Whole-proteome phylogeny of prokaryotes by feature frequency profiles: An alignment-free method with optimal feature resolution

Se-Ran Jun[a], Gregory E. Sims[a], Guohong A. Wu[a], and Sung-Hou Kim[a,b,1]

[a]Department of Chemistry, University of California, Berkeley, CA 94720; and [b]Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley CA 94720

Contributed by Sung-Hou Kim, November 13, 2009 (sent for review October 6, 2009)

We present a whole-proteome phylogeny of prokaryotes constructed by comparing feature frequency profiles (FFPs) of whole proteomes. Features are *l*-mers of amino acids, and each organism is represented by a profile of frequencies of all features. The selection of feature length is critical in the FFP method, and we have developed a procedure for identifying the optimal feature lengths for inferring the phylogeny of prokaryotes, strictly speaking, a proteome phylogeny. Our FFP trees are constructed with whole proteomes of 884 prokaryotes, 16 unicellular eukaryotes, and 2 random sequences. To highlight the branching order of major groups, we present a simplified proteome FFP tree of monophyletic class or phylum with branch support. In our whole-proteome FFP trees (*i*) Archaea, Bacteria, Eukaryota, and a random sequence outgroup are clearly separated; (*ii*) Archaea and Bacteria form a sister group when rooted with random sequences; (*iii*) Planctomycetes, which possesses an intracellular membrane compartment, is placed at the basal position of the Bacteria domain; (*iv*) almost all groups are monophyletic in prokaryotes at most taxonomic levels, but many differences in the branching order of major groups are observed between our proteome FFP tree and trees built with other methods; and (*v*) previously "unclassified" genomes may be assigned to the most likely taxa. We describe notable similarities and differences between our FFP trees and those based on other methods in grouping and phylogeny of prokaryotes.

branching order | *l*-mers | prokaryotic phylogeny | random sequence outgroup | whole-genome phylogeny

Currently, a widely accepted phylogeny and classification of prokaryotes is based on the comparison of genes that encode small subunit ribosomal RNA (SSU rRNA) (1). This method also led to the proposal of three domains of organisms (Archaea, Bacteria, and Eukaryota). The branching order of the three domains with respect to the common origin was inferred by rooting the SSU rRNA tree using anciently duplicated genes (e.g., EF-Tu/EF-G, ATPase α and β subunits) (2). However, as more gene sequences became available, taxonomic groupings and phylogenies for prokaryotes derived from alternative genes often showed conflict with those based on SSU rRNA (3–6). This conflict is more evident especially for the relationships between taxonomic groups, suggesting that the phylogeny of organisms is irresolvable through phylogenies derived from one or a few selected genes. At best, such phylogenies only reconstruct a possible evolutionary history of the selected gene or gene set—not the history of whole genomes or organisms. It is generally believed that the use of the whole genome/proteome may provide more robust information for inferring the phylogeny of organisms (3–6). This is supported by the observation that phylogenies based on progressively larger gene sets become more consistent and also less sensitive to artifacts from horizontal gene transfer (7). However, whole-genome/proteome comparison cannot be accomplished for a large population of organisms with multiple sequence alignment (MSA)-based methods because it is likely that there is only a small fraction of the total number of genes that are shared and highly homologous in all organisms compared.

The main approaches used for inferring whole-genome–based prokaryotic phylogenies can be divided into three categories depending on the sequence information used: orthologous genes, protein sequence/structure domains, or whole-genome/proteome sequences. The methods in the first category can be divided further into two classes. One class uses the content or order (8, 9) of orthologous genes, and the other builds trees from a concatenated alignment (supermatrix) (10) or by assembling/combining trees (supertree) (11) from MSAs of individual genes. In the second category, the methods are based on protein domain (12, 13) assignment, either Pfam (14) or SCOP (15) domains, of the ORF sequences [at present, Pfam domains cover approximately 50% and SCOP domains approximately 40% of all ORFs (12, 13)]. The methods in the first category require the "correct" selection of orthologous genes, and those in the second category require the assignment of protein domains at the sequence or structure level.

The methods based on whole-genome/proteome sequences, the third category, can be further divided into two classes: the pairwise alignment-based approach and the alignment-free approach. A few examples of each approach are briefly summarized below. For the alignment-based approach, Henz et al. (16) constructed a phylogeny of 91 prokaryotic genomes, in which distances were estimated from a maximum subset of nonoverlapping "high scoring segment pairs" reported by BLASTN for each pair of genomes. Chan et al. (17) derived a phylogeny of 230 prokaryotic genomes using MUMs (maximal exact substrings that are unique in the two genomes) obtained from MUMmer (18), a software package designed for pairwise alignment of genomes/proteomes. Chan et al. used only the intermediate output of MUMmer and did not use the more extensive alignment capabilities of MUMmer; thus, strictly speaking, their method is not alignment-based. Chan et al. then compared their tree with that of Henz et al. and showed that their taxonomic grouping more closely resembled that of the National Center for Biotechnology Information (NCBI).

Several methods for alignment-free genome comparison have been developed. One method by Otu et al. (19) introduced a measure based on the relative information between proteome sequences using Lempel-Ziv complexity. In other examples, Pride et al. (20) showed how well tetranucleotide-based patterns were shared in the whole genomes of related organisms; Qi et al. (21) used the frequency of fixed *k*-strings subtracted by a mutation background, which was obtained from a *k*-2 Markov model; and

Author contributions: S.-R.J., G.E.S., and S.-H.K. designed research; S.-R.J. performed research; S.-R.J., G.E.S., and G.A.W. contributed new reagents/analytic tools; S.-R.J., G.E.S., G.A.W., and S.-H.K. analyzed data; and S.-R.J., G.E.S., G.A.W., and S.-H.K. wrote the paper.

The authors declare no conflict of interest.

[1]To whom correspondence should be addressed. E-mail: SHKim@cchem.berkeley.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0913033107/DCSupplemental.

Ulitsky et al. (22) introduced another measure based on the average length of maximal exact common substrings (ACSs). Ulitsky et al. compared their tree with two proteomic trees obtained from the approaches of Otu et al. and Qi et al. They showed that the phylogeny and taxonomic groupings obtained with the ACS method agreed best with the SSU rRNA-based method.

In the alignment-free approach used in our study, the feature frequency profile (FFP) method (23), an organism is represented by its whole-proteome sequence (WPS), which consists of the amino acid sequences of all predicted proteins in the chromosome(s) of the organism. We used protein sequences rather than base sequences to minimize the adverse effects of base composition and codon preference biases on phylogeny construction. Each WPS is represented by a profile of feature (l-mer) frequencies. These feature frequency profiles are used to construct a distance matrix at some feature length (resolution) optimal for inferring phylogeny, and then a tree is built from the distance matrix. Thus, our method does not require the identification of any common orthologous genes or protein domains as in the methods of the first and second categories described previously. Our method is also different from other methods in the third category (using whole-genome/proteome sequences), in that we do not align any special portions of the whole genomes/proteomes as in the alignment-based approach, but we use the alignment-free approach, paying special attention to selecting the feature length optimally suited for inferring phylogeny.

We have constructed two kinds of trees within the range of the optimal feature lengths (see *Optimal Feature Lengths and FFP Trees* in *Materials and Methods*): (*i*) whole-proteome FFP trees for all available WPSs of prokaryotes; and (*ii*) simplified FFP trees in which each leaf corresponds to either a class or phylum. The latter trees are used to evaluate statistical support for the branching order of major groups with the jackknife monophyly index (JMI) (24) (see *Statistical Support for the Branching Order of Major Groups* in *Materials and Methods*). We then present the best simplified proteome FFP tree with largest average JMI. We compare the taxonomic groupings at several taxonomic levels with the "reference" taxonomy (based on the classification of the NCBI), as well as with the ACS tree (22) and the MUM tree (17), which were obtained by other alignment-free approaches and are among the most comprehensive whole-proteome trees of prokaryotes at present. We also compare the branching orders among the taxonomic groups in the best simplified FFP tree with those in trees constructed with the other methods.

Results and Discussion

Overall, the taxonomic groupings of WPSs by our FFP method agree very well with the reference taxonomy; almost all main groupings agree with the reference taxonomy at several taxonomic levels (domain, phylum, class, and genus), with some minor discrepancies in monophyly and grouping. However, substantial differences in the branching orders of major groups are observed between our proteome FFP tree and trees from other methods based on, for example, SSU rRNAs, orthologous genes, and protein domains.

Dataset and FFP Trees. Our dataset includes WPSs of 884 prokaryotes (26 phyla, 41 classes, and 315 genera), 16 unicellular eukaryotes, and 2 random sequences. Figure 1 presents our whole-proteome FFP tree at feature length $l = 13$ built by BIONJ (25) using all members of the entire dataset. We also built the simplified FFP tree with JMI at feature length $l = 13$ in Fig. 2 using BIONJ after excluding seven proteomes labeled with NCBI taxonomy IDs in blue, red, and green in Fig. 1 (see *Materials and Methods*). To compare our whole-proteome FFP tree with 16S rRNA-based trees, we also generated a 16S rRNA tree (Fig. S1) with BIONJ using a Jukes-Cantor model distance matrix, which was calculated from an MSA obtained from the Ribosomal Database Project 10 (RDP) (26).

Three Domains and Basal Prokaryotes. Although random sequences cannot be meaningfully aligned with any gene, they can be included in the dataset for analysis with the FFP method. In our study, we included two random sequences of lengths equal to the longest and the shortest proteomes in the dataset. As shown in Fig. 1, Archaea, Bacteria, Eukaryota, and the random sequences are clearly separated. Moreover, Archaea and Bacteria form a sister group excluding Eukaryota when the random sequences were used as an outgroup. This sister group arrangement was also supported with a rooted tree built with UPGMA (27) and by rooting a tree built by neighbor joining (NJ) (28) with random sequences. This three-domain arrangement was consistently observed even when the feature length was reduced as short as $l = 8$. This topological arrangement is inconsistent with the commonly accepted view that Eukaryota and Archaea form a sister group.

At the most basal position of Bacteria are the mesophillic Planctomycetes, which agrees with the placement by Brochier and Philippe (29). It is interesting to notice that Planctomycetes have a large genome that is often found to be enclosed in a membrane as in Eukaryotes, reproduce by budding, and lack peptidoglycan in their cell walls (30). These observations invoke an intriguing notion that these features may have been a character of the last universal common ancestor. The above observations may be consistent with an analysis by Kurland et al. (31, 32), who conjecture that the last common ancestor of Archaea and Bacteria possessed protein domains that are more similar to the Eukaryotic domain complement. The Bacterial root of our tree, Planctomycetes, seems to possess some of these primitive Eukaryotic features. It is also interesting to note that the basal position of Archaea in our tree is the new phylum, Thaumarchaeota (also mesophillic). Among Archaea, Thaumarchaeota is the only phylum to possess type IB topoisomerase genes, which are also present in Eukaryotes, some bacteria, and viruses (33).

Relationships Among Major Groups. As mentioned earlier, it is clear from Fig. 1 that the taxonomic groupings arrived at with the FFP method using the optimal feature resolution mostly agree with those of the reference taxonomy at several taxonomic levels (domain, phylum, class, and genus), although there are a few membership discrepancies (see *Membership Discrepancy* below). However, with respect to the overall relationship and branching order among major groups, many differences are observed between the FFP tree (in Fig. 2) and those based on SSU rRNA and other genome features. We notice that the intergroup relationships of our FFP tree mostly support the consensus view of several different methods, such as Fig. 4.2 of *Assembling the Tree of Life* (34) and other proposed relationships noted in the literature review (3–6). Notably, the branching order in our 16S rRNA tree, for example, was quite different from that of the FFP tree. Some of the notable differences with high JMI support are described below:

(*i*) Thaumarchaeota, a new phylum, is at the basal position of all Archaea so far sequenced;
(*ii*) Methanococci, Methanobacteria, and Methanopyri are clustered together, forming the most recent divergence in Archaea;
(*iii*) As mentioned previously, Planctomycetes is at the basal position of Bacteria;
(*iv*) Mollicutes was a class group within the Firmicutes phylum in the previous version of the NCBI taxonomy of 2007 but forms a separate phylum, Tenericues, in the current version. In our FFP tree in Fig. 2, Bacilli and Clostridia form a Firmicutes clade, as in the current revision of the NCBI

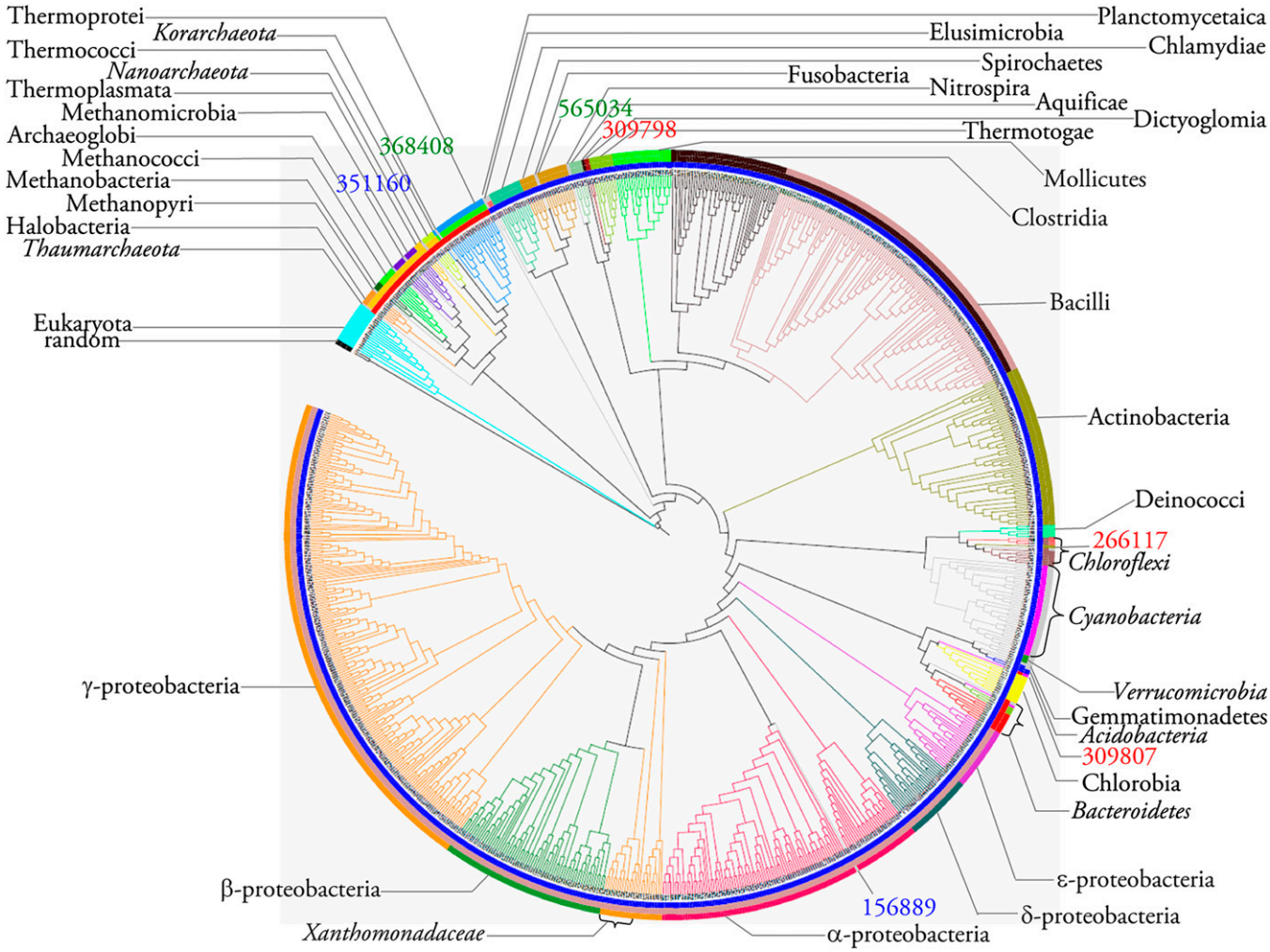

Fig. 1. The whole-proteome FFP tree at feature length $l = 13$. The proteome FFP tree is colored by the reference taxonomy: the inner circle colored by domain, the middle circle by phylum, and the outer circle and branches by class. Additionally, the proteome FFP tree is labeled with the name of groups (classes in regular font or phyla/family/genera in italic font) and NCBI taxonomy IDs mentioned in the text. The organisms labeled in blue (351160 and 156889) are "unclassified" at the class level; the position of those in red (309807, 266117, and 309798) differs from the current assignment; those in green (565034 and 368408) are excluded when constructing simplified proteome FFP trees. The ITOL online tool is used to render the tree. Branch lengths are not to scale so that the clading and tree topology can be clearly displayed.

taxonomy. Bacilli and Clostridia, but not Mollicutes, form endospores (35), which supports their close relationship;

(*v*) According to the reference taxonomy, Chlamydiae and Verrucomicrobia form a superphylum, but they are not together in our FFP tree.

Monophyly. Most of the groups are monophyletic in the FFP tree, with very few discrepancies at each taxonomic level. Operationally, we identify monophyly by the following process: first, identify a clade of a given group, and then check whether there are descendants belonging to different group(s) with at least two members mixed in the clade. Some of the notable differences follow.

Spirochaetes are divided into three genus groups: *Borrelia*, *Treponema*, and *Leptospira*. In our tree, *Leptospira* do not group with the remaining two, as shown in Fig. 2.

At the class level, γ-Proteobacteria is separated into several subgroups in Fig. 1. One subgroup that consists of a single family, Xanthomonadaceae, clusters with β-Proteobacteria.

At the genus level, (*i*) the whole-proteome FFP tree shows better resolution than the 16S rRNA tree at the genus level: among 118 genera with at least two members, 14 are not monophyletic in the FFP tree, whereas 19 genera are not monophyletic in a 16S rRNA tree built through the RDP database (26); (*ii*) the genera in Cyanobacteria are mixed with each other in Fig. 1; (*iii*) gene alignment–based evidence indicates that *Shigella* evolved from *Escherichia coli* by pseudogene formations and gene inactivations by insertion sequence expansion, resulting in a polyphyletic *Shigella* clade (36, 37). Our analysis does not include pseudogenes, and the FFP method shows a monophyletic *Shigella* clade. This also contrasts with trees based on 16S rRNAs, in which extensive mixing between *E. coli* and *Shigella* occurs, and with trees based on multigenes alignments (38), in which *E. coli* and *Shigella* are still mixed but less extensively. This suggests that 16S rRNAs as well as highly conserved genes may be evolving too slowly to resolve relationships at the finer taxonomic level.

Membership Discrepancy. As mentioned earlier, the taxonomic classification of WPSs by our FFP method agrees well with the reference taxonomy, with a few membership discrepancies. There are three organisms (labeled with NCBI taxonomy IDs in red in Fig. 1) with membership discrepancies, as shown in Fig. 1.

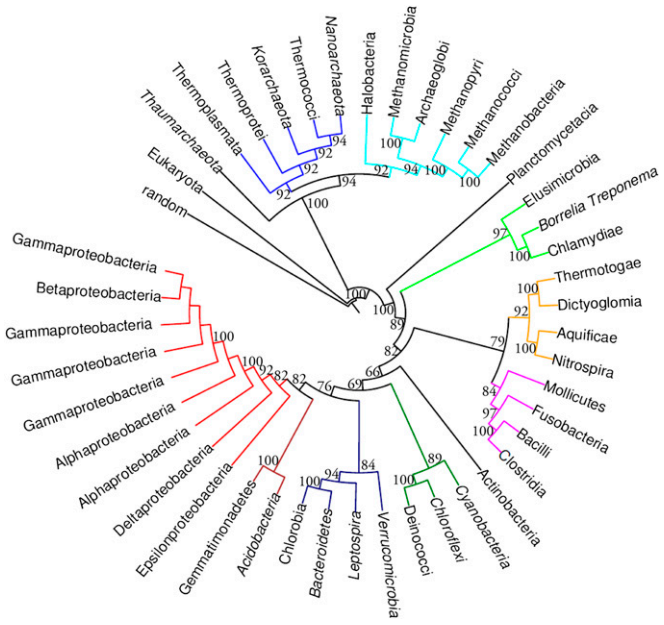

Fig. 2. The simplified proteome FFP tree at feature length l = 13. The coloring indicates "supra-class" groups, which are defined by statistical support values of >82, except for Archaea, where there are three clear clades. The numbers indicate the JMI (%). The simplified FFP tree was generated by collapsing major groups into a single leaf, and then the JMI was used to measure how robust the tree is to taxa sampling. If a clade has 100% jackknife monophyly index, it means that the clade always exists in jackknife FFP trees. Unnumbered clades labeled with the same name were removed as a unit during the jackknife test. Feature length l = 13 shows the most robustness in terms of average JMI. Labels indicate classes in regular font and phyla/genera in italic font.

Notably, *Salinibacter ruber DSM 13855* (309807) is not grouped with other Sphingobacteria (Bacteroidetes) but with Chlorobia. *S. ruber* clearly showed the closest relationship with Chlorobia based on FFP distances; the nine closest neighbors by FFPs distances were all members of Chlorobia. We suggest that a taxonomic revision of Salinibacter may be required.

New Classification. Among the organisms unclassified at the class level, our proteome FFP tree suggests the assignment of uncultured methanogenic archaeon RC-I (351160 in blue) to Methanomicrobia and *Magnetococcus sp. MC-1* (156889 in blue) to α-Proteobacteria.

Effect of Plasmids on the Branching Order of Major Groups. In our dataset, 35% of organisms have from 1 to 21 plasmid(s), and the size of plasmids can be quite significant (83.2% for *Sinorhizobium meliloti*) compared with the corresponding chromosome(s). We considered only chromosomal proteins when constructing the FFP trees in Figs. 1 and 2. However, the inclusion of additional plasmid proteins with the chromosomal proteins does not affect the tree topology at the class level. Thus, we obtain exactly the same tree as in Fig. 2 whether or not plasmids are included. Because plasmids are believed to be easily transferable between closely related species, combining plasmid proteomes together with the host genome simulates horizontal gene transfer in FFP analysis. Thus, our results suggest that our proteome FFP tree is likely to be robust to the horizontal transfer of genetic materials, and/or the plasmid transfer may be confined to the class groups of their hosts.

Comparison with Trees Based on Protein Domains and Orthologous Genes. We investigated the discrepancies between our FFP tree and trees based on protein domain organization (13) and orthologous genes (10) with respect to the branching order of taxonomic groups. We observe notable differences with respect to *Aquifex aeolicus* and *Thermotoga maritima*. *A. aeolicus* is grouped with the one of δ-Proteobacteria and *T. maritima* with Clostridia in the domain-based tree. *A. aeolicus* and *T. maritima* are grouped together, forming a relatively recent divergence in the gene-based tree. But in our FFP tree, *A. aeolicus* and *T. maritima* are clustered together near the root of Bacteria. Furthermore, the divergence order of class groups in Proteobacteria and Firmicutes (if Mollicutes is assumed to belong to Firmicutes, according to a 2007 version of the NCBI taxonomy) is also different from that of our proteome FFP tree. Our proteome FFP tree is the only one that groups Bacilli and Clostridia first, much like the current NCBI taxonomy.

Comparison with the ACS Tree. To compare our tree with the ACS tree of Ulitsky et al. (22), we assembled a dataset ("ACS dataset") that included only the prokaryotic proteomes present in the ACS dataset (note that the ACS tree file was not available from the authors). An FFP tree was built using the NJ tree-building algorithm (28; as used by Ulitsky et al.) on the ACS dataset. The grouping in the FFP tree agrees better with the reference taxonomy than the ACS tree at most taxonomic levels. For example, the ACS tree shows that *Nanoarchaeum equitans Kin4-M* and *Halobacterium sp. NRC-1* are placed among Bacteria; Spirochaetes is separated into three groups; Firmicutes is separated into three groups; and one of the ε-Proteobacteria, *Campylobacter jejuni*, is clustered with Clostridia. With regard to the branching order among major taxonomic groups, there are also many differences between the two trees.

Comparison with the MUM Tree. We also constructed a MUM dataset (17) and used the BIONJ method (25), as Chan et al. did (17), to build our proteome FFP tree. With regard to grouping and monophyly, our FFP tree and the MUM tree agree for many groups at most taxonomic levels, but with several differences. For example, in the MUM tree, *Bdellovibrio bacteriovorus* (δ-Proteobacteria) is placed within Chlamydiae; *Symbiobacterium themophilum* (Clostridia) is placed outside Firmicutes; Mollicutes is not monophyletic. There are also many differences between the two trees in branching order among major taxonomic groups.

Conclusion. The whole-proteome FFP method provides insights into the phylogeny of prokaryotes and highlights some differences from selected gene-based alignment approaches. Taken together, our observations suggest that our proteome FFP tree may also be robust at the class level to the horizontal transfer of genetic material and shows excellent resolution at most taxonomic levels. Thus, the FFP method is very useful for resolving the whole-proteome taxonomy and phylogeny of prokaryotes.

Materials and Methods

Proteome Dataset and Reference Taxonomy. For the dataset, we downloaded all available translated chromosomal amino acid sequences from NCBI (June 2009). Our dataset comprises the whole proteomes of 884 prokaryotes and 16 unicellular eukaryotes of varying size, excluding two organisms, which are *Candidatus Sulcia muelleri GWSS* (75K aa long) and *Candidatus Carsonella ruddii PV* (50K aa long) because of their short length relative to the largest proteome (6960K aa long). Additionally, two random sequences whose length corresponds to the shortest (118K aa long) and longest (6,960K aa long) proteomes in the dataset (generated via the standard Perl rand function) were added for an outgroup. A list of species used in our dataset with NCBI taxonomy ID and NCBI accession number is provided in Table S1. In cases in which an organism has multiple chromosomes, they are concatenated to form a proteome for the organism. When comparing the FFP trees with the ACS tree of Ulitsky et al. (22) or the MUM tree of Chan et al. (17), we used their datasets. The reference taxonomy was determined from the NCBI taxonomy, which is hierarchically organized on the basis of information provided primarily by sequence submitters and with supple-

mental curation by NCBI staff. According to the reference taxonomy, four organisms and Cyanobacteria are unclassified at the class level.

Tree Construction. Our FFP trees were constructed from distance matrices. There are several software tools for distance-based tree construction, such as UPGMA (27), NJ (28), and BIONJ (25). To measure the difference between tree topologies, we used the Robinson and Foulds distance (39) implemented in the treedist function in PHYLIP (40) and maximum agreement subtree implemented in the mast function in PhyloNet (41). In addition, we used the consens function in PHYLIP to construct a consensus tree with extended majority rule and the ITOL online tool (http://itol.embl.de/itol.cgi) to render the trees. We have not shown the branch lengths to scale so that the tree topology and branching order can be clearly displayed.

FFPs and FFP Distances. A general description of the FFP method has been published (23), and a description of the details more relevant to this work is given below. The proteomes of organisms are stored as a collection of individual protein sequences. The raw frequency of each feature (*l*-mer) in a sequence of length *L* is counted as follows: first, we slide a window of length *l* along a protein sequence from position 1 to $L - l + 1$ and again start sliding the window along the next protein sequence and so on until the entire proteome is scanned. The count profile, $C_l = \{c_1, c_2, \ldots, c_N\}$, where c_i is the raw frequency of the corresponding feature and $N = 20^l$ is the total number of all possible *l*-mers, is transformed into a frequency profile F_l by normalizing by the proteome length, yielding the relative abundance of each feature. Thus, an organism is represented as an FFP of its proteome. We used the Jensen-Shannon divergence (42) with FFPs to calculate dissimilarities between organisms.

Optimal Feature Lengths and FFP Trees. We have developed a procedure for selecting the optimal feature resolution through the following three steps. ***Step 1: Cumulative relative entropy.*** First, we analyzed FFPs of individual proteomes on the basis of cumulative relative entropy (CRE). We used two conditions to decide the proper feature lengths: (*i*) the maximum entropy principle (43), which states that the correct FFP maximizes the entropy subject to the observed FFPs of shorter features, and (*ii*) the capability of generating FFPs of all longer features. Under the assumption that a given sequence is a ring, the solution of the constrained optimization problem for the first condition (*i*), the expected frequency profile $\widehat{F}_{l+1} = (\widehat{f}_i)$ is represented as follows [for details, see Sadovsky (44)]: for a feature $w = a_1 a_2 \ldots a_{l+1}$,

$$\widehat{f}_w = \frac{f_{a_1 a_2 \ldots a_l} \times f_{a_2 a_3 \ldots a_{l+1}}}{f_{a_2 a_3 \ldots a_l}}.$$

Our FFPs do not satisfy the ring assumption, but the reduction in the number of features is negligible considering the proteome size.

To measure how well an expected FFP $\widehat{F}_l = (\widehat{f}_i)$ approximates an observed FFP $F_l = (f_i)$, we computed the relative entropy between two profiles:

$$RE(F_l, \widehat{F}_l) = \sum_i f_i \log_2 \frac{f_i}{\widehat{f}_i},$$

where the sum is over all features in an observed FFP. If the relative entropy is zero at feature length *l*, a FFP at feature length *l* satisfies the first condition, and a FFP at feature length *l*-1 has the ability to regenerate FFPs at feature length from 1 to *l*, but no longer than *l*. We have defined the CRE at feature length *l* by the sum of the relative entropy from *l* to infinity (we tested up to *l* = 15, beyond which the relative entropy was almost zero):

$$CRE(l) = \sum_{k=l}^{15} RE(F_k, \widehat{F}_k).$$

Because the relative entropy is nonnegative, if a sequence has zero CRE at feature length *l*, an FFP at feature length *l* has maximum entropy and all of the information of FFPs of longer features as well as shorter features, so that feature length *l* with zero CRE satisfies the two conditions (*i*) and (*ii*) we adopted. Fig. 3 is a plot of CRE curves vs. feature length *l* for six proteomes that are the smallest and largest ones chosen from Archaea, Bacteria, and Eukaryota, respectively. Generally, the curves of CRE for the other proteomes are placed between the left-most and right-most curves, and most of proteomes start having zero CRE at feature length *l* = 10.

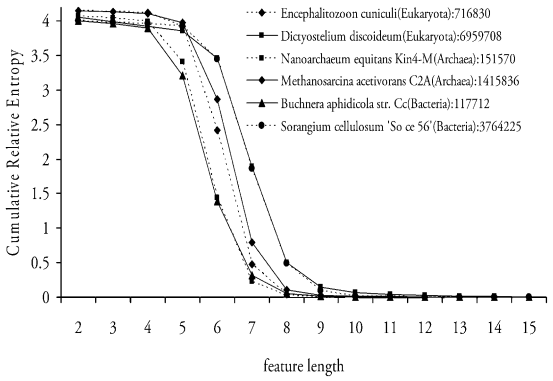

Fig. 3. Cumulative relative entropy. A plot of cumulative relative entropy vs. feature length *l* for six proteomes that are the smallest and largest ones chosen from each domain in our dataset. The proteome length and domain information of those proteomes are indicated. Among the six proteomes, *Buchnera aphidicola*, which is the smallest, starts to have zero CRE at feature length *l* = 8, and *Dictyostelium discoideum*, which is the largest, starts to have zero CRE value at feature length *l* =10.

Step 2: Tree convergence. We computed a distance matrix of FFPs for each feature length *l* and used the matrix to build a tree using BIONJ (25). We noticed that the percentage of FFP pairs that do not share any common features varies from 0 to 0.6% for *l* = 1, . . ., 15. The topology difference between a pair of trees at feature length *l* and *l* + 1 is estimated with the Robinson and Foulds distance (39). In Fig. 4, the tree topology distances become very small and remain so for feature length *l* = 9, . . ., 15, which reveals that the tree topology does not change much for feature length *l* = 9, . . ., 15. However, we noticed occasional differences in the position of major groups even among trees within the tree convergence range. Similar behavior was observed with alternative topological distances, such as the maximum agreement subtree method (41).

Step 3: Random sequence outgroup perturbation. In our dataset of proteomes, two random sequences were included as an outgroup. Because a suitable outgroup should not affect the topology of the ingroup, we compared two trees with and without the random outgroup to check whether the outgroup disrupts the tree topology of major groups. For *l* = 10 and beyond, there was no disruption.

On the basis of the three tests described above, we chose the optimal range of feature lengths from *l* = 10, . . ., 15. Note that proteome FFP trees within the range showed the same groupings at each taxonomic level but occasional differences in the position of major groups. A consensus tree constructed with extended majority rule (40) from the six FFP trees within the range showed practically the same topology as the one at feature length *l* = 13 in Fig. 1.

Statistical Support for the Branching Order of Major Groups. To determine a statistically reliable best feature length within the optimal range for the

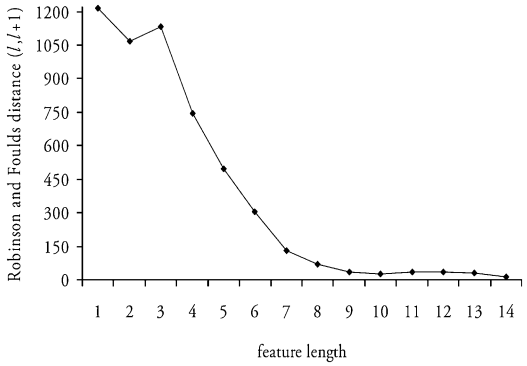

Fig. 4. Tree topology distance. Robinson and Foulds distance between trees at feature lengths *l* and *l* + 1 for $l \geq 1$. Note that tree convergence begins at *l* = 9, revealing that the tree topology does not change much as *l* increases.

branching order of major groups, we performed a jackknife test to assess sampling bias with major groups in two steps.

Step 1: Simplified FFP trees. For the simplicity of the jackknife test, we excluded seven proteomes in our dataset, which are labeled with NCBI taxonomy IDs in blue, red, and green in Fig. 1, so that the clade corresponding to the major group can be deleted in a single step during jackknife operation: two proteomes, 351160 and 156889 (in blue), are unclassified; three proteomes, 309798, 255117, and 309807 (in red), were misplaced; and two proteomes, 368408 and 565034 (in green), were outliers of Thermoprotei and Spirochaetes, respectively. Excluding these proteomes did not affect the tree topology of all major groups except for 565034, which caused perturbation in Spirochaetes. The tree in Fig. 2 is a simplified proteome FFP tree at feature length l = 13 generated by representing all members in a major group by a single leaf.

Step 2: The best feature length with jackknife monophyly index. Because unequal population sizes among different major groups may affect the tree topology, we estimated the statistical support associated with the branching order of the major groups with the JMI (24), which estimates the statistical confidence against taxa sampling bias. The FFP jackknife trees were constructed using BIONJ (25) by excluding each major group at a time. For each clade in a simplified tree, the JMI was calculated by counting how many times the clade exists among FFP jackknife trees. Feature length l = 13 was shown to be the most robust to taxa sampling, with the largest average JMI within the optimal range. The tree in Fig. 2 is a simplified proteome FFP tree at feature length l = 13 with JMI values. Unnumbered clades labeled with the same name were removed as a unit during the jackknife test. It is worthwhile noting that FFP jackknife trees showed excellent grouping in the same way as FFP trees.

ACKNOWLEDGMENTS. We thank Drs. Hiroshi Nikaido and Alex Glazer for their expert advice. This work was supported by National Institutes of Health Grant GM62412 and by the Korean Ministry of Education, Science and Technology (World Class University Project R31-2008-000-10086-0).

1. Woese CR, Fox GE (1977) Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proc Natl Acad Sci USA* 74:5088–5090.
2. Iwabe N, Kuma K, Hasegawa M, Osawa S, Miyata T (1989) Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proc Natl Acad Sci USA* 86:9355–9359.
3. McInerney JO, Cotton JA, Pisani D (2008) The prokaryotic tree of life: Past, present... and future? *Trends Ecol Evol* 23:276–281.
4. Wolf YI, Rogozin IB, Grishin NV, Koonin EV (2002) Genome trees and the tree of life. *Trends Genet* 18:472–479.
5. Delsuc F, Brinkmann H, Philippe H (2005) Phylogenomics and the reconstruction of the tree of life. *Nat Rev Genet* 6:361–375.
6. Snel B, Huynen MA, Dutilh BE (2005) Genome trees and the nature of genome evolution. *Annu Rev Microbiol* 59:191–209.
7. Gontcharov AA, Marin B, Melkonian M (2004) Are combined analyses better than single gene phylogenies? A case study using SSU rDNA and rbcL sequence comparisons in the Zygnematophyceae (Streptophyta). *Mol Biol Evol* 21:612–624.
8. Huson DH, Steel M (2004) Phylogenetic trees based on gene content. *Bioinformatics* 20:2044–2049.
9. Korbel JO, Snel B, Huynen MA, Bork P (2002) SHOT: A web server for the construction of genome phylogenies. *Trends Genet* 18:159–162.
10. Ciccarelli FD, et al. (2006) Toward automatic reconstruction of a highly resolved tree of life. *Science* 311:1283–1287.
11. Bininda-Emonds OR (2004) The evolution of supertrees. *Trends Ecol Evol* 19:315–322.
12. Yang S, Doolittle RF, Bourne PE (2005) Phylogeny determined by protein domain content. *Proc Natl Acad Sci USA* 102:373–378.
13. Fukami-Kobayashi K, Minezaki Y, Tateno Y, Nishikawa K (2007) A tree of life based on protein domain organizations. *Mol Biol Evol* 24:1181–1189.
14. Bateman A, et al. (2004) The Pfam protein families database. *Nucleic Acids Res* 32 (Database issue):D138–D141.
15. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 247: 536–540.
16. Henz SR, Huson DH, Auch AF, Nieselt-Struwe K, Schuster SC (2005) Whole-genome prokaryotic phylogeny. *Bioinformatics* 21:2329–2335.
17. Chan PY, Lam TW, Yiu SM, Liu CM A more accurate and efficient whole genome phylogeny. In Proceedings of 4th Asia-Pacific Bioinformatics Conference, Taipei, Taiwan, pp. 337–352. edited by Tao Jiang et al. London: Imperial College Press.
18. Kurtz S, et al. (2004) Versatile and open software for comparing large genomes. *Genome Biol* 5:R12.1–R12.9.
19. Otu HH, Sayood K (2003) A new sequence distance measure for phylogenetic tree construction. *Bioinformatics* 19:2122–2130.
20. Pride DT, Meinersmann RJ, Wassenaar TM, Blaser MJ (2003) Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Res* 13:145–158.
21. Qi J, Wang B, Hao BI (2004) Whole proteome prokaryote phylogeny without sequence alignment: A K-string composition approach. *J Mol Evol* 58:1–11.
22. Ulitsky I, Burstein D, Tuller T, Chor B (2006) The average common substring approach to phylogenomic reconstruction. *J Comput Biol* 13:336–350.
23. Sims GE, Jun SR, Wu GA, Kim SH (2009) Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proc Natl Acad Sci USA* 106: 2677–2682.
24. Siddall ME (1995) Another monophyly index: Revisiting the jackknife. *Cladistics* 11: 33–56.
25. Gascuel O (1997) BIONJ: An improved version of the NJ algorithm based on a simple model of sequence data. *Mol Biol Evol* 14:685–695.
26. Cole JR, et al. (2007) The ribosomal database project (RDP-II): Introducing *myRDP* space and quality controlled public data. *Nucleic Acids Res* 35 (Database issue): D169–D172.
27. Sokal RR, Michener CD (1958) A statistical method for evaluating systematic relationships. *University of Kansas Scientific Bulletin* 28:1409–1438.
28. Saitou N, Nei M (1987) The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol Biol Evol* 4:406–425.
29. Brochier C, Philippe H (2002) Phylogeny: A non-hyperthermophilic ancestor for bacteria. *Nature* 417:244.
30. Fuerst JA (2005) Intracellular compartmentation in planctomycetes. *Annu Rev Microbiol* 59:299–328.
31. Kurland CG, Collins LJ, Penny D (2006) Genomics and the irreducible nature of eukaryote cells. *Science* 312:1011–1014.
32. Kurland CG, Canbäck B, Berg OG (2007) The origins of modern proteomes. *Biochimie* 89:1454–1463.
33. Brochier-Armanet C, Gribaldo S, Forterre P (2008) A DNA topoisomerase IB in Thaumarchaeota testifies for the presence of this enzyme in the last common ancestor of Archaea and Eucarya. *Biol Direct* 3:54–61.
34. Baldauf SL, et al. (2004) In *Assembling the Tree of Life*, eds Cracraft J, Donoghue MJ (Oxford Univ Press, New York), pp 43–60.
35. Ueda K, et al. (2004) Genome sequence of *Symbiobacterium thermophilum*, an uncultivable bacterium that depends on microbial commensalism. *Nucleic Acids Res* 32:4937–4944.
36. Pupo GM, Lan R, Reeves PR (2000) Multiple independent origins of *Shigella* clones of *Escherichia coli* and convergent evolution of many of their characteristics. *Proc Natl Acad Sci USA* 97:10567–10572.
37. Escobar-Páramo P, Giudicelli C, Parsot C, Denamur E (2003) The evolutionary history of *Shigella* and enteroinvasive *Escherichia coli* revised. *J Mol Evol* 57:140–148.
38. Touchon M, et al. (2009) Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genet* 5:e1000344.
39. Robinson DF, Foulds LR (1981) Comparison of phylogenetic trees. *Math Biosci* 53: 131–147.
40. Felsenstein J (1993) *PHYLIP (Phylogeny Inference Package) version 3.5c.* (Department of Genetics, University of Washington, Seattle).
41. Than C, Ruths D, Nakhleh L (2008) PhyloNet: A software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics* 9:322–337.
42. Lin J (1991) Divergence measures based on the Shannon entropy. *IEEE Trans Inf Theory* 37:145–151.
43. Jaynes ET (1957) Information theory and statistical mechanics. *Phys Rev* 106:620–630.
44. Sadovsky MG (2003) Comparison of real frequencies of strings vs. the expected ones reveals the information capacity of macromoleculae. *J Biol Phys* 29:23–38.