Estimation of Phylogenetic Inconsistencies in the Three Domains of Life

Victor Soria-Carrasco[1] *and Jose Castresana*[1]

Department of Physiology and Molecular Biodiversity, Institute of Molecular Biology of Barcelona, CSIC, Barcelona, Spain

Discrepancies in phylogenetic trees of bacteria and archaea are often explained as lateral gene transfer events. However, such discrepancies may also be due to phylogenetic artifacts or orthology assignment problems. A first step that may help to resolve this dilemma is to estimate the extent of phylogenetic inconsistencies in trees of prokaryotes in comparison with those of higher eukaryotes, where no lateral gene transfer is expected. To test this, we used 21 proteomes each of eukaryotes (mainly opisthokonts), proteobacteria, and archaea that spanned equivalent levels of genetic divergence. In each domain of life, we defined a set of putative orthologous sequences using a phylogenetic-based orthology protocol and, as a reference topology, we used a tree constructed with concatenated genes of each domain. Our results show, for most of the tests performed, that the magnitude of topological inconsistencies with respect to the reference tree was very similar in the trees of proteobacteria and eukaryotes. When clade support was taken into account, prokaryotes showed some more inconsistencies, but then all values were very low. Discrepancies were only consistently higher in archaea but, as shown by simulation analysis, this is likely due to the particular tree of the archaeal species used here being more difficult to reconstruct, whereas the trees of proteobacteria and eukaryotes were of similar difficulty. Although these results are based on a relatively small number of genes, it seems that phylogenetic reconstruction problems, including orthology assignment problems, have a similar overall effect over prokaryotic and eukaryotic trees based on single genes. Consequently, lateral gene transfer between distant prokaryotic species may have been more rare than previously thought, which opens the way to obtain the tree of life of bacterial and archaeal species using genomic data and the concatenation of adequate genes, in the same way as it is usually done in eukaryotes.

Introduction

During the last few years, putative cases of lateral gene transfer have been reported, not only between closely related prokaryotes but also between different divisions of life (Koonin et al. 2001; Boucher et al. 2003; Mirkin et al. 2003; Kunin et al. 2005; Lerat et al. 2005). Whereas gene transfer within the same species or between closely related species is well-known and cellular mechanisms to favor it have been described, the evolutionary significance of the transfer of genes between distant species is a matter of controversy, and there is an active debate about whether the latter type of transfer is quantitatively anecdotal (Glansdorff 2000; Kurland 2005) or reflects a paradigm shift in prokaryotic evolution (Gogarten et al. 2002; Bapteste et al. 2005). An intermediate view that is emerging from large-scale analysis of prokaryotic genomes is that there exists a common core of genes vertically evolving while only a certain proportion of genes has been subject to lateral gene transfer (Ge et al. 2005; Snel et al. 2005). An important practical consequence of the views that invoke large amounts of lateral gene transfer is that the possibility to obtain a resolved tree of life of prokaryotes and a good taxonomic classification of them has been questioned (Bapteste et al. 2005; McInerney et al. 2008).

Many computational methods have been proposed to determine whether a sequence has been laterally transferred from one organism to another (Ragan et al. 2006), but phylogenetic methods seem to constitute the best way to identify this phenomenon. However, the complexity of sequence evolution makes also phylogenetic methods vulnerable to artifacts that may produce erroneous trees, particularly when comparing highly divergent sequences (Philippe and Laurent 1998; Beiko et al. 2005; Cantarel et al. 2006; Talavera and Castresana 2007). Furthermore, the difficulties in using correctly assigned orthologous genes is an additional factor that may produce gene trees that disagree with species trees. These problems make difficult the distinction between lateral gene transfer and phylogenetic errors (including orthology misassignment). As a first step to resolve this dilemma, it seems convenient to estimate the amount of phylogenetic inconsistencies in the three domains of life in a comparative manner. Then, it may be tested that, if gene transfer between distant species has been quantitatively very important during the evolution of prokaryotes, there should be, on average, many more discordant branches in phylogenetic trees of bacteria and archaea than in equivalent trees of higher eukaryotes (animals, fungi, and plants). Otherwise, if inconsistencies are similar in the different domains, phylogenetic artifacts may be a major explanation of discordant prokaryotic branches. Thus, higher eukaryotes can be used to measure the background phylogenetic error level. It should be noted that, although there are reported cases of putative lateral gene transfer events in eukaryotes, they mainly affect to protists (Andersson 2005; Keeling et al. 2005), making higher eukaryotes an appropriate group for this comparative phylogenetic analysis.

Materials and Methods

Genome Data

Archaeal and bacterial proteomes, as well as the *Arabidopsis thaliana* proteome, were obtained from GenBank (Benson et al. 2005). All metazoan proteomes were retrieved from Ensembl v.34 (Hubbard et al. 2005). Fungi species proteomes were downloaded from different sites: *Ustilago maydis*, *Aspergillus nidulans*, *Fusarium graminearum*, *Magnaporthe grisea*, and *Neurospora crassa* from Broad Institute (http://www.broad.mit.edu/annotation/fgi/); *Phanerochaete chrysosporium* and *Trichoderma reesei*

[1] Present address: Institut de Biologia Evolutiva (CSIC-UPF), Passeig Maritim de la Barceloneta, Barcelona, Spain.

Key words: phylogenetic methodology, lateral gene transfer, orthology, eukaryotes, proteobacteria, archaea.

E-mail: jcvagr@ibmb.csic.es.

Mol. Biol. Evol. 25(11):2319–2329. 2008
doi:10.1093/molbev/msn176
Advance Access publication August 12, 2008

from DOE Joint Genome Initiative (http://genome.jgi-psf.org/); *Candida albicans* from Stanford Genome Technology Center (http://www-sequence.stanford.edu/group/candida/); *Schizosaccharomyces pombe* from the *S. pombe* Sequencing Group (Sanger Institute, ftp://ftp.sanger.ac.uk/pub/yeast/pombe/Protein_data/pompep); and *Saccharomyces cerevisiae* from Saccharomyces Genome Database (ftp://ftp.yeastgenome.org/yeast/). For the eukaryotic proteomes downloaded from Ensembl, which contain different splice forms of genes, we retained only the longest splice form.

Orthology Assignment Protocol

First, a reciprocal best Blast hit (RBH) method was used to retrieve putative orthologous sets. A RBH was defined as a pair of protein sequences (x, y) in which sequence x of proteome X yields, as the top BlastP hit, sequence y in proteome Y and, reciprocally, sequence y yields sequence x as the best hit when searching back in proteome X. For this analysis, a database was built for each of the three domains, each containing the proteomes of 21 species. BlastP (Altschul et al. 1997) searches were done using default parameters except that the E value threshold was 1×10^{-4}. Proteome sequences of *A. thaliana* and *Bacillus subtilis*, as outgroups for eukaryotes and proteobacteria, respectively, and *Aeropyrum pernix*, as a member of Crenarchaeota, were used as initial proteomes to query against the other proteomes in each group. Then, only sets containing at least 14 species were retained. With this method, we obtained 1,552 putative orthologous genes for eukaryotes, 580 for proteobacteria, and 457 for archaea.

Second, we used a phylogenetic-based (PB) method of orthology assignment. In this case, we performed a BlastP (Altschul et al. 1997) homology search of a selected proteome of eukarya (*S. pombe*), bacteria (*Rickettsia prowazekii*), and archaea (*Thermoplasma acidophilum*) against the corresponding database, limiting the output to hits with an E value equal or lower than 1×10^{-4}. The proteomes used as query were initially chosen because they contained the smallest number of genes in each domain (because this orthology method is computationally intensive). The Blast output was then parsed using a script that retained proteins with a maximum of 100 hits and the presence in at least 14 species, one of them being the outgroup (in the case of archaea, a member of crenarchaeota, either *Sulfolobus*, *Aeropyrum*, or *Pyrobaculum*). These sequences were then aligned and cleaned from problematic blocks, and maximum likelihood trees were constructed from the resulting alignments (see details of phylogenetic reconstructions below). To avoid the use of genes with very complex paralogy problems, resulting trees were analyzed with the help of Bioperl (Stajich et al. 2002) phylogenetic modules, rejecting all trees in which, if there was more than one sequence per species, they did not appear as a monophyletic clade. When more than one sequence per species was present, the sequence corresponding to the shortest branch was selected. This phylogeny-based protocol avoided the use of genes with complex paralogy relationships at the same time that it was able to retain genes with lineage-specific duplications. Using this protocol, we obtained 346 sets of probable orthologous sequences of eukaryotes, 127 of bacteria, and 171 of archaea. In some alignments obtained with the PB method, there were sequences with a high proportion of gaps. In order to assess the effect of such sequences, we removed sequences in which more than 25% of aligned positions consisted of gap characters. All measurements from these alignments were very similar (within a 5% error), and only results with the first data set are reported.

To avoid the influence of the different alignment lengths, PB alignments of proteobacteria and eukaryotes were trimmed a calculated percentage from the C-terminal part of each gene to obtain the mean length of archaea, which had the shortest length (182 amino acids). The same results were obtained when the positions were removed from random places instead of taking them from the C-terminal part (data not shown). After this procedure, the average alignment length was $\sim$182 amino acids for each of the three domains of life.

To try to minimize errors due to short alignments, a high-quality data set was also constructed with all PB alignments $\geq$300 amino acids. After this filter, the number of alignments was 88, 20, and 23 for eukaryotes, proteobacteria, and archaea, respectively.

Phylogenetic Reconstructions

Alignments of the selected orthologous sequences were done with Mafft version 5.531 (Katoh et al. 2005) using a Neighbor-Joining tree as starting tree. Alignments were cleaned using Gblocks 0.91 (Castresana 2000; Talavera and Castresana 2007) with relaxed conditions in order to preserve as much information as possible: "Minimum Number Of Sequences For A Conserved Position" and "Minimum Number Of Sequences For A Flank Position" was half the number of sequences, "Minimum Number Of Contiguous Nonconserved Positions" was 5, "Maximum Number Of Contiguous Nonconserved Positions" was 10, "Minimum Length of a Block" was 5, and "Allowed Gap Positions" was "With Half." After that, maximum likelihood phylogenetic trees were constructed using Phyml version 2.4.4 (Guindon and Gascuel 2003) with the JTT model of protein evolution and four rate categories, with the Gamma distribution parameter and the proportion of invariable sites estimated by the program. Bootstrap values were obtained from 100 replicas.

For constructing the reference trees, we used only the genes from the PB data set that were present in all 21 species because these genes should have less paralogy problems than those where some species was missing. We concatenated the alignments after having cleaned them with Gblocks using default options (more stringent because there is more information in the concatenated alignments [Talavera and Castresana 2007]) except that the "Allowed Gap Positions" parameter was set to "With Half." This gave us three alignments of 17,453 positions (from 62 concatenated proteins) in eukaryotes, 15,928 positions (from 81 proteins) in proteobacteria, and 14,947 positions (from 88 proteins) in archaea. Finally, maximum likelihood trees

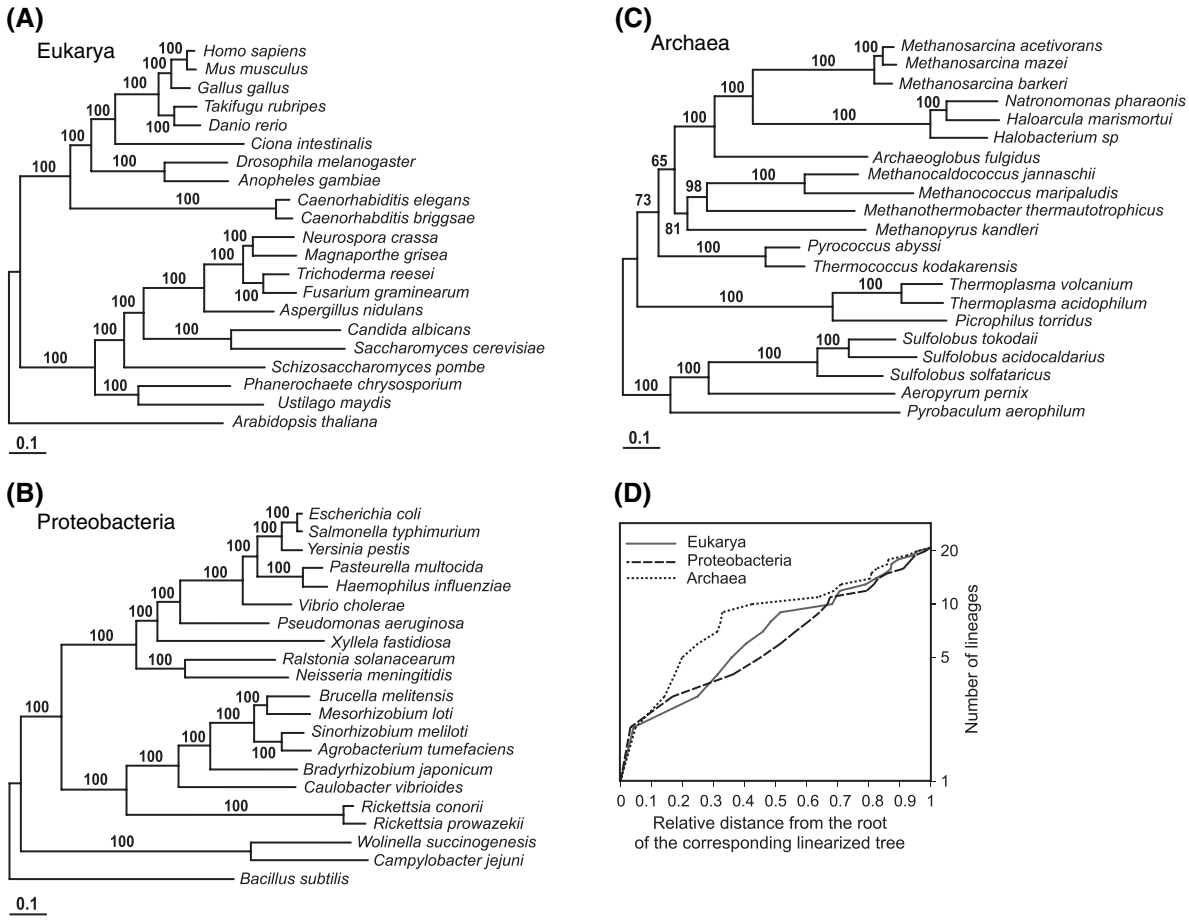

FIG. 1.—Maximum likelihood reference trees of the three domains of life. The trees correspond to (*A*) eukarya (eukaryotes), (*B*) proteobacteria, and (*C*) archaea. The scale bar has the same length in the three trees. Bootstrap support values are shown along the corresponding branches. Trees were rooted according to the known taxonomy of each group. (*D*) For all domains, the lineage accumulation plots of the corresponding linearized trees are shown. The three lineage accumulation plots are autoscaled along the *x* axis to reflect the relative differences in lineage accumulation among the three data sets. The *x* axis represents relative genetic distance or relative time from root (0) to tips (1). The *y* axis is in log scale.

were constructed from these alignments as before. The resulting eukaryotic tree places nematodes at the root, which may be due to long-branch attraction. Therefore, we also performed several measurements using a reference tree where nematodes and insects form a clade (ecdysozoans). These analyses produced very similar results to those using the reference tree obtained by ourselves.

Linearized Trees

A molecular clock was applied to each reference tree, using the nonparametric rate smoothing approach (Sanderson 1997) implemented in r8s 1.71. Because we could not remove the outgroup previous to the analysis in archaea, and to include the same number of species in all sets, we kept the outgroup in proteobacteria and eukaryotes, placing the root arbitrarily in the outgroup branch (as indicated in fig. 1). Smoothing of rate variation along the tree was performed with the penalized likelihood method using the truncated Newton algorithm (Sanderson 2002). Twenty-five smoothing factors with $\log_{10}$ from -6 to $+6$ were used for the penalized likelihood method. The smoothing factor with the lowest cross-validation score, as calculated by r8s, was used. Cumulative cladogenesis events were plotted using Genie version 3.0 (Pybus and Rambaut 2002).

Tree Distances

The symmetric distance or Robinson–Foulds (RF) distance (Robinson and Foulds 1981) between two trees was calculated with the v_treecomp program included in the Vanilla package version 1.2 (Drummond and Strimmer 2001). This program counts the number of partitions of the first tree not present in the second tree plus the number of partitions of the second tree not present in the first tree, and it divides this sum by two (other programs do not divide the total number of incongruent partitions by two). We also calculated, with a modified version of the Ktreedist program (Soria-Carrasco, Talavera, et al. 2007), the number of partitions of the gene tree that had a bootstrap bigger than 80% and that were not present in the reference tree. In this RF

distance, which we call here asymmetric distance, we did not count partitions of the reference tree that were not present in the gene tree because this number greatly increases when there are collapsed clades in the gene tree. If partitions of both the reference and the comparison tree were counted, the effect of collapsing unsupported clades of the gene tree would be completely masked, and there would be almost no differences between the RF distances with or without collapsing. For the calculation of the Edit Paths tree comparison measurements, including the edit distance and the mean length per edit, we used the EEEP program (Beiko et al. 2005; Beiko and Hamilton 2006) with a bootstrap collapse threshold of 80%, a strict reference tree ratchet (-rR), and using only weak time constraints (-uc). Mean length per edit was calculated as the total path count divided by the edit distance. When the gene tree did not contain all species, the missing species were eliminated from the reference tree previous to the computation of all the distances. All calculated topological distances were divided by the number of species in the corresponding tree before calculating the average for each domain.

Approximately Unbiased Test

The approximately unbiased (AU) test (Shimodaira 2002) was performed with the Consel package (Shimodaira and Hasegawa 2001), using the sitewise likelihoods obtained with Tree-Puzzle version 5.2 (Strimmer and von Haeseler 1996). For each alignment, we computed the test for both the gene tree and the reference tree. Rejection was set at a 0.05 significance level. A JTT model of protein evolution (Jones et al. 1992) with four rate categories and a proportion of invariable sites was used for the calculation of the likelihoods. The test was also performed with the WAG model of protein evolution (Whelan and Goldman 2001). For this purpose, gene trees were recalculated in phyml with the WAG matrix previous to performing the likelihood-based test. We obtained similar results with both evolutionary models (less than 1% differences) and only those with the JTT model are shown.

Sequence Simulations

We simulated protein sequences using Rose (Stoye et al. 1998), which allows different evolutionary rates at different positions. We first selected six representative Gblocks cleaned alignments from the alignments of all three domains (two from each domain). From these alignments, we extracted patterns of rate heterogeneity using Tree-Puzzle version 5.2 (Strimmer and von Haeseler 1996) with a discrete among-site rate heterogeneity model of 16 rates categories. For each domain, we simulated sequences along the reference tree. Because the average distance from outgroup to tips was slightly larger when calculated from individual alignments than in the reference tree, we scaled up the reference tree to simulate sequences with the exact average divergence corresponding to the individual trees. These trees were used to run 50 simulations per rate heterogeneity pattern with the Rose program, obtaining 300 sets of sequences for each domain. Simulated sequences were realigned with Mafft and cleaned with Gblocks. Alignments were trimmed a calculated percent value so that their average length were ~182 amino acids. Maximum likelihood trees were calculated as before.

Statistics

Statistical differences ($P < 0.05$) among distributions of tree distances obtained from different data sets were calculated with the Tukey–Kramer test using the JMP package version 5.1 (SAS Institute, Cary, NC).

Results

Selection of Species for Inclusion in the Analysis

In order to produce trees of similar reconstruction difficulty for each of the three domains of life, we compiled the proteomes of 21 species that spanned similar levels of genetic divergences in eukaryotes, bacteria, and archaea. For the eukaryotic trees, we used opisthokonts (metazoans and fungi) with completely sequenced genomes available plus *A. thaliana* as an outgroup. First, we constructed a reference tree from the concatenated genes available in the 21 selected species. Not knowing the precise species tree, we considered this tree as the best reference tree for eukaryotes, that is, the tree that best reflects the species phylogeny. Trees based on large concatenated protein data sets have previously been shown to provide a coherent phylogeny of species (Brown et al. 2001; Daubin et al. 2002; Ciccarelli et al. 2006; Soria-Carrasco, Valens-Vadell, et al. 2007). The maximum likelihood reference tree of eukaryotes had an average of 1.72 substitution/position from the outgroup to the ingroup species (fig. 1*A*). To obtain a similar level of genetic divergence in bacteria, we had to restrict the analysis to proteobacteria and we used *B. subtilis* as an outgroup (we call this set proteobacteria despite including a firmicute species). The average distance from the outgroup to the ingroup species was 1.64 substitutions/position, only 5% less than in eukaryotes. The proteobacterial species selected were related by a topology with a shape and branch lengths very similar to the tree of eukaryotes (fig. 1*B*). We obtained this tree by removing three species from an initial larger tree so that the remaining species were related by a tree similar to the eukaryotic tree. To test the similarity in the branching patterns, we calculated the accumulation of the number of lineages, that is, the succession of split events, in the linearized reference trees of eukaryotes and proteobacteria. Both plots were indeed very similar (fig. 1*D*), indicating that both trees should be, in principle, of a similar reconstruction difficulty, allowing a direct comparison between both data sets.

For archaea, we used, as in eukaryotes, most of the species available to us. Rooting the archaeal tree between euryarchaeota and crenarchaeota produced a similar genetic distance span as in eukaryotes and proteobacteria, with an average of 1.56 substitutions/position from the species of one group to the species of the other group (fig. 1*C*). The shape of this tree was quite different from the bacterial and eukaryotic trees, but the smaller number of archaeal

species available did not allow us to construct an alternative topology. In addition, the plot of the accumulation of the number of lineages was very different from the previous trees (fig. 1*D*). The higher slope at the beginning of the plot indicates a succession of close cladogenesis events at the base of the tree of the selected archaeal species. This is also indicated by the small bootstrap values at the base of the archaeal tree. It is known that this type of branching pattern is difficult to reconstruct (Rokas et al. 2005). Despite these differences, we used the archaeal set for comparison with the other two domains.

Overall Disagreement between Gene Trees and Species Trees in the Three Domains of Life

The use of paralogous instead of orthologous genes may be among the most important causes of disagreement between gene trees and species trees. In order to analyze the effects of orthology assignment in the three domains, we used two different criteria to compile orthologues: a reciprocal best Blast hit (RBH) method and a more rigorous PB method (see Materials and Methods). In each case, the operational definition of orthology was the same in the three species sets, and, although this is a difficult variable to control, we assume that the problems of phylogenetic inconsistency due to paralogy should be, in principle, similar across all sets.

In our first comparative analysis of phylogenetic disagreements, we used the RBH alignments after having cleaned from them the most divergent bocks. This method rendered 1,552, 580, and 457 putative orthologous genes for eukaryotes, proteobacteria, and archaea, respectively. The symmetric difference or RF distance (Robinson and Foulds 1981), which measures the number of different partitions between two trees, indicated that levels of disagreement between gene trees and the species tree were higher in the prokaryotic groups than in eukaryotes: on average, 0.20, 0.32, and 0.36 partitions per species were found for eukaryotes, proteobacteria, and archaea, respectively (fig. 2*A*; confidence intervals for these and other values are given in table 1). However, methods of orthology assignment based only on Blast are known to have low specificity and therefore to identify more incorrect orthologues (paralogues) than phylogeny-based methods, thus causing misleading species phylogenies (Chen et al. 2007).

To avoid the problems of the RBH protocol, we devised an orthology assignment protocol that identifies putative orthologues selected from gene families with a limited number of gene duplications. For this purpose, we first restricted the analysis to genes that produced less than 100 Blast hits in all species of the respective domain. Second, we only used genes that were present in at least 14 species. And third, we reconstructed phylogenetic trees of all homologues found by Blast (not only the best hits) and we only used genes where duplications were intralineage, that is, where all genes of the same species were monophyletic. Then, it was possible to select one gene per species for the orthology set. That is, we refrained from using genes with clearly ancient duplications, where orthology is more difficult to define, and we only used genes with no apparent duplication or with recent, lineage-specific duplications. The latter type of duplication does not present a priori serious orthology assignment problems because any gene from the multicopy family can represent the species in the orthology set. This restricted our analysis to 346, 127, and 171 putative orthologous genes for eukaryotes, proteobacteria, and archaea, respectively. Even if this data set is smaller, these genes should have less paralogy problems (Chen et al. 2007), and therefore they should allow us to construct more reliable trees. In addition, it is important to note that the genes of the PB method were selected so that they can experience a limited number of gene duplications and losses, meaning that they are not necessarily essential genes. We performed the rest of analyses with this set. As expected, in the PB orthologues, we found lower RF distances in all groups: 0.19, 0.24, and 0.32 partitions per species for eukaryotes, proteobacteria, and archaea, respectively (fig. 2*B*). It is interesting to note that the largest reduction in phylogenetic errors took place in proteobacteria. Thus, many phylogenetic inconsistencies usually found in this group could be simply due to paralogy problems of a very relaxed orthology protocol. However, even with data sets using the PB orthology protocol, which is very conservative, there were still many phylogenetic discrepancies in the three domains (almost no single tree is correct) and, as with the RBH set, there were more discrepancies in prokaryotes than in eukaryotes.

Another important factor known to affect phylogenetic reconstruction is alignment length. However, in the analyses above, the average alignment length was smaller in prokaryotes (214 and 182 amino acids in proteobacteria and archaea, respectively) than in eukaryotes (251 amino acids), and it is expected that shorter alignments will lead to more phylogenetic artifacts due to lack of data. Indeed, we corroborated that there is a strong negative correlation between alignment length and RF distance in a pool of the three data sets (Spearman's Rho $= -0.48$; $P < 0.0001$). To control for the influence of alignment length, alignments of proteobacteria and eukaryotes were trimmed a calculated percentage from the C-terminal part of each gene to obtain the mean length of archaea. After this procedure, the average alignment length was $\sim$182 amino acids in the three domains. The distributions of RF distances were now more similar for eukaryotes and proteobacteria, with an average of 0.23 and 0.25 partitions per species, respectively (fig. 2*C*), indicating the importance of taking alignment length into account, as shown in previous works (Beiko et al. 2005). In archaea, the number of different partitions with respect to the reference tree remained significantly higher than in the two other groups.

We also calculated, in the trimmed data set, RF distances among all pairs of trees to estimate the average number of errors between trees without using the reference tree. We obtained 0.32, 0.37, and 0.39 different clades per species for eukaryotes, proteobacteria, and archaea, respectively (fig. 2*D*). Values for archaea were now more similar to those of the other domains, indicating that the quality of the archaeal reference tree may also be inflating the RF values in this domain. It is interesting to note that most genes produced different topologies to those of other genes

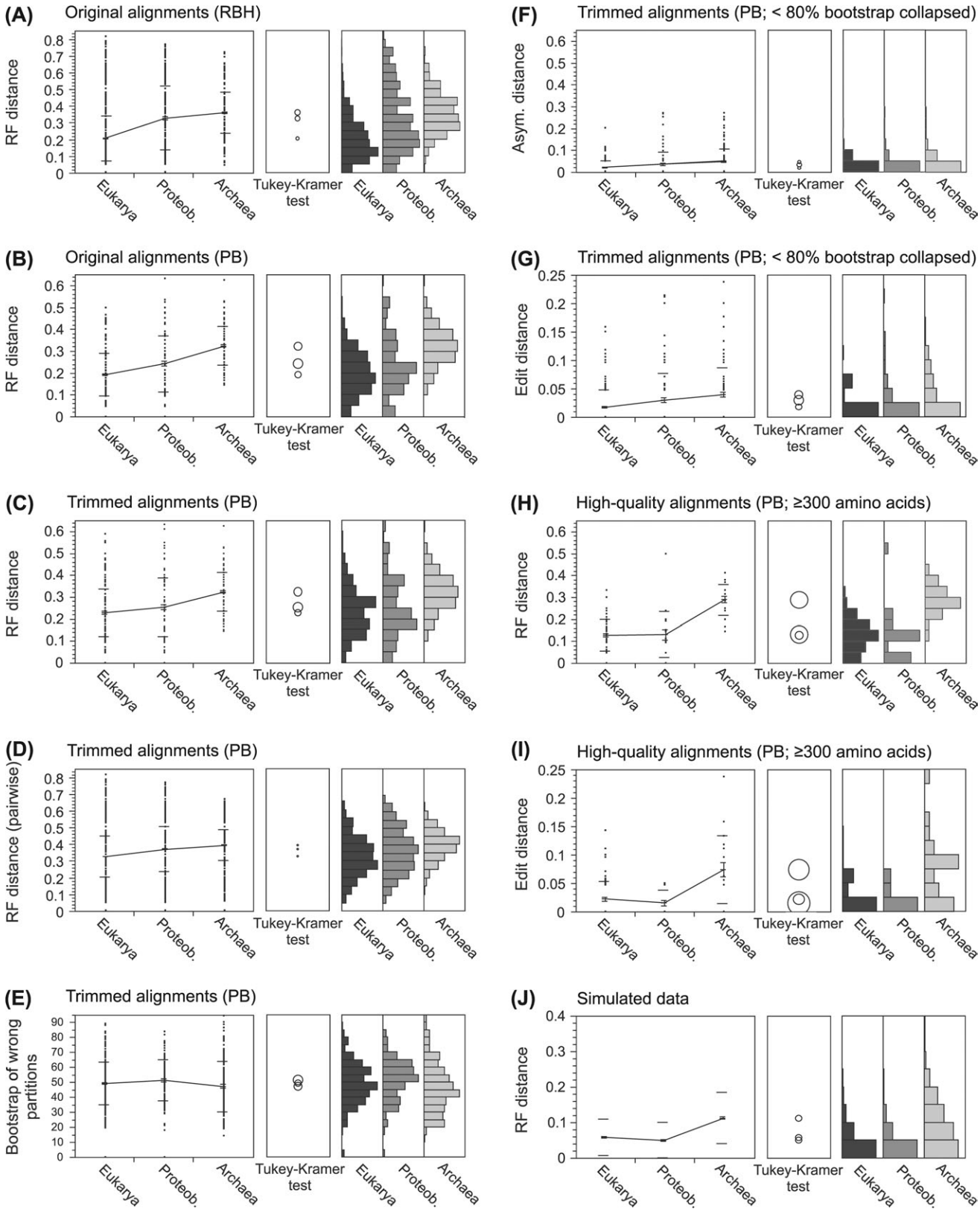
(A)
Original alignments (RBH)
RF distance
(B)
Original alignments (PB)
RF distance
(C)
Trimmed alignments (PB)
RF distance
(D)
Trimmed alignments (PB)
RF distance (pairwise)
(E)
Trimmed alignments (PB)
Bootstrap of wrong partitions
(F)
Trimmed alignments (PB; < 80% bootstrap collapsed)
Asym. distance
(G)
Trimmed alignments (PB; < 80% bootstrap collapsed)
Edit distance
(H)
High-quality alignments (PB; ≥300 amino acids)
RF distance
(I)
High-quality alignments (PB; ≥300 amino acids)
Edit distance
(J)
Simulated data
RF distance
Eukarya
Proteob.
Archaea
Tukey-Kramer test

Table 1
Means and 95% Confidence Intervals of Different Topological Tests in the Three Domains of Life

	Eukarya		Proteobacteria		Archaea	
	Mean	95% CI	Mean	95% CI	Mean	95% CI
RF distance (RBH, original)	0.20	0.20–0.21	0.32	0.31–0.34	0.36	0.34–0.37
RF distance (PB, original)	0.19	0.18–0.20	0.24	0.22–0.26	0.32	0.31–0.34
RF distance (trimmed)	0.23	0.22–0.24	0.25	0.23–0.28	0.32	0.31–0.34
RF distance (trimmed, pairwise)	0.32	0.32–0.32	0.37	0.36–0.37	0.39	0.39–0.39
Asymmetric distance (trimmed, <80% col.)	0.016	0.01–0.02	0.031	0.02–0.04	0.043	0.03–0.05
Edit distance (trimmed, <80% col.)	0.017	0.01–0.02	0.029	0.02–0.04	0.040	0.03–0.05
Length per edit (trimmed, <80% col.)	0.36	0.26–0.47	0.74	0.26–1.22	0.42	0.28–0.56
Bootstrap (trimmed, wrong partitions)	49.1	47.6–50.6	51.2	48.8–53.7	47.1	44.6–49.7
Bootstrap (trimmed, right partitions)	81.6	80.7–82.5	83.1	81.8–84.5	86.7	85.7–87.6
RF distance (≥300 aa)	0.13	0.11–0.14	0.13	0.08–0.18	0.29	0.26–0.32
RF distance (≥300 aa, pairwise)	0.18	0.18–0.18	0.20	0.18–0.22	0.34	0.34–0.35
Asymmetric distance (≥300 aa, <80% col.)	0.022	0.02–0.03	0.022	0.00–0.04	0.083	0.06–0.11
Edit distance (≥300 aa, <80% col.)	0.022	0.01–0.03	0.015	0.00–0.03	0.073	0.05–0.10
Length per edit (≥300 aa, <80% col.)	0.15	0.12–0.18	0.18	0.10–0.26	0.32	0.14–0.50
AU test (trimmed)	23.4	18.9–27.9	46.5	37.7–55.3	66.1	58.9–73.3
AU *P* values (trimmed)	0.16	0.14–0.18	0.11	0.09–0.13	0.07	0.05–0.08
AU test (≥300 aa)	14.8	7.20–22.3	25.0	4.20–45.8	87.0	72.1–102
AU *P* values (≥300 aa)	0.19	0.16–0.23	0.20	0.13–0.28	0.04	0.00–0.08
RF distance (simulated data)	0.06	0.05–0.06	0.05	0.04–0.05	0.11	0.10–0.12

NOTE.—CI, confidence interval; aa, amino acids; and col., collapsed. RF and asymmetric distances are given in partitions per species and the edit distance in edits per species. Bootstrap and AU tests are percent values.

(almost no pair of genes showed RF distances of zero in the three domains).

RF distances give us the number of different clades between two trees regardless of the relative support of such clades. To take support indices into account, we estimated the bootstrap values of all gene trees. We then calculated the distributions and average bootstrap values of those clades that were different from the reference tree and which could cause stronger inconsistencies. The average bootstrap values were very similar for the three domains of life, indicating that there were no large differences in the relative level of support of the wrong clades: 49%, 51%, and 47% for eukaryotes, proteobacteria, and archaea, respectively (fig. 2*E*). Bootstrap values for the right clades were also very similar (table 1).

To have an estimation of phylogenetic distances that directly takes clade support into account, we introduce here the asymmetric distance, which counts only partitions of the gene tree that were not present in the reference tree after collapsing nodes that had <80% bootstrap support. This reduced the distances to 0.016, 0.031, and 0.043 different gene tree partitions per species for eukaryotes, proteobacteria, and archaea, respectively (fig. 2*F*). Although prokaryotic distances are higher, all values are now very small compared with the overall RF distances due to the high number of polytomies obtained after collapsing unresolved nodes.

We also calculated the edit distance after collapsing nodes that had <80% bootstrap support. This distance measures the minimum number of branch movements or edits that are necessary in a given tree to make it identical to a reference tree. We obtained similar values to those calculated with the asymmetric distance: 0.017, 0.029, and 0.040 movements per species for eukaryotes, proteobacteria, and archaea, respectively (fig. 2*G*). These are very small values that certainly would not need to be explained by different causes than those derived from paralogy problems and tree reconstruction problems. However, the mean length per edit, which measures the number of internal nodes of the reference tree that are traversed by a given edit to reconcile the trees, was different between prokaryotes and eukaryotes: 0.36 for eukaryotes, 0.74 for proteobacteria (0.52 after removing an extreme outlier), and 0.42 for archaea.

We also performed, for the trimmed alignments, the AU test (Shimodaira 2002), which calculates the difference

←

FIG. 2.—Topological differences (divided by the number of species) in individual gene trees. (*A*) RF distances to the reference tree calculated from the original alignments obtained with the RBH method of orthology assignment. (*B*) RF distances calculated from the original alignments obtained with the PB method of orthology assignment. (*C*) RF distances to the reference tree calculated from the PB alignments trimmed to be of an average length of ~182 amino acids. (*D*) RF distances between all pairs of gene trees calculated from the PB trimmed alignments. (*E*) Bootstrap support values of clades that are not present in the reference tree, calculated from the PB trimmed alignments. (*F*) Asymmetric distances (counting only partitions of the gene tree), with 80% bootstrap nodes collapsed, calculated from the PB trimmed alignments. (*G*) Edit distances, with 80% bootstrap nodes collapsed, calculated from the PB trimmed alignments. (*H*) RF distances to the reference tree calculated from the high-quality alignments (PB, ≥300 amino acids). (*I*) Edit distances, with 80% bootstrap nodes collapsed, calculated from the high-quality alignments. (*J*) RF distances calculated from alignments of ~182 amino acids simulated along the reference trees. The left panel shows the individual distances, the mean ± standard error, and the standard deviation. A line connects the mean values of each domain. The middle panel shows a graphical representation of the Tukey–Kramer test, which is a conservative test to evaluate if means are significantly different. In the multiple comparison circles, every circle is centered at the mean value of each domain and when two means are statistically different ($P < 0.05$), the circles do not intersect or intersect by an angle <90°. The right panel shows the distribution of distances in the three domains of life. All mean values and 95% confidence intervals are given in table 1.

in the maximum likelihood values of each alignment for the gene and the species trees and reports the significance of this difference. This measures whether a gene tree as a whole is compatible with the species tree or not. We found 23.4% of eukaryotic alignments that were incompatible with the species tree, 46.5% in proteobacteria, and 66.1% in archaea (table 1). In this test, a substantial number of proteobacterial and archaeal trees were rejected in comparison with the eukaryotic trees. However, rejected trees, although generally very different from the reference tree, showed low levels of highly supported differences as measured by the asymmetric distance: 0.033, 0.058, and 0.060 partitions per species for eukaryotes, proteobacteria, and archaea, respectively. Similar values were found with the edit distance. It is interesting to note the degree of congruence between the AU test, the edit distance, and the asymmetric distance (table 1). On the other hand, the global RF distance may be a good estimation of the observed differences between two trees, and it allows to appreciate big trends such as those caused by changing orthology protocols or alignment length. However, this distance seems not to reflect small differences between data sets such as those detected by the edit and asymmetric RF distances based on supported nodes.

To try to minimize stochastic errors due to short alignments, we produced another data set where we selected those alignments with a length $\geq$300 amino acids (the high-quality data set). This restricted the number of alignments to 88, 20, and 23 in eukaryotes, proteobacteria, and archaea, respectively. The average outgroup–ingroup distance remained similar among the three domains after this selection (1.77, 1.74, and 1.61 substitutions/position for eukaryotes, proteobacteria, and archaea, respectively). The average alignment length decreased in archaea but it remained similar between eukaryotes and proteobacteria (483, 500, and 382 amino acids for eukaryotes, proteobacteria, and archaea, respectively). In this data set, all RF distances with respect to the reference tree were much smaller, indicating again the profound impact of alignment length on phylogenetic errors (fig. 2*H*). In addition, RF distances in eukaryotes and proteobacteria were identical: 0.13 partitions per species. In archaea, this distance was again the highest one, with 0.29 partitions per species. Pairwise distances were again similar for eukaryotes and both prokaryotic groups (table 1). We also calculated asymmetric distances and edit distances after collapsing nodes that had $<$80% bootstrap support, and we obtained again similar and very small values for eukaryotes and proteobacteria (table 1 and fig. 2*I*). Finally, the AU test was discordant in this data set: 14.8%, 25%, and 86.9% trees were rejected in eukaryotes, proteobacteria, and archaea, respectively (table 1). However, differences between these values were not significant. The small number of genes in this high-quality data set may affect the complete congruence of the AU test with the other measurements of tree discrepancy.

Why Are There More Inconsistencies in the Archaeal Tree?

As expected, the archaeal distances were always significantly higher than those of eukaryotes and proteobacteria. The reference tree of archaeal species has a greater accumulation of cladogenesis events at the base of the tree (fig. 1*C* and *D*) and, as indicated above, trees with such topologies are known to be very difficult to reconstruct and thus different individual genes may produce different topologies (Rokas et al. 2005). In order to actually test the inherent difficulty of reconstruction of the different trees, we simulated protein sequence alignments along the reference trees of eukaryotes, proteobacteria, and archaea. We produced 300 simulated alignments for each set, with $\sim$182 amino acids length, and we used them to reconstruct phylogenetic trees with the same methods used for the real alignments. In absolute terms, the RF distances of the simulated alignments (fig. 2*J*) were smaller than the comparable RF distances obtained from the observed alignments (fig. 2*C*), but this can be explained because we did not simulate amino acid biases, different evolutionary rates in different lineages, or orthology assignment problems, all of which may be present in the real data sets. However, in relative terms, the trees simulated along the archaeal reference tree showed a much higher RF distance (0.11 partitions per species) than the trees simulated along the eukaryotic and proteobacterial trees (0.06 and 0.05 partitions per species, respectively) (fig. 2*J*). This indicated that, indeed, the archaeal tree is more difficult to reconstruct due to the short internal branches, and this may explain the higher topological inconsistencies found in the real alignments of archaea. Thus, the best comparison of phylogenetic inconsistencies can be established in this work between the proteobacterial and opisthokont data sets.

Discussion

Several topological distances showed that, when species were related by a similar tree shape and genetic divergence, orthologues were constructed with a phylogenetic-based method, and alignment length was normalized, the magnitude of phylogenetic inconsistencies was similar in eukaryotes and proteobacteria. For example, with the RF distance, we measured 0.23 and 0.25 different clades per species for eukaryotes and proteobacteria, respectively. When clade support was taken into account, tree discrepancies were higher in proteobacteria, but in these types of tree distances, all values were very low: 0.016 different and strongly supported clades per species in eukaryotes and 0.031 in proteobacteria. These are values obtained for these specific data sets of proteobacteria and opisthokonts, and surely, these values can vary in other data sets. It seems, however, that the overall magnitude of errors in prokaryotes is not very different from eukaryotes when taking the adequate controls into account. Thus, differences in inconsistencies among domains are very far from those expected to reject the existence of the tree of life in some of the domains of life and not in the others. Although the slightly bigger number of clade differences in the specific proteobacterial data set used here could be attributed to lateral gene transfer, an alternative explanation to many of such inconsistencies could be the existence of a greater proportion of hidden paralogies (see below).

A caveat of our analysis is that we used a relatively small number of genes compared with the genes present

in a typical bacterial genome. This is, however, one of the highest number of genes that can be used (for these specific data sets) if one wants to apply a reliable orthology protocol for accurate phylogenetic inference. The use of more relaxed orthology protocols will surely identify more genes and with higher numbers of tree inconsistencies. One may think that these genes should be analyzed because they may be more likely to be transferred, but they are also more likely to present paralogy problems. Therefore, we think that, in the case of using relaxed orthology protocols, it would be more difficult to tell whether the observed inconsistencies are due to lateral gene transfer or to paralogy problems, and thus, these genes are less informative to address this particular problem.

Despite using a relatively strict orthology protocol, there was a high number of inconsistencies in eukaryotes when considering both supported and unsupported clades: 0.23 different partitions per species with respect to the reference tree, meaning that there were, on average, almost five different clades in a 21-species eukaryotic tree. All strange eukaryotic groupings we found, such as, for example, the grouping of human with chicken or insects with fungi, that would have been removed from a manual set of eukaryotic orthologues, were kept in our analysis to follow exactly the same orthology assignment protocol used for the prokaryotic analyses. Similar observations can be derived from the trees with unsupported clades collapsed. Many strong inconsistencies observed in the eukaryotic genes can be attributed, first, to erroneous orthology assignment. Second, alignment length has an enormous impact on phylogenetic reconstruction and, as shown in our analysis, alignments of small length tend to produce trees with more differences with respect to the species tree. Indeed, it has been shown that single gene phylogenies, as opposed to longer concatenated alignments, show many discrepancies in different eukaryotic groups (Huerta-Cepas et al. 2007). Third, alignment difficulties arising in the comparison of divergent sequences are also very likely to cause many inconsistencies (Lebrun et al. 2006; Talavera and Castresana 2007). And fourth, strange eukaryotic trees are also surely induced by problems in the accuracy of phylogenetic reconstruction, such as saturation of genetic distances, long-branch attraction, amino acid composition bias, or the existence of short internal branches (Philippe and Laurent 1998; Rokas et al. 2003, 2005).

All problems in eukaryotic phylogenies mentioned above are also expected to be relevant for prokaryotes, although we do not know in which relative proportion. It can then be argued that phylogenetic inconsistencies in eukaryotes are mainly due to paralogy problems, whereas prokaryotes are less affected by this problem. This could occur if, for example, gene and genome duplications and segregating gene losses happened more often during eukaryotic evolution. In this case, our results would remain consistent with frequent gene transfer within prokaryotes. Nevertheless, current data seem to indicate that, in fact, gene duplications have remained constant through all lineages of eukaryotes and prokaryotes. According to Lynch and Conery (2003), chromosomal events that result in gene duplications appear to occur at rates that are roughly proportional to those of mutations causing nucleotide substitutions, and this is reflected in estimates of gene duplication events that are of the same magnitude for eukaryotes (multicellular and unicellular) and prokaryotes. In contrast, the ability of a duplicate gene to survive increases with decreasing effective population size, that is, a duplicate gene is more likely to survive in multicellular eukaryotes than in unicellular organisms (Conery and Lynch 2001; Lynch and Conery 2003). This would mean that there is a higher rate of gene loss in prokaryotes, creating more problems of hidden paralogy, which are the most difficult to resolve. Interestingly, a greater amount of hidden paralogy in prokaryotes would also explain that the inconsistencies (edits) in prokaryotes traverse a greater part of the tree than in eukaryotes. At any rate, different data seem not to support more problems of orthology assignment in eukaryotes. Rather, paralogy problems, hidden or not, seem to strongly affect phylogenies across the three domains of life, providing a consistent and well-known phenomenon to explain many of the phylogenetic inconsistencies observed in the three domains.

Although many open questions remain, it seems that both prokaryotes and eukaryotes are affected by significant amounts of paralogy and phylogenetic artifacts. The small differences observed between the eukaryotes and the prokaryotes analyzed here are not the expected ones to reject the tree of life in one case and not in the other. In our opinion, this would mean that lateral gene transfer in prokaryotes may be much smaller than previously thought, at least between divergent species, and that vertical evolution is the predominant force also in prokaryotes. Tree inconsistencies could rather be due to phenomena that are common to all parts of the tree of life: the difficulties and challenges of defining orthologues, building accurate alignments, and reconstructing phylogenetic trees, all of which are more severe for divergent sequences and for reconstructions based on single genes (Castresana 2007; Huerta-Cepas et al. 2007). Our comparative approach illustrates the relative importance of these problems in different domains of life, but it is necessary to note that we have only used one data set per domain of life in these analyses. To understand the putative impact of lateral gene transfer in prokaryotic evolution, further comparative tests should be performed with different eukaryotic and prokaryotic groups and using of other accurate orthology assignment protocols. If lateral gene transfer was finally demonstrated not to be so predominant among prokaryotes, the concatenation of widely represented protein sequences (with the purpose of improving the phylogenetic signal) should be an adequate strategy to reconstruct the tree of life and to improve the current phylogenetic classification of bacteria and archaea (Ciccarelli et al. 2006; Soria-Carrasco, Valens-Vadell, et al. 2007) in the same way as it is usually done in eukaryotes (e.g., Dunn et al. 2008).

Acknowledgments

This work was financially supported with a research grant in bioinformatics from the Fundación BBVA (Spain). We thank Fernando González-Candelas and Aidan Budd for carefully reading an earlier version of the manuscript and giving useful suggestions.

Literature Cited

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Andersson JO. 2005. Lateral gene transfer in eukaryotes. Cell Mol Life Sci. 62:1182–1197.

Bapteste E, Susko E, Leigh J, MacLeod D, Charlebois RL, Doolittle WF. 2005. Do orthologous gene phylogenies really support tree-thinking? BMC Evol Biol. 5:33.

Beiko RG, Hamilton N. 2006. Phylogenetic identification of lateral genetic transfer events. BMC Evol Biol. 6:15.

Beiko RG, Harlow TJ, Ragan MA. 2005. Highways of gene sharing in prokaryotes. Proc Natl Acad Sci USA. 102:14332–14337.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. 2005. GenBank. Nucleic Acids Res. 33:D34–D38.

Boucher Y, Douady CJ, Papke RT, Walsh DA, Boudreau ME, Nesbo CL, Case RJ, Doolittle WF. 2003. Lateral gene transfer and the origins of prokaryotic groups. Annu Rev Genet. 37:283–328.

Brown JR, Douady CJ, Italia MJ, Marshall WE, Stanhope MJ. 2001. Universal trees based on large combined protein sequence data sets. Nat Genet. 28:281–285.

Cantarel BL, Morrison HG, Pearson W. 2006. Exploring the relationship between sequence similarity and accurate phylogenetic trees. Mol Biol Evol. 23:2090–2100.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 17:540–552.

Castresana J. 2007. Topological variation in single-gene phylogenetic trees. Genome Biol. 8:216.

Chen F, Mackey AJ, Vermunt JK, Roos DS. 2007. Assessing performance of orthology detection strategies applied to eukaryotic genomes. PLoS ONE. 2:e383.

Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P. 2006. Toward automatic reconstruction of a highly resolved tree of life. Science. 311:1283–1287.

Conery JS, Lynch M. 2001. Nucleotide substitutions and the evolution of duplicate genes. Pac Symp Biocomput. 6:167–178.

Daubin V, Gouy M, Perriere G. 2002. A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. Genome Res. 12:1080–1090.

Drummond A, Strimmer K. 2001. PAL: an object-oriented programming library for molecular evolution and phylogenetics. Bioinformatics. 17:662–663.

Dunn CW, Hejnol A, Matus DQ, et al. (18 co-authors). 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature. 452:745–749.

Ge F, Wang LS, Kim J. 2005. The cobweb of life revealed by genome-scale estimates of horizontal gene transfer. PLoS Biol. 3:e316.

Glansdorff N. 2000. About the last common ancestor, the universal life-tree and lateral gene transfer: a reappraisal. Mol Microbiol. 38:177–185.

Gogarten JP, Doolittle WF, Lawrence JG. 2002. Prokaryotic evolution in light of gene transfer. Mol Biol Evol. 19:2226–2238.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 52:696–704.

Hubbard T, Andrews D, Caccamo M, et al. (52 co-authors). 2005. Ensembl 2005. Nucleic Acids Res. 33:D447–D453.

Huerta-Cepas J, Dopazo J, Dopazo H, Gabaldón T. 2007. The human phylome. Genome Biol. 8:R109.

Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci. 8:275–282.

Katoh K, Kuma K, Toh H, Miyata T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. 33:511–518.

Keeling PJ, Burger G, Durnford DG, Lang BF, Lee RW, Pearlman RE, Roger AJ, Gray MW. 2005. The tree of eukaryotes. Trends Ecol Evol. 20:670–676.

Koonin EV, Makarova KS, Aravind L. 2001. Horizontal gene transfer in prokaryotes: quantification and classification. Annu Rev Microbiol. 55:709–742.

Kunin V, Goldovsky L, Darzentas N, Ouzounis CA. 2005. The net of life: reconstructing the microbial phylogenetic network. Genome Res. 15:954–959.

Kurland CG. 2005. What tangled web: barriers to rampant horizontal gene transfer. Bioessays. 27:741–747.

Lebrun E, Santini JM, Brugna M, Ducluzeau AL, Ouchane S, Schoepp-Cothenet B, Baymann F, Nitschke W. 2006. The Rieske protein: a case study on the pitfalls of multiple sequence alignments and phylogenetic reconstruction. Mol Biol Evol. 23:1180–1191.

Lerat E, Daubin V, Ochman H, Moran NA. 2005. Evolutionary origins of genomic repertoires in bacteria. PLoS Biol. 3:e130.

Lynch M, Conery JS. 2003. The origins of genome complexity. Science. 302:1401–1404.

McInerney JO, Cotton JA, Pisani D. 2008. The prokaryotic tree of life: past, present... and future? Trends Ecol Evol. 23:276–281.

Mirkin BG, Fenner TI, Galperin MY, Koonin EV. 2003. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. BMC Evol Biol. 3:2.

Philippe H, Laurent J. 1998. How good are deep phylogenetic trees? Curr Opin Genet Dev. 8:616–623.

Pybus OG, Rambaut A. 2002. GENIE: estimating demographic history from molecular phylogenies. Bioinformatics. 18:1404–1405.

Ragan MA, Harlow TJ, Beiko RG. 2006. Do different surrogate methods detect lateral genetic transfer events of different relative ages? Trends Microbiol. 14:4–8.

Robinson DF, Foulds LR. 1981. Comparison of phylogenetic trees. Math Biosci. 53:131–147.

Rokas A, Kruger D, Carroll SB. 2005. Animal evolution and the molecular signature of radiations compressed in time. Science. 310:1933–1938.

Rokas A, Williams BL, King N, Carroll SB. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature. 425:798–804.

Sanderson MJ. 1997. A nonparametric approach to estimating divergence times in the absence of rate constancy. Mol Biol Evol. 14:1218–1231.

Sanderson MJ. 2002. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. Mol Biol Evol. 19:101–109.

Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. Syst Biol. 51:492–508.

Shimodaira H, Hasegawa M. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics. 17:1246–1247.

Snel B, Huynen MA, Dutilh BE. 2005. Genome trees and the nature of genome evolution. Annu Rev Microbiol. 59:191–209.

Soria-Carrasco V, Talavera G, Igea J, Castresana J. 2007. The K tree score: quantification of differences in the relative branch length and topology of phylogenetic trees. Bioinformatics. 23:2954–2956.

Soria-Carrasco V, Valens-Vadell M, Peña A, Antón J, Amann R, Castresana J, Rosselló-Mora R. 2007. Phylogenetic position of *Salinibacter ruber* based on concatenated protein alignments. Syst Appl Microbiol. 30:171–179.

Stajich JE, Block D, Boulez K, et al. (21 co-authors). 2002. The Bioperl toolkit: perl modules for the life sciences. Genome Res. 12:1611–1618.

Stoye J, Evers D, Meyer F. 1998. Rose: generating sequence families. Bioinformatics. 14:157–163.

Strimmer K, von Haeseler A. 1996. Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. Mol Biol Evol. 13:964–969.

Talavera G, Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol. 56:564–577.

Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol. 18:691–699.

Andrew Roger, Associate Editor

Accepted August 1, 2008