bacteria Branch

The taxonomic distribution of all 401 bacterial genomes is listed in Table 3. We have performed an exhaustive comparison of the 401-genome bacterial branch of the CVTrees with the Bergey's taxonomy, similar to what described verbally above for the Archaea. A taxon by taxon analysis is presented in the *Supplementary Material*[14]. A summary of this detailed analysis is given in what follows.

Table 3: Taxonomic distribution of the 401 bacterial genomes.

Phylum	Classes	Orders	Families	Genera	Species	Strains
B1 (*Aquificae*)	1	1	1	1	1	1
B2 (*Thermotogae*)	1	1	1	1	1	1
B4 (*Deinococcus-Thermus*)	1	2	2	2	3	4
B6 (*Chloroflexi*)	1	1	1	1	2	2
B10 (*Cyanobacteria*)	1	3	3	8	15	19
B11 (*Chlorobi*)	1	1	1	2	4	4
B12 (*Proteobacteria*)	5	33	53	99	157	208
B13 (*Fermicutes*)	3	7	14	22	58	96
B14 (*Actinobacteria*)	3	9	15	16	31	35
B15 (*Planctomycetes*)	1	1	1	1	1	1
B16 (*Chlamydia*)	1	1	2	3	7	11
B17 (*Spirochaetes*)	1	1	2	3	7	9
B19 (*Acidobacteria*)	1	1	1	2	2	2
B20 (*Bacteroidetes*)	3	3	5	5	6	7
B21 (*Fusobacteria*)	1	1	1	1	1	1
Total 15	25	66	103	167	296	401

Only when a taxon contains two or more lower taxa it corresponds to one or more branching points in a tree. If a taxon contains more than three lower taxa these lower ones are simply juxtaposed in a taxonomy. However, in any phylogenetic tree, faithful or not, there appears a branching order among lower taxa. This adds a new dimension to the comparison of taxonomy with trees and may bring about new evolutionary knowledge. Therefore, we start with a collection of the number of taxa contained in a higher taxon at all taxonomic ranks as shown in Table 4.

Table 4: Number of taxa contained in an higher taxon (domain *Bacteria*).

i =	1	2	3	4	5	6	7	8	9	10	11	13	$\Sigma_{i>1}$
Strains/Species	242	31	14	3	3			1	1		1		54
Species/Genus	110	30	9	8	5		2	0	2	1			57
Genera/Family	69	23	5	2	1	1	1				1		34
Families/Order	41	16	7	1	1								25
Orders/Class	15	3	2			2	2					1	10
Classes/Phylum	11		3		1								4
Total	488												184

We note that the numbers in Tables 3 and 4 should be consistent by satisfying a few check sums. For example, denoting the i-th number in the j-th row of Table 4 by n_i^j, the sums $a^j = \sum_{i \geq 1} n_i^j$ and $b^j = \sum_{i \geq 1} i \times n_i^j$ 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*[14].

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*[14] 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 is 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*[14]. 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.

Cases	Number of Taxa contained		*i*>1 cases Compared with CVTrees	
	i=1	*i*>1	Consistent	Different
Strains/Species	242	54	47	7
Species/Genus	110	57	46	11
Genera/Family	69	34	26	8
Families/Order	41	25	14	11
Orders/Class	15	10	5	5
Classes/Phylum	11	4	0	4
Total	488	184	138	46

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 Fig. 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.

Bergey's Taxonomy				CVTrees			
Class	Subsection	Family	Genus	Species	Strain	Grouping	K
C_1	1	F_1	G_7	1	1	Glovi	
			G_{11}	5	5	Prom9(II)	3,4,5,6
						Promt(II)	4,5,6
						Promm	
			G_{13}	3	8	Synja(II)	3,4,5,6
						Synp6(2)	3,4,5,6
						Synpx(4)	3,4,5,6
			G_{14}	1	1	Syny3	
			$G_?$	1	1	Synel	
	3	F_1	G_{17}	1	1	Triei	
	4	F_1	G_1	1	1	Anava	
			G_8	1	1	Anasp	

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, *Oscillatorales* 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 bacteria 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[25, 26] of *P. marinus* were published, see Table 7 below:

Table 7: The five ecotypes in the genus *Prochlorococcus*.

Ecotype	Name	Abbr.	Remark
eMIT9312	*P. marinus* str. MIT 9312	Prom9	Near surface, high-light adapted
eMED4	*P. marinus* MED4	Promp	As above
eSS120	*P. marinus* subsp. marinus marinus str. CCMP1375	Proma	Deep water, low-light adapted
eNATLA2	*P. marinus* NATL2A	Promt	As above
eMIT9313	*P. marinus* str. 9313	Promm	As above

Although the names of these 5 ecotypes look like different strains in one and the 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 $K = 3 \sim 6$ agrees with the two high-light 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 Fig. 3.

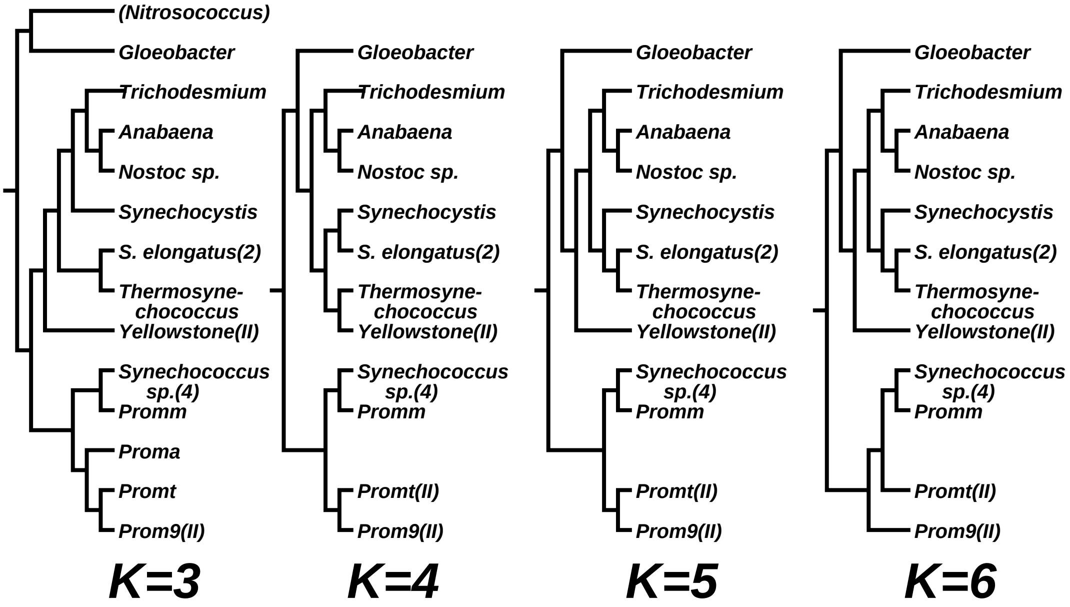

Figure 3: Convergence of all the 19 organisms in Phylum *Cyanobacteria*. For abbreviations of species in *Prochlorococcus* see Table. 7.

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

1. The 4 strains of Synechococcus sp. (Synpx(4)) do not get into its desiganted genus *Synechococcus*. Instead it joins the *Prochlorococcus* species Promm. In fact, these strains live in the same habitat as Promm and are considered "potential competitor" of the latter[25]. The grouping (Promm, Synpx(4)) persists from K = 3 to 6. It is reasonable to recognize Synpx(4) 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 $B_{10}C_1O_1F_1G_?$ from the CVTrees. Apparently it is closer to *Synechococcus elongatus*.
3. The Bergey's designation of the genus *Gloeobacter* and *Prochlorococcus* to the same family as *Synechococcus* is not justified in the CVTrees. The NCBI taxonomy of putting *Prochlorococcus* in the order *Prochlorales* and *Gloeobacter* in the order *Gloeobacteriales* might be more reasonable than the Bergey's.

2.3.3. Phylum B12 (*Proteobacteria*)

The Phylum B12 is represented by 208 organisms in the dataset. The 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 Fig. 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 on the possibility of assigning a new order to Magmc within the Alpha group.

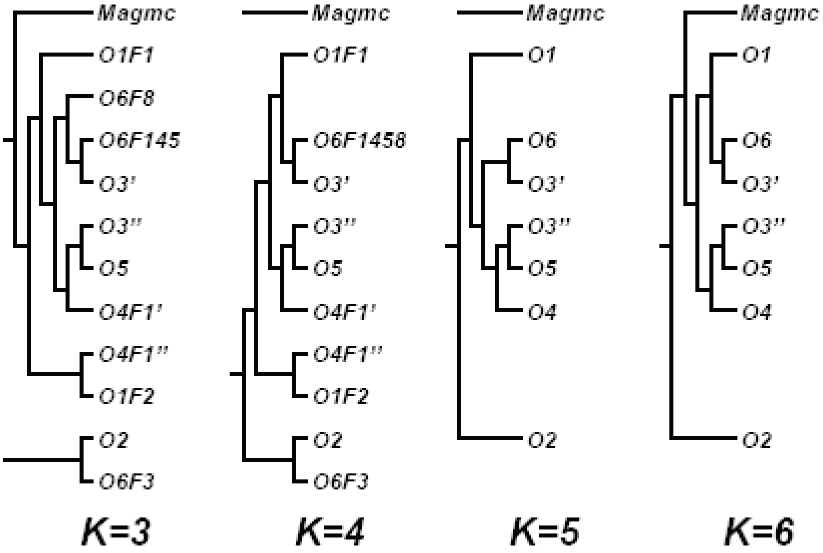

Figure 4: Convergence of orders in the *Alphaproteobacteria* group with *K*. Note that at *K* = 4 to 6 the six orders defined in the Bergey's Manual form a monophyletic cluster. The newly sequenced *Magnetococcus MC-1* is abbreviated as Magmc.

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 Fig. 5 for the order convergence with *K*.

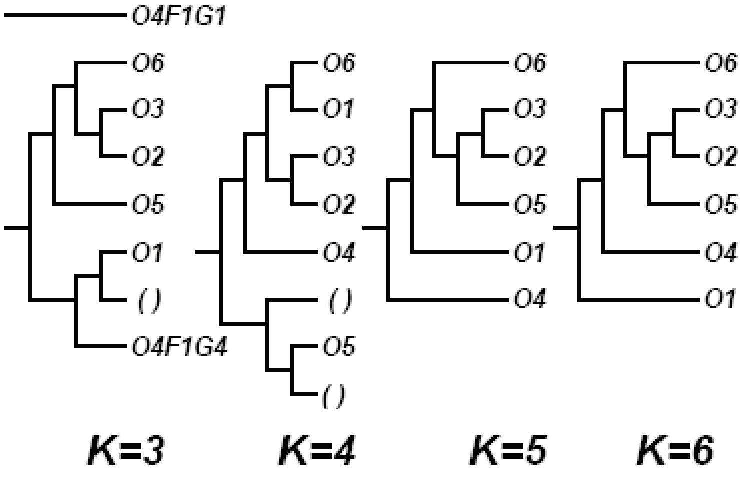

Figure 5: The convergence of the 6 orders in the *Betaproteobacteria* group. The monophyletic structure at *K* = 5 and 6 shows a satisfactory agreement of phylogeny with taxonomy. The blank parentheses at *K* = 3, 4 denote organisms from Gamma group.

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*, 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”[27]. 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 family. In all CVTrees for $K = 3 \sim 6$ 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*.

Species	Abbr.	Genome Size	Gene Number
Escherichia coli	EcoliK	4 639 675	4 237
Erwinia carotovora	Erwct	5 064 019	4 472
Photorhabdus luminescens	Pholl	5 688 987	4 683
Salmonella entrica	Salpa	4 585 229	4 093
Shigella dysenteriae	Shids	4 369 232	4 274
Sodalis glossinidius	Sodgl	4 171 146	2 432
Yersinia pestis	Yerpn	4 534 590	3 981

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 the16S 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.

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

Species	Abbr.	Genome Size	Gene Number
Baumannia cicadellinicola	Bauch	686 194	595
C. Blochmannia pennsylvanicus	Blopb	791 654	610
Buchnera aphidicola	Bucap	641 454	546
Wigglesworthia brevipalpis	Wigbr	697 724	611

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 [28]. This situation is denoted as Delta(13-3) in Fig. 6.

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 below in 3.3.7.

2.3.4. Phylum B13 (*Firmicutes*)

The phylum B13 is represented by 96 organisms in our dataset. The *Firmicutes* are a very diverse group. “Hugenholz recognizes at least four other phyla within the *Firmicutes*.” (Footnote 3 on p. 2 of the Bergey’s *Outline*[19]). 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*[14] 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 *Chlamydophla 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*[29] . For the two outliers see the *Supplementary Material*.

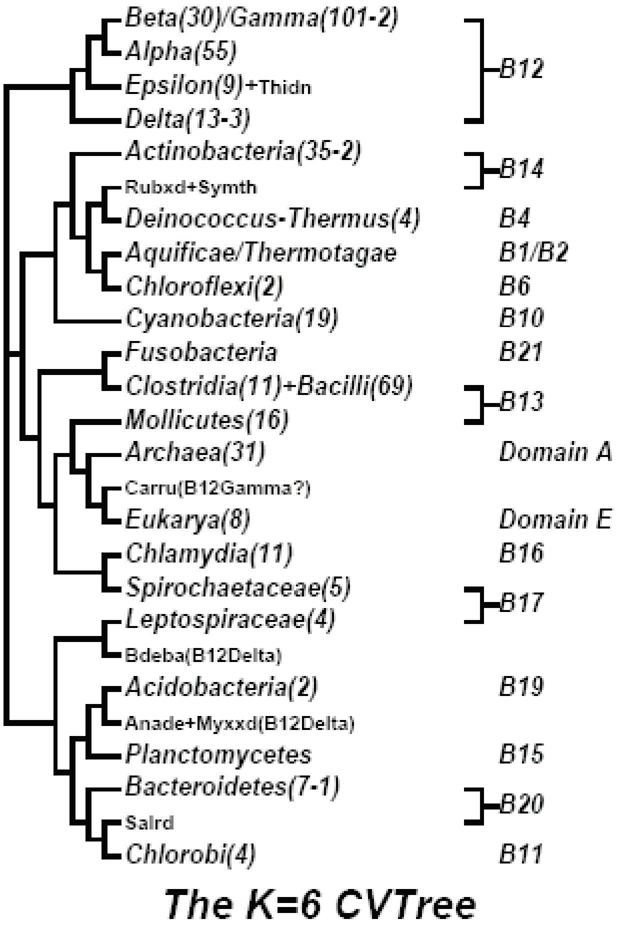

Figure 6: The highest rank CVTree at K = 6. A taxon name represents a monophyletic cluster with the number of organisms given in parentheses. For example, Gamma(101-2) is the cluster for *Gammaproteobacteria* made of 99 organisms, the 2 outliers, Garru and Thidn, appear elsewhere in the tree. Note that this is an unrooted tree and the branches are not to scale.

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[30]. 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 $K = 3 \sim 6$. 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[31] . Even the notion of prokaryote species has been challenged many times and as recent as in 2006[32]. In view of this situation the clustering of the overwhelming majority of the 432 organisms in CVTrees into a few monophyletic branches that clearly correspond to the biologist’s taxonomy is an encouraging fact. The best convergent CVTree in terms of the highest taxonomic ranks at K = 6 is given in Fig.6, that at K = 5 is given in the *Supplementary Material*.

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

1. Two outliers from *Gammaproteobacteria* is denoted as Gamma(101-2). The bacterial endosymbiont *Carsonella ruddii*[24] (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., [33] 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*: Footnote 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 Symth, did not get very far from the greater cluster of which the main body of 33 Actinobacteria appears to be a branch.
3. The 3 outliers in Delta(13-3) have been discussed before in Subsection 3.3.3 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 clustering 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[34] , 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 are 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 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 thousands 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 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*[14] 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.

Acknowledgements

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.

References

1. Zuckerkandl, E., Pauling, L., Molecules as documents of evolutionary history, J.Theor.Biol., 1965, 8: 357 – 366.
2. The AToL Project Home Page: atol.sdsc.edu
3. Driskell, A. C., Ané, C., Burleigh, J. G. et al., Prospect for building the tree of life from large

sequence databases, Science, 2004, 306: 1172 – 1174.

4. Woese, C. R., Fox, G. E., Phylogenetic structure of the prokaryotic domain: the primary kingdoms, Proc. Natl. Acad. Sci. USA, 1977, 74: 5088–5090.
5. Bergey's Manual Trust, *Bergey's Manual of Systematic Bacteriology*, 2nd Ed., vol. 1-5, New York, Springer-Verlag, 2001-2008.
6. Asai, T., Zaporojets, D., Squires, C. et al., An Escherichia coli strain with all chromosomal rRNA operons inactivated: complete exchange of rRNA genes between bacteria, Proc. Natl. Acad. USA, 1999, 96: 1971 – 1976.
7. Teichmann, S. A., Mitchison, G., Is there a phylogenetic signal in prokaryote proteins? J. Mol. Evol., 1999, 49: 98 – 107.
8. Snel, B., Huynen, M. A., Dutilh, B. E., (2005) Genome trees and the nature of genome evolution, Annu. Rev. Microbiol., 2005, 59: 191 – 209.
9. Ciccarelli, F. D., Doerks, T., von Mering, C. et al., Toward automatic reconstruction of a highly resolved tree of life, Science, 2006, 311: 1283 – 1287.
10. Qi, J., Wang, B., Hao, B.-L., Whole genome prokaryote phylogeny without sequence alignment: a *K*-string composition approach, J. Mol. Evol., 2004, 58: 1 – 11.
11. Hao, B.-L., Qi, J., Prokaryote phylogeny without sequence alignment: from avoidance signature to composition distance, J. Bioinformatics & Computat. Biol., 2004, 2: 1 –19.
12. Qi, J., Luo, H., Hao, B.-L., CVTree: a phylogenetic tree reconstruction tool based on whole genomes, Nucl. Acids Res., 2004, 32, Web Server Issue: W45 – W47.
13. The NCBI ftp-site:
    ftp://ftp.ncbi.nih.gov/genomes/Bacteria/
14. A *Supplementary Material* to this paper is downloadable from either
    http://www.itp.ac.cn/˜hao/Suppl440.pdf or
    http://tlife.fudan.edu.cn/Suppl440.pdf.
15. Felsenstein, J., PHYLIP (Phylogeny Inference Package) version 3.5c., distributed by the author at: http://evolution.genetics.washington.edu/phylip.html
16. Michel, H., The future of the molecular biosciences: consequences of the massive parallel approach, in *Sceince and Technology Development: A Retrospective View over the Past Century and a Prospective Look into the Future*, ed. Y.-X. Lu, Shanghai Education Press, 2000, p. 70.
17. Shi, X.-L., Xie, H.-M., Zhang, S.-Y., Hao, B.-L., Decomposition and reconstruction of protein sequences: the problem of uniqueness and factorizable language, J. Korean Phys. Soc., 2007, 50: 118 – 123.
18. Konstantinidis, K. T., Tiedje, J. M., Towards a genome-based taxonomy for prokaryotes, J. Bacteriol., 2005, 187: 6258 – 6264.
19. Garrity, G. M., Bell, J. A., Lilburn, T. G., *Taxonomic Outline of the Prokayotes. Bergey's Manual of Systematic Bacteriology*, 2nd Ed., Spinger-Verlag, New York, Rel. 5.0, May 2004, DOI: 10.1007/bergeysoutline200405.
20. The NCBI Taxonomy Browser:
    http://www.ncbi.nlm.nih.gov/Taxonomy/
21. The taxonomic list at EBI:
    http://www.ebi.ac.uk/genomes/bacteria.html
22. Margulis, L., Schwartz, K. V., Five Kingdoms. An Illustrated Guide to the Phyla of Life on

Earth, 3rd Ed. W. H. Freeman, 1998.

23. Woese, C.R., Kandler, O., Wheelis, M.L.,Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya, Proc. Natl. Acad. Sci. USA, 1990, 87: 4576 – 4579.
24. Nakabachi, A., Yamashita, A., Toh, H. et al., The 160-kilobase genome of the bacterial endosymbiont *Carsonella*, Science, 2006, 314: 267.
25. Z. I. Johnson, Z. I., Zinser, E. R., Coe, A. et al., Niche partitioning among *Prochlorococcus* ecotypesalong ocean-scaleenvironmentalgradients,Science,2006,311:1737 –1740.
26. M. C. Coleman, M. C., Sullivan, M. B., Martiny, A. C. et al., Genome islands and the ecology and evolution of *Prochlorococcus*, Science, 2006, 311: 1768 – 1770.
27. Woese, C. R., Olsen, G. J., Ibba, M. et al., Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process, Microbiol. & Mol. Biol. Revs., 2000, 64:202 –236.
28. Pringshein, E.G., The relationship between bateria and *Myxophyceae*, Bacterialogical Revs., 1949, 13: 47.
29. Zhi, X.-Y., Cai, M., Yang, L.-L. et al., Evidence for the establishment of the *Actinobacteria* phylum, Microbiology, 2006, 33:181 – 183 (in Chinese).
30. T. Garnier, T., Eiglmeier, K., Camus, J.-C. et al., The complete genome of *Mycobacterium bovis*, Proc. Natl. Acad. Sci. USA, 2003, 100: 7877 – 7882.
31. Murray, R. G. E., The higher taxa, or, a place for everything ···? in *Bergey's Manual of Systematic Bacteriology*, 1st Ed., vol. 4, Baltimore, Williams & Wilkins, 2329 – 2332.
32. Doolittle, W. F., Papke, R. T., Genomics and the bacterial species problem, Genome Res., 2006, 7: 116.
33. Goffeau, A., Life with 482 genes, Sceince, 1995, 270:445 – 446.
34. Burggraf, S., Stetter, K. O., Pouviere, P. et al., *Methanopyrus kandleri*: an archeal methanogen unrelated to all other known methanogens, Sys. Appl. Microbiol., 1991, 14: 346 – 381.
35. Huynen, M., Snel, B., Bork, P., Lateral gene transfer, genome surveys, and the phylogeny of prokaryotes, Science, 1999, 286: 1443a.