Chapter 6

# Whole-Genome-Based Phylogeny and Taxonomy for Prokaryotes

Guanghong Zuo and Bailin Hao

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.68563

**Abstract**

A faithful prokaryotic phylogeny should be inferred from genomic data and phylogeny determines taxonomy. The ever-growing amount of sequenced genomes makes this approach feasible and practical. Whole-genome phylogeny must be based on alignment-free methodology and should be verified by direct comparison with taxonomy at all ranks from domains down to species. When the number of genomes goes into tens of thousands, the realization of the above program also presents technical challenges. The power of a long-tested Web Server named Composition Vector Tree (CVTree) will be demonstrated on examples from mega-classification of bacteria to high resolution at and below the species level.

**Keywords:** alignment-free phylogeny, *Bacteria, Archaea*, CVTree, mega-classification, prokaryotic taxonomy

## 1. Introduction

Prokaryotes are the most successful creatures on Earth, comprising two of the three main domains of life [1, 2]: *Archaea* and *Bacteria*. It has been estimated that there are $10^{30}$ living cells [3] on the planet. Although the notion of prokaryotic species has been a subject of long debate, the estimated number of species, whatever the definition one adopts, surely exceeds $10^7$ [4]. By contrast, for the time being only less than 14,000 species names have been validly published and come with a standing in nomenclature [5]. Nevertheless, though based on such rare sampling, bacterial phylogeny and systematics have made significant progress since the late 1970s mainly due to the use of 16S rRNA sequences as molecular markers [6, 7]. However, the fact that prokaryotic phylogeny becoming congruent with taxonomy on the basis of the same kind of markers calls for independent verification. The verification should preferably

use different input data, for example, non-RNA sequences, and rely on distinct methodology, for example, an alignment-free approach. At present, convincing answers to the problem are naturally provided by making use of the ever-growing amount of genomic data.

In fact, the idea is by no means new. As early as in 1987, an "Ad Hoc" Committee on reconciliation of approaches to bacterial systematics stated in its report [8]:"There was general agreement that the complete deoxyribonucleic acid (DNA) sequence would be the reference standard to determine phylogeny and that phylogeny should determine taxonomy. Furthermore, nomenclature should agree with (and reflect) genomic information."

Furthermore, in the heyday of the human genome project Carl Woese stepped forward bravely with a manifesto for microbial genomics [9]. Woese wrote about the same time that "Genome sequencing has come of age, and genomics will become central to microbiology's future. It may appear at the moment that the human genome is the main focus and primary goal of genome sequencing but do not be deceived. The real justification in the long run is microbial genomics" [10].

The development of microbiology in subsequent years witnesses the foresight of Carl Woese. In particular, there has been continuing discussion on how to construct genome-based phylogeny and taxonomy, see, for example [11–14]. The abundance of genomic data enables the transition from comparing methodological suggestions to devising practical tools for bench microbiologists. In this chapter, we review our decade-long effort [15–20] to develop a whole-genome-based and alignment-free Composition Vector Tree (CVTree) approach and demonstrate the companion CVTree Web Server.

## 2. The CVTree approach

The CVTree approach is endowed with several distinctive features. It is based on whole genomes. It utilizes an alignment-free method for genome comparison. The resulting phylogenetic tree turns out to be essentially parameter free. The evaluation of the trees is realized by direct comparison with prokaryotic taxonomy. We elaborate these points in more detail in the subsequent text.

First of all, the feasibility of CVTree is guaranteed by the availability of the ever-growing amount of sequenced genomes. Since the first two bacterial genomes were published in 1995 [21, 22], the number of sequenced prokaryotic genomes has been accumulating rapidly. According to the GOLD Database [23], nearly 90,000 prokaryotic genomes have been deposited by the end of August 2016. The fact that more than half of the available genomes are incomplete or permanent drafts do not diminish the usefulness of these data, as nowadays annotation of genomic sequences may be easily carried out by using public-domain software or services such as IMG [24]. Moreover, the CVTree approach is insensitive to details of annotation. In principle, a whole genome contains most of phylogenetic information of an organism. Taking whole genomes as input data circumvents the selection of sequence segments or homologous proteins, thus eliminating ambiguity caused by human judgments. For example, lateral gene transfer, causing serious bias in phylogeny if based on a single or a few proteins, appears merely as a mechanism of genome evolution together with lineage-dependent gene loss.

Second, prokaryotic genomes are extremely diverse in their size and gene content. For example, the five sequenced genomes of *Mycoplasma genitalium* have a median protein count of 484 and a genome size of 0.58 Mbp [22], whereas the largest bacterial genome sequenced so far, that of *Sorangium cellulosum* So0157-2 strain, consists of 10,174 proteins and 14.8 Mbp [25]. We did not mention the highly degenerated tiny genomes of some bacterial endosymbionts, which are not recommended to be included in a phylogenetic study of free-living organisms. More than 20-fold differences in protein number and genome size preclude comparison of these genomes by sequence alignments. In other words, a whole-genome-based prokaryotic phylogeny must be built by using alignment-free comparison of genomes.

Our way of being alignment-free consists in extending the notion of amino acid frequency ($K = 1$) to an alphabet made of $20^K$ oligo-peptides of length $K$ ($K \geq 3$). By taking all the protein products encoded in a genome and counting the number of each type of the $K$-peptides by using a sliding window of width $K$, we construct a raw composition vector (CV) by arranging the counts in a lexicographical order of the $K$-peptides. A simple-minded way of using these CVs to represent species and defining species separation by the distance between CVs did not yield much meaningful results. Many researchers, along with the authors of this chapter, may have encountered this hurdle.

Upon reflection on Kimura's theory of neutral evolution [26], one realizes the necessity of subtracting a background caused by neutral mutations left in a genome. These neutral mutations have nothing to do with evolutionary process but contribute to components of the raw CVs. Since, according to Kimura, mutations occur randomly at molecular level, the neutral contributions to the $K$-peptide counts may be taken into account by invoking some statistical consideration as follows. First, collect the counts of all $K$-, ($K$-1)-, and ($K$-2)-peptides from the protein products of a genome. Then, predict the number of a given type of $K$-peptide from the counts of shorter ones by using a ($K$-2)-th Markov prediction [15, 16]. Suppose that for a certain type of $K$-peptide, the actual count coincides with the prediction. This would mean that the count of this particular $K$-peptide does not contain new phylogenetic information, as what added to the counts of ($K$-1)- and ($K$-2)-peptides is merely a statistical formula. What really matters is the difference between the actual count and the predicted number. Replacing each component of a raw CV by the corresponding difference, a "renormalized" CV is obtained. The subtraction procedure is crucial for success of the method, but we skip the mathematical details, as these can be found in previous publications, for example, in [15, 16] and [20]. We indicate that the key formula of the subtraction procedure may be derived in two independent ways, either by using the relation between joint probability and conditional probability [15, 16] or by applying the maximal entropy principle [27].

The peptide length $K$ figuring in the above description looks like a parameter. However, it does not serve as a parameter since a fixed $K$ is used for all genomes to generate a tree. The minimal value of $K = 3$ is dictated by the ($K$-2)-th Markov model itself. Larger $K$-values put emphasis on species specificity. The optimal value of $K$ depends on the total amount of amino acids letters in all the protein sequences under study, $K = 5$ and 6 being the best for *Bacteria* and *Archaea*. For a detailed discussion on the role and choice of $K$, please consult [20].

Traditionally, the quality of phylogenetic trees is evaluated by statistical resampling procedures such as bootstrap or jackknife tests. However, successfully passing these tests tells at

most the stability and self-consistency of the trees with respect to small variations in the input data, by far not the objective correctness of the branching scheme. We note that the CVTree results have passed both bootstrap and jackknife tests [28]. Furthermore, from the early CVTree constructed on 106 genomes [16] to trees based on 10,000 or more genomes, the agreement of CVTrees with taxonomy has kept improving. This fact may be taken as successfully passing larger and larger "anti-jackknife" tests. Therefore, we advocate the viewpoint that, instead of doing the time-consuming and indirect statistical resampling tests, the branching orders in a phylogenetic tree should be checked with the taxonomic hierarchy for the same set of input data. In fact, this is nothing but realization of the "general agreement" formulated in the 1987 Report of the Ad Hoc Committee [8].

A key notion in checking the agreement of phylogeny with taxonomy consists in monophyly. Introduced in 1866 by Ernst Huckael and originated from zoology, this notion requires the recognition of common ancestry, a requirement hardly satisfied by prokaryotes predominantly with asexual reproduction. Therefore, we take a pragmatic standpoint by restricting ourselves to the input dataset only. Being a reciprocal notion, monophyly applies to both taxonomy and phylogeny. An input data set comes with a reference taxonomy, in the case of CVTree, the NCBI taxonomy. If a taxon under study contains all the subordinate members inclusively, meaning that no member escapes to other taxon and no stranger from other taxon gets in, then the taxon is said to be monophyletic. Similarly, if a tree branch contains leaves representing species from one and the same taxon without strangers from other taxa, the branch is said to be monophyletic. In this sense, the genus *Clostridium* cannot be considered monophyletic in taxonomy, as many separate clusters are listed, including a big *senso stricto* group and many smaller clusters, see, for example, [29, 30]. Naturally, one cannot use the notion of monophyly to evaluate the *Clostridium* part of a tree. By the way, CVTree may help to bring the taxonomy of *Clostridium* to a better shape in the future.

## 3. The CVTree3 Web server

The underlying idea of CVTree and the corresponding algorithm described in the last section is simple in essence but hard to implement as many vectors and matrices of very high dimensions are involved. In order to help microbiologists to use this convenient tool, we have designed a public-domain Web Server called CVTree. The first CVTree Web Server was published in 2004 [31] and ceased service by now. An improved second release of 2009 [32] is still functioning [33]. However, we strongly recommend the users to try out the latest 2015 release CVTree3 [34] with many new functions added [35]. This is a much more powerful Web Server, which resides in a cluster with 64 cores and is capable to construct trees based on several thousands of genomes in a few minutes. In fact, all the descriptions in the subsequent sections refer to this latest version of Web Server.

Suffice it to type the above URL into the browser in order to enter the server without any login procedure. Leaving an email address is not obligatory but useful. As there is an online and printable help file, we skip most of the technicalities of how to use the server and concentrate on its characteristic features and typical results.

### 3.1. Input data set

The CVTree3 server is equipped with a built-in collection of genomes. For the time being, there are more than 3000 bacterial and archaeal genomes of a wide taxonomic assortment for picking up. These data are updated from time to time. Users can also upload their own genome data, 100 M compressed or not at a time. It is highly recommended to put the users' data on a wide taxonomic background no matter what kind of problem is studied. A background with broad sampling in taxonomy increases the stability of the results and allows outliers to escape to where they prefer. In order to avoid confusion, we mention in passing that many examples in this chapter are based on CVTrees built on 10,000 or more genomes.

### 3.2. Lineage information

Both built-in and user-uploaded genomes come with lineage information. For built-in genomes, the information is taken from the NCBI taxonomy [36] with minor corrections when necessary. Users should supply lineage information for the uploaded data. Lineage information for a genome looks like the following:

<D>Bacteria<K>Bacteria<P>Proteobacteria<C>Alphaproteobacteria<O>Caulobacterales<F>Caulobacteraceae<G>Caulobacter<S>Caulobacter_crescentus<T>Caulobacter_crescentus_CB15_uid5789.NCBI

where <D>, <K>, <P>, <C>, <O>, <F>, <G>, <S>, and <T> stand for Domain, Kingdom, Phylum, Class, Order, Family, Genus, Species, and sTrain, respectively (for prokaryotes, <D> and <K> do not make difference; they are kept for future extension of CVTree to Eukarya). A missing or uncertain rank carries a fixed indicator "Unclassified", for example, <F>Unclassified denotes an as-yet-not-designated family.

We note that in the early days of whole-genome phylogenetic studies, say, in 2004 [16], genomes were given abbreviations in figures and tables. With the number of genomes growing into hundreds and thousands, it is more convenient for the experts to deal with fully fledged names including strain tags, and so on, as is done in CVTree3.

### 3.3. Interactive display of trees

Because it is hard to comprehend a phylogenetic tree with many thousands of leaves, CVTree3 is equipped with an interactive display capable of collapsing or expanding branches in the tree, keeping the overall topology unchanged. For example, when there are 179 genomes assigned to the class <C>*Epsilonproteobacteria* in the input data set and they all appear in a monophyletic branch, the whole branch may be collapsed into a single leaf labeled by the class name with the total number of genomes indicated in parentheses. In this way, the number of leaves in the whole tree may be greatly reduced, while the overall structure is clearly represented. In fact, at $K = 5$ or 6 a big CVTree usually appears in a maximally collapsed form with only three branches as shown in **Figure 1**.

In **Figure 1**, all three collapsed leaves would have appeared in red, because red color is used to represent monophyletic entries. If not monophyletic, they are usually shown in blue. Other

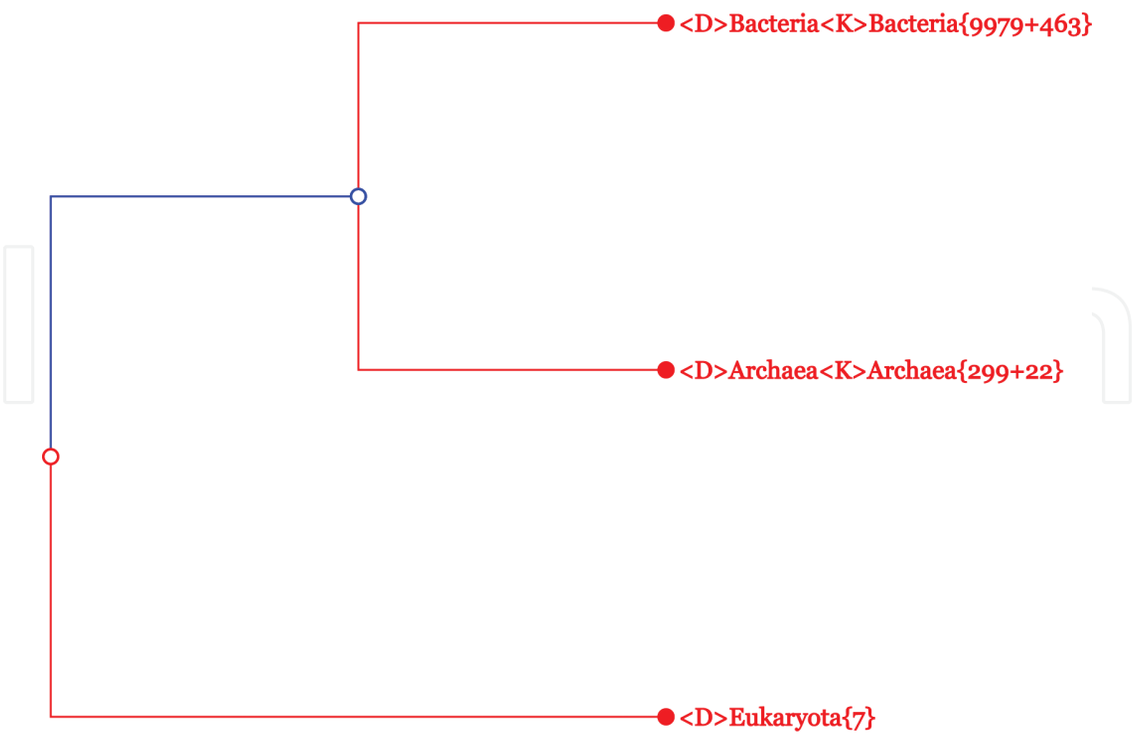


**Figure 1.** A maximally collapsed CVTree confirming the three main domains of life as suggested by Carl Woese. The numbers in the expression {9979+463} tell that there are 9979 genomes with complete lineage information and 463 with incomplete or absent information.

colors used in CVTree3 include green (taxon matching a Query) and purple (taxon with Unclassified rank). By the way, making an enquiry provides the quickest way to get to a taxon of interest. For example, typing *Epsilonproteobacteria* to replace "Query Search" in the preamble of the tree display immediately leads to a properly collapsed tree with the taxon in enquiry shown in green.

### 3.4. Lineage modification and re-collapsing a tree

A complete lineage may be incorrect. In some cases, an "unclassified" taxon name may be inferred from its neighborhood. When inspecting a tree, it may be tempting to modify some lineage information in order to reach better agreement between the tree-branching order and taxonomy. The CVTree3 Web Server provides a mechanism to demonstrate the consequences of trial lineage modifications. To this end, the user prepares a "Lineage Modification" file comprising separate lines of the following form:

old_lineage<space>new_lineage

(<space> means typing a space in between the two pieces of information). For example, there is a monophyletic genus *Aliivibrio* represented by eight genomes in CVTree. However, as a whole this cluster gets inserted into the genus *Vibrio* and thus violates the monophyly of the latter, as is shown in **Figure 2**.

Upon seeing this situation, one may wish to rename *Aliivibrio* simply as *Vibrio*. One adds the following line in the "Lineage Modification" file:

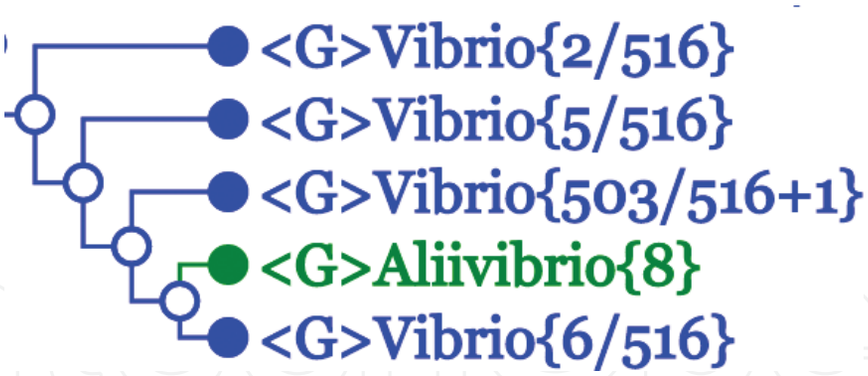


**Figure 2.** A monophyletic genus Allivibrio represented by eight genomes gets into the genus Vibrio and violates the monophyly of the latter. Vibrio{5/516} means this line represents five genomes out from a total of 516 Vibrio genomes.

<G>Aliivibrio <G>Vibrio

(sometimes more ranks must be included in order to make the modification unique) and then submit the file to the server for re-collapsing. After a while, a renewed tree appears in which the entries shown in **Figure 2** shrink to a single line <G>*Vibrio*{524}. Nevertheless, it must be emphasized that any actual lineage modification should be carried out in accordance with the International Code of Nomenclature of Prokaryotes [37] and be published in an appropriate journal. What described above only provides a trial means.

### 3.5. Report of taxa statistics

When comparing a tree containing thousands of leaves with an underlying taxonomy, one would like to check the overall "convergence," that is, how many taxa are monophyletic or non-monophyletic, at all taxonomic ranks from phyla down to species. The two previous releases of CVTree Web Server ran under a fixed peptide length *K*. It was up to the user to collect and observe the convergence of trees under different *K*. The CVTree3 Web Server, however, produces trees for several *K*-values in one run, say, for $K = 3$ to 9. This provides a new angle to evaluate the quality of the resulted trees. Obviously, it is not an easy job to accomplish if done manually. CVTree3 generates a summary table after each collapsing and re-collapsing. The summary is given as a long list arranged according to the taxonomic hierarchy. As a taxon that contains only a single species must be "embarrassingly" monophyletic, such items may be suppressed in the report. Take again the example given in the previous subsection. In the summary before doing the lineage modification, there is a monophyletic genus *Aliivibrio*{8} and a non-monophyletic genus *Vibrio*{516}. After making the modification only a monophyletic *Vibrio*{524} remains, but the total number of monophyletic genera does not change, as *Vibrio* adds to monophyletic ones but *Aliivibrio* drops out from the summary.

### 3.6. Output of print-quality sub-tree figures

Any part of a tree may be extracted to generate print-quality figures. The CVTree3 Web Server provides several formats for output. The formats include Encapsulated PostScript (.eps), Scalable Vector Graph (.svg), Portable Document File (.pdf), and Portable Network Graphics (.png). The output figures may be in the original colors or be made monochromatic.

## 4. Taxonomic resources for prokaryotes

Taxonomy has always been a work in progress. As we shall refer to taxon names and ranks repeatedly, it helps to indicate the main taxonomic resources used in our study.

### 4.1. The NCBI taxonomy database

The NCBI taxonomy database [36] carries a disclaimer after each entry that "it is not an authoritative source for nomenclature or classification." However, the NCBI taxonomy reflects much dynamic and up-to-date knowledge, as for any deposited molecule data, the authors would supply a piece of taxonomic information, not necessarily agreeing with the "generally accepted" opinion but better than none. This said, the NCBI taxonomy is taken as a start point for making a default Lineage Information file that comes with CVTree3.

### 4.2. Bergey's manual

The second edition of Bergey's Manual of Systematic Bacteriology [29], a grandiose work of more than 8600 pages, took 12 years (2001–2012) to accomplish. Upon its completion, Bergey's Manual Trust made it clear that this was the last hardcopy edition. Future editions would be electronic. In fact, the electronic Bergey's Manual of Systematics of Archaea and Bacteria, abbreviated as BMSAB, saw the light at the end of 2015 in the Wiley Online Library [30]. We note that BMSAB is organized on the basis of genera and many genus descriptions are taken from the previous volumes of Bergey's Manual [29] without any change.

### 4.3. The Prokaryotes IV

The Prokaryotes, a collective multivolume work, has its fourth edition [38] published in 2013–2014. Six volumes out of a total of 11 are devoted to classifications of *Bacteria* and *Archaea*. The taxa are divided basically by families with historical accounts as well as some emphasis on genera and species proposed after the corresponding volumes of Bergey's Manual [29]. Volumes in this fourth edition draw much information from the All-Species Living Tree project [39, 40] which is an initiative of the journal Systematics and Applied Microbiology in collaboration of a group of European microbiologists to reconstruct a single phylogenetic tree based on 16S rRNA sequences for all available type strains of *Bacteria* and *Archaea*. The latest release (*http://www.arb-silva.de/projects/living-tree/)* LTPs123 of September 2015 was based on 11,490 *Bacteria* and 449 *Archaea* sequences. In what follows, the All-Species Living Tree is abbreviated as LVTree when needed.

### 4.4. International Journal of Systematic and Evolutionary Microbiology

International Journal of Systematic and Evolutionary Microbiology (IJSEM) is the standard place to publish taxonomic proposals. Proposals published in a few journals other than IJSEM only make a small fraction of that in IJSEM. Taken altogether, about 800 new taxa appear yearly for the time being. As there is necessarily a time lag for new taxa to be recorded in BMSAB [30] or in The Prokaryotes IV [38], one must take into account data published in periodicals such

as IJSEM and alike. To this end, a timely and helpful resource is the List of Prokaryotes with Standing in Nomenclature, abbreviated as LPSN [5]. Speaking about nomenclature, one must note that a preliminary draft of the long-due revision of International Code of Nomenclature of Prokaryotes (subtitled as 2008 Revision) has appeared in IJSEM in 2015 [37].

## 5. Applications of CVTree

Now, we are prepared to summarize the success of the CVTree approach and to discuss its prospective.

### 5.1. Retrospective verifications of CVTree

First of all, taxonomic revisions published in recent years all agree with the branching orders in CVTrees without exceptions as long as the corresponding sequenced genomes are available for comparison. In particular, 16 such cases were listed in [34]. This kind of agreements may be taken as retrospective verifications of CVTree results. A recent example deals with a debate on the taxonomic placement of *Eubacterium rectale* when CVTree results support the objection to reclassify this species into a new genus *Agathobacter* [41]. Moreover, CVTree predicts that the species under debate should belong to an existing genus *Roseburia*. We mention two more examples among many. First, earlier predictions of CVTree helped to transfer *Cellvibrio gilvus* from the originally assigned class *Gammaproteobacteria* to the genus *Cellulomonas* in phylum *Actinobacteria* [19]. Second, CVTree revealed the wrong taxonomic assignment of *Burkholderia* JV3 strain and suggested to bring it to the genus *Stenotrophomonas* [19].

### 5.2. Mega-classification of Bacteria and Archaea

Cavalier-Smith [42] coined the term mega-classification for taxonomic demarcation of the ranks order, class, and higher. Up to present time, the highest taxonomic rank recognized by the International Code of Nomenclature of Prokaryotes [37] is class. A proposal to include the rank phylum in the Code appeared only quite recently [43]. With many thousands of sequenced genomes available nowadays, CVTree may help to improve the mega-classifications in many aspects. Due to space limitation of this chapter, we only briefly touch on some facts at the phylum level.

For the time being, more than 400 *Archaea* genomes have been sequenced. They are well organized at ranks above class or even order [44]. For example, the phylum *Crenarchaeota* contains a single class *Thermoprotei*; the phylum *Euryarchaeota* consists of eight to nine classes; the phylum *Thaumarchaeota* proposed a few years ago is also supported by CVTree. A few newly proposed but not yet fully established archaeal phyla may require more genomic data to confirm.

As regarding the bacterial branch, in CVTrees constructed by using 10,442 *Bacteria* genomes, an overwhelming majority of phyla comes out monophyletic without making any lineage modification or only with minor modifications (Ref. [34] where the tree was based on fewer genomes). **Table 1** compares all phyla which are monophyletic in LVTree with their counterparts in CVTree.

| Phylum | LVTree | CVTree |
|---|---|---|
| Acidobacteria | 25 | 17/19+4, Note 1 |
| Actinobacteria | 2897 | 1705/1742+26, Note 2 |
| Aquificae | 28 | 21 |
| Armatimonadetes | 3 | 2 |
| Bacteroidetes | 1240 | 649/651+14, Note 3 |
| Caldiserica | 1 | 1 |
| Chlamydiae | 13 | 131 |
| Chlorobi | 11 | 12 |
| Chloroflexi | 23 | 17 |
| Chrysiogenetes | 4 | 2 |
| Cyanobacteria | 16 | 198 |
| Deferribacteres | 11 | 6 |
| Deinococcus_Thermus | 84 | 54 |
| Dictyoglomi | 2 | 4 |
| Elusimicrobia | 6 | 2 |
| Fibrobacteres | 4 | 2 |
| Fusobacteria | 39 | 49 |
| Gemmatimonadetes | 1 | 2 |
| Ignavibacteriae | 2 | 4 |
| Lentisphaerae | 4 | 1 |
| Nitrospirae | 7 | 7 |
| Planktomycetes | 23 | 19 |
| Spirochaetes | 93 | 104, 12, 9+38, Note 4 |
| Synergistetes | 23 | 18 |
| Tenericutes | 186 | 193/203+1, Note 5 |
| Thermodesulfobacteres | 8 | 10 |
| Thermotogae | 43 | 57 |
| Verrucomicrobia | 43 | 22 |

Note 1. <F>Holophagaceae joins <O>Myxococcales in the next branch.

Note 2. <C>Coriobacteriia escapes from the main cluster of <P>Actinobacteria.

Note 3. Two genera from <F>Chitinophagaceae escape from the main cluster of the latter, separated by <P>Chlorobi and <P>Ignavibacteriae.

Note 4. <P>Spirochaetes splits into three disjoint orders separated by other phyla. See discussion below.

Note 5. <G>Acholeplasma escapes from <P>Tenericutes.

Numerals indicate the number of 16S rRNA sequences or genomes in each phylum. For the meaning of $n + m$, please see the caption of **Figure 1**.

**Table 1.** A comparison of monophyletic bacteria phyla in LVTree and CVTree.

In order to make the comparison more effective, we have transplanted many of the CVTree3 features to a LVTree Viewer [45]. Users are advised to make Query Search on the same taxon name alternately in CVTree3 Web Server and LVTree Viewer.

In spite of the "overemphasis on rRNA similarity as a single arbitrary criterion of relatedness" [42], the agreement between CVTree and LVTree at the phylum level is remarkable in **Table 1**.

Due to space limitation, we will not elaborate the Notes in **Table 1** except for making a remark on Note 4. The phylum *Spirochaetes* splits into three disjoint monophyletic clusters corresponding to the orders *Spirochaetales*, *Brachyspirales*, and *Bdellovibrionales*, each essentially containing one family. This might be the largest discrepancy between CVTree and LVTree phylogenies besides the two phyla discussed in the subsequent text.

In fact, two "big" phyla were absent in **Table 1**: *Proteobacteria* and *Firmicutes*. The phylum *Proteobacteria*, represented by the largest number of genomes, splits basically into two disjoint clusters. Most of the taxonomic uncertainties concentrate in the phylum *Firmicutes*. In fact, in the last 20 years, many new phyla have been extracted from *Firmicutes*, including *Actinobacteria* and *Tenericutes*, and the process still continues. As an example, **Figure 3** shows how *Coprothermobacter* takes the position of an independent phylum in CVTree. It was labeled as an "established phylum" in a 2004 census [46] but still listed in BMSAB [30] and The Prokaryotes IV [38] as a genus within *Firmicutes* with proviso.

### 5.3. Taxonomic position of newly sequenced genomes without proper standing in nomenclature

As the cost of sequencing, a bacterial genome drops below the expenses of average phenotyping experiments, many biological studies now start from genome sequencing. However, a substantial part of newly sequenced genomes appears without validly published names and proper lineage information. The corresponding teams are not interested and sometimes not in a position with budget and manpower to fill up the gap in compliance with the International Code of Nomenclature of Prokaryotes. After extracting the interested information, the genomes were dumped as Permanent Drafts. According to GOLD [23], Permanent Drafts make the most rapidly growing part of genomic data. If the situation persists, as Barny Whitman warned, the microbiological "literature will be once again be full of names of uncertain meaning, and the difficult work of several generations of microbial systematists will be undone" [47].

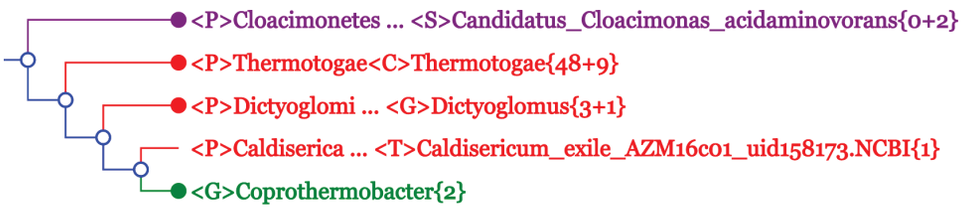


**Figure 3.** The genus *Coprothermobacter* from the phylum *Firmicutes* acquires status of a separate phylum. This figure also shows that Candidatus_Cloacimonas_acidaminovorans, an abundant bacterium in biogas reactors, actually takes the position of a phylum (formerly candidate phylum WWE1).

As an example, let us look at **Figure 4** which was cut from a CVTree based on 10,770 genomes. There is a genome with an illegal name Listeriaceae_bacterium_FSL_A5_0209. An inspection of its neighborhood hints on a plausible lineage like

<F>Listeriaceae<G>Listeria<S>Listeria_newyorkensis<T>Listeriaceae_bacterium_FSL_A5_0209

The line above assumes that it is a strain of an existing species, *L. newyorkensis*. Whether the assumption holds or not requires additional evidence by using other markers, a task often not possible or not worthwhile to do. In order to keep the possibility that this genome belongs to a new species other than *L. newyorkensis*, one may replace the <S> part of the lineage information by <S>Listeria_sp_FSL_A5_0209. This lineage modification eventually leads to a monophyletic genus *Listeria*{64} next to <G>*Brochothrix*{2}. The last two genera, taken together, make a monophyletic family *Listeriaceae*{66}. The two types of tags, "_sp_" and "_bacterium_," are frequently encountered in "illegal", that is, not validly published, names. As at the time of writing, in our genome warehouse, there are more than 6000 names that come with a tag "_bacterium_" and more than 2000 names that contain a tag "_sp_." These names may be at least partially improved by using CVTree, but not by LVTree as the latter excludes such names by design.

### 5.4. High resolution at the species level and below

Contrary to 16S rRNA sequence analysis, which does not possess high enough resolution at species and subspecies levels, CVTree approach distinguishes itself for extremely high-resolution power at infra-subspecific levels. This capability opens up new horizons in basic research as well as in applications. We briefly mention a few.

#### *5.4.1. Population genetics of prokaryotes*

Compared to Eukarya, the population genetics of prokaryotes is a much less studied subject. So far, only the clone structure of commensal *Escherichia coli* has been explored to some extent, see, for example, [48]. The major branches of *E. coli* strains in CVTree agree with the so-called phylogroups very well not only for the commensal groups A, B1, and B2 but also for pathogenic groups D and E, see Figure 5 in [34]. However, serotyping tests generate much finer divisions of *E. coli* strains and the correlation of serotypes with the branching orders in CVTrees has not been fully elucidated. In contrast to serotypes of *E. coli*, serotypes of *Streptococcus pyogenes* correlate well with CVTree branches [34].

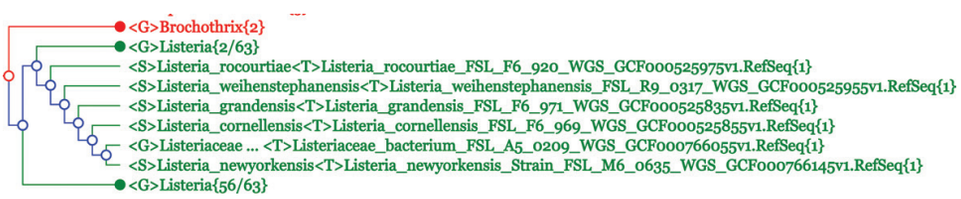


**Figure 4.** The questionable rank <G>*Listeriaceae* should be <G>*Listeria*; the illegal genome name Listeriaceae_bacterium_FSL_A5_0209 may be a new species Listeria_sp_FSL_A5_0209 requiring a formal name or a strain of the existing species Listeria_newyorkensis with strain tag FSL_A5_0209.

### *5.4.2. Distinguishing pathogenic bacteria species*

Since the late 1980s, DNA-DNA hybridization (DDH) has become the standard measure to delineate bacteria species. As the clinically distinguishable *Yersinia pseudotuberculosis* and *Y. pestis* strains could not be resolved by using DDH, it was proposed to combine them into one and the same species. The proposal, however, was rejected by the Judicial Commission "because of possible danger to public health if there was confusion regarding *Y. pestis*, the plague bacillus" [49]. In CVTree, however, the corresponding strains go to different branches causing no problem in treating them as two species [19].

Another example concerns whether *Shigella* strains are clones within *E. coli* species or make individual species in the genus *Escherichia* on equal footing as *E. coli*. Even many sequence-based analyses put the former within the branches of the latter. Consequently, there seems to be a consensus that the genus name *Shigella* is kept only for historical and clinical reasons. CVTree's resolution again exceeds many other approaches by showing that the four well-described *Shigella* species are indeed sister species of *E. coli* within the same genus *Escherichia* [50].

### *5.4.3. Biogeography of bacteria*

Geographic variations of multicellular plants and animals played an inspiring role for Charles Darwin to conceive the theory of evolution. Darwin did not mention microbes in his *Origin of Species* due to obvious limitations of his time. However, does it make sense to study geographic distribution of bacteria nowadays? The division of *Helicobacter pylori* into seven or more subpopulations carries geographic imprints which may be left by the migration of their human hosts [51]. The availability of more than 550 sequenced *H. pylori* genomes for the time being allows a much deeper study of the problem than a decade ago. A much more direct example without the intermediate influence of hosts is provided by "*Sulfolobus islandicus*" strains collected from various volcanic hot springs in Eurasian and North American continents. Genomic analyses including CVTree revealed that these clearly separated genomes should still be considered as geovars of the same species [52].

### *5.4.4. Electronic screening of bacterial strains*

Many bacterial strains, naturally occurring in environment or intentionally made mutants, are screened for pathogenicity, drug-resistance, or metabolic products. These are costly and time-consuming jobs. When a certain amount of experimental data has been accumulated, mapping of the data onto a phylogenetic tree and picking up the most promising leaves for further exploration would significantly increase the efficiency of the screening process.

## 6. Concluding remarks

Biology starts from classification. However, the discipline of taxonomy is declining as less and less young scientists enter the field. The situation is especially true in microbiology. However,

as eloquently pointed out by Barny Whitman, the supervisor of Bergey's Manual [29, 30], the solution lies in DNA sequencing and genomic analyses [47]. Recently, Whitman put forward a proposal to expand type material for naming prokaryotes to include DNA sequences [53]. With this proposal accepted by the microbiology research community, phylogeny and taxonomy of prokaryotes will ultimately become by-products of genomic analyses. Convenient and convincing phylogenomic tools such as CVTree are deemed to play an essential role in the future.

## Acknowledgements

The work described in this chapter has been supported by the National Basic Research Program of China (973 Grant Nos. 2007CB814800 and 2013CB834100) and by the State Key Laboratory of Applied Surface Physics and Department of Physics of Fudan University.

## Author details

Guanghong Zuo and Bailin Hao*

*Address all correspondence to: hao@mail.itp.ac.cn

T-Life Research Center and Department of Physics, Fudan University, Shanghai, China

## References

[1] Woese CR, Fox GE. Phylogenetic structure of the prokaryotic domains: The primary kingdoms. Proceedings of the National Academy of Sciences United States of America. 1977;**74**:5088-5090

[2] Woese CR, Kandler O, Wheelis ML. Towards a natural system of organisms: Proposal for the domains of *Archaea*, *Bacteria*, and *Eucarya*. Proceedings of the National Academy of Sciences United States of America. 1990;**87**:4576-4579

[3] Whitman WB, Coleman DC, Wiebe WJ. Prokaryotes: The unseen majority. Proceedings of the National Academy of Sciences United States of America. 1998;**95**:6578-6583

[4] Editorial. Microbiology by numbers. Nature Reviews Microbiology. 2011;**9**:628. DOI:10.1038/nrmicro2644

[5] Parte AC. LPSN-list of prokaryotic names with standing in nomenclature. Nucleic Acids Research. 2014;**42:**D617-D616. DOI:10.1093/nar/gkt1111

[6] Fox GE, Pechman KR, Woese CR. Comparative cataloging of 16S ribosomal ribonucleic acid: Molecular approach to prokaryotic systematics. International Journal of Systematic Bacteriology. 1977;**27**:44-57

[7] Fox GE, Magrum LJ, Balch WE, Wolfe RS, Woese CR. Classification of methanogenic bacteria by 16S ribosomal RNA characterization. Proceedings of the National Academy of Sciences United States of America. 1977;**74**:4537-4541

[8] Wayne LG, Brenner DJ, Colwell RR, Grimont PAD, Kandler O, Krichevsky MI, Moore LH, Moore WEC, Murray RGE, Stackebrandt E, Starr MP, Truper HG. Report of the Ad Hoc Committee on reconciliation of approaches to bacterial systematics. International Journal of Systematic Bacteriology. 1987;**77**:463-464

[9] Woese CR. A manifesto for microbial genomes. Current Biology. 1998;**8**:R781-R783

[10] Woese CR. The quest for Darwin's grail. ASM News. 1999;**65**:260-263

[11] Coenye T, Gevers D, Van de Peer Y, et al. Towards a prokaryotic genomic taxonomy. FEMS Microbiology Reviews. 2005;**29**:147-167

[12] Konstantinidis KT, Tiedje JM. Prokaryotic taxonomy and phylogeny in the genomic era: Advancements and challenges ahead. Current Opinion in Microbiology. 2007;**10**:504-509

[13] Klenk H-P, Göker M. En route to a genome-based classification of *Archaea* and *Bacteria*?. Systematic and Applied Microbiology. 2010;**33**:175-182

[14] Chun J, Rainey FA. Integrating genomics into the taxonomy and systematics of the Bacteria and Archaea. International Journal of Systematic and Evolutionary Microbiology. 2014;**64**:316-324

[15] Hao B, Qi J, Wang B. Prokaryote phylogeny based on complete genomes without sequence alignment. Modern Physics Letters B. 2003;**17**:91-94

[16] Qi J, Wang B, Hao B. Whole proteome prokaryote phylogeny without sequence alignment: A K-string composition approach. Journal of Molecular Biology. 2004;**58**:1-11. DOI:10.1007/s00239-003-2493-7

[17] Gao L, Qi J, Sun J, Hao B. Prokaryote phylogeny meets taxonomy: An exhaustive comparison of composition vector trees with systematic bacteriology. Science in China Series C: Life Sciences. 2007;**50**:587-599. DOI:10.1007/s11427-007-0084-3

[18] Li Q, Xu Z, Hao B. Composition vector approach to whole-genome-based prokaryotic phylogeny: Success and foundations. Journal of Biotechnology. 2010;**149**:115-119. DOI:10.1016/j.jbiotec.2009.12.015

[19] Hao B. CVTrees support the Bergey's systematics and provide high resolution at species levels and below. The Bulletin of BISMiS. 2011;**2**:189-196

[20] Zuo G, Li Q, Hao B. On K-peptide length in composition vector phylogeny of prokaryotes. Computational Biology and Chemistry. 2014;**53**:166-173. DOI:10.1016/j.compbiolchem.2014.08.02

[21] Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science. 1995;**269**:496-512

[22] Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult CJ, Kerlavage AR, Sutton G, et al. The minimal gene complement of *Mycoplasma genitalium*. Science. 1995;**270**:397-404

[23] Reddy TBK, Thomas A, Stamatis D, Bertsch J, Isbandi M, Jansson J, Mallajosyula J, Pagani I, Lobos E, Kyrpides N. The genomes OnLine Database v.5:a metadata management system based on a four level (meta)genome project classification. Nucleic Acids Research. 2014;**43**:D1099-D1106. DOI:10.1093/nar/gku950

[24] Markowitz VM, Chen I-MA, Palaniappan K, Chu K, Szeto E, Pillay M, Ratner A, Huang J, Woyke T, Hutermann M, Anderson I, Billis K, Varghese N, Mavromatis K, Pati A, Ivanova NM, Kyrpies NC. IMG 4 version of the integrated microbial genomes comparative analysis system. Nucleic Acids Research. 2014;**42**:D560-D567. DOI:10.1093/nar/gkt963

[25] Han K, Li Z-F, Peng R, Zhu L-P, Zhou T, Wang L-G, Li S-G, Zhang X-B, Hu W, Wu Z-H, Qin N, Li Y-Z. Extraordinary expansion of a *Sorangium cellulosum* genome from an alkaline milieu. Scientific Reports. 2013;**3**:2101. DOI:10.1038/srep02101

[26] Kimura K. The Neutral Theory of Molecular Evolution. Cambridge University Press; 1984. p. 384

[27] Hu R, Wang B. Statistically significant strings are related to regulatory elements in the promoter region of *Saccharomyces cerevisiae*. Physica A. 2001;**290**:464-474

[28] Zuo G, Xu Z, Yu H, Hao B. Jackknife and bootstrap tests for the composition vector trees. Genomics Proteomics Bioinformatics. 2010;**8**:262-267. DOI:10.1016/S1672-0229(10)60028-9

[29] Bergey's Manual Trust. Bergey's Manual of Systematic Bacteriology. 2nd ed. Vol. 1-5 New York, NY: Springer; 2001-2012. DOI:10.1007/b92997, 10.1007/978-0-387-68572-4, 10.1007/978-0-387-68233-4

[30] Whitman WB, et al., editors. Bergey's Manual of Systematics of Archaea and Bacteria. Online at John Wiley & Sons; 2015. DOI:10.1002/9781118960608

[31] Qi J, Luo H, Hao B. CVTree: A phylogenetic tree reconstruction tool based on whole genomes. Nucleic Acids Research. 2004;**32**:W45-W47. DOI:10.1093/nar/gkh362

[32] Xu Z, Hao B. CVTree update: A newly designed phylogenetic study platform using composition vectors and whole genomes. Nucleic Acids Research. 2009;**37**:W174-W178. DOI:10.1039/nar/gkp278

[33] CVTree Web Server [Internet]. 2009. Available from: http://tlife.fudan.edu.cn/cvtree [Accessed: 15-September-2016]

[34] Zuo G, Hao B. CVTree3 Web Server for whole-genome-based and alignment-free prokaryotic phylogeny and taxonomy. Genomics Proteomics Bioinformatics. 2015;**13**:321-331. DOI:10.1016/j.gpb.2015.08.004

[35] CVTree3 Web Server [Internet]. 2015. Available from: http://tlife.fudan.edu.cn/cvtree3 [Accessed: 15-September-2016]

[36] The NCBI Taxonomy Database [Internet]. 2016. Available from: https://www.ncbi.nlm.nih.gov/taxonomy [Accessed: 15-September-2016]

[37] Parker CT, Tindall BJ, Garrity GM, editors. International Code of Nomenclature of Prokaryotes. International Journal of Systematic and Evolutionary Microbiology. 2015;**1**:465-481. DOI:10.1099/ijsem.0.000778

[38] Rosenberg E, DeLong EF, Lory S, Stackebrandt E, Thompson F, editors. The Prokaryotes. 4th ed. Vol. 1-11 Berlin Heidelberg: Springer; 2013-2014. DOI:10.1007/978-3-642-31331-8, 10.1007/978-3-642-39044-9

[39] Yarza P, Richter M, Peplies J, Euzeby J, Amann R, Schleifer K-H, Ludwig W, Gloeckner FO, Rosselo-Mora R. The All-Species Living Tree project: A 16S rRNA-based phylogenetic tree of all sequenced type strains. Systematic and Applied Microbiology. 2008;**31**:241-250. DOI:10.1016/j.syapm.2008.07.001

[40] Yarza P, Sproer C, Swiderski J, Mrotzek N, Spring S, Tindall BJ, et al. Sequencing orphan species initiative (SOS): Filling the gaps in the 16S rRNA gene sequence database for all species with validly published names. Systematic and Applied Microbiology. 2013;**36**:69-73. DOI:10.1016/j.syapm.2012.12.006

[41] Zuo G, Hao B. Whole-genome-based phylogeny supports the objections against the reclassification of *Eubacterium rectale* to *Agathobacter rectalis*. International Journal of Systematic and Evolutionary Microbiology. 2016;**66**(6):2451. DOI:10.1099/ijsem.0.001407

[42] Cavalier-Smith T. The neomuran origin of archaebacteria, the negibacterial root of the universal tree and bacterial megaclassification. International Journal of Systematic and Evolutionary Microbiology. 2002;**52**:7-76

[43] Oren A, da Costa MS, Garrity GM, Rainey FA, Rosselo-Mora R, Schick B, Sutcliffe I, Trujilo ME, Whitman WB. Proposal to include the rank of phylum in the International Code of Nomenclature of Prokaryotes. International Journal of Systematic and Evolutionary Microbiology. 2015;**65**:4284-4287

[44] Zuo G, Xu Z, Hao B. Phylogeny and taxonomy of *Archaea*: A comparison of the whole-genome-based CVTree approach with 16S rRNA sequence analysis. Life. 2015;**5**:949-968. DOI: 10.3390/life5010949

[45] Zuo G, Zhi X, Xu Z, Hao B. LVTree Viewer: An interactive display for the all-species living tree incorporating automatic comparison with prokaryotic systematics. Genomics Proteomics Bioinformatics. 2016;**14**:94-102. DOI:10.1016/j.gpb.2015.12.002

[46] Schloss PD, Handeldsman J. Status of the microbial census. Microbiology and Molecular Biology Reviews. 2004;68:686-691. DOI: 10.1128/MMBR.68.4.686-691.2004

[47] Whitman WB. Intent of the nomenclatural code and recommendations about naming new species based on genomic sequences. The Bulletin of BISMiS. 2011;**2**:135-139

[48] Tenaillon O, Skurnik D, Picard B, Demamur E. The population genetics of commensal Escherichia coli. Nature Reviews Microbiology. 2010;8:207-217. DOI:10.1038/nrmicro2298

[49] Brenner DJ, Staley JT, Krieg NR. Classification of Prokaryotic Organisms and the Concept of Bacterial Speciation. Chapter 5 in Bergey's Manual [29]. Vol. 2B; 2002. pp. 27-32

[50] Zuo G, Xu Z, Hao B. *Shigella* strains are not clones of *Escherichia coli* but sister species in the genus *Escherichia*. Genomics Proteomics Bioinformatics. 2013;**11**:61-65. DOI: 10.1016/j.gpb.2012.11.002

[51] Falush D, Wirth T, Linz B, Pritchard JK, Stephens M, Kid M, et al. Traces of human migrations in *Helicobacter pylori* populations. Science. 2003;**299**:1582-1585

[52] Zuo G, Hao B, Staley JT. Geographic divergence of "*Sulfolobus islandicus*" strains assessed by genomic analyses including electronic DNA hybridization confirms they are geovars. Antonie van Leeuwenhoek. 2014;**105**:431-435. DOI: 10.1007/s10482-013-0081-4

[53] Whitman WB. Modest proposals to expand the type material for naming of prokaryotes. International Journal of Systematic and Evolutionary Microbiology. 2016;**66**:2108-2112. DOI: 10.1099/ijsem.0.000980