Syst. Biol. 55(3):441–453, 2006
Copyright © Society of Systematic Biologists
ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150600697416

Reciprocal Illumination in the Gene Content Tree of Life

E. KURT LIENAU,[1,2] ROB DESALLE,[1] JEFFREY A. ROSENFELD,[1,2] AND PAUL J. PLANET[1,3]

[1] *American Museum of Natural History, Molecular Laboratories, Central Park West at 79th Street, New York, New York 10024, USA;*
E-mail: planet@amnh.org (P.J.P.)
[2] *New York University, Department of Biology, Graduate School of Arts and Science, 100 Washington Square East, New York, New York 10003, USA*
[3] *Department of Pediatrics, Children's Hospital of New York, Columbia University, College of Physicians and Surgeons, New York, New York 10032, USA*

Abstract.—Phylogenies based on gene content rely on statements of primary homology to characterize gene presence or absence. These statements (hypotheses) are usually determined by techniques based on threshold similarity or distance measurements between genes. This fundamental but problematic step can be examined by evaluating each homology hypothesis by the extent to which it is corroborated by the rest of the data. Here we test the effects of varying the stringency for making primary homology statements using a range of similarity (e-value) cutoffs in 166 fully sequenced and annotated genomes spanning the tree of life. By evaluating each resulting data set with tree-based measurements of character consistency and information content, we find a set of homology statements that optimizes overall corroborration. The resulting data set produces well-resolved and well-supported trees of life and greatly ameliorates previously noted inconsistencies such as the misclassification of small genomes. The method presented here, which can be used to test any technique for recognizing primary homology, provides an objective framework for evaluating phylogenetic hypotheses and data sets for the tree of life. It also can serve as a technique for identifying well-corroborated sets of homologous genes for functional genomic applications. [CCM; consistency; corroboration; gene content; homology; ILD; phylogeny; presence absence; tree of life.]

Phylogenies based on gene content are now feasible with the availability of multiple, fully sequenced genomes. As in traditional systematic analyses, these phylogenies use features of organisms, in this case, the presence or absence of genes, as character data for reconstructing the branching pattern of evolution (Bansal and Meyer, 2002; Dutilh et al., 2004; Fitz-Gibbon and House, 1999; Gophna et al., 2005; Gu and Zhang, 2004; House and Fitz-Gibbon, 2002; Lake and Rivera, 2004; Snel et al., 1999; Tekaia et al., 1999; Wolf et al., 2001a, 2002). Similar recent approaches have used the presence or absence of protein domains or folds (Lin and Gerstein, 2000; Yang et al., 2005) and gene order/proximity (Wolf et al., 2001b) as data. Because of the remarkable phylogenetic breadth of available genomic data, many studies have focused on the reconstruction of the tree of life (Bansal and Meyer, 2002; Dutilh et al., 2004; Fitz-Gibbon and House, 1999; Gophna et al., 2005; Gu and Zhang, 2004; House and Fitz-Gibbon, 2002; Lake and Rivera, 2004; Snel et al., 1999; Tekaia et al., 1999; Wolf et al., 2001a, 2002). Recently, gene content data have also been used to address the origin of eukaryotes (Lake and Rivera, 2004), and there is emerging interest in using this type of data to infer functional interactions, pathways, and networks (Bowers et al., 2004; Marcotte et al., 1999; Overbeek et al., 1999; Pellegrini et al., 1999).

Despite their wide acceptance, most gene content phylogenies show significant inconsistencies with otherwise well-supported taxonomy (Clarke et al., 2002; Dutilh et al., 2004; Lake and Rivera, 2004; Yang et al., 2005). These discordances have been attributed to unspecified phylogenetic noise, nonvertical gene events (e.g., horizontal transfer) (Clarke et al., 2002; Dutilh et al., 2004; Gophna et al., 2005), use of inappropriate optimality criteria for tree construction (Dutilh et al., 2004; Gu and Zhang, 2004; Huson and Steel, 2004; Lake and Rivera, 2004), and biases introduced by comparing genomes of grossly different sizes (Bansal and Meyer, 2002; Lake and Rivera, 2004; Yang et al., 2005). Solutions offered to these problems have included elimination or downweighting of inconsistent characters (Clarke et al., 2002; Dutilh et al., 2004; Gophna et al., 2005), weighted corrections for smaller genomes (Bansal and Meyer, 2002; Yang et al., 2005), accounting for multiple gene copies in a genome (Gu and Zhang, 2004), and limiting gene presence/absence groupings to those found in a specific genome (i.e., *conditioning*) (Lake and Rivera, 2004).

A largely unexplored source of phylogenetic bias and inconsistency is the step in which the initial statements of gene content are specified. Preliminary hypotheses of homology are often made using techniques based on a similarity threshold, set as an a priori criterion, for defining whether a gene can be reasonably considered to be absent or present. Any such technique risks lumping genes together that are not true homologues and risks separating genes that are valid homologues. Such inaccurate homology statements can have large effects on tree topology. A recent study by Hughes et al. (2005) that examined six data sets determined at different similarity thresholds for 99 prokaryotic genomes found that the data sets produced contradictory phylogenetic trees that differed both in topology and in their ability to resolve ancient relationships. In the absence of an obvious criterion for choosing one data set over another, Hughes et al. (2005) concluded that their results "did not increase confidence in the applicability of gene content analyses to the resolution of prokaryotic phylogenies."

To limit erroneous homology statements, other researchers have used database search operations that accept only mutually strong BLAST scores between genes (Gophna et al., 2005; Lake and Rivera, 2004) or curated gene families based on these scores (e.g., COGs [Clusters of Orthologous Groups]; Tatusov et al., 1997). Others have developed approaches that add more sophisticated criteria to similarity searches and gene clustering techniques (Clarke et al., 2002; Gophna et al., 2005; Harlow et al., 2004; Krause et al., 2005; Park and Teichmann, 1998). However, there is no clear empirical framework or objective measure for evaluating differences between techniques or choosing between data sets.

To further explore the effects of unexamined a priori homology statements, we devised a method, based on the principle of reciprocal illumination, for evaluating primary homology statements that combines character-based and topological tests for measuring corroboration of each homology hypothesis. We used this method to test 111 gene content data sets for 166 eukaryotic and prokaryotic taxa, each constructed at different similarity thresholds for defining homologue groups. From this procedure, we derived highly corroborated data sets that produce well-resolved phylogenies that greatly ameliorate previously noted inconsistencies in gene content studies. We also compared these to data sets created with the widely used COG database (Dutilh et al., 2004), and a conditioned genome approach.

Reciprocal Illumination and Homology Testing

All phylogenetic analyses begin by making primary homology statements that tentatively describe the evolutionary relationships among attributes. These initial hypotheses can be tested and reformulated as secondary homology hypotheses by measuring the extent to which they are corroborated by other homology hypotheses on the resulting phylogenetic tree (DePinna, 1991). Assessing homology corroboration within this framework relies on the principle of reciprocal illumination (Hennig, 1966), in which each individual hypothesis is evaluated by the extent to which it agrees with the overall, favored hypothesis given all available data. This operation entails a test of each homology hypothesis (character) using, as evidence, the relationships or nodes inferred from the optimal tree. For any given data set, reciprocal illumination allows for the most severe test (sensu Popper, 1968) of each homology hypothesis based on its congruence with the total assembled information in the rest of the data set. Using reciprocal illumination to evaluate homology statements has received much attention for morphological and molecular sequence data (Brower, 1996; DePinna, 1991; Kluge, 2003; Rieppel and Kearney, 2002; Wheeler, 2001), but gene content data have not been analyzed from this perspective.

We suggest that in choosing among data sets, one should prefer those that are composed of the most severely tested, well-corroborated homology hypotheses. Statements of gene presence or absence can be tested by their agreement with the overall phylogenetic patterns suggested by all other gene presence/absence statements in the data set using measurements such as the consistency index (CI; Kluge, 1969). However, it is important to note that corroboration of a hypothesis is determined not only by consistency (i.e., the extent to which the data do not disagree) but also by information content (i.e., the amount of evidence that could possibly refute a hypothesis). The extent to which a hypothesis is refutable is determined by the boldness, or resolution of that hypothesis, and the amount of relevant information available to test it.

Methods

Data Set Construction

We used a single-linkage clustering (SLC) algorithm to create 111 data sets at differing e-value thresholds. This algorithm grouped sequences based on measurements derived from an all-against-all amino acid similarity search of 166 sequenced and annotated genomes from UniProt (Apweiler et al., 2004) using BLAT (Bansal and Meyer, 2002). The SLC algorithm groups protein sequences together if they have at least one pairwise e-value score with any other member of the group that was as good or better than each specified e-value threshold. Therefore, all proteins that are being grouped together only need one connection to any member of the group (see Fig. 1). Based on each SLC group, we constructed presence/absence matrices in which each column represents an SLC grouping and each row, a taxon. The presence in each taxon of at least one representative gene in each SLC group was coded as a "1." Absence was coded as a "0." All SLC groups with only one gene member were excluded from the analysis. A distinct data set was created for each e-value threshold.

We also used a technique based on the conditioning method introduced by Rivera and Lake (2004) to construct data sets based on mutually strong hits. In our implementation of this technique, each protein sequence from a single genome (referred to as the conditioning genome) is used to search each genome in the data set. The top three protein hits (based on e-value) from each genome are then retrieved and used, in turn, to search the conditioning genome. If any of the three protein sequences retrieves the original protein sequence as one of its own top three hits in the conditioning genome, then the genome receives a "1." This process is done for each gene in the conditioning genome against every other genome in the data set. We used three genomes, *Arabidopsis thaliana*, *Escherichia coli*, and *Archeoglobus fulgidus*, as conditioning genomes. All data sets, databases, and computer scripts are freely available from the authors upon request.

We constructed elided SLC gene-content data sets by concatenating individual matrices into a single matrix (Wheeler et al., 1995). An SSU rDNA data set was generated to compare the results of the gene content data sets with an external data set. SSU rDNA sequences were attained from the Ribosomal Database Project-II (RDP-II) (http://rdp.cme.msu.edu) for each of the species used in this analysis and aligned using default alignment settings in CLUSTALX 1.63.

Phylogenetic Analysis

For all data sets, we did heuristic searches for the most parsimonious trees using the 'ratchet' method implemented using PAUPRat (Sikes and Lewis, 2001) in conjunction with PAUP4.0b10 (Swofford, 2000). We did three

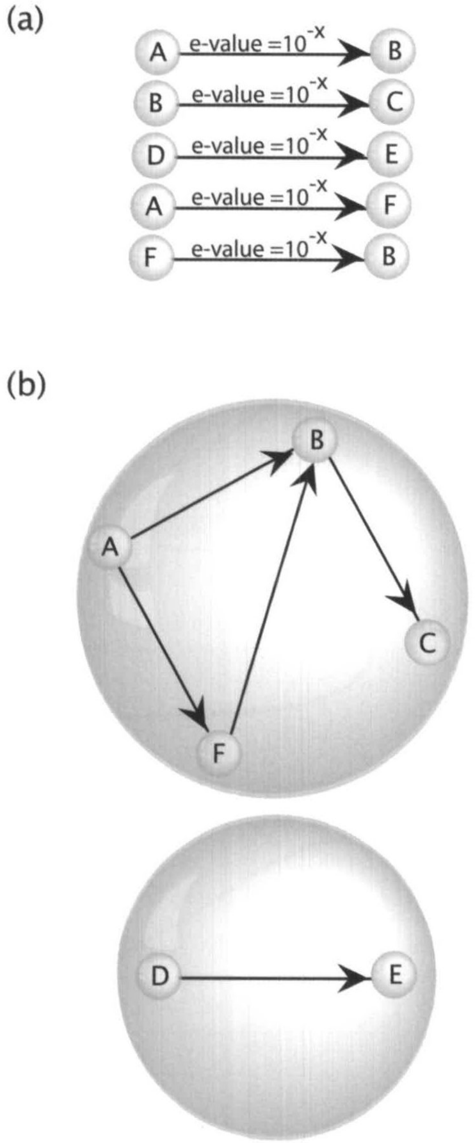


FIGURE 1. Single linkage clustering (SLC) algorithm. The algorithm combines objects (in this case, proteins) when they have at least one linkage to another member of the groups. Panel (a) shows six proteins A, B, C, D, E, and F have linkages based on measurements of similarity that are better than some threshold (here labeled 10^{-x}). These linkages are A and B, B and C, D and E, and A and F. The single linkage-clustering algorithm groups the proteins in two nonoverlapping clusters (b).

sets of 200 replicates of the ratchet method, up weighting 15%, 17%, and 21% of the characters in each set, using the tree-branch reconnection (TBR) technique and saving only one tree at each step. Using the resulting trees as starting trees, we then did 100 TBR replicates saving multiple trees (up to 1000) at each step (multrees option in PAUP). All characters and state transformations were given equal weight. To calculate confidence in the resulting trees, we generated Bremer decay indices using the program Autodecay (Eriksson, 1998) and 100 bootstrap replicates in PAUP with 100 iterations of random addition followed by TBR. For distance-based tree searches for all data sets, we used the default settings for neighbor joining in PAUP. For Bayesian phylogenetic analysis of the SSU rDNA, data we used the Metropolis-coupled Markov-chain Monte Carlo (MCMCMC) with four chains (heat 0.5) over 500,000 iterations, with a burn-in of 100 iterations, in the computer program MrBayes (Huelsenbeck and Ronquist, 2001). Default settings (4-by-4 model nucleotide substitution model, equal substitution rates, assuming equal among-site rate variation) produced trees equivalent in topology to more complex models. For LogDet/paralinear we used default settings in PAUP after converting our presence/absence matrix to contain "A" and "T" in place of "1" and "0," respectively.

Consistency and Corroboration Analysis

We computed the rescaled consistency index (RCI) (Farris, 1989), consensus fork index (Colless, 1980) and the Rohlf consensus index 1 (Rohlf, 1982) using PAUP. We computed the RCI for the optimal trees derived from each data set. The RCI is defined as RCI = $(M/S) \times [(G-S)/(G-M)]$ (where M is the minimum number of steps on the tree, S is the observed number of steps, and G is the maximum number of steps possible on the particular tree). The consensus fork index (CFI) and Rohlf CI1 were calculated using the strict consensus tree. The CFI is found by dividing the number of bifurcating nodes on the consensus tree by the maximum number of possible nodes. The Rohlf CI1 is a related metric that gives greater weight to polytomies with a greater number of dependent branches (i.e., nodes closer to the root) while also correcting for biases introduced by tree topology (Rohlf, 1982). We computed the combined corroboration metric (CCM) for each data set by multiplying the RCI by each of the topological consensus indices: $\mathrm{CCM_R} = \mathrm{RCI} \times \mathrm{RohlfCI1}$ and $\mathrm{CCM_{CF}} = \mathrm{RCI} \times \mathrm{CFI}$.

For comparison of gene content and SSU rDNA trees using the ILD test, we used the partition homogeneity function in PAUP to carry out 100 replicates of partition randomization, each with 100 replicates of random addition followed by TBR.

RESULTS AND DISCUSSION

Data Set Construction, Composition, and Phylogenetic Analysis

To construct data sets for homology assessment we used a single linkage clustering (SLC) algorithm in which a gene is included in a group if it has a similarity score above a given threshold with at least one other gene from the group (Fig. 1). A recent study by Hughes et al. (2005) also used an SLC clustering algorithm to construct presence/absence data sets at different similarity thresholds based on pairwise local (BLAST) alignments of all sequences in the analysis. These thresholds were defined by percent sequence identity and length of the BLAST alignment. For a similarity score threshold we used the

BLAST-based e-value (or expect value) that is commonly employed for nucleotide and protein sequence database searches (Altschul et al., 1997). The e-value represents the probability due to chance that a better sequence match exists in the database. It is therefore an expression of the probability of finding a given similarity score by chance alone. The e-value calculation incorporates the similarity and length of the matching sequence with the size of query sequence and the size of the database. We created 111 gene content data sets for 166 organisms, each at a different e-value threshold ranging from e-value 10^{-5} to an e-value of 10^{-300}.

Figure 2 shows the large amount of variation in the composition of these data sets. The number of proteins in each data set decreases with increasing stringency. This is due to the exclusion of "singleton" genes that cannot be assigned to any SLC group at a given threshold. The total number of characters (or SLC groups) in each data set initially increases with increasing similarity stringency as groups are broken up into smaller groups, but the number then decreases as SLC groups are divided into groups of singletons and excluded from the analysis. This trend is mirrored by a rise and fall in the number of steps in the most parsimonious trees with increasing similarity stringency. Like the total number of proteins, the maximum cluster size decreases nearly linearly, but the average cluster size decreases more rapidly and nears an asymptote of approximately four to five proteins per SLC group starting at e-values of approximately 10^{-50}. This indicates that a relatively small proportion of protein clusters contain many more members than the rest of the clusters at mid-range similarity threshold stringencies (approximately 10^{-50} to 10^{-100}). It seems likely that this small portion of proteins contains most of the information about deeper phylogenetic relationships.

For comparison to our SLC data sets, we constructed three data sets using a mutual best hits technique that incorporated the idea of conditioning genomes described by Rivera and Lake (2004). Each of these three data sets was "conditioned" by restricting gene content to genes found in a single organism. We chose one organism from each domain to condition the data (*Archeoglobus fulgidus*, *Escherichia coli*, and *Arabidopsis thaliana*). Because a previous study noted differences in outcome when genomes of different sizes are used (Lake and Rivera, 2004), we also chose these genomes to represent a spectrum of genome size: *Archeoglobus fulgidus* (2420 protein encoding genes), *Escherichia coli* (4237 protein encoding genes), and *Arabidopsis thaliana* (25,498 protein encoding genes).

In addition, we obtained a previously analyzed data set derived from the COG database (Dutilh et al., 2004). COGs, or clusters of orthologous groups, are determined by a procedure in which all sets (or cliques) of three genes drawn from any genome in the database that are all similar to one another are identified. Then any cliques that share at least two genes in common are merged to make the final clusters (Tatusov et al., 1997). Current implementations for finding COGs add processes that remove paralogous genes from the same COG. In addition, the

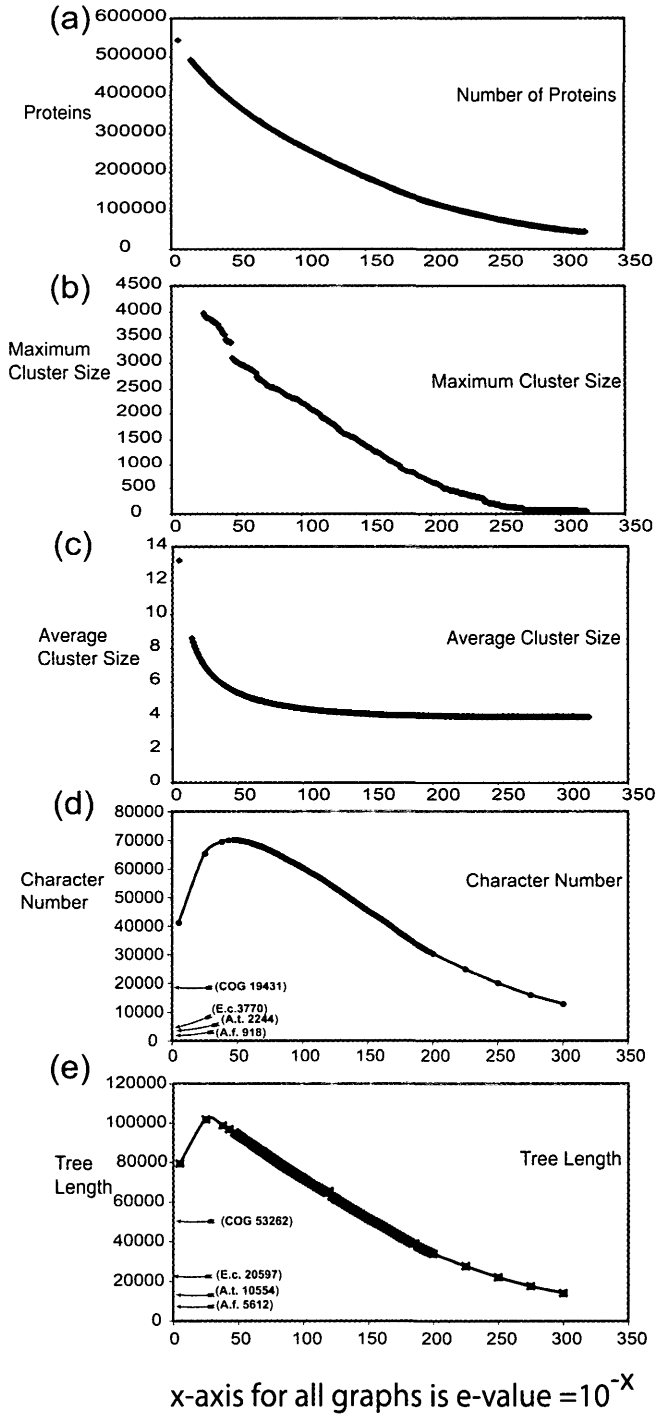


FIGURE 2. SLC data set calulations at differing simliarity threshold stringencies. Graphs are as follows: (a) the total number of proteins used to make each data set; (b) the size (number of proteins) of the largerst gene cluster in each data set; (c) the average size (number of proteins) of a gene cluster in each data set; (d) the number of characters in each analysis; (e) the number of steps in the optimal tree(s) found for each data set. Graph (b) starts at an e-value of 10^{-25} rather than 10^{-5} because of scaling (the value at 10^{-5} (193,760 proteins) for this graph were much higher than the range of the rest of the data). The total number of proteins for each data set (a) and the maximum number of proteins in a single cluster (b) diminish regularly with increased similarity threshold stringency, whereas the average number of proteins in a cluster (c) nears a minimum number at a threshold of around 10^{-100}.

COG databases are manually corrected, inspected, and curated (Tatusov et al., 2001).

We aggressively searched for the optimal tree(s) for each data set. Notably, for almost all of the data sets analyzed here, the ratchet tree search technique of Nixon (1999) found more parsimonious trees than were found using less strategic searches available in PAUP, such as multiple rounds of random addition followed by TBR branch swapping.

Internal Consistency

Consistency for each character (homology hypothesis) in a phylogenetic data set often is assessed using the consistency index (CI), which is the ratio of the minimum possible number of changes in a given character to the observed number of changes on the tree. To correct this value for artificial inflation when there are only a few possible steps for a character, Farris introduced the rescaled consistency index (RCI) (Farris, 1989). Ensemble consistency indices, the average CI (or RCI) of all characters in the data set (Kluge, 1969), represent the overall degree to which homology statements within the data set do not disagree.

Using the optimal tree (s) as a guide we calculated the ensemble RCI for each SLC data set. As expected, the ensemble RCI increased as the similarity threshold became more stringent (Fig 3a), showing that higher levels of overall internal consistency are obtained at higher similarity stringencies.

In contrast, the RCI for the data set derived from the COG database was remarkably low, which is consistent with previous observations of a high level of noise in this data set (Dutilh et al., 2004). RCI scores for all three data sets made using our conditioned genome technique were even lower, suggesting that these data sets also have high levels of discordant phylogenetic signal (Fig. 3a). This observation may be due to the fact that our conditioning technique does not explicitly exclude potentially paralogous genes from these data sets.

It is unclear why the SLC data sets had such dramatically higher consistency index scores even at low similarity threshold levels. One possible explanation is that the SLC technique requires only one point of similarity for each gene to be assigned to a group. Genes, therefore, never have any similarity above the threshold with any gene outside the group to which they are assigned. Thus, SLC guarantees that there will be no ambiguity in the placement of genes in a gene homology hypothesis, and each gene will be represented only once in a given matrix. Duplicating the same gene in different homology hypotheses could lead to an artificial increase in homoplasy, and therefore a decrease in consistency.

External Congruence

Visual inspection of trees produced from data sets with lower stringency similarity thresholds showed significant deviations from widely accepted taxonomic relationships, whereas trees from data sets with mid-level similarity stringency thresholds (e-values from 10^{-50} to 10^{-100}) united most members of well-accepted taxonomic groups (Fig. 4). Trees produced from even higher levels of similarity stringency produced poorly resolved consensus trees that give little taxonomic information. To further explore these *gestalt* taxonomic observations, we compared our data sets to independently derived phylogenetic data from the gene for small subunit ribosomal DNA (SSU rDNA).

We compared the topology of each of the most parsimonious SLC gene content trees to trees generated with the SSU rDNA data set using tree consensus indices as measurements of tree similarity (Fig. 5). Tree consensus indices measure agreement of trees as a function of the number of fully resolved nodes in the consensus tree obtained from all trees examined. Thus, tree consensus indices are convenient measures of the topological agreement among multiple trees, and can accommodate more than one optimal tree for comparisons.

Topological tree comparison allowed us to compare gene content trees to SSU rDNA trees constructed using different optimality criteria (Bayesian likelihood and parsimony). These comparisons offered a partial control for the possibility that our analysis was being confounded by misleading biases that may affect different phylogenetic inference techniques (Felsenstein, 1978; Kolaczkowski and Thornton, 2004). Although all gene content data sets had low topological agreement with both the Bayesian- and parsimony-derived SSU rDNA trees, a clear trend emerged with middle-range gene content data sets (e-values from 10^{-50} to 10^{-100}) yielding trees that conflicted least with the SSU rDNA trees (Fig. 5a, b). The decline in consensus indices for trees constructed at higher similarity stringencies is largely due to the poor resolution of trees from gene content data sets themselves (Fig. 3) and is, therefore, not only reflective of disagreement with the SSU rDNA derived topology.

To further control for possible methodological biases in our tree comparisons, we compared our parsimony trees to distance analyses. Several recent studies have suggested that parsimony analysis of gene content data sets is more prone to gross taxonomic/classification errors compared to distance-based techniques such as LogDet (Dutilh et al., 2004; Lake and Rivera, 2004; Rivera and Lake, 2004; but see also Fitz-Gibbon and House, 1999; House and Fitz-Gibbon, 2002; Huson and Steel, 2004). It is important to note that these errors have not been attributed to long-branch attraction (LBA). They are instead likened to the effects of nucleotide compositional bias in DNA sequences (Lake and Rivera, 2004). To test if the taxonomic inconsistencies we observed were due simply to the use of parsimony-based analysis, we compared gene content and SSU rDNA tree topologies derived from distance (neighbor joining, LogDet) and parsimony techniques. We found that trees from the SSU rDNA data set were almost always more similar, based on topological tree comparison indices (data not shown), to the most parsimonious gene content trees than they were to gene content trees made using distance methods. This finding held true regardless of the optimality

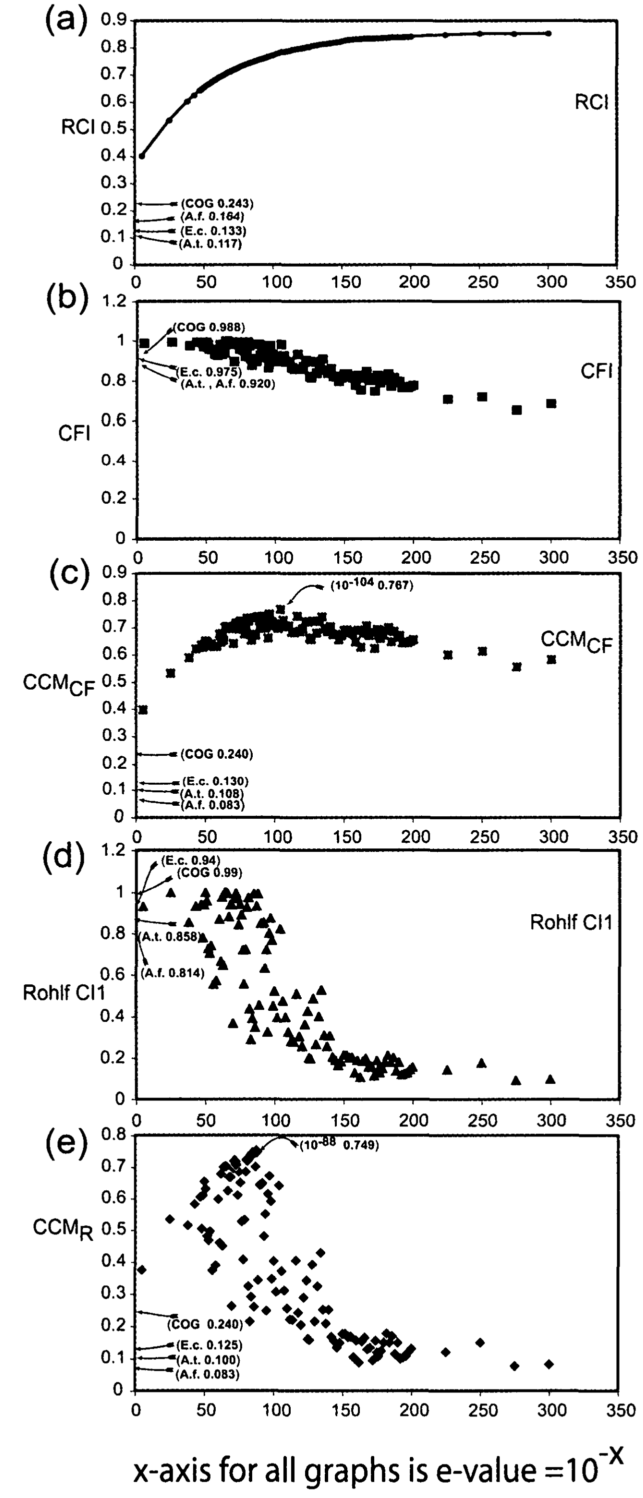


FIGURE 3. Character consistency, topological consistency, and combined indices at differing similarity threshold stringencies maximal values, if present, are indicated with a curved arrow with the corresponding e-value in parentheses. Values for the COG data set (Dutilh et al., 2004) and the three conditioned data sets inspired by Rivera and Lake (2004) are indicated with a straight arrow. Conditioned data sets are designated with initials of genus and species of the conditioning genome (A.f. = *Archeoglobus fulgidus*; A.t. = *Arabidopsis thaliana*; E.c. = *Escherichia coli*). The scores for the COG data set are consistent with previous findings that showed very high levels of noise (homoplasy) in this data set (Dutilh et al., 2004). The scores for the conditioned data sets showed high levels of homoplasy as well. Graphs are as follows: (a) the rescaled consistency index (RCI); (b) the consensus fork index (CFI); (c) the combined corroboration metric (CCM) that uses the CFI as the topological component (CCM_{CF}); (d) the Rohlf consistency index 1 for consensus trees (Rohlf CI1); (e) the CCM that uses the Rohlf CI1 as the topological component (CCM_R). The RCI increases with increasing stringency. Both the Rohlf CI1 and the CFI indices show a trend toward lower values at higher stringencies. Note that the Rohlf CI1 and CCM_R vary significantly over consecutive e-values, suggesting that small differences between e-values can have drastic effects on these indices. CCM_R scores of data sets used in the elided matrix (top 5% CCM_R) are shown in red in panel (e).

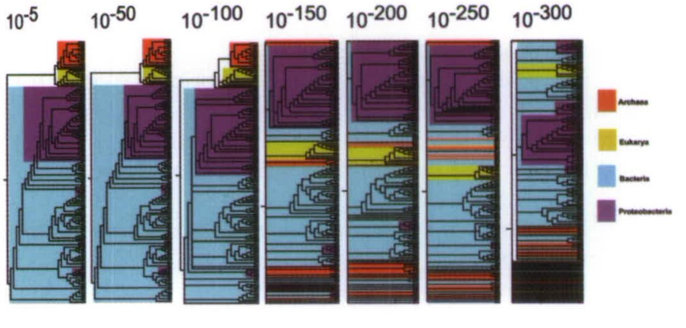


FIGURE 4. Tree topologies from a range of e-value stringencies. Strict consensus trees derived from low to high similarity threshold are shown from left to right. Taxa belonging to the Archaea are colored orange; Eukarya, yellow; Bacteria, blue. Taxa belonging to the proteobacteria are purple to illustrate taxonomic differences between trees at the level of phylum. Note that as stringency increases, trees become increasingly unresolved, and at lower stringencies many taxa are misclassified.

criterion used for the SSU rDNA tree or the e-value of the gene content data set. Thus, for these data sets and to the extent that the SSU rDNA analysis is a valid representation of bacterial phylogeny, parsimony analysis seems to be no more prone to taxonomic errors than distance-based analyses.

Phylogenetic character corroboration by external data sources can be assessed using the incongruence length difference (ILD) test, which measures the extent and statistical significance of character disagreement between separate or partitioned data sets (Farris, 1994). To test our data sets in this framework, we concatenated our gene content data sets with the SSU rDNA sequences from each species. We then used the ILD test to measure the incongruence between gene content data sets and the SSU rDNA alignment. In every case tested, the SSU rDNA data set was statistically incongruent with the gene content data sets ($P < 0.01$), but, mirroring the internal character consistency indices above, the ILD showed less absolute and relative incongruence at lower e-values (Fig. 5c, d).

It is important to note that we do not consider the SSU rDNA data set, or indeed any single gene data set, alone to be sufficient for inferring the phylogeny of the tree of life. However, data sets composed of SSU rDNA are widely regarded as containing reliable phylogenetic information for at least some relationships in the tree of life, and many of the major bacterial groupings are

←

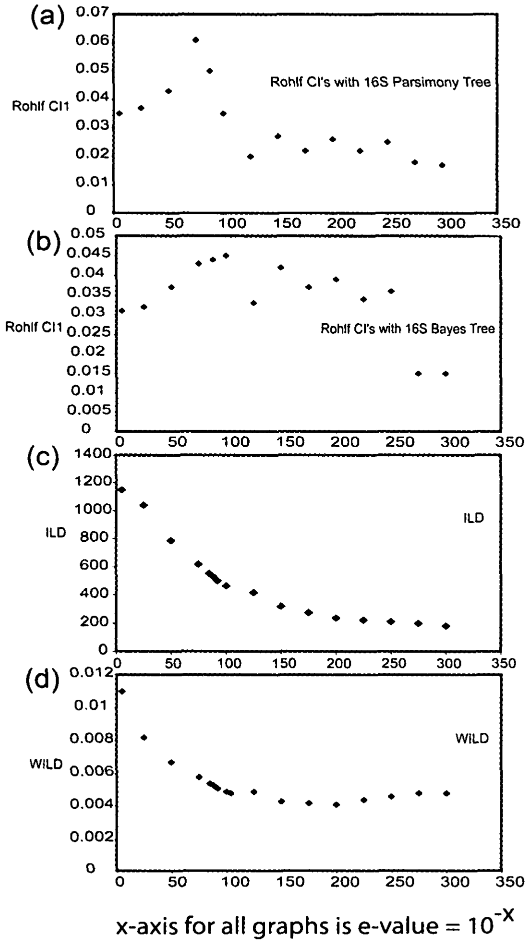


FIGURE 5. Calculations of external corroboration of SLC gene content data sets. Panels a and b show scores of Rohlf consensus index (y-axis) plotted against e-value thresholds (x-axis) used for construction of SLC gene content data sets. Each Rolf consensus index is derived from the consensus tree of the optimal tree(s) from the SLC gene content data set combined with the optimal topology for the SSU rDNA data set. The results are shown for comparison to SSU rDNA trees using unweighted parsimony (panel a) and Bayesian likelihood analysis (panel b). Panels (c) and (d) show measurements of.data set incongruence between the SSU rDNA data set and the SLC gene content data sets (y-axis) plotted against the e-value thresholds (x-axis) used for construction of SLC gene content data sets. Panel (c) shows the raw score for the ILD, and panel (d) shows the ILD score normalized by dividing by the total number of steps in the optimal tree.

defined largely based on their SSU rDNA sequences. We therefore regard the topological and character congruence with the SSU rDNA data sets as mutually informative indications of phylogenetic relationships.

Information Content

Both the ILD and character consistency indices showed that higher stringency similarity thresholds create data sets that have less conflict both internally and with external data. However, using these measures alone in assessing corroboration of hypotheses is misleading. Several researchers have noted that the ILD test can be biased in cases where one data set (or partition) contributes more information than the other (Darlu and Lecointre, 2002; Dolphin et al., 2000; Dowton and Austin, 2002; Hipp et al., 2004; Lee, 2001; Yoder et al., 2001). It is possible that by weakening the contribution of the gene content data set by increasing the similarity threshold stringency may lead to false congruence in the ILD test.

Likewise, because consistency indices address the extent to which characters do not disagree, loss of information could yield high, but misleading, consistency indices. Hypotheses can be highly consistent when they are weakly defined or only have limited information available to test them. Information loss can be reflected in multiple, discordant, optimal trees, and the inability to resolve certain phylogenetic relationships. In reciprocal illumination, the loss of resolution is critically detrimental, because loss of nodes entails loss of evidence for testing each homology hypothesis.

Indeed, inspection of the consensus trees produced from higher stringency data sets showed a lack of resolution in many nodes, suggesting that as similarity threshold stringency and consistency increased, information about certain relationships was lost (Fig. 3). To quantify this loss of resolution, we used tree topology congruence indices, the consensus fork index (CFI) (Colless, 1980) and the rohlf consensus index 1 (Rohlf CI1) (Rohlf, 1982), which measure tree disagreement as a function of the number of uncertain relationships (polytomies) in consensus trees. These measures showed that at stringencies above approximately 10^{-100}, the consensus trees were less and less resolved (Fig. 3b, d).

Decreasing resolution at lower e-values (higher stringencies) despite their higher CI and RCI values illuminates the inability of character consistency indices to alone measure corroboration of homology hypotheses. Despite the high consistency of homology hypotheses in these data sets, the information content is low, and, therefore, the overall amount of corroboration and the severity of test is weak.

The Combined Corroboration Metric

Because of the opposing trends in measurements of phylogenetic consistency and resolution with increasing similarity stringency, we sought to maximize both character consistency (based on the RCI) and information content (based on topological consensus indices) to find the data set with the highest level of internal corroboration. To accomplish this task, we searched for the maximum value of the product of the character and topological indices (referred to here as the combined corroboration metric, or CCM). The character consistency component of the CCM measures the extent to which the data do not disagree, whereas the topological component of the CCM addresses the extent to which the data could possibly refute each homology hypothesis based on the resolution of the optimal tree. Thus, the CCM score is an aggregate measure of the overall amount of corroboration (and the severity of test) for homology hypotheses in the data set.

We used two CCMs; one incorporates the CFI (CCM_{CF}), which weights all polytomies in consensus trees equally, and the other is based on the Rohlf CI1 (CCM_R), which emphasizes polytomies in those relationships closer to the root of the tree. Although each of these measures indicated a different optimal data set (Fig. 3c, e), both of these data sets produced trees that were in remarkable agreement with accepted taxonomic relationships.

In general, data sets constructed using SLC showed much higher levels of corroboration as determined by the CCM. This may stem from the much higher levels of character consistency found in these data sets. The COG data set and the data sets constructed using the conditioning technique produced trees were as highly resolved as many of the SLC-derived trees. However, these data sets also had high levels of character conflict that were reflected in low consistency values. Therefore, the CCM scores for these data sets are significantly lower.

One interesting property of the SLC data sets produced at any e-value is that they included many more informative characters than data sets constructed using the COG gene families or the conditioned data sets, which, along with the increased consistency of these data sets, makes them especially attractive for testing phylogenetic hypotheses using as much available evidence as possible. SLC also offers a computationally tractable and straightforward method for titrating the similarity threshold in a search for optimal CCM scores.

Importantly, the CCM provides a measure for comparing future alterations and improvements to any homology recognition method, and more sophisticated homology recognition techniques may yield data sets with even higher levels of corroboration (e.g., Clarke et al., 2002; Gophna et al., 2005; Harlow et al., 2004; Krause et al., 2005; Park and Teichmann, 1998).

The Tree of Life

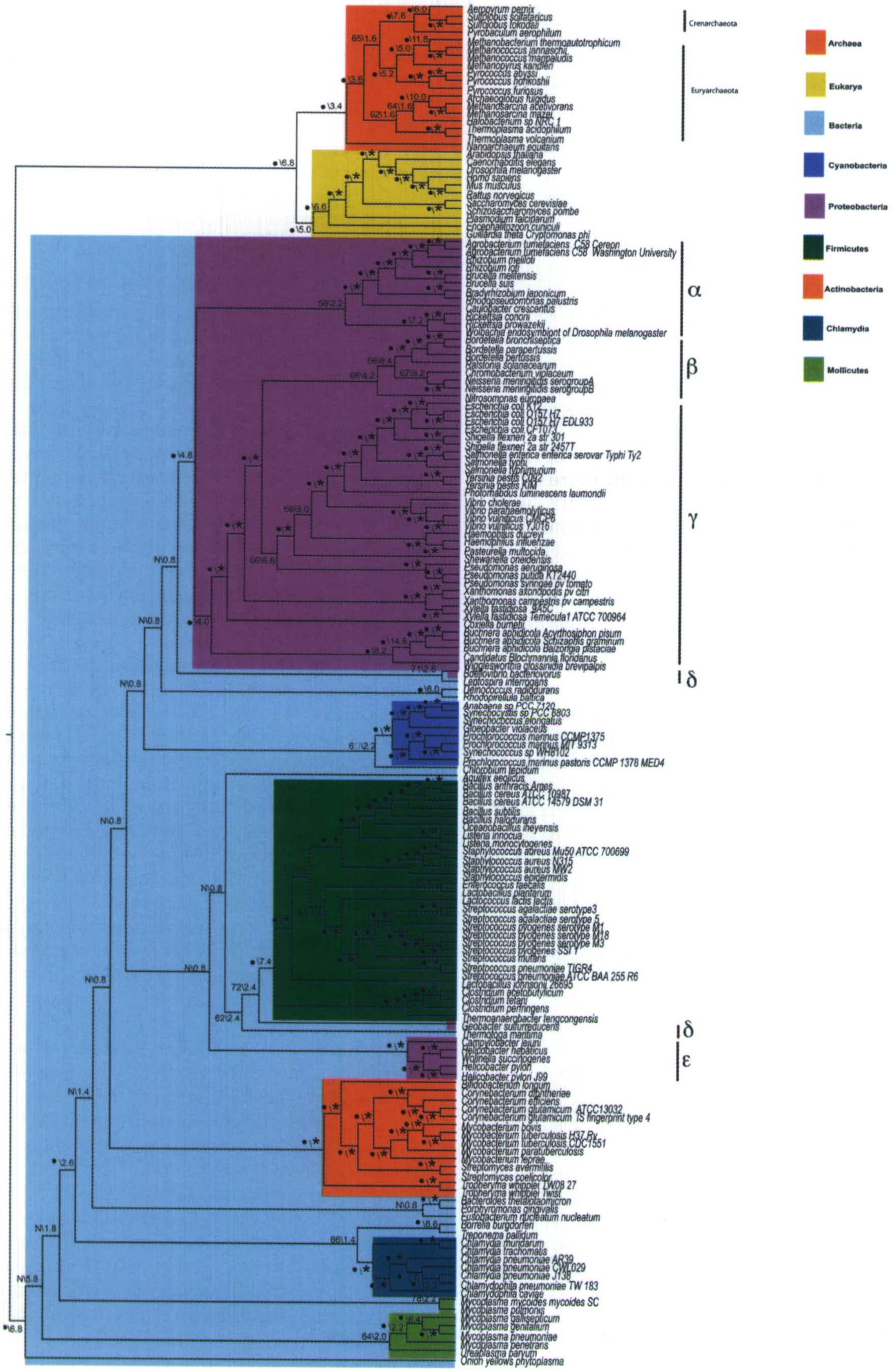
Because tree-of-life studies are mostly concerned with more ancient basal relationships, we chose the CCM_R to continue our analysis. To account for trivial differences in the highest CCM_R scores, we concatenated the data sets representing the top 5% of CCM_R scores for a combined analysis. In this technique, often referred to as elision (Wheeler et al., 1995), data sets constructed using different parameters are combined in the same analysis with the goal of decreasing the effect of homology statements that are not robust to parametric changes. Patterns of homology that persist in each data set are up-weighted by being represented more than once in the matrix, whereas patterns that only appear in one, or a few, are down-weighted. In nucleotide alignments, elision is used to ameliorate the effect of alignment ambiguous positions. Here we used this technique to emphasize gene homology hypotheses that persisted in all of our most highly corroborated data sets. Elision of several data sets may also help to account for differences in optimal similarity criteria for delineating different gene families.

The single most parsimonious phylogeny derived from the elided data set is presented in Figure 6. This tree has high bootstrap and Bremer decay support and greatly improves taxonomic inconsistencies seen in other gene content phylogenies, such as the cosegregation of the reduced genome proteobacteria (e.g., Rickettsia and Buchnera), a phenomenon referred to as "big genome attraction" (BGA) (Lake and Rivera, 2004). It also shows strong Bremer (6.8) and bootstrap (100%) support for the monophyly of Archaea, Eukarya, and Bacteria, lending further support to the three Domain organization of life (Woese et al., 1990).

Archaea.—In the optimal phylogeny presented here, *Nanoarchaeum equitans* is the most basal branch of the Archaeal clade (Fig. 5). This result is consistent with phylogenies based on SSU rDNA sequence data (Huber et al., 2002), transcription and translation machinery (Brochier et al., 2005a), and on a concatenated data set of 35 ribosomal protein sequences (Waters et al., 2003). Arguments for the ancient divergence of this lineage have also been based on the presence of genes that are "split" such that different portions of a functional gene product are found and transcribed in different parts of the genome (Randau et al., 2005; Waters et al., 2003). High gene density and a paucity of operons and pseudogenes may also argue for the antiquity of this lineage (Waters et al., 2003). This hypothesis has been challenged by studies based on protein BLAST comparisons and phylogenetic analysis of other protein encoding genes, which suggest that *Nanoarchaeum equitans* might be a highly derived Euryarchaeal lineage and represents the result of genome reduction (Brochier et al., 2005b).

The Archaeal domain is then split into two clades. The first of these clades contains members of both of the traditional major subdivisions of the Archaea, the Crenarchaeota, and the Euryarcheota. The second clade is composed exclusively of Euryarcheota. Thus, based on our tree, the Euryarcheota are a paraphyletic group. In

FIGURE 6. The most parsimonious tree from elision of data sets that had the top 5% CCM_R values (produced at e-values 1×10^{-88}, 1×10^{-85}, 1×10^{-81}, 1×10^{-73}, 1×10^{-72}). The tree has a length of 401,170 steps (CI = 0.814, RI = 0.905, RCI = 0.737) and is completely resolved (therefore CCM_R= RCI = 0.737). Support values are shown on each branch as follows: Bremer decay indices/bootstrap percentage values. A dot indicates bootstrap values greater than 80%. An asterisk indicates Bremer support values greater than 20. Some Bremer indices are shown as fractions because each value was divided by 5 to account for the concatenation of five distinct data sets in the elision. Colors of taxonomic groups are the same as in Figure 4, with additional colors indicating the Firmicutes (green), Cyanobacteria (darker blue), Actinobacteria (light green), Mycoplasmales (yellow-green), and Chlamydiales (blue-green). Proteobacterial subgroups are labeled with their corresponding Greek letter. Note the inclusion of the reduced genome Rickettsiales in the alpha protebacterial clade, and Buchnera within the proteobacteria. Also note that epsilon and some delta subgroup members are not monophyletic with other proteobacteria.


Aeropyrum pernix
Sulfolobus solfataricus
Sulfolobus tokodaii
Pyrobaculum aerophilum
Methanobacterium thermoautotrophicum
Methanococcus jannaschii
Methanococcus maripaludis
Methanopyrus kandleri
Pyrococcus abyssi
Pyrococcus horikoshii
Pyrococcus furiosus
Archaeoglobus fulgidus
Methanosarcina acetivorans
Methanosarcina mazei
Halobacterium sp NRC 1
Thermoplasma acidophilum
Thermoplasma volcanium
Nanoarchaeum equitans
Arabidopsis thaliana
Caenorhabditis elegans
Drosophila melanogaster
Homo sapiens
Mus musculus
Rattus norvegicus
Saccharomyces cerevisiae
Schizosaccharomyces pombe
Plasmodium falciparum
Encephalitozoon cuniculi
Guillardia theta Cryptomonas phi
Agrobacterium tumefaciens C58 Cereon
Agrobacterium tumefaciens C58 Washington University
Rhizobium meliloti
Rhizobium loti
Brucella melitensis
Brucella suis
Bradyrhizobium japonicum
Rhodopseudomonas palustris
Caulobacter crescentus
Rickettsia cononi
Rickettsia prowazekii
Wolbachia endosymbiont of Drosophila melanogaster
Bordetella bronchiseptica
Bordetella parapertussis
Bordetella pertussis
Ralstonia solanacearum
Chromobacterium violaceum
Neisseria meningitidis serogroupA
Neisseria meningitidis serogroupB
Nitrosomonas europaea
Escherichia coli K12
Escherichia coli O157 H7
Escherichia coli O157 H7 EDL933
Escherichia coli CFT073
Shigella flexneri 2a str 301
Shigella flexneri 2a str 2457T
Salmonella enterica enterica serovar Typhi Ty2
Salmonella typhi
Salmonella typhimurium
Yersinia pestis C092
Yersinia pestis KIM
Photorhabdus luminescens laumondii
Vibrio cholerae
Vibrio parahaemolyticus
Vibrio vulnificus CMCP6
Vibrio vulnificus YJ016
Haemophilus ducreyi
Haemophilus influenzae
Pasteurella multocida
Shewanella oneidensis
Pseudomonas aeruginosa
Pseudomonas putida KT2440
Pseudomonas syringae pv tomato
Xanthomonas axonopodis pv citri
Xanthomonas campestris pv campestris
Xylella fastidiosa 9A5C
Xylella fastidiosa Temecula1 ATCC 700964
Coxiella burnetii
Buchnera aphidicola Acyrthosiphon pisum
Buchnera aphidicola Schizaphis graminum
Buchnera aphidicola Baizongia pistaciae
Candidatus Blochmannia floridanus
Wigglesworthia glossinidia brevipalpis
Bdellovibrio bacteriovorus
Leptospira interrogans
Deinococcus radiodurans
Rhodopirellula baltica
Anabaena sp PCC 7120
Synechocystis sp PCC 6803
Synechococcus elongatus
Gloeobacter violaceus
Prochlorococcus marinus CCMP1375
Prochlorococcus marinus MIT 9313
Synechococcus sp WH8102
Prochlorococcus marinus pastoris CCMP 1378 MED4
Chlorobium tepidum
Aquifex aeolicus
Bacillus anthracis Ames
Bacillus cereus ATCC 10987
Bacillus cereus ATCC 14579 DSM 31
Bacillus subtilis
Bacillus halodurans
Oceanobacillus iheyensis
Listeria innocua
Listeria monocytogenes
Staphylococcus aureus Mu50 ATCC 700699
Staphylococcus aureus N315
Staphylococcus aureus MW2
Staphylococcus epidermidis
Enterococcus faecalis
Lactobacillus plantarum
Lactococcus lactis lactis
Streptococcus agalactiae serotype3
Streptococcus agalactiae serotype 5
Streptococcus pyogenes serotype M1
Streptococcus pyogenes serotype M18
Streptococcus pyogenes serotype M3
Streptococcus pyogenes SSI 1
Streptococcus mutans
Streptococcus pneumoniae TIGR4
Streptococcus pneumoniae ATCC BAA 255 R6
Lactobacillus johnsonii 26695
Clostridium acetobutylicum
Clostridium tetani
Clostridium perfringens
Thermoanaerobacter tengcongensis
Geobacter sulfurreducens
Thermotoga maritima
Campylobacter jejuni
Helicobacter hepaticus
Wolinella succinogenes
Helicobacter pylori
Helicobacter pylori J99
Bifidobacterium longum
Corynebacterium diphtheriae
Corynebacterium efficiens
Corynebacterium glutamicum ATCC13032
Corynebacterium glutamicum IS fingerprint type 4
Mycobacterium bovis
Mycobacterium tuberculosis H37 Rv
Mycobacterium tuberculosis CDC1551
Mycobacterium paratuberculosis
Mycobacterium leprae
Streptomyces avermitilis
Streptomyces coelicolor
Tropheryma whipplei TW08 27
Tropheryma whipplei Twist
Bacteroides thetaiotaomicron
Porphyromonas gingivalis
Fusobacterium nucleatum nucleatum
Borrelia burgdorferi
Treponema pallidum
Chlamydia muridarum
Chlamydia trachomatis
Chlamydia pneumoniae AR39
Chlamydia pneumoniae CWL029
Chlamydia pneumoniae J138
Chlamydophila pneumoniae TW 183
Chlamydophila caviae
Mycoplasma mycoides mycoides SC
Mycoplasma pulmonis
Mycoplasma gallisepticum
Mycoplasma genitalium
Mycoplasma pneumoniae
Mycoplasma penetrans
Ureaplasma parvum
Onion yellows phytoplasma
Crenarchaeota
Euryarchaeota
α
β
γ
δ
δ
ε
Archaea
Eukarya
Bacteria
Cyanobacteria
Proteobacteria
Firmicutes
Actinobacteria
Chlamydia
Mollicutes

conjunction with the high support for the three-domain structure of life, this result suggests that certain features exclusively shared between the crenarcheota (or eocytes) and the Eukarya (Lake et al., 1984; Rivera and Lake, 1992) are either symplesiomorphic or inherited via nonvertical inheritance events such as horizontal transfer.

Eukarya.—The most basal member of Eukarya in our optimal tree is the cryptophyte *Guillardia theta*. The next most basal lineage is a microsporidium, *Encephalitozoon cuniculi*, followed by the apicomplexan *Plasmodium falcipurans*. The classification of microsporidia as basal eukaryotes is also supported by phylogenetic and comparative analysis of SSU rDNA (Curgy et al., 1980; Ishihara and Hayashi, 1968; Leipe et al., 1993; Vossbrinck et al., 1987; Vossbrinck and Woese, 1986), phylogenies based on elongation factors (Kamaishi et al., 1996a; Kamaishi et al., 1996b), and by the absence of mitochondria in these organisms. However, our result is not consistent with the more recent view that microsporidia are highly derived, close relatives of fungi, which is supported by phylogenies based on tubulin genes (Edlund et al., 1996; Keeling, 2003; Keeling and Doolittle, 1996), the presence of mitochondrial genes in the nuclear genome of four microsporidia (Germot et al., 1997; Hirt et al., 1997; Katinka et al., 2001; Peyretaillade et al., 1998), phylogenies based on the RNase PolII gene (Hirt et al., 1999), and a phylogeny derived from genome-wide sampling of genes thought to evolve slowly (Thomarat et al., 2004).

Bacteria.—Our tree agrees with most of the traditional phylum designations in the Bacteria domain. The Cyanobacteria and Actinobacteria each form a monophyletic group. Interestingly, the Firmicutes form a monophyletic group that excludes the Mollicutes (mycoplasma). This result is also shared with other gene content phylogenies (Hughes et al., 2005; Tekaia et al., 1999; Wolf et al., 2001a) and some SSU rDNA analyses (e.g., Brochier et al., 2002) but is not supported by phylogenies based on other single genes (Wolf et al., 2004) or concatenated gene alignments (Brochier et al., 2002; Brown et al., 2001; Eisen, 1995). Considering the small size of Mollicutes genomes used here, this result may signal that our analysis is still partially prone to the effects of BGA.

The Proteobacteria are mostly grouped together with the exception of the members of the δ and ε subdivisions. Whereas the δ subdivision members do not form a clade, the ε subdivision species are found together at a position just basal to the Firmicutes clade. This result, which has been reported in many other gene content analyses (Dutilh et al., 2004; Gu and Zhang, 2004; House and Fitz-Gibbon, 2002; Hughes et al., 2005; Tekaia et al., 1999; Wolf et al., 2001a; Yang et al., 2005), and is sometimes seen in SSU rDNA trees as well, suggests that the classification of these organisms may need to be reassessed. The remainder of the proteobacterial clade is divided into two monphyletic groups representing the α subdivision, which includes the reduced genome *Rickettsia* group, and the β and γ subdivision, which includes the reduced genome *Buchnera* group.

The phylogeny presented here and several other gene content phylogenies (Dutilh et al., 2004; House and Fitz-Gibbon, 2002; Tekaia et al., 1999; Wolf et al., 2001a; Yang et al., 2005) do not agree with the traditional SSU rDNA-based trees in the placement of thermophiles in the most basal position of the tree—a subject that bears heavily on reconstructions of the nature of the last universal common ancestor. Indeed, the placement of these taxa near to the root of the universal tree is controversial (Brochier and Philippe, 2002; Di Giulio, 2003a, 2003b), and there seems to be no firm consensus.

Optimizing Homology Hypotheses Using Corroboration

The concept of reciprocal illumination underlies several analytical techniques in systematic biology. In one such technique, referred to as *successive weighting* or *successive approximations* (Farris, 1969), fixed, primary homology hypotheses are iteratively reweighted based on their consistency with the overall phylogenetic hypothesis. Homology hypotheses that are consistent with the overall favored hypothesis are given greater weights and the analysis is run again. This process is repeated until the analysis converges on a single tree or a set of tree topologies. Another technique, Goloboff's implied weights (Goloboff, 1993, 1997), applies a similar operation during tree searching. In these techniques, the reciprocal operation is used to down-weight inconsistent primary homology hypotheses, which, nonetheless, remain unchanged or statically defined.

The rationale behind choosing the data set with the highest CCM is related to the process of reciprocal illumination in successive weighting. However, instead of differentially weighting homology hypotheses, our method effectively recodes homology statements based on a new global homology recognition criterion, to find a more corroborated set of homology hypotheses. From this perspective, our technique can be conceived of as an operation that precedes techniques that rely on the differential weighting of characters to achieve consistency. Further, because our technique does not explicitly down-weight homology hypotheses that are not consistent with the favored hypothesis, potential information in these inconsistent characters is retained and can contribute to the analysis.

Another related technique, referred to as *sensitivity analysis* (Wheeler, 1995), optimizes congruence based on weighting schemes or alignment parameters that minimize incongruence between data partitions. Unlike measures of character consistency, *sensitivity analysis* does not specifically assess each homology hypothesis but instead tests corroboration between *sets* of homology hypotheses. This aspect has been criticized as decreasing the overall amount of corroboration for each individual homology hypothesis (Grant and Kluge, 2003). Sensitivity analysis may also be problematic when partitions are arbitrarily or incorrectly defined (Siddall and Kluge, 1999).

The success of our homology-testing framework in finding data sets that correct taxonomic inconsistencies suggests that weakly corroborated homology statements may be as much of a problem as other suspected causes of bias in gene content data sets, such as disparate genome sizes or inappropriate optimality criteria. We prefer the data sets with optimal CCM scores because they offer the least refuted explanation of as much data as possible, and, therefore, the best possible tree of life hypothesis given the data at hand. Using the most corroborated data sets should limit false homology statements, and may also act to decrease bias presented by horizontal gene transfer and phylogenetic noise.

Our results show that a consistent, highly supported tree of life can be produced without subjectively eliminating or differentially weighting data in the analysis. The results also underscore the utility of the principle of reciprocal illumination in evolutionary biology. The CCM provides a means of using this principle to objectively select a similarity stringency threshold appropriate for the type of data and the taxa involved. Indeed, this empirical framework can be used to assess any primary homology hypothesis, including those based on morphological and sequence data.

Additionally, our technique can be used as an objective basis for delineating which genes belong in a gene family and which similar genes should be excluded. It can, therefore, be used to validate the homology of genes that are aligned in any sequence-based phylogenetic analysis. Because they limit false homology, highly corroborated data sets also may improve techniques that rely on the validity of the primary homology statements to infer functional networks.

Acknowledgements

We thank Bas Dutilh for his updated COG matrix; Anthony Deo, Indra Neil Sarkar, A. Phillips, David Figurski, and Mark Siddall for helpful discussion; and Hans-Erik G. Aronson and the AMDeC Bioinformatics Core Facility at the Columbia Genome Center, Columbia University, for computational resources. This work was funded by grants from the US National Institutes of Health (to David Figurski, RD, and PJP). EKL and JAR are partially supported by a training grant from the US National Institutes of Health to New York University. EKL and JAR are further supported by a McKracken Fellowship of NYU.

References

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Apweiler, R., A. Bairoch, and C. H. Wu. 2004. Protein sequence databases. Curr. Opin. Chem. Biol. 8:76–80.

Bansal, A. K., and T. E. Meyer. 2002. Evolutionary analysis by whole-genome comparisons. J. Bacteriol. 184:2260–2272.

Bowers, P. M., S. J. Cokus, D. Eisenberg, and T. O. Yeates. 2004. Use of logic relationships to decipher protein network organization. Science 306:2246–2249.

Brochier, C., E. Bapteste, D. Moreira, and H. Philippe. 2002. Eubacterial phylogeny based on translational apparatus proteins. Trends Genet. 18:1–5.

Brochier, C., P. Forterre, and S. Gribaldo. 2005a. An emerging phylogenetic core of Archaea: Phylogenies of transcription and translation machineries converge following addition of new genome sequences. BMC Evol. Biol. 5:36.

Brochier, C., S. Gribaldo, Y. Zivanovic, F. Confalonieri, and P. Forterre. 2005b. Nanoarchaea: Representatives of a novel archaeal phylum or a fast-evolving euryarchaeal lineage related to Thermococcales? Genome Biol. 6:R42.

Brochier, C., and H. Philippe. 2002. Phylogeny: A non-hyperthermophilic ancestor for bacteria. Nature 417:244.

Brower, A. 1996. Three steps of homology assesment. Cladistics 12:265–272.

Brown, J. R., C. J. Douady, M. J. Italia, W. E. Marshall, and M. J. Stanhope. 2001. Universal trees based on large combined protein sequence data sets. Nat. Genet. 28:281–285.

Clarke, G. D., R. G. Beiko, M. A. Ragan, and R. L. Charlebois. 2002. Inferring genome trees by using a filter to eliminate phylogenetically discordant sequences and a distance matrix based on mean normalized BLASTP scores. J. Bacteriol. 184:2072–2080.

Colless, D. 1980. Congruence between morphometric and allozyme data for Menidia species—A reappraisal. Sys. Zool. 29:289–299.

Curgy, J., J. Vavra, and C. Vivares. 1980. Presence of ribosomal RNAs with prokaryotic properties in Microsporidia. Biol. Cell 38:49–52.

Darlu, P., and G. Lecointre. 2002. When does the incongruence length difference test fail? Mol. Biol. Evol. 19:432–437.

DePinna, M. 1991. Concepts and tests of homology in the cladistic paradigm. Cladistics 7:367–394.

Di Giulio, M. 2003a. The ancestor of the Bacteria domain was a hyperthermophile. J. Theor. Biol. 224:277–283.

Di Giulio, M. 2003b. The universal ancestor and the ancestor of bacteria were hyperthermophiles. J. Mol. Evol. 57:721–730.

Dolphin, K., R. Belshaw, C. D. Orme, and D. L. Quicke. 2000. Noise and incongruence: Interpreting results of the incongruence length difference test. Mol. Phylogenet. Evol. 17:401–406.

Dowton, M., and A. D. Austin. 2002. Increased congruence does not necessarily indicate increased phylogenetic accuracy—The behavior of the incongruence length difference test in mixed-model analyses. Syst. Biol. 51:19–31.

Dutilh, B. E., M. A. Huynen, W. J. Bruno, and B. Snel. 2004. The consistent phylogenetic signal in genome trees revealed by reducing the impact of noise. J. Mol. Evol. 58:527–539.

Edlund, T., J. Li, G. Visvesvara, M. Vodkin, G. McLaughlin, and S. Katiyar. 1996. Phylogenetic analysis of beta-tubulin sequences from amitochondrial protozoa. Mol. Phylogenet. Evol. 5:359–367.

Eisen, J. A. 1995. The RecA protein as a model molecule for molecular systematic studies of bacteria: Comparison of trees of RecAs and 16S rRNAs from the same species. J. Mol. Evol. 41:1105-1123.

Farris, J. 1989. The retention index and the rescaled consistency index. Cladistics 5:417–419.

Farris, J. 1994. Testing the Significance of Incongruence. Cladistics 10:315–319.

Farris, J. S. 1969. A successive approximations approach to character weighting. Syst. Zool. 18:374–385.

Felsenstein, J. 1978. Cases in which parsimony and compatibility methods will be positively misleading. Syst. Zool. 27:401–410.

Fitz-Gibbon, S. T., and C. H. House. 1999. Whole genome-based phylogenetic analysis of free-living microorganisms. Nucleic Acids Res. 27:4218–4222.

Germot, A., H. Philippe, and H. Le Guyader. 1997. Evidence for loss of mitochondria in Microsporidia from a mitochondrial-type HSP70 in Nosema locustae. Mol. Biochem. Parasitol. 87:159–168.

Goloboff, P. 1993. Estimating character weights during tree search. Cladistics 9:83–91.

Goloboff, P. 1997. Self-weighted optimization: Tree searches and character state reconstructions under implied transformation costs. Cladistics 13:225–245.

Gophna, U., W. F. Doolittle, and R. L. Charlebois. 2005. Weighted genome trees: Refinements and applications. J. Bacteriol. 187:1305–1316.

Grant, T., and A. Kluge. 2003. Data exploration in phylogenetic inference: Scientific, heuristic, or neither. Cladistics 19:379–418.

Gu, X., and H. Zhang. 2004. Genome phylogenetic analysis based on extended gene contents. Mol. Biol. Evol. 21:1401–1408.

Harlow, T. J., J. P. Gogarten, and M. A. Ragan. 2004. A hybrid clustering approach to recognition of protein families in 114 microbial genomes. BMC Bioinformatics 5:45.

Hennig, W. 1966. Phylogenetic systematics. Translated by D. Dwight Davis and Rainer Zangerl. Urbana, University of Illinois Press.

Hipp, A. L., J. C. Hall, and K. J. Sytsma. 2004. Congruence versus phylogenetic accuracy: Revisiting the incongruence length difference test. Syst. Biol. 53:81–89.

Hirt, R. P., B. Healy, C. R. Vossbrinck, E. U. Canning, and T. M. Embley. 1997. A mitochondrial Hsp70 orthologue in Vairimorpha necatrix: Molecular evidence that microsporidia once contained mitochondria. Curr. Biol. 7:995–998.

Hirt, R. P., J. M. Logsdon, Jr., B. Healy, M. W. Dorey, W. F. Doolittle, and T. M. Embley. 1999. Microsporidia are related to Fungi: Evidence from the largest subunit of RNA polymerase II and other proteins. Proc. Natl. Acad. Sci. USA 96:580–585.

House, C. H., and S. T. Fitz-Gibbon. 2002. Using homolog groups to create a whole-genomic tree of free-living organisms: An update. J. Mol. Evol. 54:539–547.

Huber, H., M. J. Hohn, R. Rachel, T. Fuchs, V. C. Wimmer, and K. O. Stetter. 2002. A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. Nature 417:63–67.

Huelsenbeck, J. P., and F. Ronquist. 2001. MrBayes: Bayesian inference of phylogenetic trees. Bioinformatics 17:754–755.

Hughes, A., V. Ekollu, R. Friedman, and J. Rose. 2005. Gene family content-based phylogeny of prokaryotes: The effect of criteria for inferring homology. Syst. Biol. 54:268–276.

Huson, D. H., and M. Steel. 2004. Phylogenetic trees based on gene content. Bioinformatics 20:2044–2049.

Ishihara, R., and Y. Hayashi. 1968. Some properties of ribosomes from the sporoplasm of Nosema bombycis. J. Invert. Pathol. 11:377–385.

Kamaishi, T., T. Hashimoto, Y. Nakamura, Y. Masuda, F. Nakamura, K. Okamoto, M. Shimizu, and M. Hasegawa. 1996a. Complete nucleotide sequences of the genes encoding translation elongation factors 1 alpha and 2 from a microsporidian parasite, *Glugea plecoglossi*: Implications for the deepest branching of eukaryotes. J. Biochem. (Tokyo) 120:1095–1103.

Kamaishi, T., T. Hashimoto, Y. Nakamura, F. Nakamura, S. Murata, N. Okada, K. Okamoto, M. Shimizu, and M. Hasegawa. 1996b. Protein phylogeny of translation elongation factor EF-1 alpha suggests microsporidians are extremely ancient eukaryotes. J. Mol. Evol. 42:257–263.

Katinka, M. D., S. Duprat, E. Cornillot, G. Metenier, F. Thomarat, G. Prensier, V. Barbe, E. Peyretaillade, P. Brottier, P. Wincker, F. Delbac, H. El Alaoui, P. Peyret, W. Saurin, M. Gouy, J. Weissenbach, and C. P. Vivares. 2001. Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. Nature 414:450–453.

Keeling, P. J. 2003. Congruent evidence from alpha-tubulin and beta-tubulin gene phylogenies for a zygomycete origin of Microsporidia. Fungal Genet. Biol. 38:298–309.

Keeling, P. J., and W. F. Doolittle. 1996. Alpha-tubulin from early-diverging eukaryotic lineages and the evolution of the tubulin family. Mol. Biol. Evol. 13:1297–1305.

Kluge, A. 1969. Quantitative phyletics and the evolution of Anurans. Syst. Zool. 18:1–32.

Kluge, A. 2003. The repugnant and the mature in phylogenetic inference: A temporal similarity and historical identity. Cladistics 19:356–368.

Kolaczkowski, B., and J. W. Thornton. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature 431:980–984.

Krause, A., J. Stoye, and M. Vingron. 2005. Large scale hierarchical clustering of protein sequences. BMC Bioinformatics 6:15.

Lake, J. A., E. Henderson, M. Oakes, and M. W. Clark. 1984. Eocytes: A new ribosome structure indicates a kingdom with a close relationship to eukaryotes. Proc. Natl. Acad. Sci. USA 81:3786–3790.

Lake, J. A., and M. C. Rivera. 2004. Deriving the genomic tree of life in the presence of horizontal gene transfer: Conditioned reconstruction. Mol. Biol. Evol. 21:681–690.

Lee, M. S. 2001. Uninformative characters and apparent conflict between molecules and morphology. Mol. Biol. Evol. 18:676–680.

Leipe, D. D., J. H. Gunderson, T. A. Nerad, and M. L. Sogin. 1993. Small subunit ribosomal RNA+ of Hexamita inflata and the quest for the first branch in the eukaryotic tree. Mol. Biochem. Parasitol. 59:41–48.

Lin, J., and M. Gerstein. 2000. Whole-genome trees based on the occurrence of folds and orthologs: Implications for comparing genomes on different levels. Genome Res. 10:808–818.

Marcotte, E. M., M. Pellegrini, H. L. Ng, D. W. Rice, T. O. Yeates, and D. Eisenberg. 1999. Detecting protein function and protein-protein interactions from genome sequences. Science 285:751–753.

Nixon, K. C. 1999. The Parsimony Ratchet, a new method for rapid parsimony analysis. Cladistics 15:407–414.

Overbeek, R., M. Fonstein, M. D'Souza, G. D. Pusch, and N. Maltsev. 1999. The use of gene clusters to infer functional coupling. Proc. Natl. Acad. Sci. USA 96:2896–2901.

Park, J., and S. A. Teichmann. 1998. DIVCLUS: An automatic method in the GEANFAMMER package that finds homologous domains in single- and multi-domain proteins. Bioinformatics 14:144–150.

Pellegrini, M., E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates. 1999. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. Proc. Natl. Acad. Sci. USA 96:4285–4288.

Peyretaillade, E., V. Broussolle, P. Peyret, G. Metenier, M. Gouy, and C. P. Vivares. 1998. Microsporidia, amitochondrial protists, possess a 70-kDa heat shock protein gene of mitochondrial evolutionary origin. Mol. Biol. Evol. 15:683–689.

Popper, K. R. 1968. The logic of scientific discovery, 3d edition. Hutchinson, London.

Randau, L., R. Munch, M. J. Hohn, D. Jahn, and D. Soll. 2005. Nanoarchaeum equitans creates functional tRNAs from separate genes for their 5'- and 3'-halves. Nature 433:537–541.

Rieppel, O., and M. Kearney. 2002. Similarity. Biol. J. Linn. Soc. 75:59–82.

Rivera, M. C., and J. A. Lake. 1992. Evidence that eukaryotes and eocyte prokaryotes are immediate relatives. Science 257:74–76.

Rivera, M. C., and J. A. Lake. 2004. The ring of life provides evidence for a genome fusion origin of eukaryotes. Nature 431:152–155.

Rohlf, F. 1982. Consensus indices for comparing classifications. Math. Biosci. 59:131–144.

Siddall, M. E., and A. G. Kluge. 1999. Letter to the editor. Cladistics 15:429–440.

Snel, B., P. Bork, and M. A. Huynen. 1999. Genome phylogeny based on gene content. Nat. Genet. 21:108–110.

Tatusov, R. L., E. V. Koonin, and D. J. Lipman. 1997. A genomic perspective on protein families. Science 278:631–637.

Tatusov, R. L., D. A. Natale, I. V. Garkavtsev, T. A. Tatusova, U. T. Shankavaram, B. S. Rao, B. Kiryutin, M. Y. Galperin, N. D. Fedorova, and E. V. Koonin. 2001. The COG database: New developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res. 29:22–28.

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

Thomarat, F., C. P. Vivares, and M. Gouy. 2004. Phylogenetic analysis of the complete genome sequence of *Encephalitozoon cuniculi* supports the fungal origin of microsporidia and reveals a high frequency of fast-evolving genes. J. Mol. Evol. 59:780–791.

Vossbrinck, C. R., J. V. Maddox, S. Friedman, B. A. Debrunner-Vossbrinck, and C. R. Woese. 1987. Ribosomal RNA sequence suggests microsporidia are extremely ancient eukaryotes. Nature 326:411–414.

Vossbrinck, C. R., and C. R. Woese. 1986. Eukaryotic ribosomes that lack a 5.8S RNA. Nature 320:287–288.

Waters, E., M. J. Hohn, I. Ahel, D. E. Graham, M. D. Adams, M. Barnstead, K. Y. Beeson, L. Bibbs, R. Bolanos, M. Keller, K. Kretz, X. Lin, E. Mathur, J. Ni, M. Podar, T. Richardson, G. G. Sutton, M. Simon, D. Soll, K. O. Stetter, J. M. Short, and M. Noordewier. 2003. The genome of Nanoarchaeum equitans: Insights into early archaeal evolution and derived parasitism. Proc. Natl. Acad. Sci. USA 100:12984–12988.

Wheeler, W. 1995. Sequence alignment, parameter sensitivity, and the phylogenetic analysis of molecular data. Syst. Biol. 44:321–341.

Wheeler, W. 2001. Homology and the optimization of DNA sequence data. Cladistics 17:S3–S11.

Wheeler, W. C., J. Gatesy, and R. DeSalle. 1995. Elision: A method for accommodating multiple molecular sequence alignments with alignment-ambiguous sites. Mol. Phylogenet. Evol. 4:1–9.

Woese, C. R., O. Kandler, and M. L. Wheelis. 1990. Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya. Proc. Natl. Acad. Sci. USA 87:4576–4579.

Wolf, M., T. Muller, T. Dandekar, and J. D. Pollack. 2004. Phylogeny of Firmicutes with special reference to Mycoplasma (Mollicutes) as inferred from phosphoglycerate kinase amino acid sequence data. Int. J. Syst. Evol. Microbiol. 54:871–875.

Wolf, Y. I., I. B. Rogozin, N. V. Grishin, and E. V. Koonin. 2002. Genome trees and the tree of life. Trends Genet 18:472–479.

Wolf, Y. I., I. B. Rogozin, N. V. Grishin, R. L. Tatusov, and E. V. Koonin. 2001a. Genome trees constructed using five different approaches suggest new major bacterial clades. BMC Evol. Biol. 1:8.

Wolf, Y. I., I. B. Rogozin, A. S. Kondrashov, and E. V. Koonin. 2001b. Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. Genome Res. 11:356–372.

Yang, S., R. F. Doolittle, and P. E. Bourne. 2005. Phylogeny determined by protein domain content. Proc. Natl. Acad. Sci. USA 102:373–378.

Yoder, A. D., J. A. Irwin, and B. A. Payseur. 2001. Failure of the ILD to determine data combinability for slow loris phylogeny. Syst. Biol. 50:408–424.

First submitted 17 May 2005; reviews returned 23 August 2005; final acceptance 11 January 2006
Associate Editor: Frank Anderson