Does a tree-like phylogeny only exist at the tips in the prokaryotes?

Christopher J. Creevey¹, David A. Fitzpatrick¹, Gayle K. Philip¹, Rhoda J. Kinsella¹, Mary J. O’Connell¹, Melissa M. Pentony¹, Simon A. Travers¹, Mark Wilkinson² and James O. McInerney¹,²*

¹Bioinformatics and Pharmacogenomics Laboratory, Department of Biology, National University of Ireland Maynooth, Maynooth, Co. Kildare, Ireland
²Department of Zoology, The Natural History Museum, Cromwell Road, London SW7 5BD, UK

The extent to which prokaryotic evolution has been influenced by horizontal gene transfer (HGT) and therefore might be more of a network than a tree is unclear. Here we use supertree methods to ask whether a definitive prokaryotic phylogenetic tree exists and whether it can be confidently inferred using orthologous genes. We analysed an 11-taxon dataset spanning the deepest divisions of prokaryotic relationships, a 10-taxon dataset spanning the relatively recent γ-proteobacteria and a 61-taxon dataset spanning both, using species for which complete genomes are available. Congruence among gene trees spanning deep relationships is not better than random. By contrast, a strong, almost perfect phylogenetic signal exists in γ-proteobacterial genes. Deep-level prokaryotic relationships are difficult to infer because of signal erosion, systematic bias, hidden paralogy and/or HGT. Our results do not preclude levels of HGT that would be inconsistent with the notion of a prokaryotic phylogeny. This approach will help decide the extent to which we can say that there is a prokaryotic phylogeny and where in the phylogeny a cohesive genomic signal exists.

Keywords: phylogenetic supertrees; prokaryotic phylogeny; taxonomic congruence; phylogenomics; molecular evolution

1. INTRODUCTION

Small subunit ribosomal RNA (SSU rRNA) gene sequences have revolutionized our understanding of prokaryote phylogeny, but it is unclear to what extent ‘universal trees’ based on these data also reflect phylogenetic histories of other genes. The recent sequencing of three strains of Escherichia coli revealed that only 39.2% of proteins are common to all three strains (Blattner et al. 1997; Hayashi et al. 2001; Welch et al. 2002), implying relatively recent, extensive horizontal gene transfer (HGT), duplications and/or loss. If HGT has been common or pervasive in prokaryotic evolution, producing many gene trees that are incongruent when interpreted as species trees, then the very idea of a prokaryotic phylogenetic tree may be questionable.

Conclusive support for a prokaryotic tree, rather than a bush or a network, would be obtained if a larger number of gene trees than would be expected by chance were congruent with a single phylogeny. As the level of congruence among gene trees decreases, the plausibility that prokaryotic phylogeny is adequately described by a tree decreases. Recent evidence of coherent phylogenetic signals from multiple genes in some closely related groups (Daubin et al. 2001) suggests that HGT has little effect on genome phylogenies (Kurland et al. 2003). Here we use supertree methods to measure agreement among gene trees and to test the hypothesis of a prokaryotic phylogenetic tree at both shallow and deeper levels.

Several methods of constructing supertrees have been devised (Baum 1992; Purvis 1995; Semple & Steel 2000; Wilkinson et al. 2001) and a variety of supertrees have been constructed using phylogenetic trees based on molecular and/or morphological data (e.g. Purvis 1995; Daubin et al. 2001; Pisani et al. 2002). These studies have generally assumed that input trees are in sufficient agreement as to yield a meaningful supertree. Here we use supertree construction to investigate agreement among gene trees, and to ask whether or not there really is an underlying phylogeny that can be accurately represented by a tree diagram (Nakhleh et al. 2004). We compared results from recently evolved groups (γ-proteobacteria) and for deeper branches of prokaryotes. In agreement with other researchers, we find gene tree congruence at the tips and extensive conflict at deeper levels. The results demonstrate the difficulty of inferring deep phylogeny, and are consistent with the hypothesis that deep bacterial phylogeny is more of a network than a tree.

2. METHODS

(a) Gene tree construction

Information on genome sequences used in this study is available in electronic Appendix A. Homologous sequences were identified by performing ‘all against all’ searches of a database using the BLASTP algorithm (Altschul et al. 1997) with a cut-off E-value of 10⁻⁷. Only those homologous families where every member found every other member (and nothing else) were retained. Gene trees were then only constructed from single gene families (with at least four members). This conservative approach has been designed to
minimize the inadvertent analysis of paralogues. The protein sequences of each of these families were then aligned separately using ClustalW, v. 1.81 (Thompson et al. 1994) (using the default settings). Maximum likelihood (ML) trees were constructed using the quartet puzzling approach implemented in TreePuzzle (Schmidt et al. 2002). The Whelan and Goldman (WAG matrix) model of substitution was used (Whelan & Goldman 2001), assuming a uniform rate of heterogeneity with amino acid frequencies estimated, and the resulting quartets that appeared greater than 50% of the time were included in the final tree. Neighbour joining trees were constructed with PROTDIST (using the Jones, Taylor and Thornton (JTT) matrix (Jones & Taylor 1994) and assuming one category of substitution rates) and Neighbor (using the neighbour-joining algorithm) from the PHYLIP package (Felsenstein 1993).

(b) Most similar supertree analysis (MSSA)
A supertree containing all the leaves found in the gene trees was proposed. Considering each gene tree in turn, the supertree was pruned until both trees possessed the same leaf set. A simple tree-to-tree distance was used to evaluate similarity between the pruned supertree and gene tree. For each pair of leaves we counted the number of nodes, separating them on each tree and took the absolute difference. The sum of these pairwise differences gives the dissimilarity of the trees. To normalize for large tree bias (Purvis 1995) the sum was divided by the total number of comparisons. In this way, a proposed supertree was assigned a score of zero if, for all gene trees, its sub-tree on the gene-trees leaf set was identical to the gene tree. Higher scores indicate increasing dissimilarity. This scoring system was used as an optimality criterion for choosing among alternative supertrees. Numerous other tree-to-tree distance or fit measures could be used to define optimal supertrees (Thorley & Wilkinson 2003). The present method is most similar to the average consensus procedure with branch lengths all set at unity (Lapointe & Cucumel 1997).

For the analysis of the datasets in this study, either exhaustive or heuristic searches of all possible tree topologies were performed to find the supertree with the minimum difference score when compared to the gene trees. In the case of heuristic searches, sub-tree pruning and regrafting (SPR) as described in PAUP (Swofford 2002) was used to traverse supertree space.

(c) ‘Yet another permutation tail probability’ test
We developed a randomization method to test the null hypothesis that phylogenetic signal in the gene trees was no better than random. We have called this the ‘yet another permutation tail probability’ (YAPTP) test (Faith & Cranston 1991; Wilkinson 1998). For each YAPTP replicate, each gene tree was replaced with a randomly chosen topology for the same leaf set. This removes any congruent phylogenetic signal between the randomized gene trees, while leaving the numbers, sizes of gene trees, the frequency with which any particular taxon was found across the gene trees, and the frequency of co-occurrence of any group of taxa within gene trees unaltered. A heuristic search of tree space (with 10 random additions of the SPR algorithm) was then done and the score of the best supertree was recorded. This was repeated 100 times. We reject the null hypothesis that the gene trees contain no more phylogenetic signal than expected by chance alone if the score for the raw data is not bettered by any of the 100 sets of randomly permuted gene trees (z ≈ 0.01).

(d) Idealized data
Ideally, all gene trees would be completely compatible with a single supertree. To compare the behaviour of our data to perfect data, we generated fully compatible gene trees (an ideal dataset). For each original gene tree, pruning the best supertree of all but those taxa present in the original gene tree produced a corresponding ideal tree. Thus the set of ideal trees fit the best supertree perfectly and also replicate the taxonomic composition, frequency of co-occurrence, and extent of overlap in the original gene trees. An exhaustive search of supertree space was performed using the sets of ideal trees and the scores of all the supertrees were calculated.

(e) Bootstrap analysis
To assess the support for internal branches on a supertree, a bootstrap analysis was performed. Individual gene trees were resampled with replacement, until a new dataset was created with the same number of gene trees as the original. A heuristic search of tree space was done for each pseudoreplicate and the results, reported here as bootstrap proportions (BP), were summarized using a majority-rule consensus tree.

(f) Jackknife analysis
To compare support between the 10- and 11-taxon datasets, we used jackknifing to sample an equal number of gene trees from the larger 10-taxon dataset as are in the smaller 11-taxon dataset. The gene trees for both the datasets were sorted into categories based on their number of taxa (table 1). As the 11-taxon dataset had an extra category than the γ-proteobacteria dataset, for the 11-taxon dataset, the categories containing gene trees with 10 taxa and 11 taxa were combined into a single category. Within each category of gene trees for the γ-proteobacteria, individual gene trees were then resampled with replacement until a new dataset was created with the same number of gene trees as the same category from the 11-taxon dataset. This was necessary as each dataset had differing numbers of sizes of trees (table 1), and to show that support for a phylogeny was not a result of a larger amount of information in one dataset. A heuristic search of tree space was performed for
each pseudoreplicate and results of the bootstrap analysis were summarized using a majority-rule consensus tree.

(g) Shimodaira–Hasegawa tests

For every gene tree with a different topology to the appropriately pruned supertree, a Shimodaira–Hasegawa (SH) test was performed. This was done using TREE-PUZZLE (Schmidt et al. 2002). The pruned supertree and gene tree were both compared using the underlying alignment from which the gene tree was derived.

(h) Software availability

Software for all these analyses is available at http://bioinf.may.ie/software/clann/.

3. RESULTS

For the 61 genomes study, we identified 1117 single gene families of four or more taxa (with a combined length of 306 638 aligned amino acid positions) and inferred corresponding trees (see supplementary tables S1, S2 and S3 for more details). One hundred supertree analyses (each with 10 random starting points using the SPR algorithm to search supertree space) were conducted on bootstrap resamplings of the gene trees and are summarized in the majority-rule consensus supertree in figure 1.

From the (10-taxon) γ-proteobacterial dataset we identified 618 single gene families with four or more taxa (with a combined length of 185 678 alignment positions). Gene trees (see supplementary tables S4 and S5 for more information) were constructed using ML, and an exhaustive search of supertree space (2 027 025 trees) was performed for both raw, idealized and one instance of permuted gene trees (figure 2a). The unrooted phylogenetic supertree shown in figure 2a is the single optimal supertree. The distribution of scores for the 100 best trees from the YAPTP test is centred on 667 (± 68), whereas the best score from the raw gene trees is 240, with only 0.001% of the trees from the idealized gene trees receiving a better score. This result agrees with earlier studies (Lerat et al. 2003; Canback et al. 2004).

In the third analysis, using 11 genomes to span deep prokaryotic relationships, 198 single gene families with four or more taxa were identified (with a combined length of 70 318 aligned positions). For each alignment, ML phylogenetic analyses (as implemented in TREE-PUZZLE (Schmidt et al. 2002)) were done, yielding 198 gene trees (see supplementary tables S6 and S7 for more details). Blue diamonds in figure 2b represent the distribution of supertree scores (ranging from 203 to 280) following an exhaustive search of 34 459 425 supertrees uniting all 11 taxa. The histogram centred on a score of 207 (± 9) represents the distribution of scores of the best supertrees following 100 iterations of the YAPTP test. The best supertree constructed from the raw trees received a score (203), which is well within the distribution of the 100 YAPTP test scores. The agreement among gene trees is not greater than expected by chance alone. The red distribution in figure 2a,b represents the distribution of supertree scores for a single repetition of the YAPTP test. In figure 2a this (red) distribution is extremely dissimilar to the blue distribution from the raw gene trees. This is in contrast to the same distribution in figure 2b, which is extremely similar to the distribution of the raw gene trees. The green distribution in figure 2b indicates the results following an exhaustive search of tree space using idealized gene trees that are completely compatible with the best supertree for the raw gene trees (for which the best supertree has a score of zero). This distribution is very different to the supertree-score distribution for the raw gene trees. The unrooted phylogenetic supertree shown in figure 2b is the single optimal supertree.

Given that there are different numbers of gene trees and different numbers of candidate supertrees evaluated in the exhaustive searches, the numerical values on the graphs in figure 2a,b are not directly comparable. However, if both sets of gene trees were equivalent in terms of phylogenetic signal, then the shapes of the graphs should be similar. It is obvious that there are substantial differences between the two graphs. Whereas the γ-proteobacterial dataset yields distributions of supertree scores for the raw and ideal gene trees that are strikingly similar, this is not the case for the 11-taxon dataset.

The scores received by each individual gene tree when compared to the pruned best supertree are shown in figure 3. The range of scores varies from 0 for trees that are completely compatible with the supertree, to 2.4 for those trees that are most incompatible with the supertree. The bar on the left of each histogram indicates those gene trees that are completely compatible with the corresponding supertree. Figure 3b indicates that many trees are completely compatible with the (γ-proteobacterial) supertree in figure 2a (332 incompatible, 286 compatible). In addition, the data in figure 3d indicate that randomly permuting the dataset has a very adverse affect on the compatibility between the supertree and gene trees (580 incompatible, 38 compatible). Furthermore, of the 332 gene trees that differed from the supertree, SH tests revealed that only 56 (9% of all gene trees) described their underlying alignments significantly better than did the supertree. Of the remaining 276 datasets, the pruned supertree better described six.

By contrast, figure 3a shows that more gene trees are incompatible with the (11-taxon) supertree in figure 2b than are compatible with it (165 incompatible, 33 compatible). The situation only changes slightly when the dataset is randomly permuted (figures 3c, 183 incompatible, 15 compatible). For the 165 gene trees that differed in topology from the appropriately pruned supertree, SH tests (see §2) revealed that 74 (44%) fitted their underlying alignments significantly better than did the pruned supertree. Of the remaining 91 datasets, 88 were not significantly different and for three datasets, the supertree topology was better.

The results of bootstrap analyses of the 11-taxon dataset and jackknife analyses of the 10-taxon dataset are shown on the internal branches in figure 2a,b respectively. In agreement with the analyses of gene-tree score distributions and the comparisons with idealized and randomized data, the γ-proteobacterial dataset showed strong support for all internal branches, whereas the deep-level phylogeny had low levels of support for most branches (the average being 44%), with the most well-supported branch having a BP value of 80.

4. DISCUSSION

Support for relationships from the 61-taxon dataset seems to be restricted to the tips of the phylogeny. Many
Figure 1. This majority-rule consensus tree summarizes the results of the bootstrap analysis of the 61-taxon dataset. Any relationship with less than 50% BP support was defined as unresolved. The numbers represent the percentage BP support received by the internal branch labelled, whereas those resolved relationships without labels had greater than 95% BP support.
(presumably relatively recent) relationships receive 100% BP support, while other (potentially more ancient) relationships remain unresolved. The same pattern emerges from comparison of results for the smaller datasets, with very good support (mean BP = 97) for relatively recent relationships and very poor support (mean BP = 44) for deeper relationships. The failure to reject the null hypothesis, that the set of single gene-family trees derived from complete genomic data lack phylogenetic signal, dramatically underscores the difficulty of inferring ancient divergences from the early history of life (Philippe & Germot 2000; Brown 2001; Lake & Rivera 2004).

Why is inferring deep prokaryotic phylogeny so difficult? Deep divergences give more time for the accumulation of
multiple hits that erode phylogenetic signal. Failure to pass the YAPTP test is consistent with complete erosion of phylogenetic signal but the results of the SH tests suggest that a substantial proportion (44%) of those trees which disagree with the optimal supertree have significantly better support for an alternative. Multiple hits have undoubtedly increased the difficulty of inferring deep prokaryote phylogeny but rather than no signal at all, there appear to be some weak but conflicting signals in the deep gene trees. The nature of these signals merits further study.

Deep divergences also provide more time for the evolution of rate and base composition heterogeneities that can lead to systematic biases in phylogeny estimates. We have made no attempt to examine gene trees or alignments for evidence of systematic biases and cannot rule out their importance here, though we note that any systematic biases are insufficient to lead to pass a randomization test.

The lack of strong support for a single deep-level phylogeny may also be caused by the sparseness of our samples. Of an estimated six million species of prokaryotes (Curtis et al. 2002), we have only used 11. Perhaps greater sampling is required to break long branches and tease apart the signal from the noise. This remains a possibility for further study but our analysis of 61 genomes failed to resolve deeper branches with any greater confidence.

Another scenario could be the inadvertent inclusion of paralogous gene families. However, for hidden paralogy to be able to explain the data, there is a requirement for a duplication event to occur. Then, because we used single gene families, paralogous genes must subsequently survive at least two speciation events and then the three resulting species must independently lose a copy of the gene family, and furthermore, the copies that are lost must be different paralogues in at least two cases. In addition, because we have the requirement that these gene families do not have a paralogue in any completed genome, there must be at least one other taxon where there is a single homologue. Although not impossible, this is a relatively unparsimonious scenario.

The analysis presented here is also compatible with (but not sufficient to prove) the recently espoused notion of the Darwinian threshold (Woese 2002). In this scenario, the absence of a single phylogenetic signal for deep-level relationships is possibly a result of HGT, while the identification of a core phylogeny in the \( \gamma \)-proteobacteria indicates much less frequent confounding events. The contrastingly strong phylogenetic signal in the \( \gamma \)-proteobacteria supports the hypothesis that modern prokaryotes are more compartmentalized and less likely to engage in such widespread gene transfer. This might provide the context in which to evaluate the observations of many independent gene acquisitions in different strains of \( E. \) coli (Blattner et al. 1997; Hayashi et al. 2001; Welch et al. 2002). In our analyses, we require that a single-gene family is present in at least four different genomes. Because of this requirement, such genes are relatively unlikely to be transient acquisitions. This could be taken as evidence to suggest that the independently acquired genes in different \( E. \) coli strains are likely to be ephemeral. Although gene acquisition is a natural and continuous process, gene retention may not be so
easy, and there may be a gradient in terms of the propensity of any gene to be retained in a genome (Kurland et al. 2003). However, if retention of acquired genes was common, then we could not hope to recover the species tree that we see in our analysis.

It has recently been shown that the SSU rRNA gene can be forcibly exchanged between bacterial species (Asai et al. 1999), thereby raising the question of whether or not this can happen in nature. The $\gamma$-proteobacterial supertree from our analysis is remarkably similar to a tree that is derived from the SSU rRNA gene (data not shown), even though this gene was not included in any dataset. Therefore the SSU rRNA gene is unlikely to be a frequent subject of inter-species transfer and retention, at least in the $\gamma$-proteobacteria. It is not sensible to repeat this analysis for the deep-level phylogeny.

We have made no attempt to discriminate between informational and operational genes, despite the suggestion that there are fundamental differences in their rates of HGT (Jain et al. 1999). Supertree analyses from whole genomes should provide a powerful means of testing such hypotheses.

5. CONCLUSION

We have developed a method of interrogating sets of phylogenetic trees for evidence of compatibility, similarity, signal and noise. We have shown here, using this simple phylogenetic approach, that the compatibility between strongly-supported individual gene trees spanning the major divisions of prokaryotic diversity is very low. This suggests that early prokaryotic evolution cannot be represented effectively with a single organismal phylogeny. Although we cannot discriminate with absolute certainty between high levels of orthologous replacement (HGT), hidden paralogy or lack of phylogenetic signal at the base of the prokaryotic tree, our findings in this study are not a result of short amino acid alignments (table 1) or sparse sampling (as shown from the similar weakly supported individual gene trees spanning the 61-taxon study).

We have demonstrated that the method we have employed can be used to investigate genome-based phylogenies and also to detect underlying signal in the gene trees. This is a very promising approach to reconstruct a tree of life. For all datasets, the same set of rules was applied, but the results were considerably different. The conclusion therefore, appears to be that it is difficult to invest a great deal of confidence in a deep-level prokaryotic phylogeny if 84% of the gene trees conflict with it. For more recent relationships, we can be much more confident in the tree, given that almost half the orthologues are in complete agreement.

C.J.C. and J.O.M. were supported by grants from the Health Research Board of Ireland (grant number RP 124/2000) and the Higher Education Authority of Ireland (RYPTI Cycle III). M.W. was supported by BBSRC grant 40/G18385. Thanks to Peter Foster for reading an earlier version of this manuscript, and for the commentary provided by three anonymous reviewers.

REFERENCES


As this paper exceeds the maximum length normally permitted, the authors have agreed to contribute to production costs.

Visit www.journals.royalsoc.ac.uk and navigate through to this article in *Proceedings: Biological Sciences* to see the accompanying electronic appendix.