**Abstract**

**Background:** The availability of multiple complete genome sequences from diverse taxa prompts the development of new phylogenetic approaches, which attempt to incorporate information derived from comparative analysis of complete gene sets or large subsets thereof. Such attempts are particularly relevant because of the major role of horizontal gene transfer and lineage-specific gene loss, at least in the evolution of prokaryotes.

**Results:** Five largely independent approaches were employed to construct trees for completely sequenced bacterial and archaeal genomes: i) presence-absence of genomes in clusters of orthologous genes; ii) conservation of local gene order (gene pairs) among prokaryotic genomes; iii) parameters of identity distribution for probable orthologs; iv) analysis of concatenated alignments of ribosomal proteins; v) comparison of trees constructed for multiple protein families. All constructed trees support the separation of the two primary prokaryotic domains, bacteria and archaea, as well as some terminal bifurcations within the bacterial and archaeal domains. Beyond these obvious groupings, the trees made with different methods appeared to differ substantially in terms of the relative contributions of phylogenetic relationships and similarities in gene repertoires caused by similar life styles and horizontal gene transfer to the tree topology. The trees based on presence-absence of genomes in orthologous clusters and the trees based on conserved gene pairs appear to be strongly affected by gene loss and horizontal gene transfer. The trees based on identity distributions for orthologs and particularly the tree made of concatenated ribosomal protein sequences seemed to carry a stronger phylogenetic signal. The latter tree supported three potential high-level bacterial clades: i) Chlamydia-Spirochetes, ii) Thermotogales-Aquificales (bacterial hyperthermophiles), and iii) Actinomycetes-Deinococcales-Cyanobacteria. The latter group also appeared to join the low-GC Gram-positive bacteria at a deeper tree node. These new groupings of bacteria were supported by the analysis of alternative topologies in the concatenated ribosomal protein tree using the Kishino-Hasegawa test and by a census of the topologies of 132 individual groups of orthologous proteins. Additionally, the results of this analysis put into question the sister-group relationship between the two major archaeal groups, Euryarchaeota and Crenarchaeota, and suggest instead that Euryarchaeota might be a paraphyletic group with respect to Crenarchaeota.

**Conclusions:** We conclude that, the extensive horizontal gene flow and lineage-specific gene loss notwithstanding, extension of phylogenetic analysis to the genome scale has the potential of uncovering deep evolutionary relationships between prokaryotic lineages.
Background

The determination of multiple, complete genome sequences of bacteria, archaea and eukaryotes has created the opportunity for a new level of phylogenetic analysis that is based not on a phylogenetic tree for selected molecules, for example, rRNAs, as in traditional molecular phylogenetic studies [1,2], but (ideally) on the entire body of information contained in the genomes. The most straightforward version of this type of analysis, to which we hereinafter refer to as 'genome-tree' building, involves scaling-up the traditional tree-building approach and analyzing the phylogenetic trees for multiple gene families (in principle, all families represented in many genomes), in an attempt to derive a consensus, 'organismal' phylogeny [3–5]. However, because of the wide spread of horizontal gene transfer and lineage-specific gene loss, at least in the prokaryotic world, comparison of trees for different families and consensus derivation may become highly problematic [6,7]. Probably due to all these problems, a pessimistic conclusion has been reached that prokaryotic phylogeny might not be reconstructable from protein sequences, at least with current phylogenetic methods [4].

With the complete genome sequences at hand, it appears natural to seek for alternatives to traditional, alignment-based tree-building in the form of integral characteristics of the evolutionary process. Probably the most obvious of such characteristics is the presence-absence of representatives of the analyzed species in orthologous groups of genes, and recently, at least three groups have employed this approach to build genome trees, primarily for prokaryotes [8–10]. An alternative way to construct a genome tree involves using the mean or median level of similarity among all detectable pairs of orthologs as the measure of the evolutionary distance between species [11]. Yet another possibility involves building species trees by comparing gene orders. This approach had been pioneered in the classical work of Dobzhansky and Sturtevant who used inversions in Drosophila chromosomes to construct an evolutionary tree [12]. Subsequently, mathematical methods have been developed to calculate rearrangement distances between genomes, and, using these, phylogenetic trees have been built for certain small genomes, such as plant mitochondria and herpesviruses [13,14]. These approaches, however, are applicable only to genomes that show significant conservation of global gene order, which is manifestly not the case among prokaryotes [15–17]. Even relatively close species such as, for example, Escherichia coli and Haemophilus influenzae, two species of the γ-subdivision of Proteobacteria, retain very little conservation of gene order beyond the operon level (typically, two-to-four genes in a row), and essentially none is detectable among evolutionarily distant bacteria and ar chaea [15,16,18]. Very few operons, primarily those coding for physically interacting subunits of multiprotein complexes such as certain ribosomal proteins or RNA-polymerase subunits, are conserved across a wide range of prokaryotic lineages [15,16]. On the other hand, pairwise comparisons of even distantly related prokaryotic genomes reveal considerable number of shared (predicted) operons, which creates an opportunity for a meaningful comparative analysis [19][20,21].

The critical issue with all these approaches to genome tree building is to what extent each of them reflects phylogeny and to what extent they are affected by other evolutionary processes, such as lineage-specific gene loss and horizontal gene transfer. Comparative analyses have strongly suggested that these phenomena make major contributions to genome evolution, at least in prokaryotes [7,22–25]. These phenomena have the potential to severely affecting phylogenetic tree topology, particularly when similar sets of genes are lost indifferent lineages because of similar environmental pressures, or when a preferential trend of horizontal gene flow exists between different lineages. The possibility even has been discussed that the amount of lateral gene exchange is such that it invalidates the very principle of representing the evolution of species as a tree; instead, the only adequate representation of evolutionary history could be a complex network [6][25]. Genome-trees seem to be the last resort for the species tree concept. Unless phylogenetic signal can be revealed by at least some approaches based on genome-wide comparisons, the conclusion seems imminent that this concept should be abandoned and replaced by a more complex representation of evolution.

Here, we compare the topologies produced with five, largely independent approaches to genome-tree building: i) presence-absence of genomes in Clusters of Orthologous Groups of proteins (COGs); ii) conservation of local gene order (pairs of adjacent genes) among prokaryotic genomes; iii) distribution of percent identity between apparent orthologs; iv) sequence conservation in concatenated alignments of ribosomal proteins; v) comparative analysis of multiple trees reconstructed for representative protein families. We find that, while the presence-absence approach is most heavily affected by gene loss and horizontal transfer, the other four methods reveal stronger phylogenetic signals. Although the topologies of the trees constructed with different approaches were only partially compatible, three previously unnoticed high-level clades among bacteria were revealed with notable consistency. We suggest that, in spite of all the complexity brought about by horizontal gene transfer and lineage-specific gene loss, these groups reflect certain evolutionary reality, i.e. the trajectory of evolution for a relatively stable gene core. It appears that this is the
only meaningful way to treat the notion of a species tree: as the history of a relatively large ensemble of genes, not a comprehensive representation of the history of entire genomes.

Results

New criteria for genome-tree construction

To our knowledge, conserved gene pairs and distributions of identity level between orthologs have not been used previously as the basis for phylogenetic tree construction. Therefore we start by describing the relevant results of prokaryotic genome comparison in somewhat greater detail.

Conserved gene pairs in prokaryotic genomes

The results of the present analysis of conserved gene pairs are consistent with the notion of the fluidity of prokaryotic gene order caused by extensive recombination. Only 17 invariant genes pairs were detected, all of which consists of genes for ribosomal proteins and RNA polymerase subunits. The remaining 4586 gene pairs were missing in at least one genome. The number of gene pairs represented in three, four and a greater number of genomes decayed rapidly, with highly conserved pairs forming the tail of the distribution (Fig. 1). The 95% quantile of this distribution (excluding the highly conserved pairs) was found to fit the geometric model with a high statistical significance (Fig. 1). This is compatible with random, independent loss of gene pairs during evolution suggesting that, with the caveat of horizontal transfer, the number of gene pairs shared by three genomes could reflect the evolutionary distance between them.

The number of conserved gene pairs present in individual prokaryotic genomes varied from 208 for M. genitalium to 2314 for P. aeruginosa (Table 1). Analysis of the co-occurrence of gene pairs among the prokaryotic genomes shows high values of the Jaquard coefficient, which reflect partial conservation of gene order (see legend to Table 1), for closely related species, for example, 0.32 for E. coli and H. influenzae and 0.35 for M. thermoautotrophicum and M. jannaschii (Table 1). The value of this coefficient varied from 0.16 to 0.66, with a mean of 0.26, for archaea, and from 0.04 to 0.87, with a mean of 0.16, for bacteria. In contrast, for archaeal-bacterial comparisons, the values varied from 0.04 to 0.18, with the average of 0.08 (Table 1). These observations appear to indicate that the distribution of conserved gene pairs among prokaryotic genomes carries a phylogenetic signal.

Distributions of identity percentage between probable orthologs from complete prokaryotic genomes

Figure 2 shows a sampling of the distributions of identity percentage between pairs of apparent orthologs identified as reciprocal best hits from a range of genome pairs separated by varying phylogenetic distances. Most of the distributions are clearly unimodal, and the distributions for pairs of phylogenetically distant genomes, such as those from different major bacterial lineages or bacteria versus archaea, have their modes within a relatively narrow range around 33% identity (Figure 2).

The use of reciprocal best hits is a conservative way to identify the set of probable orthologs between pairs of genomes because some of the orthologs are missed due to complex relationships between groups of paralogs. Nevertheless, all genome-to-genome comparisons included at least 100 (for the smallest genomes such as the mycoplasmas), and typically, a considerably greater number of protein pairs ([11] and data not shown). This suggests that parameters of the distributions of the similarity level between probable orthologs identified in this fashion could potentially serve as useful measures of the evolutionary distance between genomes.

Genome trees constructed with three different approaches

Genome trees were generated using the approaches described under Material and Methods. All the trees showed a clear separation of the two major prokaryotic domains, Bacteria and Archaea (Fig. 3,4,5). Several terminal bifurcations that reflect clustering of relatively close species, such as three mycoplasmas (M. genitalium, M. pneumoniae and U. urealyticum), two spirochetes (B. burgdorferi and T. pallidum), and H. pylori and C. jejuni, are also reproduced in all trees (Fig. 3,4,5). This retention of both the deepest and the terminal branchings shows that all types of data used for tree con-
Table 1: Shared gene pairs in prokaryotic genomes.

|       | Aer  | Sus  | Arf  | Pyh  | Pya  | Mej  | Met  | Has  | Tha  | Thv  | Esc  | Vic  | Hai  | Pam  | Buc  | Psa  | Xyf  | Nei  | Cac  | Mel  | Rip  | He  | Caj  | Bas  | Bah  | Lal  | Sta  | Stp  | Myp  | Myg  | Urn  | Myt  | SyP  | Der  | Bob  | Trp  | Chp  | Cht  | Aqa  | Thm  |
|-------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
|       | Aepre |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
|       | 495   | 298  | 263  | 207  | 241  | 123  | 172  | 238  | 212  | 227  | 198  | 61   | 102  | 112  | 138  | 120  | 93   | 92   | 146  | 219  | 53   | 87   | 116  | 190  | 200  | 128  | 149  | 96   | 55   | 54   | 51   | 164  | 82   | 172  | 64   | 63   | 65   | 75   | 148  |
|       |       |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |

The diagonal (bold) shows the total number of conserved gene pairs in each genome. The upper triangle of the matrix shows the raw number of gene pairs shared by each pair of genomes and the lower triangle shows the value of 100\*Oij, where Oij is the Jaccard coefficient calculated as \( Q_{ij} = C_{ij} / (N_i + N_j - C_{ij}) \) where \( C_{ij} \) is the number of gene pairs shared by genomes \( i \) and \( j \), and \( N_i \) and \( N_j \) are the total numbers of conserved gene pairs in genomes \( i \) and \( j \), respectively [51].
struction contained at least a crude phylogenetic signal. However, beyond these obvious aspects of topology, and in particular with respect to clustering of distantly related bacteria and archaea, the trees produced with different approaches showed significant differences, which appear to reflect the relative contributions of phenotypic and phylogenetic signals. A quantitative comparison of the tree topologies using the symmetric distance method showed that the presence-absence tree was most different from the trees made by the other methods (Table 2).

**Presence-absence of genomes in COGs**

The topology of the parsimony tree built using this criterion appears to reflect primarily the phenotypes of the respective organisms (Fig. 3). This is most clearly manifest in the two major bacterial clusters that appear in this tree, each with a strong bootstrap support:

i) bacteria with large genomes, namely *E. coli*, *B. subtilis*, *Synechocystis* sp., *Deinococcus radiodurans* and *Mycobacterium tuberculosis*, and free-living bacteria with small genomes, *A. aeolicus* and *T. maritima*

ii) parasites with small genomes (mycoplasmas, spirochetes, chlamydia and rickettsia)

Parasites with moderate-sized genomes (*H. influenzae*, *N. meningitidis*, and *P. multocida*; *H. pylori* and *C. jejuni*) formed two distinct groups. Thus, well-established phylogenetic relationships between free-living and parasitic bacteria, such as those within the Proteobacteria (*E. coli-H. influenzae-P. multocida-N. meningitidis*) and within low-GC Gram-positive bacteria (*B. subtilis-mycoplasmas*), are not reflected accurately in this tree topology. The two free-living bacteria with small genomes, the hyperthermophiles *A. aeolicus* and *T. maritima*, did not join either the free-living or the parasitic bacterial cluster, despite their small number of genes similar to that in bacterial parasites (Fig. 3). That these bacteria do not group with the parasites despite similar genome sizes, suggests that it is not the number of genes per se, but rather the degree of genome degradation and the loss of coherent sets of genes that affect the topology of the presence-absence tree. The inclusion of the parasites *M. tuberculosis* and *Pseudomonas aeruginosa* in the cluster of bacteria with large genomes probably reflects the recent origin of parasitism in these lineages. It is further notable that, in this tree, the two representative of *Crenarchaeota* (*A. pernix* and *S. solfataricus*) do not comprise a sister group of the Euryarchaeota (the remaining archaeal species), but rather for an branch within the Euryarchaeal cluster (see discussion below).

In previous studies that employed similar approaches to genome-tree building, phylogenetically reasonable clades were observed after a simple omission of parasitic species [8,9]. Such an operation could be applied to the tree shown in Fig. 3, indeed resulting in the correct recovery of the proteobacterial and Gram-positive bacteri-
al lineages. However, it seems that, because known natural groups could be reproduced by this approach only after omission of certain species on the basis of independent prior knowledge, this method hardly can be useful for delineating new, phylogenetically sound clades.

Conserved gene pairs

The topology of the tree based on gene pair conservation seems to carry a stronger phylogenetic signal than the gene presence-absence tree because it correctly groups together related free-living and parasitic bacteria despite major differences in gene repertoires (Fig. 4). The bacterial side of this tree consists of three major clades: i) proteobacterial clade that, in addition to bona fide Proteobacteria, includes also A. aeolicus, M. tuberculosis, D. radiodurans, and Synechoystis sp, ii) Gram-positive clade that additionally includes T. maritima, and iii) an unexpected clade that unites spirochetes and chlamydia. In the archaeal domain, the two species of the Crenarchaeota did not form a clade, but instead were present as separate branches interspersed with euryarchaeal species. To further assess the robustness of the obtained tree, we varied the parameters of the included conserved pairs by allowing distances between the genes comprising a pair from 0 to 5 and changing the minimal number of genomes, in which a conserved gene pair had to be present, from 2 to 4. These changes did not significantly affect the tree topology (data not shown). The topology of a neighbor-joining tree constructed by using the number of gene pairs shared by two genomes to calculate the evolutionary distance between them was similar to the topology of the maximum parsimony tree (Table 2 and data not shown).

At least some unusual aspects of this tree’s the topology could be explained by horizontal transfer of operons between particular bacterial and archaeal lineages. Specifically, it has been noticed previously that T. maritima shares a considerable number of genes and operons with Gram-positive bacteria, to the exclusion of other bacteria [21]; this seems to be compatible with the position of T. maritima with the Gram-positive cluster. Similarly, considerable horizontal gene transfer appear to have occurred between the Sulfolobus and Thermoplasma lineages, which cluster together in the archaeal part of this tree. The presence of extra species in the proteobacterial cluster is more surprising because no obvious trend for operon transfer between these bacteria and bona fide Proteobacteria has been noticed during systematic genome comparisons; however, a considerable

Figure 4

Maximum parsimony tree (Dollo parsimony) of prokaryotes based on presence-absence of gene pairs in genomes. The designations are as in Fig. 3.

Figure 5

Distance tree constructed using the median of the percent identity distribution between probable orthologs for evolutionary distance calculation. An E-value cut-off of 0.001 was used to identify bidirectional best hits between proteins encoded in all pairs of genomes. Distances were calculated using the logarithmic formula. The designations are as in Fig. 3.
number of shared gene pairs was detected during the present analysis (Table 1). Artifacts of tree construction could also contribute to these associations. In contrast, the spirochete-chlamydia clade might reflect a deep phylogenetic relationship (see discussion below).

**Parameters of percent identity distributions between orthologs**

Different characteristics of the distributions of percent identity between the probable orthologs, such as the mean, the median, the mode and various quantiles, were used to calculate distances between genomes and construct phylogenetic trees. Trees built with different cut-off values for symmetrical best hits, four different formulas for the evolutionary distance calculation (see Materials and Methods) and different parameters of the distributions showed essentially the same topology, with strong bootstrap support for most of the clades (Fig. 5 and data not shown). The complete proteobacterial and Gram-positive bacterial clusters were recovered in this tree as well as the unexpected grouping of chlamydia with spirochete noticed above in the tree based on conserved gene pairs (Fig. 4,5). Also similarly to the previous two trees, the Crenarchaeae grouped with *Thermoplasma* within the archaeal part of the tree. Beyond these groupings, the tree appeared conservative in the sense that the unassigned bacterial species formed separate branches near the root of the bacterial subtree. The closest to the root were the two hyperthermophiles, *A. aeolicus* and *T. maritima*, which is compatible with the standard view of their phylogenetic position [1,26].

**Alignment-based approaches to the construction of a species tree**

The above three approaches involve construction of genome trees "par excellence", i.e. based on integral characteristics of genomes (or, more precisely, gene sets) that are not directly related to more traditional, alignment-based measures, which are usually employed for calculating evolutionary distances or for parsimony analysis.

These genome tree raise several interesting phylogenetic questions, for example, do spirochetes and chlamydia indeed share a common ancestor, and are Euryarchaeota, in fact, a paraphyletic group with respect to the Crenarchaeota. However, the reliability of the conclusions drawn from the topology of these trees remains uncertain. Therefore we decided to complement these genome-oriented approaches with more traditional ones applied on a large scale.

**Concatenated alignments of ribosomal proteins**

The alignments of the 32 ribosomal proteins conserved in all bacterial and archaeal species were concatenated head-to-tail and treated as a single alignment containing 4821 columns. The underlying assumption is that the genes coding for ribosomal proteins that function as components of a large macromolecular complex are unlikely to undergo horizontal transfer, which tends to confound comparisons of the tree topologies for other protein families and would invalidate the concatenation approach. The resulting maximum-likelihood tree contains the complete proteobacterial and Gram-positive bacterial clusters as well as the spirochete-chlamydia cluster noticed in the genome-trees. In addition to the spirochetes-chlamydia clade, the following non-trivial affinities were detected with strong bootstrap support: i) a cluster of the two hyperthermophiles, *A. aeolicus* and *T. maritima*, ii) a cluster including *D. radiodurans*, *Synechocystis*, and *M. tuberculosis*, which, at a deeper level, joined the Gram-positive bacterial branch (Fig. 6). Similar tree topologies were obtained when the ribosomal protein data were analyzed using the neighbor-joining method and when bacterial phylogeny was analyzed separately by using a concatenated alignments of 51 ribosomal proteins shared by all bacteria (data not shown). Notably, in the quantitative comparison of tree topologies, the tree made of concatenated ribosomal protein alignments showed the closest similarity to the genome-tree based on the distributions of percent identity between orthologs (Table 2).

---

**Table 2: Symmetric distances between genome-trees constructed with different methods.**

| Method                        | Symmetric distance<sup>a</sup> |
|-------------------------------|-------------------------------|
| Gene presence-absence         | 52                            |
| Conserved gene pairs          | 54                            |
| Identity distributions        | 44                            |
| Concatenated ribosomal proteins | 56                       |

<sup>a</sup>Number of different partitions of the total of 74 partitions.
The reliability of the observed non-trivial groupings was further examined by using a maximum likelihood approach (the Kishino-Hasegawa test). For each clade (usually, species) forming the group to be tested, trees with alternative topologies were manually constructed by joining the clade in question to every other major group in the tree. For example, to assess the support for the spirochetes-chlamydia grouping, spirochetes were placed, sequentially, with *Thermotoga*, *Aquifex*, the *Thermotoga-Aquifex* branch, ε-proteobacteria, the αβγ-proteobacterial branch, Proteobacteria, the *Deinococcus-Synechocystis-Mycobacterium* cluster, the low G+C Gram-positive cluster, the branch that unites the latter two clusters, and between bacteria and archaea (to the bacterial root). The same alternatives were tested for chlamydia. Alternative topologies were compared either directly, using the ProtML program, or were subjected to local rearrangement first. In cases when the topology did not revert to the original one, the final, "optimized" topology was used for the comparison. These tests showed high stability of the *Thermotoga-Aquifex* and *Deinococcus-Synechocystis-Mycobacterium* groupings (no competing topologies with likelihood within 1 SD unit from the original; Fig. 7,8, Table 3,4,5,6). The affinity of the *Deinococcus-Synechocystis-Mycobacterium* with Gram-positive bacteria also was supported, although an alternative topology, with this cluster joining Proteobacteria could not be ruled out (Fig. 9, Table 7). Assessment of the spirochete-chlamydia grouping revealed two competing topologies, albeit unusual ones. Specifically, moving ε-proteobacteria from the proteobacterial branch to the spirochete branch or, alternatively, moving spirochetes with ε-proteobacteria and simultaneously moving chlamydia to the bacterial root results in statistically
acceptable topologies (Fig. 10; Table 8, 9). Also, a minor rearrangement of the topology within the euryarchaeal branch allowed for a reasonable alternative to the topology in Fig. 8 (euryarchaeal paraphyly), with the Crenarchaeae-Euryarchaeae radiation at the archaeal root (Fig. 11, Table 10).

**Figure 8**
The Kishino-Hasegawa test for the Deinococcus-Mycobacterium-Synechocystis clade. The identical scheme of producing alternative topologies was used for each of the three species. For example for Deinococcus (see Table 4) the green ovals (## 2 to 13) indicate alternative placements of Deinococcus with Mycobacterium and Synechocystis occupying the original position and the blue ovals (## 14 to 25) indicate alternative placements of the Mycobacterium-Synechocystis pair with Deinococcus left in the original position. The same was done with Mycobacterium versus Deinococcus-Synechocystis pair (Table 5) and Synechocystis versus Deinococcus-Mycobacterium pair (Table 6).

**Figure 9**
The Kishino-Hasegawa test for the unification of the Deinococcus-Mycobacterium-Synechocystis clade with Gram-positive bacteria. See Table 7.

A census of protein families
Another approach to the "species tree" problem involves analysis of phylogenetic trees for as many individual protein families as possible, in an attempt to identify a prevailing topology or at least common phylogenetic patterns. A survey of the COG data set identified 132 COGs, each of which included a large number of bacterial
and archael species, but no or few paralogs and thus appeared to be amenable to a large-scale phylogenetic analysis (Table 11). Maximum-likelihood trees were constructed for each of these COGs, and a breakdown of nearest neighbors was derived for species and groups involved in each of the non-trivial or questionable branchings discussed above (Crenarchaea, Thermotoga, Aquifex, Deinococcus, Mycobacterium, Synechocystis, spirochetes, chlamydia, and ε-proteobacteria). In each case, a wide spread of topologies was observed, but the grouping that is observed in the concatenated ribosomal proteins tree was encountered most often, although, for example, for the spirochete-chlamydia cluster, the lead over other topologies was slim (Fig. 13,14,15).

Discussion and Conclusions
The trees constructed with each of the four approaches employed here reflect both the phylogenetic signal and the phenotypic (life style) similarities or differences between organisms, but the relative contributions of these
Table 3: Testing non-trivial groupings from the concatenated ribosomal protein tree with the Kishino-Hasegawa test

(A) 
*Aquifex*-*Thermotoga*

| #  | Likelihood  | ΔL | σΔL | RELL-BP |
|----|-------------|----|-----|---------|
| 1  | -242983.7   | best | N/A  | 0.9251  |
| 2  | -243174.6   | -190.9 | 38.5 | 0.0000  |
| 3  | -243185.0   | -201.3 | 38.0 | 0.0000  |
| 4  | -243146.1   | -162.5 | 32.1 | 0.0000  |
| 5  | -243267.6   | -283.9 | 49.0 | 0.0000  |
| 6  | -243293.3   | -309.7 | 49.0 | 0.0000  |
| 7  | -243218.8   | -235.2 | 41.8 | 0.0000  |
| 8  | -243301.0   | -317.3 | 45.7 | 0.0000  |
| 9  | -243315.4   | -331.8 | 45.0 | 0.0000  |
| 10 | -243242.8   | -259.1 | 40.0 | 0.0000  |
| 11 | -243005.7   | -22.0  | 12.2 | 0.0227  |
| 12 | -243196.1   | -212.4 | 39.2 | 0.0000  |
| 13 | -243201.5   | -217.9 | 38.8 | 0.0000  |
| 14 | -243157.9   | -174.3 | 32.2 | 0.0000  |
| 15 | -243318.8   | -335.1 | 49.8 | 0.0000  |
| 16 | -243355.8   | -372.1 | 48.3 | 0.0000  |
| 17 | -243247.1   | -263.4 | 42.1 | 0.0000  |
| 18 | -243236.5   | -252.8 | 51.2 | 0.0000  |
| 19 | -243232.0   | -248.3 | 51.1 | 0.0000  |
| 20 | -243207.0   | -223.3 | 45.0 | 0.0000  |
| 21 | -243002.6   | -19.0  | 12.8 | 0.0522  |

Table 4:

(B) 
*Deinococcus* *radiodurans*

| #  | Likelihood  | ΔL | σΔL | RELL-BP |
|----|-------------|----|-----|---------|
| 1  | -242983.7   | best | N/A  | 0.8239  |
| 2  | -243091.1   | -107.4 | 40.8 | 0.0002  |
| 3  | -243122.6   | -138.9 | 43.0 | 0.0000  |
| 4  | -243135.8   | -152.1 | 43.1 | 0.0000  |
| 5  | -243088.1   | -104.5 | 36.3 | 0.0000  |
| 6  | -243037.3   | -53.7  | 41.0 | 0.0775  |
| 7  | -243064.0   | -80.4  | 40.2 | 0.0020  |
| 8  | -243024.5   | -40.9  | 31.8 | 0.0574  |
| 9  | -243030.9   | -47.3  | 19.0 | 0.0011  |
| 10 | -243017.6   | -33.9  | 20.8 | 0.0090  |
| 11 | -243052.1   | -68.4  | 30.3 | 0.0010  |
| 12 | -243070.6   | -86.9  | 37.3 | 0.0013  |
| 13 | -243066.1   | -82.5  | 40.4 | 0.0122  |
| 14 | -243143.1   | -159.4 | 39.6 | 0.0000  |
| 15 | -243151.3   | -167.7 | 43.3 | 0.0000  |
| 16 | -243186.7   | -203.0 | 42.2 | 0.0000  |
two types of information appear to differ substantially. The gene presence-absence analysis seemed to be dominated by the phenotypic signal, primarily that from gene loss. The tree based on conserved gene pairs appeared to combine phylogenetic information with major effects of horizontal transfer of operons. In contrast, the trees based on the distributions of the identity level of orthologs appear to be more meaningful phylogenetically as indicated by the recovery of established high-level phylogenetic groups of bacteria, such as Proteobacteria and Gram-positive bacteria. The ability to correctly identify these major bacterial subdivisions and the absence of obviously wrong groupings confer credibility to non-trivial clades present in these trees, in particular the spirochete-chlamydia clade. The same logic applied to the tree made of concatenated ribosomal protein sequences,

Table 4: (Continued)

|   | Likelihood | ΔL | σΔL | RELL-BP |
|---|------------|----|-----|---------|
| 17 | -243102.9  | -119.3 | 36.6 | 0.0001  |
| 18 | -243167.6  | -184.0 | 37.8 | 0.0000  |
| 19 | -243155.4  | -171.7 | 38.9 | 0.0000  |
| 20 | -243065.3  | -81.7  | 29.6 | 0.0007  |
| 21 | -243017.6  | -33.9  | 20.8 | 0.0121  |
| 22 | -243030.9  | -47.3  | 19.0 | 0.0006  |
| 23 | -243068.3  | -84.7  | 29.9 | 0.0009  |
| 24 | -243103.9  | -120.3 | 36.7 | 0.0000  |
| 25 | -243135.2  | -151.5 | 39.0 | 0.0000  |

Table 5:

(C) *Mycobacterium tuberculosis*

| #  | Likelihood | ΔL | σΔL | RELL-BP |
|----|------------|----|-----|---------|
| 1  | -242983.7  | best | N/A | 0.8589  |
| 2  | -243160.7  | -177.0 | 46.5 | 0.0000  |
| 3  | -243192.5  | -208.9 | 48.9 | 0.0000  |
| 4  | -243216.2  | -232.5 | 48.6 | 0.0000  |
| 5  | -243140.3  | -156.6 | 44.1 | 0.0000  |
| 6  | -243146.6  | -163.0 | 48.1 | 0.0000  |
| 7  | -243153.9  | -170.3 | 48.0 | 0.0000  |
| 8  | -243071.4  | -87.7  | 41.0 | 0.0013  |
| 9  | -243023.4  | -39.7  | 34.2 | 0.0443  |
| 10 | -243037.2  | -53.5  | 33.3 | 0.0052  |
| 11 | -243098.4  | -114.7 | 39.6 | 0.0000  |
| 12 | -243126.1  | -142.4 | 44.5 | 0.0000  |
| 13 | -243146.5  | -162.8 | 46.9 | 0.0000  |
| 14 | -243087.0  | -103.3 | 52.5 | 0.0010  |
| 15 | -243128.5  | -144.8 | 54.7 | 0.0000  |
| 16 | -243150.8  | -167.2 | 54.1 | 0.0000  |
| 17 | -243079.5  | -95.9  | 49.1 | 0.0014  |
| 18 | -243136.6  | -153.0 | 50.5 | 0.0000  |
| 19 | -243152.6  | -168.9 | 49.7 | 0.0000  |
| 20 | -243062.9  | -79.3  | 41.5 | 0.0012  |
| 21 | -243037.2  | -53.5  | 33.3 | 0.0059  |
| 22 | -243023.4  | -39.7  | 34.2 | 0.0327  |
| 23 | -243047.8  | -64.1  | 43.2 | 0.0209  |
| 24 | -243062.5  | -78.8  | 49.7 | 0.0192  |
| 25 | -243076.6  | -93.0  | 51.9 | 0.0080  |
Table 6:

(D)
Synechocystis sp.

| #  | Likelihood | ΔL  | σΔL | RELL-BP |
|----|------------|-----|-----|---------|
| 1  | -242983.7  | best| N/A | 0.9617  |
| 2  | -24318.5   | -134.8| 47.5 | 0.0000  |
| 3  | -243077.9  | -94.3| 51.4 | 0.0265  |
| 4  | -243115.8  | -132.1| 50.9 | 0.0000  |
| 5  | -243084.6  | -101.0| 46.4 | 0.0031  |
| 6  | -243184.5  | -200.8| 46.3 | 0.0000  |
| 7  | -243208.1  | -224.4| 45.5 | 0.0000  |
| 8  | -243135.7  | -152.1| 38.1 | 0.0000  |
| 9  | -243072.3  | -88.6| 32.1 | 0.0006  |
| 10 | -243083.7  | -100.0| 31.6 | 0.0000  |
| 11 | -243099.4  | -115.7| 40.6 | 0.0000  |
| 12 | -243102.5  | -118.8| 45.2 | 0.0003  |
| 13 | -243097.2  | -113.5| 47.6 | 0.0030  |
| 14 | -243204.5  | -220.8| 48.0 | 0.0000  |
| 15 | -243279.8  | -296.2| 49.3 | 0.0000  |
| 16 | -243288.4  | -304.7| 49.4 | 0.0000  |
| 17 | -243194.3  | -210.7| 42.9 | 0.0000  |
| 18 | -243180.8  | -197.1| 49.5 | 0.0000  |
| 19 | -243177.5  | -193.8| 49.4 | 0.0000  |
| 20 | -243090.1  | -106.4| 41.4 | 0.0038  |
| 21 | -243083.7  | -100.0| 31.6 | 0.0000  |
| 22 | -243072.3  | -88.6| 32.1 | 0.0010  |
| 23 | -243129.2  | -145.5| 38.5 | 0.0000  |
| 24 | -243165.6  | -181.9| 45.1 | 0.0000  |
| 25 | -243195.5  | -211.9| 47.6 | 0.0000  |

Table 7:

(E)
The Demococcus-Mycobacterium-Synechocystis clade

| #  | Likelihood | ΔL  | σΔL | RELL-BP |
|----|------------|-----|-----|---------|
| 1  | -242983.7  | 0.0 | <best| 0.7280  |
| 2  | -243065.3  | -81.7| 34.7 | 0.0000  |
| 3  | -243122.4  | -138.8| 37.2 | 0.0000  |
| 4  | -243148.7  | -165.1| 35.8 | 0.0000  |
| 5  | -243053.8  | -70.1| 28.9 | 0.0001  |
| 6  | -243103.1  | -119.5| 33.6 | 0.0000  |
| 7  | -243096.4  | -112.7| 34.1 | 0.0001  |
| 8  | -243003.1  | -19.4| 23.2 | 0.1697  |
| 9  | -243010.5  | -26.9| 21.5 | 0.0560  |
| 10 | -243028.9  | -45.3| 31.2 | 0.0419  |
| 11 | -243054.3  | -70.7| 34.7 | 0.0042  |
which included two other non-trivial bacterial groupings, *Aquifex-Thermotoga* and *Synechocystis-Mycobacterium-Deinococcus*, the latter joining the Gram-positive branch. Furthermore, extensive testing of alternative topologies using the Kishino-Hasegawa test largely supported these new bacterial branches. The nature of this support becomes clearer when one examines the results of the protein family census. Each of the potential new clades was indeed most common among the observed topologies, but in no case, was the excess of this topology overwhelming. Taken together, these results seem to shed light on the very notion of a "species tree". It appears that, at best, a species tree can be viewed as a
Table 10:

(H) Crenarchaeota and Euryarchaeota

|   | Likelihood | ΔL | σΔL | RELL-BP |
|---|------------|----|-----|---------|
| 1 | -242983.7  | best | N/A | 0.5840  |
| 2 | -242993.2  | -9.5 | 33.7 | 0.4160  |

The numbers correspond to those in Fig. 5. The likelihood difference with the first (original) topology. The standard deviation of the above. The bootstrap probability of the given topology estimated with RELL method [49].

Table 11: COGs used for the comparative analysis of Maximum Likelihood trees for individual protein families

| COG    | spec | prot | Name                                                                 |
|--------|------|------|----------------------------------------------------------------------|
| COG0012| 40   | 41   | Predicted GTPase                                                      |
| COG0013| 40   | 40   | Alanyl-tRNA synthetase                                               |
| COG0016| 40   | 40   | Phenylalanyl-tRNA synthetase alpha subunit                           |
| COG0018| 40   | 41   | Arginyl-tRNA synthetase                                              |
| COG0020| 37   | 40   | Undecaprenyl pyrophosphate synthase                                  |
| COG0048| 39   | 39   | Ribosomal protein S12                                                |
| COG0049| 40   | 41   | Ribosomal protein S7                                                 |
| COG0051| 40   | 40   | Ribosomal protein S10                                                |
| COG0052| 40   | 40   | Ribosomal protein S2                                                 |
| COG0060| 40   | 40   | Isoleucyl-tRNA synthetase                                            |
| COG0061| 37   | 40   | Predicted kinase                                                     |
| COG0064| 30   | 30   | Asp-tRNAAsn/Glu-tRNAGln amidotransferase B subunit (PET 112 homolog) |
| COG0072| 40   | 40   | Phenylalanyl-tRNA synthetase beta subunit                            |
| COG0080| 40   | 41   | Ribosomal protein L1                                                 |
| COG0081| 40   | 40   | Ribosomal protein L1                                                 |
| COG0082| 33   | 33   | Chorismate synthase                                                  |
| COG0085| 40   | 40   | DNA-directed RNA polymerase beta subunit/140 kD subunit (split gene in Mjan, Mthe, Aful) |
| COG0087| 40   | 40   | Ribosomal protein L3                                                 |
| COG0088| 40   | 40   | Ribosomal protein L4                                                 |
| COG0090| 40   | 40   | Ribosomal protein L2                                                 |
| COG0091| 40   | 40   | Ribosomal protein L22                                                |
| COG0092| 40   | 40   | Ribosomal protein S3                                                 |
| COG0093| 39   | 39   | Ribosomal protein L1                                                 |
| COG0094| 40   | 40   | Ribosomal protein L5                                                 |
| COG0096| 40   | 40   | Ribosomal protein S8                                                 |
| COG0097| 40   | 40   | Ribosomal protein L6                                                 |
| COG0098| 40   | 40   | Ribosomal protein S5                                                 |
| COG0099| 40   | 40   | Ribosomal protein S13                                                |
| COG0100| 39   | 39   | Ribosomal protein S11                                                |
| COG0101| 38   | 38   | Pseudouridylate synthase (tRNA psi55)                                |
| COG0102| 40   | 40   | Ribosomal protein L1                                                 |
| COG0103| 40   | 40   | Ribosomal protein S9                                                 |
| COG0104| 31   | 31   | Adenylosuccinate synthase                                            |
| COG0105| 33   | 33   | Nucleoside diphosphate kinase                                         |
| COG0126| 39   | 39   | 3-phosphoglycerate kinase                                             |
| COG0127| 34   | 35   | Xanthosine triphosphate pyrophosphatase                              |
| COG0128| 33   | 35   | 5-enolpyruvlyshikimate-3-phosphate synthase                           |
| COG0130| 37   | 37   | Pseudouridine synthase                                               |
| COG0134| 30   | 30   | Indole-3-glycerol phosphate synthase                                  |
| COG0135| 30   | 30   | Phosphoribosylantranilate isomerase                                  |
| COG0143 | 40 | 41 | Methionyl-tRNA synthetase |
|---------|----|----|----------------------------|
| COG0148 | 39 | 43 | Enolase                     |
| COG0149 | 39 | 41 | Triosephosphate isomerase   |
| COG0151 | 30 | 30 | Phosphoribosylamine-glycine ligase |
| COG0152 | 30 | 32 | Phosphoribosylamineimidazolesuccinocarboxamide (SAICAR) synthase |
| COG0159 | 31 | 31 | Tryptophan synthase alpha chain |
| COG0162 | 40 | 43 | Tyrosyl-tRNA synthetase     |
| COG0164 | 35 | 35 | Ribonuclease HII             |
| COG0166 | 35 | 35 | Glucose-6-phosphate isomerase |
| COG0167 | 32 | 37 | Dihydroorotate dehydrogenase |
| COG0169 | 33 | 39 | Shikimate 5-dehydrogenase    |
| COG0171 | 35 | 37 | NAD synthase                |
| COG0172 | 40 | 40 | Seryl-tRNA synthetase        |
| COG0173 | 30 | 30 | Aspartyl-tRNA synthetase     |
| COG0178 | 31 | 33 | Excinuclease ATPase subunit  |
| COG0180 | 40 | 43 | Tryptophanyl-tRNA synthetase |
| COG0190 | 33 | 33 | 5,10-methylene-tetrahydrofolate dehydrogenase |
| COG0193 | 30 | 30 | Peptidyl-tRNA hydrolase      |
| COG0197 | 40 | 40 | Ribosomal protein L16/L10E   |
| COG0198 | 40 | 40 | Ribosomal protein L24        |
| COG0200 | 40 | 40 | Ribosomal protein L5         |
| COG0201 | 40 | 41 | Preprotein translocase subunit SecY |
| COG0202 | 40 | 40 | DNA-directed RNA polymerase alpha subunit/40 kDa subunit |
| COG0203 | 30 | 30 | Ribosomal protein L1         |
| COG0215 | 38 | 39 | Cysteinyl-tRNA synthetase    |
| COG0216 | 30 | 30 | Protein chain release factor A |
| COG0221 | 31 | 31 | Inorganic pyrophosphatase    |
| COG0222 | 30 | 30 | Ribosomal protein L7/L12     |
| COG0223 | 30 | 34 | Methionyl-tRNA formyltransferase |
| COG0231 | 40 | 46 | Translation elongation factor P | translation initiation factor elf-5A |
| COG0233 | 30 | 30 | Ribosome recycling factor    |
| COG0237 | 39 | 41 | Dephospho-CoA kinase         |
| COG0242 | 30 | 35 | N-formylmethionyl-tRNA deformylase |
| COG0244 | 40 | 40 | Ribosomal protein L10        |
| COG0250 | 40 | 43 | Transcription antiterminator |
| COG0256 | 40 | 40 | Ribosomal protein L1         |
| COG0258 | 40 | 47 | 5'-3' exonuclease (including N-terminal domain of Poli) |
| COG0261 | 30 | 30 | Ribosomal protein L21        |
| COG0264 | 30 | 30 | Translation elongation factor Ts |
| COG0272 | 30 | 31 | NAD-dependent DNA ligase (contains BRCT domain type II) |
| COG0275 | 30 | 30 | Predicted 5-adenosylmethionine-dependent methyltransferase involved in cell envelope biogenesis |
| COG0284 | 32 | 32 | Orotidine-5'-phosphate decarboxylase |
| COG0290 | 30 | 30 | Translation initiation factor IF3 |
| COG0292 | 30 | 30 | Ribosomal protein L20        |
| COG0294 | 33 | 36 | Dihydropterotate synthase    |
| COG0305 | 30 | 31 | Replicative DNA helicase     |
| COG0313 | 30 | 31 | Predicted methyltransferases |
| COG0319 | 30 | 30 | Predicted metal-dependent hydrolase |
| COG0335 | 30 | 30 | Ribosomal protein L1         |
| COG0336 | 30 | 30 | tRNA-(guanine-N1)-methyltransferase |
| COG0340 | 32 | 34 | Biotin-(acetyl-CoA carboxylase) ligase |
| COG0343 | 35 | 36 | Queuine/archaeosine tRNA-ribozyme transferase |
| COG0351 | 31 | 34 | Hydroxymethylpyrimidine/phosphomethylpyrimidine kinase |
| COG0359 | 30 | 30 | Ribosomal protein L9         |
| COG0441 | 40 | 43 | Threonyl-tRNA synthetase     |
| COG0442 | 40 | 40 | Prolyl-tRNA synthetase       |
| COG0452 | 32 | 32 | Phosphopantothenoylcytochrome synthetase/decarboxylase |
| COG0461 | 33 | 34 | Orotate phosphoribosyltransferase |
| COG0462 | 37 | 40 | Phosphoribosylpyrophosphate synthetase |
| COG0481 | 30 | 30 | Membrane GTPase LepA         |
| COG0495 | 40 | 41 | Leucyl-tRNA synthetase       |
| COG0504 | 38 | 38 | CTP synthase (UTPaminonona-lyase) |
prevailing phylogenetic trend, which, as far as deep branchings are concerned, may not even apply to a majority of the genes in a genome.

The potential new, deep relationships between bacterial lineages revealed during this analysis should be considered preliminary and treated with caution. Nevertheless, an evolutionary affinity between Cyanobacteria (Synechocystis) and Actinomycetes (Mycobacterium) appears plausible, particularly given the presence, in these bacterial groups, of well-developed and partly similar signal transduction systems [27]. The connection between two hyperthermophilic bacteria, Aquifex and Thermotoga, also has obvious biological meaning, although, in this case, particular caution is due, given the possibility of preferential horizontal gene exchange between these organisms that inhabit similar environments. However, the strong support for this grouping obtained in the analysis of concatenated ribosomal proteins argues against horizontal transfer as the primary cause for the observed topology. Although recent studies on the phylogeny of ribosomal proteins suggest some horizontal transfer events, these seem to be largely restricted to bacteria-specific ribosomal proteins. In the universal set of ribosomal proteins, only one, S14, showed clear signs of horizontal transfer [28]. The potential deep phylogenetic connections uncovered during this analysis call for detailed genome comparisons in search of potential shared derived characters, such as unique protein domain architectures, that could support the new clades.

The major bacterial lineages are poorly resolved in rRNA-based trees [2,29] and those built using alignments of RNA polymerase subunits [30] and translation elongation factors [29,31]. In the currently accepted taxonomy, which is based primarily (but not exclusively) on 16S RNA phylogenetic analysis, bacterial lineages that are suggested by this analysis to form higher-level clusters, tend to form primary nodes under Bacteria (Chlamydiales, Spirochetales, Cyanobacteria, the Thermus-Deinococcus group, Aquificales, Thermotogales). Thus, the genome trees primarily suggest (however tentatively) new unifications based on deep phylogenetic connections, rather than split already established clades. A notable exception is the traditional unification of Actinomycetes, or High G+C gram-positive bacteria (repre-
sented here by *Mycobacterium*, with low G+C Gram-positive bacteria (the *Bacillus*-*Clostridium* group) under Firmicutes (Gram-positive bacteria). Such a connection was not supported by any of the trees analyzed here, and it is also poorly, if at all, supported by the latest consensus trees for 16S RNA, 23S RNA and translation factor EF-Tu [29]. Therefore it seems likely that the Firmicutes clade, at least in its present composition, does not exist.

The new clade that might replace it consists of low-GC Gram-positive bacteria and the potential Actinomycetes-Deinococcales-Cyanobacteria group (Fig. 6). All methods of tree analysis applied here also challenge the traditional division of the archaeal kingdom into *Euryarchaeota* and *Crenarchaeota*, suggesting instead that *Euryarchaeota* could be a paraphyletic group with respect to *Crenarchaeota*, or in other words, that *Crenarchaeota* might have evolved from within the *Euryarchaeota*. However, the existence of a statistically supported alternative topology, with a sister-group relationship between *Euryarchaeota* and *Crenarchaeota* allows for the possibility that the apparent paraphyly of *Euryarchaeota* is an artifact caused by rapid evolution in some *Euryarchaeal* lineages, such as *Halobacterium* and *Thermoplasma*.

An independent phylogenetic study of concatenated ribosomal proteins has been recently published [32]. The main specific conclusion reported in this study was the apparent association of *Synechocystis* with Gram-positive bacteria, although instability of the tree topology de-
dependent on the subset of sites used for analysis was noticed. Another recent study addressed the issue of a global tree through phylogenetic analysis of 14 concatenated sets of orthologous proteins, for which no strong evidence of horizontal transfer was available [33]. Notably, some of the unexpected groupings within the bacterial domain reported in this study coincide or overlap with those described here, namely, a spirochete-chlamydial clade and a Deinococcales-Cyanobacteria clade. The grouping of the latter clade with Actinomycetes, the unification of the Deinococcales-Cyanobacteria-Actinomycetes clade with Gram-positive bacteria and the grouping of the two bacterial hyperthermophiles were not reproduced in the work of Brown and co-workers. The differences between the results of the two studies could owe to the differences between data sets analyzed, the methods used or, most likely, both. We should note that the present study engaged a substantially broader data set and more diverse methods for tree construction. We believe, however, that in terms of the potential contribution of genome-wide phylogenetic analysis to phylogenetic taxonomy, the areas where different methods and independent analyses by different groups converge might be more important than the areas of discrepancy. It appears that potential new clades revealed in such independent studies are strong candidates for new, high-level taxa.

The results of the present study suggest that genome trees based on new, integral criteria do not provide substantial advantages in phylogenetic reconstruction over more traditional, alignment-based methods expanded to the genomic scale. In fact, the latter seem to be more sensitive in detecting potential deep evolutionary relationships and this is expected to further improve with the increasing number of completely sequenced genomes becoming available for analysis. We believe, however, that this conclusion does not necessarily indicate that genome trees, such as those based on representation of genomes in orthologous sets or conservation of gene pairs, are useless. In addition to revealing some new phylogenetic affinities, they are capable of alerting researchers to other evolutionary phenomena, such as loss of similar gene sets in different organisms and preferential horizontal gene exchange between certain lineages.

**Material and Methods**

**Sequence data**

The sequences of the proteins encoded in complete genomes were extracted from the Genome division of the
Entrez retrieval system [34]. The analyzed genomes included those of 30 bacteria: Aquifex aeolicus (Aquae), Bacillus halodurans (Bacha), Bacillus subtilis (Basu), Borrelia burgdorferi (Borbu), Buchnera sp. (Busp), Campylobacter jejuni (Camje), Caulobacter crescentus (Caucr), Chlamydia trachomatis (Chltr), Chlamydia pneumoniae (Chlpn), Deinococcus radiodurans (Deira), Escherichia coli (Esco), Haemophilus influenzae (Haein), Helicobacter pylori (Helpy), Lactococcus lactis (Lacl), Mesorhizobium loti (Meslo), Mycoplasma genitalium (Myge), Mycoplasma pneumoniae (Mycopn), Mycobacterium tuberculosis (Myctu), Neisseria meningitidis (Neime), Pasteurella multocida (Pasmu), Psedomonas aeruginosa (Pseae), Rickettsia prowazekii (Ripcr), Staphylococcus aureus (Stau), Streptococcus pyogenes (Strpy), Synecocystis PCC6803 (SynPC), Thermotoga maritima (Them), Treponema pallidum (Trep), Ureaplasma urealyticum (Ureur), Vibrio cholerae (Vibch), Xylella fastidiosa (Xylfa), and ten archaea: Aeropyrum pernix (Aerpe), Archaeoglobus fulgidus (Arcul), Halobacterium sp. (Halsp), Methanobacterium thermoautotrophicum (Metth), Methanococcus jannaschii (Metcja), Pyrococcus horkoshii (Pyrho), Pyrococcus abyssi (Pyrab), Sulfolobus solfataricus (Sulso), Thermoplasma acidophilum (Theac), Thermoplasma volcanium (Thevo).

Phylogenetic tree construction

 Parsimony trees based on the presence-absence of conserved gene pairs in prokaryotic genomes

 The database of Clusters of Orthologous Groups of proteins (COGs) was used as the source of information on orthologous genes in prokaryotic genomes [35,36]. Briefly, the COGs were constructed from the results of all-against-all BLAST [37] comparison of proteins encoded in complete genomes by detecting consistent groups of genome-specific best hits (BeTs). The COG construction procedure does not rely on any preconceived phylogenetic tree of the included species except that certain obviously related genomes (for example, two species of mycoplasmas or pyrococci) were grouped prior to the analysis, to eliminate strong dependence between BeTs. In order to avoid spurious occurrence of the same gene pair, only gene pairs conserved in three or more genomes were considered. A pair of genes from two COGs was considered to be conserved if the respective genomes were adjacent in at least one genome and were separated by no more than two genes in at least two additional genomes. This relaxed definition of a conserved gene pair was adopted to take into account the high level of recombination in prokaryotic genomes. From the data on the presence-absence of each conserved gene pair in the analyzed genomes (excluding pairs of closely related species: E. coli-Buchnera sp., H. influenzae-P. multocida, C. trachomatis-C. pneumoniae, P. horkoshii-P. abyssi, M. genitalium-M. pneumoniae-U. urealyticum, H. pylori – C. jejuni, T. acidophilum-T. volcanium), a 0/1 matrix analogous to the one used for the presence-absence of individual genes was constructed, and a tree was built using Dollo parsimony [38]. A parsimony method was chosen for this analysis because the presence-absence of a conserved gene pair in a genome can be naturally treated in terms of character states. The Dollo model is based on the assumption that each derived character state (in this case, the presence of a gene pair) originates only once, and homoplasies exist only in the form of reversals to the ancestral condition (absence of a gene pair) [38]. In other words, parallel or convergent gains of the derived condition are assumed to be highly unlikely. The Dollo parsimony method is not sensitive to gene loss which is extremely common in evolution of prokaryotes, but the results can be affected by independent acquisition of the same gene pair by different genome via horizontal gene transfer. Phylogenetic analysis was performed by using the PAUP 4.0 program [39], with 1000 bootstrap replicates performed to assess the reliability of the tree topology. In addition, the tree topology was analyzed using the neighbor-joining method [40].

 Parsimony trees based on the representation of genomes in orthologous gene sets

 The information on orthologous genes in prokaryotic genomes and the yeast genome was derived from the COGs as in the previous approach, and the orthology data were similarly represented as a 0/1 matrix of presence-absence of the analyzed genomes in the COGs. A Dollo parsimony tree was constructed and the reliability of its topology was assessed using the bootstrap method as described above.

 Distance trees based on distributions of identity percentage between orthologous protein sequences

 The sequences of all proteins encoded in the analyzed genomes were compared to each other using the gapped BLASTP program [37]. Reciprocal, genome-specific BeTs were collected at different expectation (E) value cutoffs (0.01, 0.001, 0.0001, 0.00001). This method for identification of probable orthologs is, in principle, similar to the method employed in COG construction, but differs in that there is no requirement for the formation of triangles of consistent BeTs. The result of this procedure is a conservative selection of orthologous pairs because the cases of lineage-specific duplication that result in non-symmetrical BeTs are excluded and so are orthologous pairs with very low sequence similarity. However, the limitation of the COG system, namely the requirement that each orthologous group is represented in at least three genomes, is avoided. The distributions of identity percentage among the reciprocal best hits were derived for each pair of species. The mean, mode, medi-
Quantitative comparison of tree topologies
To compare tree topologies quantitatively, the symmetric distance between trees [50] was computed using the TREEDIST program of the PHYLIP package (version 3.6a). Briefly, each of the two compared trees is divided by each internal branch into two partitions. The symmetric distance is the number of partitions that are found in one tree but not the other.

Acknowledgements
We thank M. Nei for simulating discussions about the Dollo parsimony analysis, J. Felsenstein for alerting us of the inclusion of the TREEDIST program in PHYLIPI3.6a and D. Leipe for discussions on taxonomy.

References
1. Woese CR: Bacterial evolution. Microbiol Rev 1987, 51:221-271
2. Olsen GJ, Woese CR, Overbeek R: The winds of (evolutionary) change: breathing new life into microbiology. J Bacteriol 1994, 176:1-6
3. Doolittle RF, Fung DF, Tsang S, Cho G, Little E: Determining divergence times of the major kingdoms of living organisms with a protein clock. Science 1996, 271:470-477
4. Teichmann SA, Mitchison G: Is there a phylogenetic signal in prokaryote protein? J Mol Evol 1999, 49:98-107
5. S莳herz-Pontier T, Andersson SG: A phylogenomic approach to microbial evolution. Nucleic Acids Res 2001, 29:545-552
6. Doolittle WF: Phylogenetic classification and the universal tree. Science 1999, 284:2124-2129
7. Doolittle WF: Lateral genomics. Trends Cell Biol 1999, 9:58-59
8. Snel B, Bork P, Huynen MA: Genome phylogeny based on gene content. Nat Genet 1999, 21:108-110
9. Fitz-Gibbon ST, House CH: Whole genome-based phylogenetic analysis of free-living microorganisms. Nucleic Acids Res 1999, 27:4218-4222
10. Tekaia F, Dujon B: Pervasiveness of gene conservation and persistence of duplicates in cellular genomes. J Mol Evol 1999, 49:591-600
11. Grishin NV, Wolf YI, Koonin EV: From complete genomes to measures of substitution rate variability within and between proteins. Genome Res 2000, 10:991-1000
12. Dobzhansky T, Sturtevant AH: Inversions in the chromosomes of Drosophila pseudoobscura. Genetics 1938, 23:38-64
13. Hannewall S, Chappey C, Koonin EV, Pezner PA: Genome sequence comparison and scenarios for gene rearrangements: a test case. Genomics 1995, 30:299-311
14. Sankoff D, Blanchette M: Phylogenetic invariants for genome rearrangements. J Comput Biol 1999, 6:431-445
15. Mushegian AR, Koonin EV: Gene order is not conserved in bacterial evolution. Trends Genet 1996, 12:289-290
16. Dandekar T, Snel B, Huynen M, Bork P: Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem Sci 1998, 23:324-328
17. Huynen Mj, Snel B: Gene and context: integrative approaches to genome analysis. Adv Prot Chem 2000, 54:345-379
18. Tatusov RL, Mushegian AR, Bork P, Brown NP, Hayes WS, Bordovskyl M, Rudd KE, Koonin EV: Metabolism and evolution of Haemophilus influenzae deduced from a whole-genome comparison with Escherichia coli. Curr Biol 1996, 6:279-291
19. Overbeek R, Fonstein M, D’Souza M, Pusl GD, Maltsev N: The use of gene clusters to infer functional coupling. Proc Natl Acad Sci USA 1999, 96:2896-2901
20. Fujibuchi W, Ogata H, Matsuda H, Kanehisa M: Automatic detection of conserved gene clusters in multiple genomes by graph comparison and P-quasi grouping. Nucleic Acids Res 2000, 28:4029-4036
21. Wolf YI, Rogozin IB, Kondrashov AS, Koonin EV: Genome alignment, evolution of prokaryotic genome organization and prediction of gene function using genomic context. Genome Res 2001
22. Koonin EV, Mushegian AR, Galperin MY, Walker DR: Comparison of archaeal and bacterial genomes: computer analysis of pro-
tein sequences predicts novel functions and suggests a chi-
meric origin for the archaea. Mol Microbiol 1997, 25:619-637
23. Aravind L, Tatusov RL, Wolf YI, Walker DR, Koonin EV: Evidence
for massive gene exchange between archael and bacterial
hyperthermophiles. Trends Genet 1998, 14:442-444
24. Nelson KE, Clayton RA, Gill SR, Gwinn ML, Dodson RJ, Haft DH,
Hickey KE, Peterson JD, Nelson WC, Ketchum KA, McDonald L, Ut-
terback TR, Doolittle FR, Linher KD, Garrett MM, Stewart AM, Cotton
MD, Pratt MS, Phillips CA, Richardson D, Heidelberg J, Sutton GG,
Fleischmann RD, Eisen JA, Fraser CM, et al: Evidence for lateral
gene transfer between Archaea and bacteria from genome
sequence of Thermotoga maritima. Nature 1999, 399:323-329
25. Martin W: Mosaic bacterial chromosomes: a challenge en
route to a tree of genomes. Bioessays 1999, 21:99-104
26. Pace NR: A molecular view of microbial diversity and the bio-
sphere. Science 1997, 276:734-740
27. Ponting CP, Aravind L, Schultz J, Bork P, Koonin EV: Eukaryotic sig-
nal and domain homologues in archaea and bacteria. Ancient
ancestry and horizontal gene transfer. J Mol Biol 1999, 289:729-745
28. Brochier C, Philippe H, Moreira D: The evolutionary history of ri-
bosomal protein RpSI4: horizontal gene transfer at the heart of the ribosome. Trends Genet 2000, 16:529-533
29. Ludwig W, Strunk O, Klugbauer S, Klugbauer N, Weizenegger M,
Neumaier J, Bachleitner M, Schleifer KH: Bacterial phylogeny
based on comparative sequence analysis. Electrophoresis 1998,
19:554-568
30. Gruber TM, Bryant DA: Molecular systematic studies of eubact-
eria, using sigma70-type sigma factors of group 1 and group
J Bacterial 1997, 179:1734-1747
31. Baldauf SL, Palmer JD, Doolittle WF: The root of the universal
tree and the origin of eukaryotes based on elongation factor
phylogeny. Proc Natl Acad Sci USA 1996, 93:7749-7754
32. Hansmann S, Martin W: Phylogeny of 33 ribosomal and six oth-
er proteins encoded in an ancient gene cluster that is con-
served across prokaryotic genomes: influence of excluding
poorly alignable sites from analysis. Int J Syst Evol Microbiol 2000,
50 Pt 4:1655-1663
33. Brown JR, Douady CJ, Italia MJ, Marshall WE, Stanhope MJ: Universal
trees based on large combined protein sequence data sets.
Nat Genet 2001, 28:281-285
34. Tatusova TA, Karsh-Mitsrachi I, Ostell JA: Complete genomes in
WWW Entrez: data representation and analysis. Bioformat-
ics 1999, 15:536-543
35. Tatusov RL, Koonin EV, Lipman DJ: A genomic perspective on
protein families. Science 1997, 278:631-637
36. Tatusov RL, Galperin MY, Natale DA, Koonin EV: The COG data-
bases: a tool for genome-scale analysis of protein functions
and evolution. Nucleic Acids Res 2000, 28:33-36
37. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lip-
man DJ: Gapped BLAST and PSI-BLAST: a new generation of protein
database search programs. Nucleic Acids Res 1997, 25:3389-3402
38. Farris JS: Phylogenetic analysis under Dollo’s Law. Syst 1977,
26:77-88
39. Swoford DL: PAUP: phylogenetic analysis using parsimony (and other
methods). Sunderland, MD: Sinauer; 1998
40. Saitou N, Nei M: The neighbor-joining method: a new method for
reconstructing phylogenetic trees. Mol Biol Evol 1987, 4:406-425
41. Grishin NV: Estimation of the number of amin acid substitu-
tions per site when the substitution rate varies among the sites.
J. Mol. Evol. 1995, 41:675-679
42. Grishin NV: Estimation of evolutionary distances from protein
spatial structures. J Mol Evol 1997, 45:359-369
43. Feng DF, Doolittle RF: Converting amino acid alignment scores
into measures of evolutionary time: a simulation study of
various relationships. J Mol Evol 1997, 44:361-370
44. Felsenstein J: Inferring phylogenies from protein sequences by
parsimony, distance, and likelihood methods. Methods Enzymol
1996, 266:418-427
45. Notredame C, Higgins DG, Heringa J: T-Coffee: A novel method
for fast and accurate multiple sequence alignment. J Mol Biol
2000, 302:205-217
46. Fitch WM, Margoliash E: Construction of phylogenetic trees. Sci-
ence 1967, 155:279-284
47. Jones DT, Taylor WR, Thornton JM: The rapid generation of mu-
tation data matrices from protein sequences. Comput Appl Bio-
saci 1992, 8:275-282
48. Adachi J, Hasegawa M: MOLPHY: Programs for Molecular Phylogenetics.
Tokyo: Institute of Statistical Mathematics; 1992
49. Kishino H, Miyata T, Hasegawa M: Maximum likelihood inference
of protein phylogeny and the origin of chloroplasts. J. Mol. Evol
1990, 31:151-160
50. Robinson DF, Foulds LR: Comparison of phylogenetic trees.
Math. Biosci 1981, 53:131-147
51. Sneath PHA, Sokal RR: Numerical Taxonomy. San Francisco: W. H. Free-
man; 1973