J Mol Evol (2004) 58:1–11
DOI: 10.1007/s00239-003-2493-7

JOURNAL OF MOLECULAR EVOLUTION

Whole Proteome Prokaryote Phylogeny Without Sequence Alignment: A *K*-String Composition Approach

Ji Qi,[1,2] Bin Wang,[1] Bai-lin Hao[1,2]

[1] The Institute of Theoretical Physics, Academia Sinica, Beijing 100080, China
[2] The T-Life Research Center, Fudan University, Shanghai 200433, China

Received: 29 January 2003 / Accepted: 9 May 2003

Abstract. A systematic way of inferring evolutionary relatedness of microbial organisms from the oligopeptide content, i.e., frequency of amino acid *K*-strings in their complete proteomes, is proposed. The new method circumvents the ambiguity of choosing the genes for phylogenetic reconstruction and avoids the necessity of aligning sequences of essentially different length and gene content. The only "parameter" in the method is the length *K* of the oligopeptides, which serves to tune the "resolution power" of the method. The topology of the trees converges with *K* increasing. Applied to a total of 109 organisms, including 16 Archaea, 87 Bacteria, and 6 Eukarya, it yields an unrooted tree that agrees with the biologists' "tree of life" based on SSU rRNA comparison in a majority of basic branchings, and especially, in all lower taxa.

Key words: Prokaryote — Phylogeny — Archaea — *K*-strings — Compositional distance — Tree of life

Introduction

The advent of molecular phylogeny (Zuckerkandl and Pauling 1965) and the progress in protein and DNA sequencing thenceforth have greatly deepened the understanding of evolution. This development has provided a new tool for the classification of microbial organisms since morphological and metabolic features that may be used to infer phylogenetic relationships are rather limited for microbes compared to more complex forms of life. The justification of the endosymbiont origin of mitochondria and chloroplast as well as the division of life into the three main domains (Archaea, Bacteria, and Eukarya) is surely a major achievement of molecular phylogeny. However, contrary to general expectations, the increasing availability of complete microbial genomes has cast doubt (Doolittle 1999) instead of adding details to the phylogenetic tree which was based on the comparison of Small Subunit (SSU) rRNA sequences (Woese and Fox 1977) or other conserved genes, e.g., the elongation factor (Baldauf et al. 1996).

It turns out that different genes may tell different stories. For example, the gene coding for MHGCoA reductase puts Arcfu (species names and their abbreviations are listed in the Appendix), a definite archaean, in the Bacteria (Doolittle 2000). In addition, the tendency of the two hyperthermophilic bacteria, Aquae and Thema, to be put into Archaea, have intensified the debate on whether there has been widespread lateral or horizontal gene transfer among species (Aravind et al. 1998; Doolittle 1999; Ragan 2001). And this in turn calls into question the basic existence of the "tree of life." In only 3 years commentary on this controversial situation has escalated from suggestions that the tree of life has been "shaken" (Pennisi 1998) to some calling it time to "uproot" the tree of life (Pennisi 1999; Doolittle 2000). At least, it is now a consensus that one should

Correspondence to: Ji Qi; *email:* qiji@itp.ac.cn

not equate a tree inferred from a particular gene to the real tree of life.

In the meantime there have been several attempts to infer prokaryote phylogeny from complete genomes. This includes the gene content (Snel et al. 1999; Huynen et al. 1999; Tekaia et al. 1999), the presence/absence of genes in clusters of orthologs (Wolf et al. 2001), the supertree (Daubin et al. 2001), the conserved gene pairs (Wolf et al. 2001), and some other methods (e.g., Fitz-Gibbon and House 1999). While almost all these methods yield the trifurcation of the three main domains of life, the major branchings within Archaea and Bacteria remain poorly resolved. Furthermore, these methods eventually rely on sequence alignments and, in some cases, need fine-tuning and adjustment. So far there are no widely accepted ways to infer phylogenetic relationships from complete genome data. There is an urgent need to develop new phylogenetic methods utilizing the ever-increasing amount of molecular data, in particular, the complete genomes of organisms.

In this paper we describe an entirely new and essentially simple method that leads to results comparable with the latest classification in systematic bacteriology as reflected in the 2001 edition of *Bergey's Manual of Systematic Bacteriology* and summarized in the *Taxonomic Outline of Prokaryote Genera* (Garrity et al. 2001).

The traditional approach to construct molecular phylogenetic trees can hardly be applied to complete genomes: it does not make sense to align two complete genomes since every species has its own gene content and gene order, not to mention the different sizes of the genomes. In order to bypass the difficulty in using the whole genome data we propose to determine the evolutionary distance between organisms by counting the number of oligopeptide strings of a fixed length K in the collection of their protein sequences without doing sequence alignment. An essential step in our approach is the subtraction of a random background. Our method does not contain "free parameters," as there was neither choice of genes nor multialignment of sequences, which would implicitly depend on score matrices and other factors.

Materials and Methods

Genome Data Sets

We have included all prokaryote complete genomes that were publicly available by the end of December 2002. There are two available sets of prokaryote complete genomes. Those in GenBank (Benson et al. 2003) are the original data submitted by their authors. Those at the National Center for Biotechnological Information (NCBI) (Wheeler et al. 2003) are reference genomes curated by NCBI staff. Since the latter represents the approach of one and the same group using the same set of tools, it may provide a more consistent background for comparison. Therefore, we used all the translated amino acid sequences (the .faa files with NC_accession numbers) from NCBI. Six Eukaryotes were added for reference. The list of all genomes used is given in the Appendix. If a genome consists of more than one chromosome, we collected all the translated sequences. Altogether 103 organisms from 87 prokaryotic species distributed in 61 genera, 49 families, 41 orders, 24 classes, and 13 phyla are represented in our trees.

Frequency or Probability of Appearance of K*-Strings*

Comparison of G + C content or amino acid composition has long been a standard practice in analyzing biological sequences. By extending single-nucleotide or single-amino acid counting to longer strings, one increases the "resolution power" of the analysis, takes into account short-term correlations in the sequences, and enhances the species specificity of some sequence features. Among early work along this line we mention the use of dinluceotide relative abundance as a genomic signature (Karlin and Burge 1995). Given a DNA or amino acid sequence of length L, we count the number of appearances of (overlapping) strings of a fixed length K in the sequence. The counting may be performed for a complete genome or for a collection of translated amino acid sequences. There is a total of N possible types of such strings: $N = 4^K$ for DNA and $N = 20^K$ for amino acid sequences.

For concreteness consider the case of one protein sequence of length L. Denote the frequency of appearance of the K-string $\alpha_1\alpha_2\ldots\alpha_K$ by $f(\alpha_1\alpha_2\ldots\alpha_K)$, where each α_i is 1 of the 20 amino acid single-letter symbols. This frequency divided by the total number $(L - K + 1)$ of K-strings in the given protein sequence may be taken as the probability $p(\alpha_1\alpha_2\ldots\alpha_K)$ of appearance of the string $\alpha_1\alpha_2\ldots\alpha_K$ in the protein:

$$p(\alpha_1\alpha_2\ldots\alpha_K) = \frac{f(\alpha_1\alpha_2\ldots\alpha_K)}{(L-K+1)} \tag{1}$$

The collection of such frequencies or probabilities reflects both the result of random mutations and selective evolution in terms of K-strings as "building blocks."

Subtraction of Random Background

Mutations happen in a more or less random manner at the molecular level, while selections shape the direction of evolution. Neutral mutations lead to some randomness in the K-string composition. In order to highlight the selective diversification of sequence composition, one must subtract a random background from the simple counting results. This is done as follows.

Suppose we have done direct counting for all strings of length $(K - 1)$ and $(K - 2)$. The probability of appearance of K-strings is predicted by using a Markov model:

$$p^0(\alpha_1\alpha_2\ldots\alpha_K) = \frac{p(\alpha_1\alpha_2\ldots\alpha_{K-1})p(\alpha_2\alpha_3\ldots\alpha_K)}{p(\alpha_2\alpha_3\ldots\alpha_{K-1})} \tag{2}$$

The superscript 0 on p^0 indicates the fact that it is a predicted quantity. We note that the denominator comes from the frequency of $(K - 2)$-strings. This kind of Markov model prediction has been used in biological sequence analysis for a long time (Brendel et al. 1986). It can be justified by virtue of a maximal entropy principle with appropriate constraints (Hu and Wang 2001).

Composition Vectors and Distance Matrix

It is the difference between the actual counting result p and the predicted value p^0 that really reflects the shaping role of selective evolution. Therefore, we collect

$$a(\alpha_1\alpha_2\ldots\alpha_K) = \begin{cases} \frac{p(\alpha_1\alpha_2\ldots\alpha_K) - p^0(\alpha_1\alpha_2\ldots\alpha_K)}{p^0(\alpha_1\alpha_2\ldots\alpha_K)} & \text{when } p^0 \neq 0 \\ 0 & \text{when } p^0 = 0 \end{cases} \quad (3)$$

for all possible strings $\alpha_1\alpha_2\ldots\alpha_K$ as components to form a composition vector for a species. To simplify the notations further, we write a_i for the ith component corresponding to string type i, where i runs from 1 to $N = 20^K$. Putting these components in a fixed order, we obtain a composition vector for species A:

$$A = (a_1,\ a_2, \cdots,\ a_N)$$

Likewise, for species B we have a composition vector

$$B = (b_1,\ b_2, \cdots,\ b_N)$$

In principle there are three ways to construct the composition vectors. First, one may use the whole genome sequence. Second, one may just collect the coding sequences in the genome. Third, one makes use of the translated amino acid sequences from the coding segments of DNA. As mutation rates are higher and more variable in noncoding segments and protein sequences change at a more or less constant rate, one expects that the third choice is the best and the second is better than the first. We tried all three choices and the requirement of consistency served as a criterion. By consistency we mean that the topology of the trees constructed with growing K should converge. This is best realized with phylogenetic relations obtained from protein sequences. Therefore, in what follows we concentrate on results based on amino acid sequences.

The correlation $C(A, B)$ between any two species A and B is calculated as the cosine function of the angle between the two representative vectors in the N-dimensional space of composition vectors:

$$C(A,B) = \frac{\sum_{i=1}^{N} a_i \times b_i}{\left(\sum_{i=1}^{N} a_i^2 \times \sum_{i=1}^{N} b_i^2\right)^{\frac{1}{2}}} \quad (4)$$

The distance $D(A, B)$ between the two species is defined as

$$D(A,B) = \frac{1 - C(A,B)}{2} \quad (5)$$

Since $C(A, B)$ may vary between -1 and 1, the distance is normalized to the interval (0, 1). The collection of distances for all species pairs comprises a distance matrix.

Tree Construction

The emphasis of the present work is to provide a new way to infer evolutionary distances between species from the whole genome data without doing sequence alignment. Once a distance matrix has been calculated it is straightforward to construct phylogenetic trees by following the standard procedures. We use the neighbor-joining (NJ) method (Saitou and Nei 1987) in the PHYLIP package (Felsenstein 1993) for all $K \geq 2$ trees. The Fitch method is not feasible when the number of species is as large as 109. We did not use such algorithm as the maximal likelihood since it is not based on distance matrices alone. The final phylogenetic trees are drawn using the DRAWTREE software in the PHYLIP package.

Statistical Test of the Trees

For our new approach we have to devise statistical tests for the resulting trees. We used both bootstrap-type and jackknife-type tests.

In carrying out the bootstrap test, we randomly drew sequences from the protein pool of a species. Some amino acid sequences would be drawn repeatedly, while others might be skipped at all. We picked up the same number of sequences as the number of proteins in the genome. On average about 70% of proteins were kept with some repetitions and 30% skipped at each calculation. Bootstrap values were produced by the CONSENSUS program in the PHYLIP package.

A positive interpretation of the bootstrap calculations consists in that it is not necessary to have the actual complete proteomes to reconstruct the phylogenetic tree. Suffice it to have a majority of the protein sequences.

The jackknife-type test has been done by dropping one species at a time from the calculation. The three-kingdom division persists in all $K = 5$ and $K = 6$ cases. This is an expected result, as we have gone from 21 to 37 to 51 to 72 to 84 to 109 species over the years and the main feature of the trees has remained the same.

Results

A phylogenetic tree based on counting the number of amino acid strings of length $K = 6$ is shown in Fig. 1. The red dot in Fig. 1 denotes the trifurcation point of the main domains. In Fig. 2 we show the result for a total of 200 bootstrap calculations for the 109 species on a $K = 5$ tree. The number of appearances of a branch is marked by a color oval: red for 191–200, yellow for 181–190, green for 171–180, and blue for under 170. The trifurcation point is surrounded by three red ovals: the Eukarya, Archaea, and Bacteria branches all appeared 200 times. Most of the major branches in the tree appear more than 170 times, but there are lower counts in a few branches. Most of the branchings on these two trees agree with each other and we analyze the deviations in the Discussion. An inspection of these trees and comparison with the $K = 1$ to 4 trees (not shown) reveals the following.

At the overall level, the division of life into the three main domains Bacteria, Archaea, and Eukarya is a clean and prominent feature. No mixing among domains takes place on all trees for $K \geq 5$.

At the finest level, different strains of the same species, different species of the same genus, and different genera of the same family all come together as they should.

At the intermediate level, the division of *Proteobacteria* into alpha, beta, gamma, and epsilon groups, the separation of *Actinobacteria* from *Firmicutes*, and the division of Archaea into *Crenarchaeota* and *Euryarchaeota* all come out correctly, with very few outliers. We return to these outliers in the Discussion.

In general, our phylogenetic trees support the SSU rRNA tree of life in its overall structure and in many details. It is remarkable that our trees and the SSU rRNA tree were based on nonoverlapping parts of the genomic data, namely, the RNA segments and the protein-coding sequences, and they were obtained by using entirely different ways of inferring distances between species, nevertheless, they lead to basically consistent results. Since our method

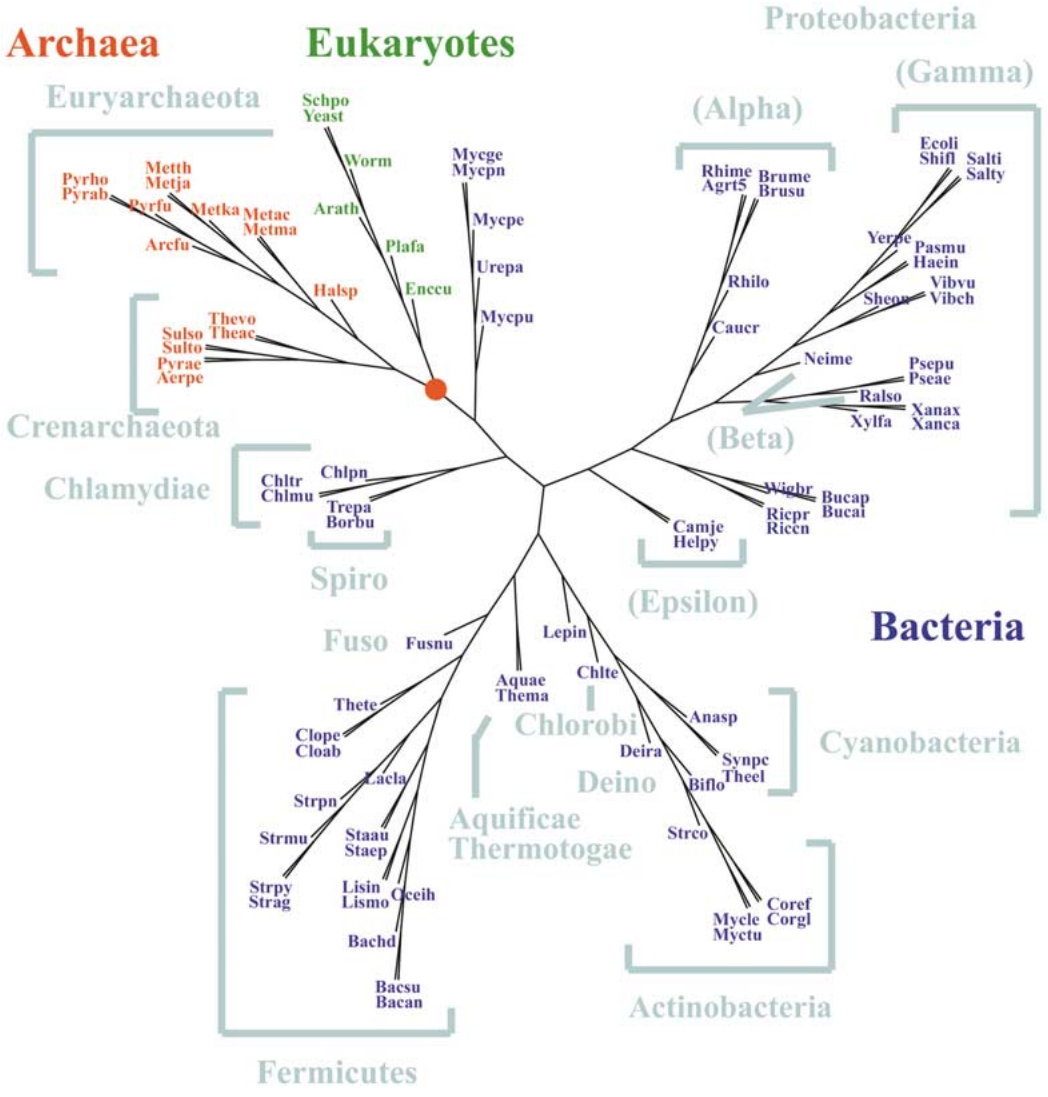

Fig. 1. A $K = 6$ phylogenetic tree for 109 organisms. Different strains within the same species were represented by one strain so there are 93 species shown in the tree. The *red dot* denotes the trifurcation point of the three domains. Archaea, Bacteria, and Eukarya are indicated by *red*, *blue*, and *green*, respectively. All 13 prokaryotic phylum names are placed close to the corresponding branches. For the largest characterized phylum, *Proteobacteria*, the class/group names are given in *parentheses*. Note that this is an unrooted tree and the branches are not to scale.

does not contain "free" parameters and "fine-tuning," it may provide a quick reference in prokaryote phylogenetics whenever the proteome of an organism is available, a situation that will become commonplace in the near-future.

Discussion

Detailed Comparison with Bergey's Manual of Systematic Bacteriology

The most comprehensive taxonomic information of prokaryotes has been collected in the two editions *Bergey's Manual of Systematic Bacteriology* (Bergey's Manual Trust 1984–1989, 2001). However, until recently the segmental results of molecular phylogeny have not reached a status to be compared with *Bergey's Manual* in a systematic way. Now, equipped with the new method and phylogenetic trees of 103 prokaryotes from 61 genera, we are in a position to do this for the first time. Although only the first of the five volumes of the latest edition of *Bergey's Manual* seen the light, fortunately there is an electronic version of a *Taxonomic Outline of Prokaryote Genera* (Garrity et al. 2001) with their lineage from phylum, class, order, and family down to genus listed explicitly.

The NCBI Taxonomy (Wheeler et al. 2003), although declared as "not a phylogenetic or taxonomic authority," agrees with the latest edition of *Bergey's Manual* for most of the species studied in this work. Therefore, we take *Bergey's Manual* as a primary

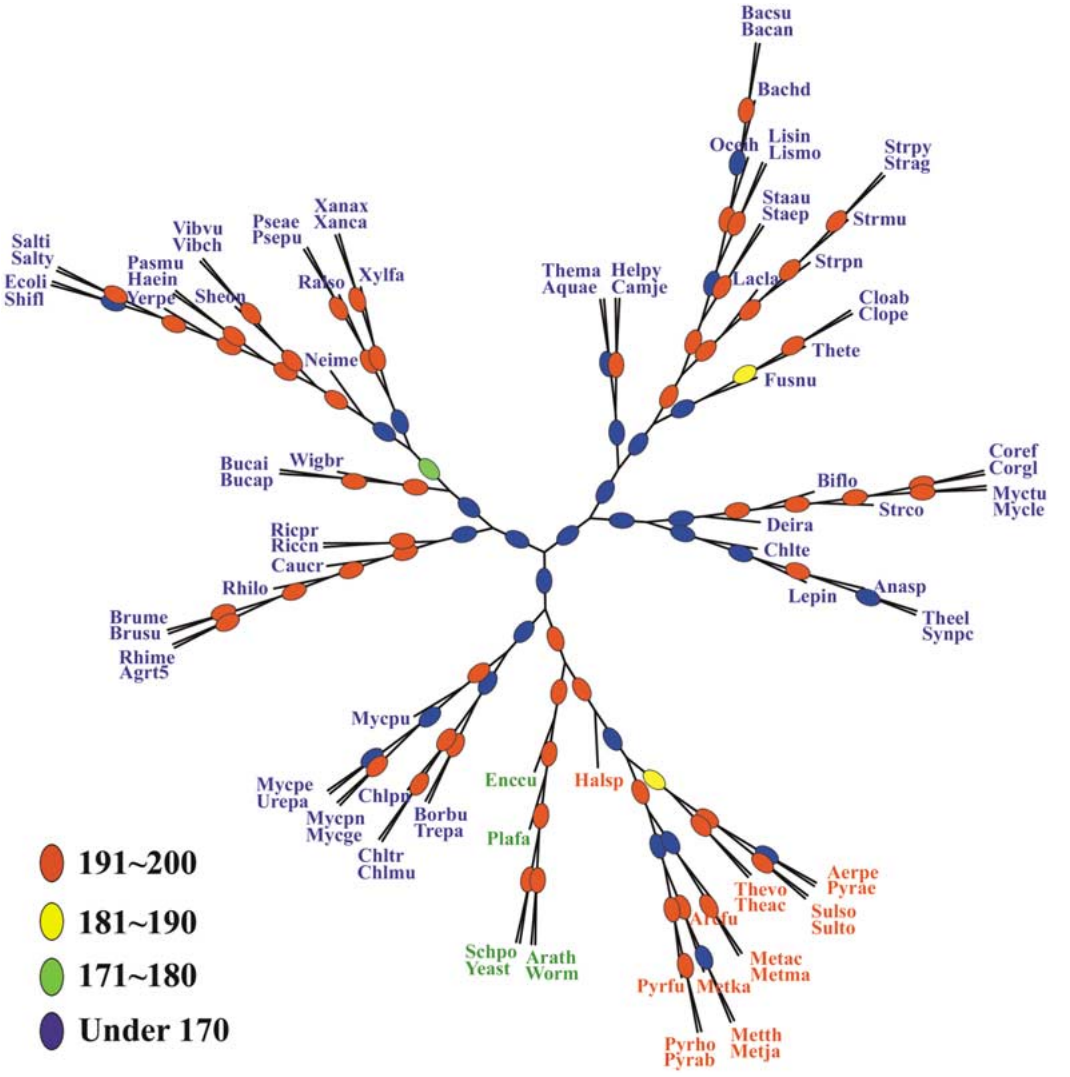

Fig. 2. A $K = 5$ tree with bootstrap numbers marked on the branches. *Colored ovals* show the range of appearances of the branches: 191–200 (*red*), 181–190 (*yellow*), 171–180 (*green*), under 170 (*blue*). A total of 200 bootstrap calculations was performed.

source and compare our trees with the taxonomic scheme of the Manual. We note that the classifications in the second edition of *Bergey's Manual* "follow a phylogenetic framework based on analysis of the nucleotide sequence of the SSU rRNA, rather than a phenotypic structure" (see Garrity et al. 2000, Preface).

Our analysis below also shows the convergence of the tree topology with increasing K. Even at the single-amino acid level ($K = 1$ and composition vector of dimension 20), many species within one genus have already clustered together. At the dipeptide level ($K = 2$ and composition vectors of dimension 400), the major groupings on the tree start to bear resemblance to the SSU rRNA tree of life. For example, 15 of 16 Archaea were grouped together, with only Halsp standing out but the three thermophilic bacteria Aquae, Thema, and Thete still mixed up with Archaea. The branchings changed slightly at $K = 3$ and 4. The topology of the phylogenetic trees becomes stable for $K = 5$ and 6. As for Eukaryotes, Yeast, Schpo, Arath, and Worm stay together throughout $K = 1$ to 6, never mixing with the prokaryotes. Enccu and Plafa stay outside at $K = 1$ and 2, but join the Eukaryotes at $K \geq 3$.

At the lowest taxonomic level, 11 bacterial species are represented by the complete genomes of two or more different strains. When there are two strains in a species they always stay together as K increases from 1 to 6, never mixing with other organisms. When there are three or more strains in a species, their relative locations stabilized at $K \geq 5$.

There are 20 genera that contain more than two species. Among these 20 the species in 15 genera, including *Pyrococcus*, containing three species, do stay together from $K = 1$ to 6, although some may migrate together before taking a final position at larger K. The remaining five genera stably converge

Table 1. Comparison of some specific features in different whole genome approaches

	Wolf et al. (2001)				Daubin et al. (2001)	Present authors
Number of species	10 A + 30 B[a]				7A + 26B + 4E	16A + 87B + 6E
Method	Presence–absence of genomes in COGs	Conservation of gene pairs	Identity percentage between probable orthologs	Concatenated alignment of ribosomal proteins	Supertree	*K*-string composition vector at $K = 5$ and $K = 6$
Three domains	Yes (only 2 domains studied)				Yes	Yes
Halsp off A1 and A2	Yes	No	Yes	Yes	No	Yes, at $K = 5$
Epsilon off B12	Yes	Yes	Yes	No, but at edge of B12	No, but at edge of B12	Yes, at $K = 5$; no, but at edge for $K = 6$
Mollicutes off B13	Yes	Yes	No	No	No	Yes
B10/B14/B4	No	No	No	Yes	Yes	Yes
B1/B2	No	No	No	Yes	Yes	Yes
B16/B17[b]	No	Yes	Yes	Yes	No	Yes

[a] Archaea, Bacteria, and Eukarya are indicated by A, B, and E, respectively.
[b] Except for Lepin.

at larger *K*. The only exception comes with *Mycoplasmataceae*, where the genus *Ureaplasma* gets mixed into the genus *Mycoplasma* from $K = 2$ to 6. This also leads to a problem at the next family–genus level. We cannot tell whether this hints at a classification problem of *Ureaplasma*.

There are eight families that are represented by more than two genera. It makes sense to look at the interrelationship among the genera within one and the same family. The convergence for $K \geq 5$ is evident. However, there are classification exceptions compared to *Bergey's Manual*. In the largest characterized family, *Enterobacteriaceae* (B12.3.13.1), the genera *Buchnera* and *Wigglesworthia* always form a small subgroup outside the gamma group, while the other four genera of the gamma group, *Escherichia*, *Salmonella*, *Shigella*, and *Yersinia*, always stay together. In addition, the two genera from the beta group are mixed with the gamma group. As mentioned before, *Ureaplasma*, though remaining in *Mycoplasmataceae* from $K = 2$ to $K = 6$, gets mixed with *Mycoplasma*. These exceptions set aside, all the lower taxa from families down to different strains in a species do converge at $K = 5$ and $K = 6$.

The Problem of Higher Taxa

The problem comes with some higher taxa at the phylum or class level. This is not surprising, as even in more mature fields such as the systematics of plants and animals the disagreement among taxonomists is mainly associated with the placement of higher taxa. If around 1974 the taxonomic standing of the whole prokaryote group was still a problem, the problem around 1989 was already how to place the higher prokaryotic taxa, as vividly described by R.G.E. Murray (1989). Today the situation has not been improved very much. In a taxonomic list such as that of Garrity et al. (2001) many classes are juxtaposed under a phylum without an evolutionary relationship indicated, and many orders are juxtaposed under a class without showing which ones are more ancient. In a phylogenetic tree, no matter how one got it, an evolutionary branching scheme, correct or not, is always associated with the taxa.

In the "Taxonomic Outline" (Garrity et al. 2001) all prokaryotes are divided into 2 Archaea phyla (A1, A2) and 23 Bacteria phyla (B1 to B23). Among the 25 phyla, 13 are represented in our trees. Summarizing the results of comparison with *Bergey's Manual* and anticipating the results of comparison with other whole genome approaches (see Table 1), we make the following observations on the grouping of higher prokaryotic taxa.

1. The two phyla *Aquificae* (B1) and *Thermotoga* (B2) group together.
2. The three phyla *Actinobacteria* (B14), *Deinococcus* (B4), and *Cyanobacteria* (B10) group together as ((B14, B4), B10). This is supported by some other whole genome approaches (Wolf et al. 2001).
3. The *Chlamydiae* (B16) and the *Spirochetes* (B17 except for Lepin) also group together. This was observed also by Wolf et al. (2001).
4. The *Epsilonproteobacteria* (Class V in B12) seems to be a stranger to the phylum *Proteobacteria* (B12). It either leaves B12 or stands at the edge of B12 in our trees and in some other approaches (Wolf et al. 2001).

Comparison with Other Whole Genome Approaches

There have been several attempts to use whole genome data to construct prokaryote phylogenetic trees. Early papers in the late 1990s treated only a small number of species: 4A + 8B +1E (Snel et al. 1999), 4A + 14B + 5E (Tekaia et al. 1999), 4A + 6B + 1E (Fitz-Gibbon and House 1999), and 5A + 16B + 2E (Huynen et al. 1999). (Here and in Table 1, A, B, and E stand for Archaea, Bacteria, and Eukarya, respectively.) Though all the inferred trees could resolve the three main domains of Archaea, Bacteria, and Eukarya, they could not bring about much information on major groupings of the higher taxa due to the limited number of organisms.

In late 2001 a few papers appeared dealing with 30 to 40 species (Daubin et al. 2001; Wolf et al. 2001). A few specific features show up repeatedly in different approaches as well as in our trees so they can no longer be considered incidental. We summarize these features in Table 1. One of the methods in (Wolf et al. 2001) was based on ribosomal proteins. Although it could not be classified as a "whole genome" approach, we keep it for comparison,

Summarizing the data collected in Table 1 and comparing them with our trees at all string lengths from one to six, we list all placement problems in our results.

1. Only two genera from the beta group of *Proteobacteria* (Neime and Ralso) are present in our data. They are separated and mixed into the gamma group in both $K = 5$ and $K = 6$ trees.
2. As mentioned before, the gamma group split into two subgroups.
3. The *Rickettsia* from the alpha group joins the smaller gamma group at $K = 6$ but stays within the whole alpha group at $K = 5$.
4. *Leptospira* stands outside the other two *Spirochetes* (B17), which are located closer to the *Chlamydiae*.
5. Aside from the mixing-up of Urepa, the four *Mycoplasma* tend to stay outside *Firmicutes* (B13) in our and some other trees. As all four species belong to the same order, *Mycoplasmatales*, we cannot say whether this is a feature for the whole class *Mollicutes* or is restricted to *Mycoplasmatales* only.
6. There are two problems associated with Archaea. Is Halsp an outlier of both *Crenarchaeota* (A1) and *Euryarchaeota* (A2) or does it belongs to A2? Different approaches disagree. The placement of *Thermoplasma* has been a problem in archaean taxonomy. In *Bergey's Manual* it came under *Euryarchaeota*, but in the book *Five Kingdoms* (Margulis and Schwartz 1998) it was attributed to *Crenarchaeota*. On both problems we have to await the opinions of bacteriologists.

As one anonymous referee pointed out, some of these placement problems might be related to small genome size. Indeed, this was true for *Buchnera* and *Wigglesworthia* in the gamma group, *Rickettsia* in the alpha group, and *Mycoplasma* in the *Firmucutes* (B13). *Chlamydia* and the two *Spirochetes* that are separated from Lepin also have small genomes. On the other hand, our method applied to small chloroplast genomes alone (Chu et al. 2003) has led to meaningful results. Since similar problems have been encountered in some other whole genome approaches (see Table 1), they call for further study.

On Justification of the K-*String Approach*

The feasibility of our approach may be better understood from a "*K*-string picture of evolution," i.e., a coarse-grained view of what is embodied in the central dogma by looking at the evolution process on the protein level without digging into the underlying coding–transcription–translation–machinery. In the primordial soup the polypeptides which became proteins as we see nowadays must be short and of a limited variety. If one could collect overlapping *K*-strings, say, for $K = 6$, from these ancestral species, they must have taken only a small portion of the $20^6 = 64{,}000{,}000$ points of the "six-string space." Later on, these polypeptides evolved by growth, fusion, and mutation. The set of "taken" points diffused in the "*K*-string space." This viewpoint is close to the view "new proteins can evolve by recombining preexisting polypeptide domains" (Alberts et al. 1994). It is worth mentioning that the six-string space has not yet saturated at present. A search of the 101,602 protein sequences in SWISS-PROT database Rel. 40 (2000) showed that all these proteins have taken only less than 26% of the six-string types. If one looks at individual prokaryote species, this contrast appears to be even more remarkable: EcoliK has taken less than 2% and Mycge less than 0.3% of the six-string types.

The possibility of using long and sparse representative vectors to represent organisms is an advantage for tree construction in the sense of reaching a higher resolution of species and avoiding saturation of the representative vectors.

A related problem is how unique the reconstruction of a protein sequence from the collection of its constituent *K*-strings would be. If unique, a protein would be equally well represented by its primary amino acid sequence and by the collection of *K*-strings with long enough *K*. Our preliminary results (Hao et al. 2001) have shown that at $K = 6$ an overwhelming majority of protein sequences from a real database does have a unique reconstruction. Although uniqueness of reconstruction for a single

Table A1. Archaea names, abbreviations, and NCBI accession numbers, ordered by their Bergey code

Species/strain	Abbrev.	Accession No.	Bergey code
Pyrobaculum aerophilum	Pyrae	NC_003364	A1.1.1.1.1
Aeropyrum pernix K1	Aerpe	NC–000854	A1.1.2.1.3
Sulfolobus solfataricus	Sulso	NC_002754	A1.1.3.1.1
Sulfolobus tokodaii	Sulto	NC_003106	A1.1.3.1.1
Methanobacterium thermoautotrophicus	Metth	NC_000916	A2.1.1.1.1
Methanococcus jannaschii	Metja	NC_000909	A2.2.1.1.1
Methanosarcina acetivorans strain C2A	Metac	NC_003552	A2.2.3.1.1
Methanosarcina mazei Goel	Metma	NC_003901	A2.2.3.1.1
Halobacterium sp. NRC-1	Halsp	NC_002607	A2.3.1.1.1
Thermoplasma acidophilum	Theac	NC_002578	A2.4.1.1.1
Thermoplasma volcanium	Thevo	NC_002689	A2.4.1.1.1
Pyrococcus abyssi	Pyrab	NC_000868	A2.5.1.1.3
Pyrococcus furiosus	Pyrfu	NC_003413	A2.5.1.1.3
Pyrococcus horikoshii	Pyrho	NC_000961	A2.5.1.1.3
Archaeoglobus fulgidus	Arcfu	NC_000917	A2.6.1.1.1
Methanopyrus kandleri AV19	Metka	NC_003551	A2.7.1.1.1

protein does not mean the same for a collection of many proteins, this result, nevertheless, speaks in favor of the compositional approach.

On Lateral Gene Transfer

Before concluding the paper we would like to comment on the effect of lateral gene transfer. Analyzing the controversies in tree constructions caused by the steady inflow of genomic data, W. Ford Doolittle (1999) was one of the first to postulate that there were extensive lateral gene transfers among microbial organisms. According to C. Woese (2000) lateral transfer events have not only taken place in evolution, but also served "the major, if not sole, evolutionary source of true innovation." However, the extent of lateral transfer has been increasingly restricted to smaller and smaller gene pools of closer and closer related species (Woese 1998). Since our method does not rely on the choice of one or another gene, lateral gene transfer might not affect our approach very much. On the contrary, it may even contribute positively to group together closely related species among which exchange of genetic material might have taken place. In other words, some aspects of lateral gene transfer might have been partly incorporated into the K-string approach. Anyway, the presence of lateral gene transfer does not preclude the possibility of tracing an essential part of evolutionary history by using whole genome data.

Limitations and Improvements of the Present Approach

The use of complete genomes is both a merit and a demerit of the method, although our bootstrap results show that the availability of most, but not necessarily all, of the proteome might be good enough to reproduce the topology of the trees.

Concentrating on the topology of the trees in the first place, we did not scale the branch lengths on the tree. However, these lengths should reflect evolution rates in terms of K-string composition changes. The calibration of branch lengths is further complicated by the overlapping nature of the K-strings when $K \geq 2$.

However, as a new method the K-string composition approach needs more justification and we intend to test it by including new complete genomes, especially those of Eukaryotes, and by applying it to numerically simulated data.

Appendix

The list of all prokaryotic genomes used in our study is given in Tables A1 and A2. The species are listed in accordance with their "Bergey Code" in order to make comparison of the trees with *Bergey's Manual* easier. The Bergey Code is a shorthand of the classification given in the 2001 edition of *Bergey's Manual of Systematic Bacteriology* (Garrity et al. 2001). For example, *Lacococcus lactis* is listed under Phylum BXIII (*Firmicutes*)—Class III (*Bacilli*)—Order II (*Lactobacillales*)—Family VI (*Streptococcaceae*)—Genus II (*Lactococcus*). We changed all Roman numerals to Arabic and wrote the lineage as B13.3.2.6.2, dropping the taxonomic units and the Latin names.

The six eukaryotes included are *Saccharomyces cerevisiae* (Yeast; NC_001133–48), *Caenorhabitidis elegans* (worm; NC_003279–84), *Arabidopsis thaliana* (Arath; NC_003070.71.74.75.76), *Encephalitozoon cuniculi* (Enccu; NC_003242.29–38), *Plasmodium falciparum* (Plafa; NC_000521.910.4314–18.25–31), and *Schizosaccharomyces pombe* (Schpo; NC_003421.23.24).

Table A2. Bacterium names, abbreviations, and NCBI accession numbers, ordered by their Bergey code

Species/strain	Abbrev.	Accession	Bergey code
Aquifex aeolicus	Aquae	NC_000918	B1.1.1.1.1
Thermotoga maritima	Thema	NC_000853	B2.1.1.1.1
Deinococcus radiodurans R1	Deira	NC–001263–64	B4.1.1.1.1
Thermosynechococcus elongatus BP-1	Theel	NC_004113	B10.l.?[a]
Cyanobacterium synechocystis PCC6803	Synpc	NC_000911	B10.1.1.1.14
Cyanobacterium nostoc sp. PCC7120	Anasp	NC_003272	B10.1.4.1.8
Chlorobium tepidum TLS	Chlte	NC_002932	B11.1.1.1.1
Rickettsia conorii	Riccn	NC_003103	B12.1.2.1.1
Rickettsia prowazekii	Ricpr	NC_000963	B12.1.2.1.1
Caulobacter crescentus	Caucr	NC_002696	B12.1.5.1.1
Agrobacterium tumefaciens C58	Agrt5	NC_003062–63	B12.1.6.1.2
Agrobacterium tumefaciens C58 UWash	Agrt5W	NC_003304–05	B12.1.6.1.2
Sinorhizobium meliloti 1021	Rhime	NC_003047	B12.1.6.1.6
Brucella melitensis	Brume	NC_003317–18	B12.1.6.3.1
Brucella suis 1330	Brusu	NC_004310.11	B12.1.6.3.1
Mesorhizobium loti	Rhilo	NC_002678	B12.1.6.4.6
Ralstonia solanacearum	Ralso	NC_003295–96	B12.2.1.2.1
Neisseria meningitidis MC58	NeimeM	NC_003112	B12.2.4.1.1
Neisseria meningitidis Z2491	NeimeZ	NC_003116	B12.2.4.1.1
Xanthomonas axonopodis citri 306	Xanax	NC_003919	B12.3.3.1.1
Xanthomonas campestris ATCC 33913	Xanca	NC_003902	B12.3.3.1.1
Xylella fastidiosa	Xylfa	NC_002488	B12.3.3.1.9
Pseudomonas aeruginosa PA01	Pseae	NC_002516	B12.3.9.1.1
Pseudomonas putida KT2440	Psepu	NC_002947	B12.3.9.1.1
Shewanella oneidensis MR-1	Sheon	NC_004347	B12.3.10.1.7
Vibrio cholerae	Vibch	NC_002505–06	B12.3.11.1.1
Vibrio vulnificus CMCP6	Vibvu	NC_004459.60	B12.3.11.1.1
Buchnera aphidicola Sg	Bucap	NC_004061	B12.3.13.1.5
Buchnera sp. APS	Bucai	NC_002528	B12.3.13.1.5
Escherichia coli CFT073	EcoliC	NC_004431	B12.3.13.1.13
Escherichia coli K12	EcoliK	NC_000913	B12.3.13.1.13
Escherichia coli O157:H7	EcoliO	NC_002695	B12.3.13.1.13
Escherichia coli O157:H7 EDL933	EcoliE	NC_002655	B12.3.13.1.13
Salmonella typhi	Salti	NC_003198	B12.3.13.1.32
Salmonella typhimurium LT2	Salty	NC_003197	B12.3.13.1.32
Shigella flexneri 2a strain 301	Shifl	NC_004337	B12.3.13.1.34
Wigglesworthia brevipalpis	Wigbr	NC_004344	B12.3.13.1.38
Yersinia pestis strain C092	YerpeC	NC_003143	B12.3.13.1.40
Yersinia pestis KIM	YerpeK	NC_004088	B12.3.13.1.40
Pasteurella multocida PM70	Pasmu	NC_002663	B12.3.14.1.1
Haemophilus influenzae Rd	Haein	NC_000907	B12.3.14.1.3
Campylobacter jejuni	Camje	NC_002163	B12.5.1.1.1
Helicobacter pylori 26695	Helpy	NC_000915	B12.5.1.2.1
Helicobacter pylori J99	Helpj	NC_000921	B12.5.1.2.1
Clostridium acetobutylicum ATCC824	Cloab	NC_003030	B13.1.1.1.1
Clostridium perfringens	Clope	NC_003366	B13.1.1.1.1
Thermoanaerobacter tengcongensis	Thete	NC_003869	B13.1.2.1.8
Mycoplasma genitalium	Mycge	NC_000908	B13.2.1.1.1
Mycoplasma penetrans	Mycpe	NC_004432	B13.2.1.1.1
Mycoplasma pneumoniae	Mycpn	NC_000912	B13.2.1.1.1
Mycoplasma pulmonis UAB CTIP	Mycpu	NC_002771	B13.2.1.1.1
Ureaplasma urealyticum	Urepa	NC_002162	B13.2.1.1.4
Oceanobacillus iheyensis	Oceih	NC_004193	B13.3.1.1.?[a]
Bacillus anthracis A2012	Bacan	NC_003995	B13.3.1.1.1
Bacillus halodurans	Bachd	NC_002570	B13.3.1.1.1
Bacillus subtilis	Bacsu	NC_000964	B13.3.1.1.1
Listeria innocua	Lisin	NC_003212	B13.3.1.4.1
Listeria monocytogenes EGD-e	Lismo	NC_003210	B13.3.1.4.1
Staphylococcus aureus Mu50	StaauM	NC_002758	B13.3.1.5.1
Staphylococcus aureus MW2	StaauW	NC_003923	B13.3.1.5.1
Staphylococcus aureus N315	StaauN	NC_002745	B13.3.1.5.1
Staphylococcus epidermidis ATCC 12228	Staep	NC_004461	B13.3.1.5.1
Streptococcus agalactiae 2603 V/R	StragV	NC_004116	B13.3.2.6.1

(Continued)

Table A2. Continued

Streptococcus agalactiae NEM316	StragN	NC_004368	B13.3.2.6.1
Streptococcus mutans UA159	Strmu	NC_004350	B13.3.2.6.1
Streptococcus pneumoniae R6	StrpnR	NC_003098	B13.3.2.6.1
Streptococcus pneumoniae TIGR4	StrpnT	NC_003028	B13.3.2.6.1
Streptococcus pyogenes MGAS8232	Strpy8	NC_003485	B13.3.2.6.1
Streptococcus pyogenes MGAS315	StrpyG	NC_004070	B13.3.2.6.1
Streptococcus pyogenes SF370	StrpyS	NC_002737	B13.3.2.6.1
Lactococcus lactis sp. IL1403	Lacla	NC_002662	B13.3.2.6.2
Corynebacterium efficiens YS-314	Coref	NC_004369	B14.(1.5).(1.7).1.1
Corynebacterium glutamicum	Corgl	NC_003450	B14.(1.5).(1.7).1.1
Mycobacterium leprae TN	Mycle	NC_002677	B14.(1.5).(1.7).4.1
Mycobacterium tuberculosis CDC1551	MyctuC	NC_002755	B14.(1.5).(1.7).4.1
Mycobacterium tuberculosis H37Rv	MyctuH	NC_000962	B14.(1.5).(1.7).4.1
Streptomyces coelicolor A3(2)	Strco	NC_003888	B14.(1.5).(1.11).1.1
Bifidobacterium longum NCC2705	Biflo	NC_004307	B14.(1.5).2.1.1
Chlamydia muridarum	Chlmu	NC_002620	B16.1.1.1.1
Chlamydia trachomatis	Chltr	NC_000117	B16.1.1.1.1
Chlamydophila pneumoniae AR39	ChlpnA	NC_002179	B16.1.1.1.2
Chlamydophila pneumoniae CWL029	ChlpnC	NC_000922	B16.1.1.1.2
Chlamydophila pneumoniae J138	ChlpnJ	NC_002491	B16.1.1.1.2
Borrelia burgdorferi	Borbu	NC_001318	B17.1.1.1.2
Treponema pallidum	Trepa	NC_000919	B17.1.1.1.9
Leptospira interrogans serovar lai strain 56601	Lepin	NC_004342.43	B17.1.1.3.2
Fusobacterium nucleatum ATCC 25586	Fusnu	NC_003454	B21.1.1.1.1

[a] Not available in Garrity et al. (2001).

Acknowledgments. The authors thank Drs. Yang Zhong (Fudan University) and Hongya Gu (Peking University) for discussion and comments. We also thank an anonymous referee who pointed out the problem with small genomes. The use of the 64 CPU IBM Cluster at Peking University is also gratefully acknowledged. This work was supported in part by grants from the Special Funds for Major State Basic Research Project of China, the Innovation Project of CAS, and Major Innovation Research Project "248" of Beijing Municipality.

References

Alberts B, et al. (1994) Molecular biology of the cell, 3rd ed. Garland, New York, p 121 and references therein

Aravind L, Tatusov RL, Wolf YI, Walker DR, Koonin EV (1998) Evidence for massive gene exchange between archaeal and bacterial hyperthermophiles. Trends Genet 14:442–444

Baldauf SL, Palmer JD, Doolittle WF (1996) The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny. Proc Natl Acad Sci USA 93:7749–7754

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2003) GenBank. Nucleic Acid Res 31:23–27 (Available at ftp://ncbi.nlm.nih.gov/genbank/genomes/Bacteria/)

Bergey's Manual Trust (1984–1989) Bergey's manual of systematic bacteriology, 1st ed, Vols 1–4. Williams & Wilkins, Baltimore

Bergey's Manual Trust (2001) Bergey's manual of systematic bacteriology, 2nd ed, Vol 1. Springer-Verlag, New York

Brendel V, Beckmann JS, Trifonov EN (1986) Linguistics of nucleotide sequences: Morphology and comparison of vocabularies. J Biomol Struct Dyn 4:11–21

Chu K, Qi J, Yu Z, Anh VO (2003) Origin and phylogeny of chloroplasts: A simple correlation analysis of complete genomes. Mol Biol Evol (in press)

Daubin V, Gouy M, Perriere G (2001) Bacterial molecular phylogeny using supertree approach. Genome Inform 12:155–164

Doolittle WF (1999) Phylogenetic classification and the universal tree. Science 284:2124–2128

Doolittle WF (2000) Uprooting the tree of life. Sci Am February pp 90–95

Felsenstein J (1993) PHYLIP (phylogeny inference package), version 3.5c. Distributed by the author at http://evolution.genetics.washington.edu/phylip.html

Fitz-Gibbon ST, House CH (1999) Whole genome-based phylogenetic analysis of free-living microorganism. Nucleic Acid Res 27:4218–4222

Garrity GM, Winters M, Searles DB (2001) Taxonomic outline of the prokaryotic genera. In: Bergey's manual of systematic bacteriology, 2nd ed, Rel 1.0 (Available at http://www.cme.msu.edu/bergeys/april2001-genus.pdf)

Hao BL, Xie HM, Zhang SY (2001) Compositional representation of protein sequences and the number of Eulerian loops. (Available at arXiv:physics/0103028 at: http://lanl.arXiv.org/)

Hu R, Wang B (2001) Statistically significant strings are related to regulatory elements in the promoter region of *Saccharomyces cerevisiae*. Physica A 290:464–474

Huynen MA, Snel B, Bork P (1999) Lateral gene transfer, genome surveys, and the phylogeny of prokaryotes. Science 286:1441

Karlin S, Burge C (1995) Dinucleotide relative abundance extremes: A genomic signature. Trends Genet 11:283–290

Margulis LM, Schwartz KV (1998) Five kingdoms, 3rd ed. WH Freeman, New York, p 60

Matte-Tailliez O, Brochier C, Forterre P, Philippe H (2002) Archaeal phylogeny based on ribosomal proteins. Mol Biol Evol 19:631–639

Murray RGE (1989) The higher taxa, or, a place for everything…? In: Williams ST, Sharpe ME, Holt JG (eds) Bergey's manual of systematic bacteriology, Vol 4. Williams and Wilkins, Baltimore, pp 2329–2332

Pennisi E (1998) Genome data shake tree of life. Science 280:672–674

Pennisi E (1999) Is it time to uproot the tree of life? Science 284:1305–1308

Ragan MA (2001) Detection of lateral gene transfer among microbial genomes. Curr Opin Gen Dev 11:620–626

Saitou N, Nei M (1987) The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

Snel B, Bork P, Huynen MA (1999) Genome phylogeny based on gene content. Nature Genet 21:108–110

Tekaia F, Lazcano A, Dujon B (1999) The genomic tree as revealed from whole genome proteome comparisons. Genome Res 9:550–557

Tomb JF, White O, Kerlavage AR, et al. (1997) The complete genome sequence of the gastric pathogen *Helicobacter pyroli*. Nature 388:539–547

Wheeler DL, Church DM, Federhen S, et al. (2003) Database resources of the National Center for Biotechnology. Nucleic Acids Res 31 28–33 (The NCBI-curated prokaryote genomes are available at ftp://ftp.ncbi.nih.gov/genomes/Bacteria/. The NCBI Taxonomy Browser is located at http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html

Woese CR (1998) The universal ancester. Proc Natl Acad Sci USA 95:6854–6859

Woese CR (2000) Interpreting the universal phylogenetic tree. Proc Natl Acad Sci USA 97:8392–8396

Woese CR, Fox GE (1977) Phylogenetic structure of the prokaryotic domain: The primary kingdoms. Proc Natl Acad Sci USA 74:5088–5090

Wolf YI, Rogozin IB, Grishin NV, Tatusov RL, Koonin EV (2001) Genome trees constructed using five different approaches suggest new major bacterial clades. BMC Evol Biol 1:8 (Available at http://www.biomedcentral.com/1471-2148/18)

Zuckerkandl E, Pauling L (1965) Evolutionary divergence and convergence in proteins. In: Bryson V, Vogel HJ (eds) Evolving genes and proteins. Academic Press, New York, pp 97–166