Genome Phylogenies Indicate a Meaningful α-Proteobacterial Phylogeny and Support a Grouping of the Mitochondria with the Rickettsiales

David A. Fitzpatrick, Christopher J. Creevey, and James O. McInerney

Department of Biology, National University of Ireland, Maynooth, County Kildare, Ireland

Placement of the mitochondrial branch on the tree of life has been problematic. Sparse sampling, the uncertainty of how lateral gene transfer might overwrite phylogenetic signals, and the uncertainty of phylogenetic inference have all contributed to the issue. Here we address this issue using a supertree approach and completed genomic sequences. We first determine that a sensible α-proteobacterial phylogenetic tree exists and that it can confidently be inferred using orthologous genes. We show that congruence across these orthologous gene trees is significantly better than might be expected by random chance. There is some evidence of horizontal gene transfer within the α-proteobacteria, but it appears to be restricted to a minority of genes (~23%) most of whom (~74%) can be categorized as operational. This means that placement of the mitochondrion should not be excessively hampered by interspecies gene transfer. We then show that there is a consistently strong signal for placement of the mitochondrion on this tree and that this placement is relatively insensitive to methodological approach or data set. A concatenated alignment was created consisting of 15 mitochondrion-encoded proteins that are unlikely to have undergone any lateral gene transfer in the timeline under consideration. This alignment infers that the sister group of the mitochondria, for the taxa that have been sampled, is the order Rickettsiales.

Introduction

Proteobacteria or the purple bacteria are one of the largest known divisions within prokaryotes. Based on analysis of the evolution of 16S rRNA sequences, the proteobacteria division was first circumscribed by Carl Woese and coworkers (Woese et al. 1985; Woese 1987). Subsequent 16S rRNA analysis has led to the proteobacteria being divided into five subdivisions (α, β, δ, γ, and ε; Woese 1987).

The proteobacterial group is an important division as it includes known animal, human, and plant pathogens. Furthermore, this division has played a crucial role in eukaryotic cell origin and evolution through endosymbiosis. The hypothesis pertaining to an endosymbiotic origin of the mitochondria was first proposed in the 19th century (Altman 1890) and later popularized by Margulis (1981). Today, there is strong evidence to support the hypothesis that present-day mitochondria were once free-living α-proteobacteria (Andersson et al. 1998; Lang, Gray, and Burger 1999; Ogata et al. 2001). However, the exact position of the mitochondria within the α-proteobacteria is still debated (Wu et al. 2004). A number of analyses has placed the mitochondria among the order Rickettsiales (Gupta 1995; Lang, Gray, and Burger 1999), while others propose that mitochondria are more closely related to the Rickettsiaceae family and *Rickettsia prowazekii* in particular to the exclusion of the *Ehrlichia* and *Anaplasma* genera (Karlin and Brocchieri 2000; Emelyanov 2003). On completion of the *Wolbachia pipientis* wMel genome sequence, Wu et al. (2004) reported strong support for a grouping of *Wolbachia* and *Rickettsia* as a sister group to the exclusion of the mitochondria. A recent analysis that utilized 55 individual and 31 concatenated genes encoded in *Reclinomonas americana* and *Marchantia polymorpha* mitochondrial genomes by Esser et al. (2004) concluded that based on the available data "*Rhodospirillum rubrum* comes as close to mitochondria as any α-proteobacterium investigated" (Esser et al. 2004).

Any attempt to determine which extant α-proteobacterium is the sister group of the mitochondria depends on the hypothesis that there is a robust and meaningful α-proteobacterial phylogeny. This in turn implies that within this group horizontal gene transfer (HGT) has been so infrequent that a meaningful species phylogeny can be derived. The alternative that HGT is rampant and no phylogeny exists would mean that attempting to ascertain the sister group to the mitochondria is pointless. Up until now single-gene phylogenies (especially small-subunit rRNA-based ones) have established many of the accepted relationships between bacteria. Single-gene analyses are dependent on the gene having an evolutionary history that reflects that of the entire organism. A better approach would be to combine information from many genes, and to this end, supertree analysis is a logical method to employ as it aims to combine information from many individual genes. Another approach would be to concatenate genes. Concatenation of sequence data encounters problems when there is some evidence of HGT. These problems are comparable to those encountered when reconstructing phylogenies with genes that have recombined (Schierup and Hein 2000). Conversely, because of the way in which supertrees summarize taxonomic congruence, they limit the impact of individual gene transfers on the global topology (Escobar-Paramo et al. 2004). It is then possible using supertree methods to conduct a posteriori analyses of congruence between input trees and supertree.

The primary reason for the prerequisite of a robust supertree is to account for possible HGT events. HGT among bacteria has been shown to be a major source of genetic variation, and reported incidences of HGT are continually increasing (Martin et al. 1998; Wolf, Aravind, and Koonin 1999; de la Cruz and Davies 2000; Brown 2003; Kinsella et al. 2003). By investigating congruence across ortholog-derived phylogenetic trees, we can gauge how often HGT has occurred. If we knew how much error we might expect, then we would know how much confidence we might invest in our conclusions. Recently, there have been suggestions that HGT, in general, has not been so severe that it has wiped out phylogenetic signal among closely related species (Daubin, Gouy, and Perriere 2001; Daubin, Lerat,

Key words: α-proteobacterial phylogeny, supertree, mitochondria, removing fast-evolving sites.

E-mail: james.o.mcinerney@may.ie.

Mol. Biol. Evol. 23(1):74–85. 2006
doi:10.1093/molbev/msj009
Advance Access publication September 8, 2005

and Perriere 2003), but ortholog trees contain enough conflict that early prokaryotic evolution cannot be represented effectively with a unique phylogenetic tree (Creevey et al. 2004).

In this study, we examine the nature and extent of HGT among the α-proteobacteria, we identify widely distributed genes that appear to have a history that does not include HGT, we use these genes in order to infer the sister group relationship between the mitochondria and members of the α-proteobacteria, and we discuss the implications of these inferences.

Materials and Methods

Sequence Data

The bacterial database used in this analysis consisted of 13 completely sequenced α-proteobacteria genomes, i.e., *Sinorhizobium meliloti*, *Agrobacterium tumefaciens*, *Wolbachia*, *R. prowazekii*, *Brucella melitensis*, *Caulobacter crescentus*, *Mesorhizobium loti*, *Anaplasma marginale*, *Ehrlichia ruminantium*, *Bradyrhizobium japonicum*, *Rhodopseudomonas palustris*, *Bartonella quintana*, and *Silicibacter pomeroyi* and the partial sequence data of *Novosphingobium aromaticivorans*, *Rhodobacter sphaeroides*, and *R. rubrum*. *Magnetospirillum magnetotacticum* MS-1 homologs were used in three incidences in place of *R. rubrum* (explained later). Sequence data for the partial genomes were supplied by Department of Energy Joint Genome Institute (http://www.jgi.doe.gov/).

In addition to the α-proteobacterial genomes, we used the complete mitochondrial genomes of *R. americana*, *Malawimonas jakobiformi*, and *M. polymorpha* as well as the genome sequences of *Neisseria meningitidis* MC58 and *Escherichia coli* 0157:H7 which were used as outgroups.

α-Proteobacterial Phylogeny

For supertree construction, homologous sequences were identified by performing an all-against-all search of the bacterial database using the BlastP algorithm (Altschul et al. 1997) with a cutoff expectation (E) value of 10^{-7}. Only those homologous families where every member found every other member (and nothing else) and were distributed across at least four taxa, and when present, were always in single copy, were retained. This conservative approach has been designed to minimize the inadvertent analysis of paralogs or spliced genes. In total, 539 families met these criteria. The individual protein families were then aligned using ClustalW 1.81 (Thompson, Higgins, and Gibson 1994) using the default settings. All alignments were corrected for obvious alignment ambiguity using the alignment editor Se-Al 2.0a11, and permuation-tail probability (PTP) tests were carried out to test for an evolutionary signal that was significantly better than random noise ($P < 0.01$). We found that 121 alignments failed the PTP test and the remaining 418 alignments were used for phylogenetic analysis.

Appropriate protein models were selected for each of the 418 single-gene families within the supertree data set using the software program MODELGENERATOR (http://bioinf.nuim.ie/software/modelgenerator). MODELGENERATOR chooses the best amino acid substitution model from a total of 80 possible models. Bootstrap resampling was carried out 100 times on each alignment, and using the most appropriate of the available models, phylogenetic inferences were constructed using the software program PHYML (Guindon and Gascuel 2003). The results of these analyses were summarized using the majority-rule consensus method.

Supertree Reconstruction

In order to represent the variability within the data set, the best trees from each of the 100 bootstrap replicates from all 418 single-gene families were combined as the source data for a supertree analysis (a total of 41,800 input trees). The best supertree was found following a heuristic search of tree space using the most similar supertree analysis (MSSA) (Creevey et al. 2004) as implemented in CLANN 2.0.2 (Creevey and McInerney 2005). Using CLANN, 100 bootstrap resamplings were also carried out on the 41,800 source tree data set to assess the level of support for the internal branches. We tested for the presence of signal in the data set by performing the Yet Another Permutation-Tail-Probability test (YAPTP) randomization test (Creevey and McInerney 2005), which tests the null hypothesis that congruence between the input trees is no better than random (Creevey and McInerney 2005). In order to examine the behavior of perfect data, we generated input trees that were completely congruent with the supertree (idealized data set) using the "generatetrees" command in CLANN. For a detailed explanation of the above methods, see Creevey et al. 2004.

Shimodaira–Hasegawa Tests

To assess whether differences in topology between the supertree and individual gene trees are no greater than expected by chance, we performed Shimodaira–Hasegawa (SH) tests (Shimodaira and Hasegawa 1999) using Tree-Puzzle 5.1 (Schmidt et al. 2002). Gene trees varied in size from 4 to 16 taxa; therefore, when there were fewer than 16 sequences in any input data set, the supertree was appropriately pruned so that it contained the same leaf set that was found in the individual gene trees. The underlying amino acid alignment from which the input tree was derived was used for each SH test.

Mitochondrial Relationships

Using BlastP, with a cutoff E value of 10^{-7}, the homologs that were common to *R. americana*, *M. jakobiformi*, and *M. polymorpha* mitochondria were used to find homologs among the bacteria (Altschul et al. 1997). In some instances, more than one match per genome was found. In these cases, the best match to the mitochondrial query was retained. This approach allows each α-proteobacterium to be as similar to mitochondria as possible at the level of sequence similarity and follows the procedure of Esser et al. (2004). Appropriate homologs from *cox1*, *cox2*, and *cox3* genes could not be located for *R. rubrum*; therefore, following the procedure taken by Esser et al. (2004), *M. magnetotacticum* homologs were used instead, given that

Rhodospirillum and *Magnetospirillum* are almost always well-supported sister taxa (Esser et al. 2004). In total, 28 gene families were found to contain all 21 taxa (*N. meningitidis* and *E. coli* outgroups, 3 mitochondria, and 16 α-proteobacteria). Individual mitochondrion-encoded protein families were aligned using ClustalW 1.81 (Thompson, Higgins, and Gibson 1994) using the default settings. Gaps created in the amino acid alignments were inserted into the nucleotide sequences to produce codon-based nucleotide alignments using the program putgaps (http://bioinf.nuim.ie/software/putgaps). All alignments were corrected for obvious alignment ambiguity using the alignment editor Se-Al 2.0a11.

Phylogenetic relationships for the 28 mitochondrion-encoded genes (mitochondrial data set) were reconstructed using a variety of methods. Inference of relationships was carried out using the Bayesian approach implemented in the MrBayes v3.0B4 software (Huelsenbeck and Ronquist 2001). Codon-based nucleotide alignments were analyzed with the assumption that among-site rate variation could be described using a discrete approximation to the gamma distribution (four categories of sites), and a proportion of the alignment was constrained to be invariant. The shape parameter of the gamma distribution and the hypothesized proportion of invariant sites were allowed to vary through the Markov Chain Monte Carlo (MCMC) chain. In total, four MCMC chains were run for 1.5 million generations; phylogenetic trees were sampled every 100th generation. Plots of the likelihood values revealed that the burn-in phase took at most 200,000 generations. Therefore, 200,000 trees were discarded for all alignments. Clade probabilities for each phylogeny were determined using the sumt command of MrBayes v3.0B4. Maximum likelihood (ML) phylogenies based on amino acid sequences derived from the codon-based nucleotide alignments were reconstructed using PHYML, with the model of sequence substitution and estimated parameters governing rate variation across sites determined by MODELGENERATOR. Additionally, neighbor-joining (NJ) trees based on nucleotide LogDet distances (Lockhart et al. 1994) were reconstructed using PAUP* 4.0b10 (Swofford 1998). Constant and third codon positions were removed from all alignments for the LogDet analyses. Support for relationships were determined by carrying out 100 bootstrap resamplings of the data, constructing phylogenetic trees, and summarizing the results using a majority-rule consensus method. Phylogenetic trees derived from the equivalent protein alignments were also reconstructed using ML and Bayesian methods, and in all cases nucleotide- and protein-derived trees were comparable. We could not reconstruct phylogenetic trees using the LogDet transformation on protein sequences for the majority of our 28 mitochondrion-encoded genes because there were instances where the determinants were zero and the distances undefined (see Thollesson 2004).

Concatenated Mitochondrial Data Set

A concatenated alignment was constructed consisting only of those mitochondrion-encoded genes whose inferred phylogeny is not significantly different (according to the SH test) to the proposed α-proteobacteria supertree. The reasoning behind this strategy is that these genes probably have not undergone HGT. This concatenated alignment consisted of 15 genes (table 1) and in total had 3,039 aligned amino acid positions; removal of gapped positions reduced the alignment to 1,781 amino acids. Phylogenetic relationships for the taxa within this alignment were derived from LogDet distances and the ML criterion. ML trees were reconstructed using PHYML using the appropriate model of sequence substitution according to the hierarchial likelihood ratio test implemented in MODELGENERATOR. An NJ tree based on amino acid LogDet distances (Lockhart et al. 1994) was reconstructed using LDDIST (Thollesson 2004), the fraction of invariant sites was estimated by the method of Sidow, Nguyen, and Speed (1992), and these were excluded. Support for groups on trees were determined using the bootstrap resampling technique. Splits implied by sites in the concatenated alignment were found using the LogDet distances described above and Neighbor-Net (Bryant and Moulton 2004). These were represented as a planar graph using the SplitsTree software (Huson 1998). Phylogenetic networks permit the representation of conflicting signal or alternative phylogenetic histories (Fitch 1997). When the evolutionary history of genes is not treelike, it is necessary to use phylogenetic networks (Bryant and Moulton 2004). Even when the underlying history is treelike, parallel evolution and sampling error may make it difficult to determine a unique tree. In these cases, networks provide a valuable tool for representing ambiguity or for visualizing a space of feasible trees (Bryant and Moulton 2004).

We also subjected the alignment to a spectral analysis. Spectral analysis is a quantitative method that is suited for

Table 1
Mitochondrion-Encoded Proteins Whose Phylogeny Is not Significantly Different ($P > 0.05$) to the Proposed α-Proteobacterial Supertree According to SH Test

Gene	LogDet	BP	ML	BP	Bayesian	Prob	Sites
cox3	Rick + og	72	**Rick + Bar**	80	Rick + Bar	0.95	396
rpl5	Rick	84	Rick + og	52	**Rick**	1	200
rps14	Rick	70	Rick[a]	70	**prow**	0.93	102
rps11	Rick	87	**Rick**	71	Rick	0.97	132
nad3	Rick	68	**Rick**[a]	97	Rick[a]	0.97	130
rps2	**Rick**	62	Rick	60	Rick	1	336
rps3	**Rick**	88	prow	80	prow	0.99	253
rps8	Rick + og	97	**Rick + og**	51	prow	1	133
rps12	Rick	98	**Rick**	96	Rick	1	123
rpl6	Rick	82	**Rick**	74	Rick	0.95	189
rps1	Rick	79	**Rick**	ch	Rick	1	591
rps13	Rick + og	70	**Rick**	82	Rick + og	0.97	121
rps19	Rick	73	**Rick**	96	Rick	0.99	94
nad4l	(Rick + og)	all	Rick + Bar	72	**Rick**	0.99	102
rpl16	Rick + og	ch	**Rick**	97	Rick	0.97	137

NOTE.—Rick, Rickettsiales; prow, *Rickettsia prowazeki*; Bar, *Bartonella quintana*; og, *Escherichia coli* + *Neisseria meningitidis*; BP, bootstrap support values; Prob, Bayesian clade probabilities. The optimum topology according to the SH test are in boldface. The sister taxa of the mitochondrion according to LogDet, ML, and Bayesian reconstruction methods are shown in columns 2, 4, and 6, respectively. Corresponding bootstrap support values and Bayesian clade probabilities are shown in columns 3, 5, and 7.

[a] Except *Anaplasma marginale*.

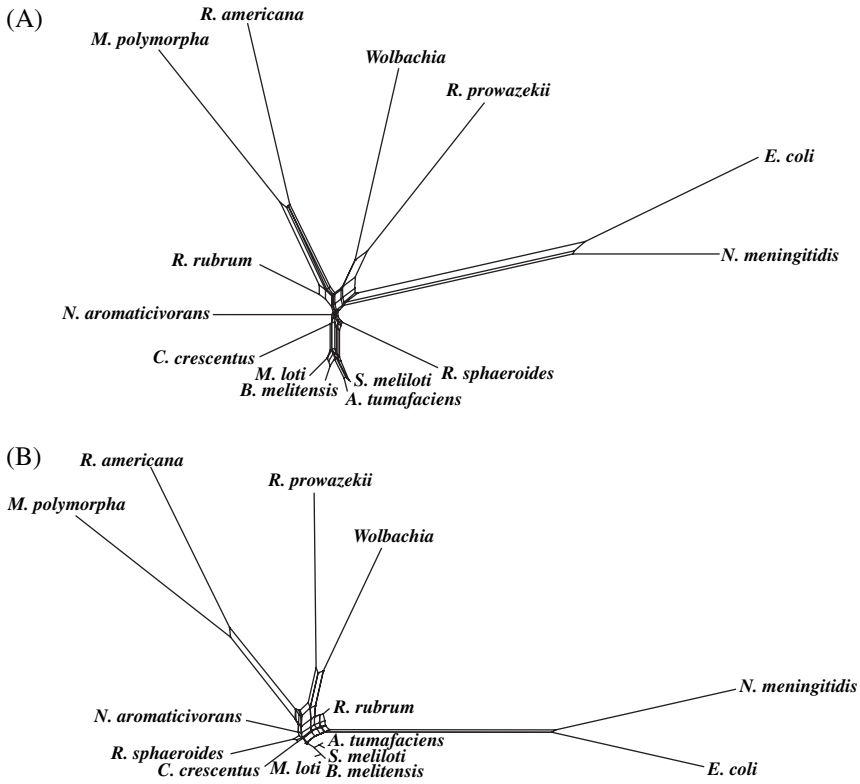

FIG. 1.—Comparison of two phylogenetic NeighborNets. The underlying alignment consists of 31 concatenated mitochondrion-encoded genes and is identical in size and content to the alignment analyzed by Esser et al. (2004). The alignment has had categories of fast-evolving sites removed until the alignment passes a χ^2 test ($P < 0.05$) for amino acid homogeneity. The top phylogenetic net (*A*) is derived from an amino acid alignment with fast-evolving sites categorized using the method of Hansmann and Martin (2000). The second phylogenetic net (*B*) has had fast-evolving sites categorized by a gamma distribution removed. Following the methodology of Esser et al. (2004), distances were determined using the LogDet transformation and invariant sites have been removed. Network (*A*) is identical to that presented by Esser et al. (2004). In network (*B*) the Rickettsiales are inferred to be the sister taxon of the mitochondrion. Differences in network topologies are solely the result of different methods used to strip fast-evolving sites.

testing alternative phylogenetic hypotheses when relationships among taxa are controversial (McCormack et al. 2000; Winchell et al. 2002). A spectral analysis on protein LogDet distances creates a complete array of bipartitions and splits in the data which includes every possible grouping among the taxa used. The strength of this method is that it works independently of any one particular tree (Lento et al. 1995).

According to Tree-Puzzle 5.1, the concatenated alignment fails the χ^2 test for homogenous amino acid composition ($P < 0.05$). The highly variable sites were iteratively removed from the concatenated alignment using two different methods in order to give the largest possible alignment in which all sequences passed the χ^2 test for homogeneity of amino acid composition. The reason for utilizing two methods was because when we analyzed the data of Esser et al. (2004) we found an alternative phylogenetic network to the one they proposed (fig. 1). The reason for this alternative inference was due to the fact that we had used a different method to remove fast-evolving sites.

The first method described by Hansmann and Martin (2000) scores sites depending on the number of amino acid character states that are found at that site. This effectively assumes that there is a star phylogeny with equal branch lengths and is the method used by Esser et al. (2004). Sites with high scores were iteratively removed until a compositionally homogenous alignment of 574 amino acids was found. The second method is implemented in Tree-Puzzle 5.1 and assumes that there are eight categories of sites. Rate variation across these sites is assumed to follow a discrete gamma distribution. Again, fast-evolving categories of sites were iteratively removed until an alignment of 651 amino acid positions that passed the χ^2 test for homogeneity of amino acid composition was found for the 15-gene data set. Phylogenetic hypotheses for the homogeneous alignments were generated using the ML framework described earlier.

To account for problems associated with deep level relationships, the amino acid alignment was recoded into the six Dayhoff groups: C, STPAG, NDEQ, HRK, MILV, and FWY. This is based on the notion that amino acid substitutions within the six groups will be common and noisy, while changes between groups will be rarer and so have less saturation (Hrdy et al. 2004). The recoded alignment was analyzed using the Bayesian criterion implemented in MrBayes v3.0B4 (Huelsenbeck and Ronquist 2001). The model used has a 6×6 general time-reversible rate matrix. Among-site variation in evolutionary rate was modeled with an allowance for a proportion of sites to be invariable and four categories of variable sites with their rates modeled by a discrete approximation to the gamma distribution. The analysis used an MCMC chain that ran for 2 million

generations sampled every 100th generation; the first 5,000 samples were discarded as the burn-in. Support for groups on the tree were determined using the *sumt* command of MrBayes, which produces a majority-rule consensus tree generated from bipartitions found most frequently during the MCMC runs.

Results

α-Proteobacteria Supertree

To test whether there is a meaningful α-proteobacterial phylogeny, a data set consisting of 16 α-proteobacterial genomes or 60,640 individual coding sequences was utilized. We identified 418 single-gene families that met our criteria of being present in single copy in at least four genomes and that have phylogenetic signal according to the PTP test. Amino acid sequences were aligned resulting in a combined length of 140,314 aligned positions. Bootstrap resampling followed by phylogenetic inference resulted in 100 ML trees for each family. Using these 41,800 trees as input data for supertree analysis, heuristic searches of supertree space were carried out and the optimal supertree was identified (fig. 2). One hundred bootstrap resamplings of the input trees was carried out, and the support values of the internal branches were placed on the supertree in figure 2.

How Compatible Are the Input Trees?

An idealized data set was created where the topology of the input trees was changed to make these trees completely compatible with the supertree (ideal data), and also the input trees were completely randomized 100 different times in order to carry out the YAPTP test. The scores of 10,000 random supertrees were then obtained based on the real (unperturbed), ideal, and randomized input trees, and the distributions of these supertree scores were assessed. In addition, we sought to determine how many perturbations per tree were needed to give the ideal data the kinds of characteristics that we observe in the real data and how many perturbations might be required to completely randomize the input trees. In this way, we determined how close the real data was to ideal data and how close it was to random data.

The range of supertree scores for the real data varied from 33,196 (best) to 92,027 (worst) (fig. 2). In general, the scores of the worst supertrees from all three data sets were similar. The distribution of the scores of the optimal supertrees from the YAPTP test is centered on 70,400 (±300). This indicates that congruence across the input trees is greater than expected by random chance (Creevey et al. 2004). The ideal data generated in this study represents a scenario where there is complete compatibility between the input trees and supertree in figure 2. The distribution of tree scores for this ideal data set was heavily skewed to the left, with a skewness value of −0.8. The real data were slightly more skewed to the left with a skewness value of −0.11. The randomized data had a skewness value of −0.02, which is close to zero, as expected. All three distributions had a standard deviation (SD) of 0.02.

We then took the ideal data set and applying the subtree pruning regrafting (SPR) method on the input trees

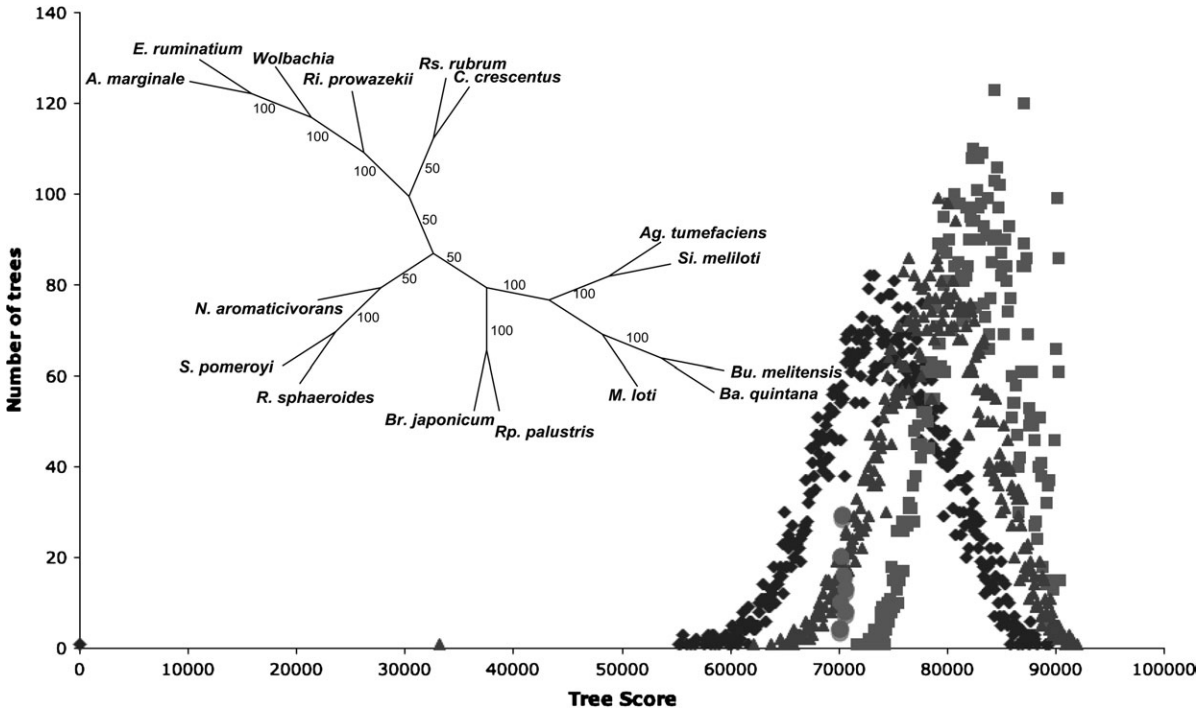

FIG. 2.—The shape of supertree space. In diamonds are the scores for 10,000 random supertrees when scored against an ideal source tree data set. The best tree from the ideal distribution has a score of 0 and the skewness of the ideal distribution is −0.08. The triangles are the scores of 10,000 random supertrees when compared to the real source tree data set. The skewness of this distribution is −0.11. Heuristic searches of supertree space using the real source data revealed that the best supertree had a score of 33,196, and when bootstrapped, nearly every internal branch of the tree had 100% bootstrap support value. The squares are the scores of 10,000 random supertrees when compared to a randomized set of source trees, this distribution had a skewness of −0.02. The SD of the skewness for the ideal, real, and randomized data was 0.02. The circles demonstrate the results of 100 YAPTP tests.

(equivalent to introducing an HGT event), we incrementally perturbed the ideal data set. These SPR events were introduced, and at each stage the best supertree was found and its score for the data set assessed until the score, according to the MSSA method, was the same as either the real or randomized data. This gave an indication of the number of perturbations required to make ideal data as noisy as the real (or random) data set. Each of these simulations was carried out 100 times. From the 100 simulations, the mean number of SPR events per tree required to transform ideal data into the equivalent of the real data set was 1.6 (with an SD of 0.07). To transform the ideal data into data with an equivalent level of incongruence as random data required 5.0 HGT events per tree (with an SD of 0.28).

Taken together, it can be seen that the real data are significantly different than random data (according to the YAPTP test) and by two measures it is more similar to ideal data than it is to random data.

Of the 418 input trees, SH tests indicated that 98 (23%) whose topology could not be induced by the appropriately pruned supertree, described their underlying alignments significantly better ($P < 0.05$) than did the supertree. The differences in topology could not be explained by chance alone. For the remaining 320 trees, there was no significant difference between the proposed supertree and underlying tree (24% of the these are identical to the proposed supertree). HGT events and hidden paralogies are among the possible reasons for the 98 families having an evolutionary history that is significantly different to that implied by the supertree. In an effort to quantify the types of genes that display significantly different topologies to the supertree, we categorized genes as being either operational or informational in function based on the criteria used by Jain, Rivera, and Lake (1999). According to the complexity hypothesis (Jain, Rivera, and Lake 1999), we would expect to see operational genes being exchanged at a higher rate than informational genes. Informational genes are members of large complex systems thereby making horizontal transfer of these gene products less probable (Jain, Rivera, and Lake 1999). We found 87 informational genes and 96 operational genes, the remaining genes have yet to be assigned a function (i.e., conserved hypothetical) or belong to a category that is neither operational nor informational. In total, 18 of the 87 informational genes (~20%) were found to have a topology that is significantly different to the proposed supertree, while 50 of the 96 operational gene (~51%) trees differ significantly to the supertree according to the SH test. Assuming that systematic or stochastic errors in phylogenetic reconstruction of the input trees is not function dependent and also that the supertree is truly representative of the relationships among the α-proteobacteria, this would indicate that the operational genes within this data set experience significantly ($P < 0.001$) more horizontal transfers or other anomalous events than informational genes.

Mitochondrial Origin

Individual Mitochondrion-Encoded Genes

The overlap between the 28 mitochondrion-encoded genes and the 301 genes that are compatible with the supertree was 15 genes. Phylogenetic hypotheses based upon these 15 mitochondrion-encoded genes were constructed for the codon-based nucleotide alignments using LogDet, ML, and Bayesian methods. All three methods revealed a general consensus regarding the possible sister group of mitochondria (table 1). The majority of these 15 gene trees placed the members of the Rickettsiales as the sister group of the mitochondria. We carried out a variety of SH tests using the Whelan and Goldman (2000) model of amino acid substitution. For each gene, the underlying alignment and the three proposed topologies derived from the above three methods were used for each gene family. An additional "constrained" tree where *R. rubrum* was placed as a sister group to the mitochondria (recreated using MrBayes) was also included in the analysis. The reasoning behind this approach is that *R. rubrum* has been suggested as a possible sister group to the mitochondria (Esser et al. 2004). According to the SH test, the constrained tree was significantly ($P > 0.05$) worse than the three alternative competing topologies in all cases (results not shown) and was therefore not considered for further analysis. Examining the optimum topologies according to the SH test reveals that in total 11 trees placed the Rickettsiales as the sister group to the mitochondria. One tree placed the Rickettsiales (except for *A. marginale*) as the sister group to the mitochondrion, another inferred that the Rickettsiales combined with *B. quintana* were the sister group, another tree placed the Rickettsiales and the outgroup together, and then another split separated this group and the mitochondria from the rest of the taxa. The final tree placed the mitochondria with *R. prowazekii* as sister taxa (table 1).

Concatenated Alignment

Concatenation of the 15 mitochondrion-encoded genes of interest gave an alignment of 3,039 amino acids in length or 1,781 amino acids when positions containing gaps were removed. Concatenation of sequence data has the benefit of reducing stochastic effects related to short sequence length. According to the amino acid compositional heterogeneity test, there was severe amino acid bias within the data set; therefore, the assumption of stationary amino acid frequencies (Gu and Li 1996) was violated. To account for these biases, the LogDet transformation (Lockhart et al. 1994) was used to infer the relationships among the α-proteobacteria and the mitochondria for the data set containing no gaps. NJ trees constructed using LogDet distances derived from the amino acid alignment place the Rickettsiales as the sister group to the mitochondria with 99% bootstrap support (fig. 3*A*). Phylogenetic hypotheses were also inferred using the program PHYML using the appropriate model of sequence substitution selected by MODELGENERATOR. The ML topology was identical to the LogDet topology and branch lengths and bootstrap support values were also comparable (fig. 5*A*). The NeighborNet of LogDet protein distances for this alignment also places the Rickettsiales as the closest α-proteobacterial group to the mitochondria (fig. 3*B*). Comparing the supertree in figure 2 to the tree built from the concatenated sequence data (fig. 3*A*), we see that the main orders such as the Rickettsiales, the Rhizobiales, and the Rhodobacterales are strongly supported clades in both cases. In our supertree,

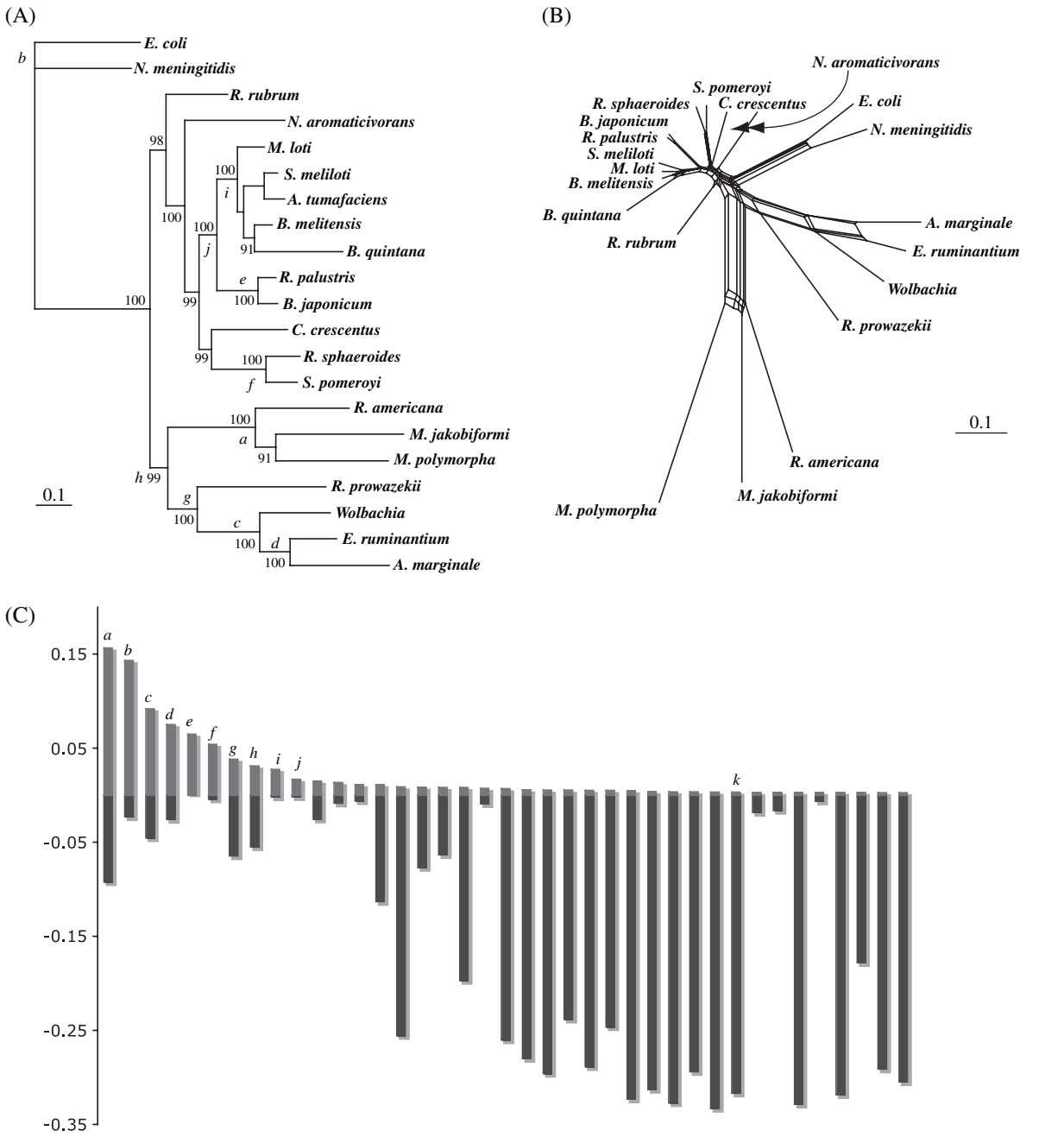

FIG. 3.—The ML phylogeny for the concatenated alignment is shown in (*A*). Bootstrap supports are shown for selected nodes. The inferred topology is very similar to an inference based on LogDet distances (see fig. 5*A*). The Rickettsiales group beside the mitochondrion with 99% bootstrap support. A NeighborNet of the concatenated alignment is also displayed in (*B*); this phylogenetic net is based on protein LogDet distances. The Rickettsiales appear to share a split with the mitochondrion. A lento plot (*C*) that utilizes the aforementioned LogDet distances for the concatenated alignment is also displayed. The top 40 supported splits (excluding individual taxa) are displayed. Bars above the *x* axis represent frequency of support for each split. Bars below the *x* axis represent the sum of all conflicts against the corresponding split above the *x* axis. Letters above columns represent particular splits in the data, and these have been mapped onto the phylogenetic tree for display purposes. Attention is drawn to split *h* which includes the mitochondrion and the Rickettsiales and split *k* which includes *Rhodospirillum rubrum* and the mitochondrion. The conflict for split *k* far outweighs any support.

C. crescentus and *R. rubrum* are grouped as sister taxa with poor support (50% bootstrap support). *Novosphingobium aromaticivorans* is grouped beside the Rhodobacterales, also with 50% bootstrap support. These findings are not entirely surprising as these bacteria are lone members of the orders Caulobacterales, Rhodospirillales, and Sphingomonadales, respectively, in our tree. Contrastingly, the concatenated alignment places these orphan taxa on the tree with very high levels of support. For example, *C. crescentus* is grouped beside the Rhodobacterales with 99% bootstrap support.

A spectral analysis of LogDet distances was performed using Spectrum (http://taxonomy.zoology.gla.ac.uk/~mac/spectrum) to further investigate the phylogenetic signal

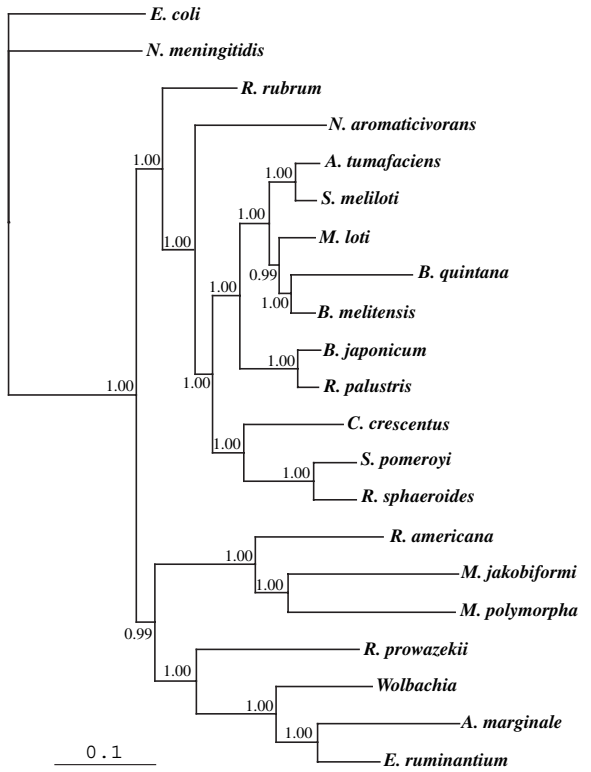

FIG. 4.—Inferred phylogeny shows a Bayesian consensus tree of the 15-gene concatenated alignment. The analysis uses a general time-reversible substitution matrix estimated from amino acid sequences recoded into the six Dayhoff groups. Among-site rate heterogeneity was modeled using a four-category gamma correction with a fraction of invariant sites. The analysis used the MCMC strategy from MrBayes; parameters such as the composition and substitution rate matrix were free. Posterior probabilities for selected branches are shown at nodes.

present in our concatenated alignment. Lento et al. (1995) have shown that a spectral analysis of LogDet distances seems to reduce the amount of random noise compared with a normal character-based spectral analysis. As Spectrum can only accommodate 20 taxa, we removed *A. tumefaciens* for this part of the analysis. With 20 taxa, the total number of possible splits is 2^{20-1} or 542,288. Of these, 101 splits are supported by some evidence and the 40 most highly supported are represented by the bars shown in the top half of figure 3*C*. Attention is drawn to the split marked *h*, the taxa within this split include the Rickettsiales and the mitochondria. The split marked *k* represents the grouping of the mitochondria and *R. rubrum* as sister taxa. This relationship is not observed in either our phylogenetic tree (fig. 3*A*) or phylogenetic network (fig. 3*B*). Furthermore, the degree of conflict for this split far outweighs the level of support. For completeness, we also performed the above analyses on the alignment with gap positions still present. The results of these analyses were in agreement with our observations in the gapless alignment.

Looking at our phylogenetic tree and network (fig. 3*A* and 3*B*), it is obvious that the taxa with long internal branches cluster together—the mitochondria, the Rickettsiales, and the outgroups. We carried out two tests to investigate whether this grouping is a systematic error (long-branch attraction, also known as the Felsenstein zone) or the grouping is real. The first test simply involved the systematic analyses of several data sets that only had a single long-branch taxon included. This test was to see if the long-branch taxa were still attaching to the tree in the same place, even when other long-branch taxa were absent. In all three cases, the long-branched taxa attached to the *R. rubrum* branch with high bootstrap support (fig. 5).

The second test of long-branch attraction involved the removal of the long-branch taxa from the tree (see fig. 6) and their pseudorandom reinsertion 100 times. The taxa were inserted into the tree as long as the point of insertion was not beside another long-branch taxon. This resulted in the construction of phylogenetic trees with interspersed long and short branches, but where no long-branch taxa were neighbors. Once the taxon-insertion process was completed, the concatenated alignment was used to determine appropriate branch lengths for all taxa. Naturally, the branches were quite long for the inserted taxa. We then simulated the evolution of a set of sequences for this artificial tree using Seq-Gen (Rambaut and Grassly 1997), and using the same methods of analyses that we have used previously on the concatenated data set, we reanalyzed the simulated sequences to see how often the long-branch taxa were incorrectly placed beside each other (i.e., what was the strength of the long-branch attraction in a data set of this nature). We found that even though there could be slight discrepancies between the tree used to simulate the data and the tree derived from the simulated data, there was never a single case where the long-branch taxa were clustered together. Although this test cannot be taken as definitive proof that long-branch attraction alone is not responsible for their clustering in the real data set, we at least show that for data sets with similar characteristics (composition, degree of overall evolutionary change, sequence length, mixture of long, and short branches) this effect is not severe enough to result in completely artifactual topologies.

We also recoded the amino acid data into the six groups of chemically related amino acids and analyzed the data using the Bayesian criteria. This approach has the effect of shortening long branches and homogenizing the amino acid composition among sequences (Hrdy et al. 2004). From this analysis, we found that the mitochondrion and Rickettsiales are again strongly supported (fig. 4).

Removing the fast-evolving sites from the 15-gene concatenated alignment using the method of Hansmann and Martin (2000) until the reduced alignment passes the amino acid compositional heterogeneity test resulted in an alignment of 574 amino acids in length. The NeighborNet of LogDet protein distances for this shortened alignment places the Rickettsiales as the closest α-proteobacterium group to the mitochondria. Bootstrap resampling was employed to assess levels of support for groups. This showed that the Rickettsiales and the mitochondria formed a clade with relatively low bootstrap support (65%, results not shown). Using the gamma distribution to model site-rate variation in evolutionary rate and removing fast-evolving sites until the data passed the χ^2 test of compositional homogeneity resulted in an alignment of 651 amino acids in length. The NeighborNet of LogDet protein distances for this shortened alignment places the Rickettsiales beside the mitochondria as they share a split; a bootstrap analysis

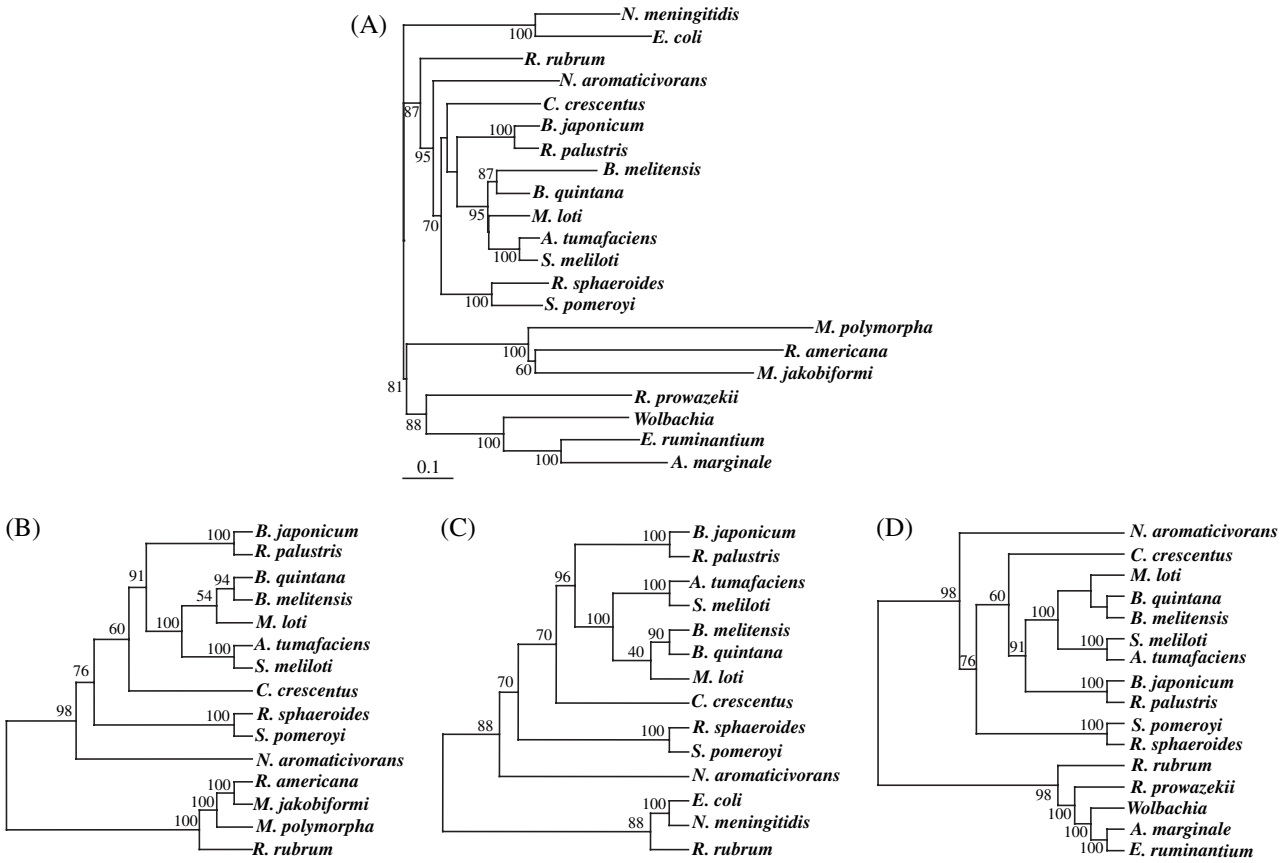

FIG. 5.—Main phylogram (*A*) shows the inferred topology of the protein concatenated alignment; distances were derived from LogDet Transformation. Two of the three long branches are removed from the alignment and the tree is redrawn. The grouping of the remaining long branch to the free-living α-proteobacteria is noted; trees are rooted around the remaining long branch for convenience. (*B*) The mitochondrion groups beside *Rhodospirillum rubrum*. (C) The outgroups sit beside *R. rubrum*. (*D*) The Rickettsiales group beside *R. rubrum*.

again revealed weak support for this finding (52%, results not shown). It should be pointed out that the topologies for the full and shortened alignment are identical; bootstrap support values for the shorter alignments are lower, however. As we systematically removed categories of fast-evolving sites from the alignment, the bootstrap support values decreased across the tree but the overall branching structure never changed.

Discussion

To investigate the sister group relationship between the α-proteobacteria and the mitochondria, it is imperative that there is a meaningful α-proteobacterial species phylogeny. We use a supertree-based approach for simultaneous determination of species-level phylogeny and analysis of HGT for 16 α-proteobacteria derived from 418 gene trees. The best supertree receives a better score (33,196) than any tree from the randomization procedure (YAPTP) which means the individual gene trees have a significantly higher level of agreement with each other than would be expected by chance. The score distributions from real and idealized gene trees (as judged by skewness values) are remarkably similar. It requires an average of 1.6 SPR branch swaps per tree to give ideal data a similar score to the real data, whereas it requires 5.0 SPR swaps per tree to convert ideal data into random data. We find that when we prune the optimal supertree to the size of each individual gene tree, 77% of the topologies are either identical to the supertree or, where the topologies are different, this difference is not significant (i.e., according to the SH test, these differences are no greater than we might expect from sampling effects, $P < 0.05$). This is compatible with the hypothesis that for these data sets the topology induced by the supertree in figure 1 can adequately describe their evolution. Conversely, this finding is incompatible with a hypothesis that HGT has completely randomized genome evolution. The remaining 23% of trees that show a significant difference could be doing so as a result of systematic biases in reconstructing phylogenetic relationships for individual gene trees, HGT, or hidden paralogy. It is unlikely, though, that HGT accounts for all 23% of the genes that are incompatible, given the difficulty of inferring orthology with absolute certainty. We have shown that a significantly (χ^2 test, $P < 0.001$) higher percentage of operational genes disagree significantly with the supertree, a finding in agreement with the complexity hypothesis. A bootstrap analysis of the input orthologous trees also lends support to a robust α-proteobacteria phylogeny as the majority of splits are strongly supported. Taken together, we suggest that the α-proteobacteria have a meaningful phylogeny.

Recently, *R. rubrum* has been suggested to be the sister group of the mitochondria based on a concatenated

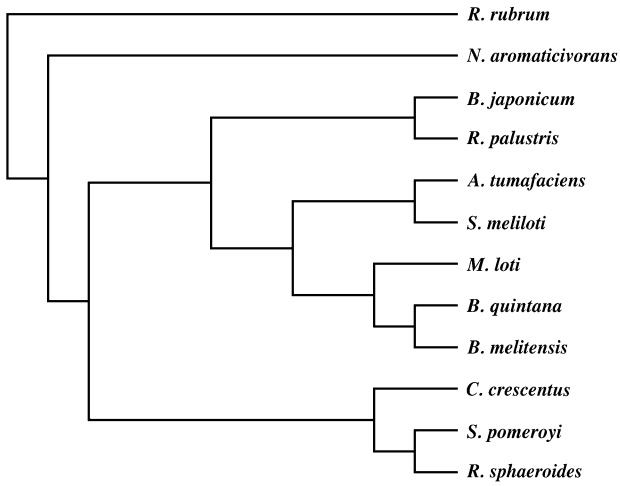

FIG. 6.—This phylogram depicts a reduced constrained tree. This tree only contains the short-branched taxa. Long-branched taxa are randomly attached to the constrained tree. Relative branch lengths for the new enlarged constrained tree were determined in PHYML using the original concatenated alignment. This procedure was repeated 100 times to give 100 random constrained trees.

alignment of 31 mitochondrion-encoded genes (Esser et al. 2004). Our analysis utilized 15 mitochondrial genes that have a phylogeny that is not significantly different (SH test, $P < 0.05$) to the proposed supertree. Using this approach, we have minimized the obfuscating influence of including genes whose history might not be compatible with one another. Our data point to the conventional sister group relationship of Rickettsiales and the mitochondrion. For completeness, we performed our analyses on a larger alignment containing all 28 mitochondrion-encoded proteins, and the conclusions of these analyses agree with the findings of our initial analysis (results not shown). When we reexamined the 31-gene sequence alignment of Esser et al. (2004), we found alternative topologies depending on what method was utilized to remove fast-evolving sites. When fast-evolving sites were stripped in the same manner as their analysis, we recreated identical results to those authors (fig. 1*A*). Alternatively, when fast-evolving sites categorized by a gamma distribution are removed, the Rickettsiales and not *R. rubrum* are inferred to be the sister group to the mitochondrion (fig. 1*B*). These results raise concerns regarding the method used to strip out fast-evolving sites. The method described by Hansmann and Martin (2000) strips sites out based on the number of character states found at a particular site, while the discrete gamma method uses ML estimation to place sites into different site categories. Clearly, these methods categorize sites very differently, and the merits of each will need further study. In our analysis, we have more extensive taxon sampling, and also we erred on the side of caution and removed fast-evolving sites using both methods described above until we found alignments that passed a χ^2 test of amino acid compositional heterogeneity. Phylogenetic inferences of both resultant alignments yielded similar conclusions in that the Ricketsiales are inferred to be the sister taxon of the mitochondrion with varying degrees of support (65% and 52% bootstrap support, respectively). Higher levels of support for these inferences were observed when categories of fast-evolving sites were present in the alignment. As the fast-evolving sites were removed and the alignments became shorter with a higher proportion of invariant and slow-evolving sites, we lost phylogenetic signal.

Further evidence that the mitochondria share a common ancestor with the Rickettsiales to the exclusion of the other taxa comes from individual mitochondrial proteins. All 15 individual trees indicate a sister group relationship between the mitochondria and at least one of the bacteria from this order. It has been suggested that phylogenetic inferences based on highly conserved mitochondrial proteins such as cytochrome oxidase subunits *cox3* have been suggested as better indicators of relationships than fast-evolving mitochondrial genes that may infer incorrect relationships as an artifact of long-branch attraction (Lang, Gray, and Burger 1999). Taking the inferred evolutionary history of *cox3* alone, we can see that this gene would suggest that the mitochondria and Rickettsiales are each others' closest relatives (table 1) with bootstrap support for this relationship being relatively strong (~80%).

A better understanding of the origin and evolution of the mitochondrion will only be possible when a more representative sample of the α-proteobacteria is fully sequenced. It is entirely possible that there are many more mitochondrion-like bacterial species yet to be identified. Currently, only 0.4% of all existing bacterial species have been identified and formally described let alone sequenced. Another consideration in such an analysis must be the drawbacks associated with the use of eukaryote mitochondrial DNA and the high rate of nucleotide substitution associated with these genes. These problems can be exacerbated when we use mitochondrion sequences in studies with bacterial homologs that have divergence times of approximately 1.5–2 billion years (Sicheritz-Ponten, Kurland, and Andersson 1998). The end result of such an analysis can be severely affected by long-branch attraction. Throughout this analysis, we were mindful of these problems and endeavored to take steps such as the use of LogDet distance matrices to help prevent misleading results, and we investigated the position of long-branched taxa in the absence of other long branches. Another approach that is analogous to a transversion analysis of nucleotide sequences is to recode the input alignment into the six Dayhoff categories and subsequently infer their phylogeny. This approach has the advantage that we could use a Bayesian approach and optimize a general time-reversible substitution matrix, rather than assuming an ad hoc amino acid substitution matrix. We investigated such an approach on the 15-gene alignment (gapped positions removed) and found that the grouping of the mitochondrion and Rickettsiales are again strongly supported.

Esser et al. (2004) make the point that the overall fermentative physiology of *R. rubrum* is quite similar to eukaryotes that lack mitochondria or contain anaerobic mitochondria. Similarly, it can also be argued that there is a strong affinity between mitochondria and *Rickettsia* for the genes coding for components of the Krebs cycle (Andersson et al. 1998), the respiratory chain (Gray, Burger, and Lang 1999), and the translation system (Sicheritz-Ponten, Kurland, and Andersson 1998).

A close look at our phylogenetic trees indicates that although the Rickettsiales are usually placed beside the mitochondria, *R. rubrum* is almost always only one branch-swap away from the mitochondria. Therefore, it is not implausible that the mitochondrion endosymbiont possessed features that resemble both *R. rubrum* and its fermentative physiology and the modern-day Rickettsiales and their respiratory chain and Krebs cycle.

Acknowledgments

We wish to thank the comments of three anonymous reviewers. We would also like to thank Gayle Philip and Jennifer Commins for proofreading the manuscript. D.A.F. was supported by a Higher Education Authority grant (Programme for Research in Third Level Institutes—PRTLI Cycle II) and C.J.C. by PRTLI Cycle III.

Literature Cited

Altman, R. 1890. Die Elementarorganismen und Ihre Beziehungen Zur Den Zellen. Verlag von Veit, Keipzig, Germany.

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. (Online) **25**:3389–3402.

Andersson, S. G. E., A. Zomorodipour, J. O. Andersson, T. Sicheritz-Ponten, C. M. U. Alsmark, R. M. Podowski, A. K. Näslund, A.-C. Eriksson, H. H. Winkler, and C. G. Kurland. 1998. The genome sequence of Rickettsia prowazekii and the origin of mitochondria. Nature **396**:133–140.

Brown, J. R. 2003. Ancient horizontal gene transfer. Nat. Rev. Genet. **4**:121–132.

Bryant, D., and V. Moulton. 2004. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. Mol. Biol. Evol. **21**:255–265.

Creevey, C. J., D. A. Fitzpatrick, G. K. Philip, R. J. Kinsella, M. J. O'Connell, M. M. Pentony, S. A. Travers, M. Wilkinson, and J. O. McInerney. 2004. Does a tree-like phylogeny only exist at the tips in the prokaryotes? Proc. R. Soc. Lond. B. Biol. Sci. **271**:2551–2558.

Creevey, C. J., and J. O. McInerney. 2005. Clann: investigating phylogenetic information through supertree analyses. Bioinformatics **21**(3):390–392.

Daubin, V., M. Gouy, and G. Perriere. 2001. Bacterial molecular phylogeny using supertree approach. Genome Inform. Ser. Workshop Genome Inform. **12**:155–164.

Daubin, V., E. Lerat, and G. Perriere. 2003. The source of laterally transferred genes in bacterial genomes. Genome Biol. **4**:R57.

de la Cruz, I., and I. Davies. 2000. Horizontal gene transfer and the origin of species: lessons from bacteria. Trends Microbiol. **8**:128–133.

Emelyanov, V. V. 2003. Mitochondrial connection to the origin of the eukaryotic cell. Eur. J. Biochem. **270**:1599–1618.

Escobar-Paramo, P., A. Sabbagh, P. Darlu, O. Pradillon, C. Vaury, E. Denamur, and G. Lecointre. 2004. Decreasing the effects of horizontal gene transfer on bacterial phylogeny: the *Escherichia coli* case study. Mol. Phylogenet. Evol. **30**:243–250.

Esser, C., N. Ahmadinejad, C. Wiegand et al. (12 co-authors). 2004. A genome phylogeny for mitochondria among alpha-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. Mol. Biol. Evol. **21**:1643–1660.

Fitch, W. M. 1997. Networks and viral evolution. J. Mol. Evol. **44**(Suppl. 1):S65–S75.

Gray, M. W., G. Burger, and B. F. Lang. 1999. Mitochondrial evolution. Science **283**:1476–1481.

Gu, X., and W. H. Li. 1996. Bias-corrected paralinear and LogDet distances and tests of molecular clocks and phylogenies under nonstationary nucleotide frequencies. Mol. Biol. Evol. **13**:1375–1383.

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

Gupta, R. S. 1995. Evolution of the chaperonin families (Hsp60, Hsp10 and Tcp-1) of proteins and the origin of eukaryotic cells. Mol. Microbiol. **15**:1–11.

Hansmann, S., and W. Martin. 2000. Phylogeny of 33 ribosomal and six other proteins encoded in an ancient gene cluster that is conserved across prokaryotic genomes: influence of excluding poorly alignable sites from analysis. Int. J. Syst. Evol. Microbiol. **50**(Pt. 4):1655–1663.

Hrdy, I., R. P. Hirt, P. Dolezal, L. Bardonova, P. G. Foster, J. Tachezy, and T. M. Embley. 2004. Trichomonas hydrogenosomes contain the NADH dehydrogenase module of mitochondrial complex I. Nature **432**:618–622.

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

Huson, D. H. 1998. SplitsTree: analyzing and visualizing evolutionary data. Bioinformatics **14**:68–73.

Jain, R., M. C. Rivera, and J. A. Lake. 1999. Horizontal gene transfer among genomes: the complexity hypothesis. Proc. Natl. Acad. Sci. USA **96**:3801–3806.

Karlin, S., and L. Brocchieri. 2000. Heat shock protein 60 sequence comparisons: duplications, lateral transfer, and mitochondrial evolution. Proc. Natl. Acad. Sci. USA **97**:11348–11353.

Kinsella, R. J., D. A. Fitzpatrick, C. J. Creevey, and J. O. McInerney. 2003. Fatty acid biosynthesis in *Mycobacterium tuberculosis*: lateral gene transfer, adaptive evolution, and gene duplication. Proc. Natl. Acad. Sci. USA **100**:10320–10325.

Lang, B. F., M. W. Gray, and G. Burger. 1999. Mitochondrial genome evolution and the origin of eukaryotes. Annu. Rev. Genet. **33**:351–397.

Lento, G. M., R. E. Hickson, G. K. Chambers, and D. Penny. 1995. Use of spectral analysis to test hypotheses on the origin of pinnipeds. Mol. Biol. Evol. **12**:28–52.

Lockhart, P., M. Steel, M. Hendy, and D. Penny. 1994. Recovering evolutionary trees under a more realistic model of sequence. Mol. Biol. Evol. **11**:605–612.

Margulis, L. 1981. Symbiosis in cell evolution. 1st edition. Freeman, New York.

Martin, K., G. Morlin, A. Smith, A. Nordyke, A. Eisenstark, and M. Golomb. 1998. The tryptophanase gene cluster of *Haemophilus influenzae* type b: evidence for horizontal gene transfer. J. Bacteriol. **180**:107–118.

McCormack, G. P., B. F. Keegan, J. O. McInerney, and R. Powell. 2000. Re: spectral analysis of echinoderm small subunit ribosomal RNA gene sequence data. Mol. Phylogenet. Evol. **15**:327–329.

Ogata, H., S. Audic, P. Renesto Audiffren et al. (8 co-authors). 2001. Mechanisms of evolution in *Rickettsia conorii* and *R. prowazekii*. Science **293**:2093–2098.

Rambaut, A., and N. C. Grassly. 1997. Seq-Gen: an application for the monte carlo simulation of DNA sequence evolution along phylogenetic trees. CABIOS **13**:235–238.

Schierup, M. H., and J. Hein. 2000. Consequences of recombination on traditional phylogenetic analysis. Genetics **156**: 879–891.

Schmidt, H. A., K. Strimmer, M. Vingron, and A. von Haeseler. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics **18**:502–504.

Shimodaira, H., and M. Hasegawa. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol. Biol. Evol. **16**:1114–1116.

Sicheritz-Ponten, T., C. G. Kurland, and S. G. Andersson. 1998. A phylogenetic analysis of the cytochrome b and cytochrome c oxidase I genes supports an origin of mitochondria from within the Rickettsiaceae. Biochim. Biophys. Acta **1365**:545–551.

Sidow, A., T. Nguyen, and T. P. Speed. 1992. Estimating the fraction of invariable codons with a capture-recapture method. J. Mol. Evol. **35**:253–260.

Swofford, D. L. 1998. PAUP*: phylogenetic analysis using parsimony (*and other methods). Sinauer Associates, Mass.

Thollesson, M. 2004. LDDist: a Perl module for calculating Log-Det pair-wise distances for protein and nucleotide sequences. Bioinformatics **20**:416–418.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. (Online) **22**:4673–4680.

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

Winchell, C. J., J. Sullivan, C. B. Cameron, B. J. Swalla, and J. Mallatt. 2002. Evaluating hypotheses of deuterostome phylogeny and chordate evolution with new LSU and SSU ribosomal DNA data. Mol. Biol. Evol. **19**:762–776.

Woese, C. R. 1987. Bacterial evolution. Microbiol. Rev. **51**: 221–271.

Woese, C. R., E. Stackebrandt, T. J. Macke, and G. E. Fox. 1985. A phylogenetic definition of the major eubacterial taxa. Syst. Appl. Microbiol. **6**:143–151.

Wolf, Y. I., L. Aravind, and E. V. Koonin. 1999. Rickettsiae and Chlamydiae: evidence of horizontal gene transfer and gene exchange. Trends Genet. **15**:173–175.

Wu, M., L. V. Sun, J. Vamathevan et al. (27 co-authors). 2004. Phylogenomics of the reproductive parasite *Wolbachia pipientis* wMel: a streamlined genome overrun by mobile genetic elements. PLoS Biol. **2**:E69.

Martin Embley, Associate Editor

Accepted August 30, 2005