Quartet Mapping and the Extent of Lateral Transfer in Bacterial Genomes

Vincent Daubin and Howard Ochman

Department of Biochemistry and Molecular Biophysics, University of Arizona, Tucson

Several recent analyses have used quartet-based methods to assess the congruence among phylogenies derived for large sets of genes from prokaryotic genomes. The principal conclusion from these studies is that lateral gene transfer (LGT) has blurred prokaryotic phylogenies to such a degree that the darwinian scheme of treelike evolution might be abandoned in favor of a net or web. Here, we focus on one of these methods, quartet mapping, and show that its application can lead to overestimation of the extent of inferred LGT in prokaryotes, particularly when applied to distantly related taxa.

Introduction

Prokaryotes have long been known to acquire and exchange genes; however, not until complete genome sequences became available has the full extent and impact of lateral gene transfer (LGT) been assessed (Doolittle 1999; Eisen 2000; Ochman, Lawrence, and Groisman 2000). Virtually any gene, even those assuming roles in key cellular processes (Brochier, Philippe, and Moreira 2000; Brown et al. 2001), might be subject to LGT, and this poses a challenge to reconstructing the evolutionary history among taxa. In macroscopic eukaryotes, a lack of concordance between gene trees can often be evaluated in light of independent evidence of phylogenetic relationships, as derived from fossils or morphology. Unfortunately, few such opportunities exist in prokaryotes, and therefore it is difficult to trust most gene trees a priori.

One approach proposed to circumvent some of the difficulties in reconstructing relationships among prokaryotes is to compare all gene trees, with the hope that a consensus tree will emerge. Concomitantly, any conflicting phylogenies—an expected result of LGT—can be identified. Such attempts have reached different conclusions about the degree of LGT shaping prokaryotic genomes (Jain, Rivera, and Lake 1999; Nesbo, Boucher, and Doolittle 2001; Brochier et al. 2002; Daubin, Gouy, and Perrière 2002; Zhaxybayeva and Gogarten 2002; Daubin, Moran, and Ochman, 2003). Among them, the application of the quartet-mapping method (Strimmer and Von Haeseler 1997; Nieselt-Struwe and Von Haeseler 2001) has provided striking support for the pervasiveness of LGT, leading some authors to abandon the notion of treelike evolution in prokaryotes.

Quartet mapping is a likelihood method originally proposed as a tool for analyzing the phylogenetic content of an *n*-sequence alignment by extracting all combinations of four sequences (‘‘quartets’’) and evaluating the likelihood of the three possible topologies for each quartet (Strimmer and Von Haeseler 1997; Nieselt-Struwe and Von Haeseler 2001). Following these authors, the posterior probability (p_i) of each of the three possible topologies (T_1, T_2 and T_3) for a given quartet is computed as follows:

$$p_i = \frac{L_i}{L_1 + L_2 + L_3}$$

where L_i is the likelihood of the tree T_i. Note that this approach considers likelihoods, not log-likelihoods (as provided in most phylogeny programs). A quartet is considered strongly supported when p_i is greater than 0.99 (or sometimes 0.90). The *n*-sequence alignment is deemed informative when a substantial proportion of the constituent quartets statistically support any of the topologies.

In their original paper, Strimmer and Von Haesler (1997) applied this method to test the ability of ribosomal DNA to retrieve the relationships of myriapods and chelicerates within the Arthropoda by using quartets of different representatives of these groups. In this case, the grouping of myriapods and chelicerates was supported by over 90% of the quartets, suggesting that the alignment was sufficiently informative to resolve the relationships among these higher taxa. However, more than 7% of the quartet alignments supported an alternative topology, indicating that depending on the particular sequences included in a quartet, a substantial proportion of quartet alignments will artifactually generate different trees.

The quartet-mapping method has been recently modified to analyze the extent of LGT in prokaryotes (Nesbo, Boucher, and Doolittle 2001; Zhaxybayeva and Gogarten 2002). Here, rather than evaluating the phylogenetic information contained in a multiple sequence alignment of many taxa for a single gene, these studies employ groups of four fully sequenced genomes to assess the congruency of hundreds of orthologous gene alignments with one another. In comparisons including distantly related prokaryotic lineages, each of the three possible quartet topologies was supported by approximately equal numbers of genes, which was taken as an indication that LGT has occurred at such a high frequency that no consensus tree exists. However, previous analyses have shown that quartets, even in recent groups such as mammals, frequently give strong support to arbitrary topologies, depending on the sequences taken to represent each mammalian order (Philippe and Douzery 1994; Adachi and Hasegawa 1996). Moreover, the original application of quartet mapping to arthropods showed that genes considered as reliable phylogenetic markers, such as ribosomal DNA, can yield a significant proportion of alignments supporting alternate topologies, even in the

Key words: bacterial phylogeny, maximum likelihood, quartet trees, lateral gene transfer.

E-mail: daubin@email.arizona.edu.

Mol. Biol. Evol. 21(1):86–89. 2004
DOI: 10.1093/molbev/msg234
Molecular Biology and Evolution vol. 21 no. 1

absence of LGT. Therefore, given the antiquity of bacterial phyla, protein sequences that exhibit higher levels of divergence than rDNA might produce results that could be incorrectly attributed to LGT. In this paper, we show that the application of quartet mapping can impart too much confidence to trees that garner no statistical support by other tests, and thus, the amount of LGT inferred from this method can be overestimated.

Materials and Methods

To examine the consequences of applying quartet-mapping methods to assess the statistical support for alternate topologies and thereby infer the amounts of LGT, we analyzed two groups of four sequenced bacterial genomes: (1) closely related enteric bacteria, consisting of *Salmonella enterica* serovar *typhi* Ty2 (NC_004631), *S. enterica* serovar *typhimurium* LT2 (NC_003197), and *Escherichia coli* K12 (NC_000913), with *Yersinia pestis* KIM (NC_004088) as the outgroup; and (2) the more distantly related Streptococci, consisting of *Streptococcus pyogenes* MGAS315 (NC_004070), *S. mutans* UA159 (NC_004350), and *S. pneumoniae* R6 (NC_003098), with *Lactococcus lactis* IL1403 (NC_002662) as the outgroup. We performed a Blastp query (Altschul et al. 1997) of each protein from one genome against a database consisting of all proteins from the species constituting a particular group. We retained as orthologous genes those having only one match per genome with a bit score greater than 30% of the bit score of the query sequence against itself. Analysis of the distribution of the Blast scores shows that a threshold of 30% is low enough to retrieve even ancient homologs (results not shown). Although we cannot totally exclude cases of hidden paralogy because of independent losses of genes, applying this method greatly reduces the risks of including paralogs in the alignment because only gene families without duplications are retained. The orthologs (1,703 for the enterics and 685 for the Streptococci) were aligned at the protein and nucleotide sequence levels with ClustalW (Higgins, Thompson, and Gibson 1996). (Sequence alignments are available from the authors upon request.) The differences in log-likelihood and their standard error (SE) between the three topologies were evaluated in Puzzle 5.1 (Strimmer and Von Haeseler 1996) (default options) for each of the quartet alignments using the Kishino-Hasegawa (KH) and the Shimodaira-Hasegawa (SH) tests (Kishino and Hasegawa 1989; Shimodaira and Hasegawa 1999) at the 5% level of significance.

Results and Discussion

Problems associated with the reliability of quartet phylogenies have been identified by Philippe and Douzery (1994) for parsimony methods and by Adachi and Hasegawa (1996) for maximum likelihood. Each demonstrated that, by this approach, different taxonomic samples could lead to very different, strongly supported, relationships between mammalian orders. Adachi and Hasegawa (1996) found that the impact of this apparent incongruence between gene trees could be greatly reduced by evaluating the SE of likelihood estimates, as determined by the KH method (Kishino and Hasegawa 1989). When considering the SE, the topologies having the highest likelihood are not, in some cases, significantly better than alternate topologies. A modification of the KH test, the SH test (Shimodaira and Hasegawa 1999), has been recently developed to allow comparisons of multiple topologies.

Most applications of quartet mapping ignore these statistics and adopt a fixed threshold, whereby one topology is judged as being strongly supported if its likelihood is equal to or greater than 100-fold higher than the sum of the two other topologies (see equation 1). The application of fixed thresholds is particularly problematic because the SE of the log-likelihood estimates are dependent on the particular alignment and often much higher than the fixed thresholds applied in most studies. The SE of log-likelihood estimates typically reach values greater than 5 ($e^5 \sim 148$), and therefore, for a given alignment, the application of a threshold of 100 could confer a posterior probability greater than 0.99 to a topology that is not statistically different from the two others.

To assess the disparity between fixed and variable (i.e., dependent on SE) thresholds, we computed the difference between the log-likelihood of the best-supported topology and that of each of the two other topologies, as well as the SE of the log-likelihood estimates for families of orthologous genes from each of the two quartets of bacterial taxa (enterics and Streptococci). The dotted line in figure 1 shows the fixed threshold typically applied in the quartet-mapping method; therefore, all alignments represented by points above this line would be considered significantly supporting one of the three possible topologies. The solid line represents the variable threshold that takes the SE into account (at the 5% level). All points residing in the shaded area between the solid and dotted lines represent topologies that are falsely supported when using a fixed threshold.

In the case of closely related species (i.e., the enteric bacteria), the difference between the two methods is small because most alignments show significant support for one of the topologies under whichever criterion is applied; hence, only a few points are in the shaded area. However, when considering more distantly related species (i.e., the Streptococci), the difference is striking. Among the 685 alignments performed for Streptococci, the KH test and the SH test gave statistical support to one topology over the two other in 18.5% (127) and 12% (82) of cases, respectively. Thus, the vast majority of the quartet alignments do not support any topology because of difficulties in building phylogenies using small taxon sampling when divergences are high. In contrast, the quartet-mapping method gives support to one topology in 55.6% (376) of the alignments. Therefore, the quartet-mapping method overestimates the support accorded to 37% of the topologies, and it is likely that all topologies, both correct and artifactual, receive more support from quartet mapping than from the other tests. In Streptococci, the alignments that are considered informative by quartet mapping, but not by the other tests, are significantly shorter ($P < 0.0001$), and the resulting trees have a higher ratio of external to internal branch lengths ($P < 0.0001$).

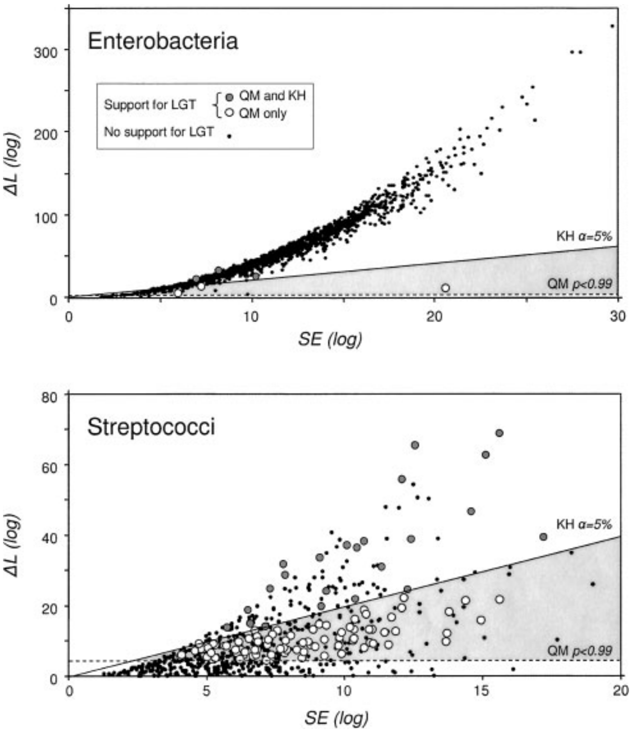

FIG. 1.—Relationship between the difference in the log-likelihood of the trees (ΔL) and the standard error (SE) of the likelihood estimates for the Enterobacteria and the Streptococci. The dotted line in each panel shows the threshold of ΔL typically considered in the quartet-mapping (QM) method ($P < 0.99$), whereas the solid line shows the criterion used in the KH test ($\alpha = 5\%$, $\Delta L = 1.96*SE$). Gray circles represent LGT topologies supported by the two methods (QM and KH), and open circles represent LGT topologies supported by quartet mapping (QM) only. Black points represent alignments for which there is no evidence of transfer by either method.

Table 1
Comparison of the Numbers of Quartet Alignments Supporting an Alternate Topology Inferred by the Three Methods

	Quartet Mapping ($P < 0.99$)	KH Test	SH Test	Total Alignments
Streptococcus	123 (18.0%)	28 (4.1%)	23 (3.4%)	685
Enterobacteria	6 (0.4%)	3 (0.2%)	3 (0.2%)	1703

These factors are both known to be responsible for reconstruction artifacts, most notably long-branch attraction, which can mimic the effects of LGT.

These results have broad implications for the interpretation of genome evolution in bacteria in that evidence of conflicting topologies among bacterial genes has often been interpreted as resulting from LGT. In fact, the impact of LGT deduced from methods based on fixed and variable thresholds lead to different conclusions about the extent of LGT affecting bacterial genomes (table 1): the number of alignments ''significantly'' supporting an alternate topology can be severalfold higher using quartet mapping when compared with the results of both the KH test and the SH test.

For both the enteric bacteria and the Streptococci, the KH test and SH test yield a small number of alignments that support alternative topologies (<5%). The fact that in some cases these genes are clustered on the bacterial chromosome (e.g., an operon encoding three subunits of glutamyl-tRNA amidotransferase in Streptococci) suggests that these tests can identify some probable cases of LGT. However, one cannot take all LGT topologies supported by the KH test and SH test as true cases of LGT, because both tests are probably also sensitive to artifacts caused by long-branch attraction. Whereas quartet mapping would lead to the conclusion that Streptococci have undergone relatively high rates of orthologous replacement by LGT, the KH test and SH test show that the extent of LGT is much lower. Knowledge that quartet phylogenies can yield such artifacts calls into question some of the conclusions about the extent of LGT affecting prokaryotic genomes based on such analyses.

As revealed by the difference in the results for the enteric bacteria and for the Streptococci (fig. 1), the estimation of LGT using quartet mapping is greater when comparing more distantly related species. Moreover, within a cluster of species, genes yielding phylogenies with short internal branches tend to support the hypothesis of LGT based on quartet mapping, which leads to the overestimation of LGT even when considering bacterial species of the same genus. Although our analyses viewed the Streptococci, which differ by at most 6% in rDNA sequences, as being relatively distantly related, previous studies have applied quartet-mapping methods to representatives of different phyla, which are substantially more divergent and thus more likely to yield a higher proportion of artifactual LGT. Considering quartets of such distantly related species increases the amount of phylogenetic incongruence caused by reconstruction artifacts, misalignment and sampling bias (Moreira and Philippe 2000; Eisen, 2000; Zwickl and Hillis 2002). As in bacteria, such tree incongruencies have been observed in numerous plant and animal groups, but these inconsistencies have led molecular phylogeneticists to question and refine their methods rather than to invoke lateral gene transfer. The antiquity of Bacteria introduces an exceptional challenge to the reconstruction of molecular phylogenies, and although LGT is certainly a major force molding prokaryotic genomes, its inference must rely on careful phylogenetic analysis with large taxon sampling and appropriate tests of incongruence.

Literature Cited

Adachi, J., and M. Hasegawa. 1996. Instability of quartet analyses of molecular sequence data by the maximum likelihood method: the Cetacea/Artiodactyla relationships. Mol. Phylogenet. Evol. **6**:72–76.

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

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

Brochier, C., H. Philippe, and D. Moreira. 2000. The evolutionary history of ribosomal protein RpS14: horizontal gene transfer at the heart of the ribosome. Trends Genet. **16**:529–533.

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

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

Daubin, V., N. Moran, and H. Ochman. 2003. Phylogenetics and the cohesion of bacterial genome. Science **301**:829–832.

Doolittle, W. F. 1999. Lateral genomics. Trends Cell Biol. **9**:M5–M9.

Eisen, J. A. 2000. Horizontal gene transfer among microbial genomes: new insights from complete genome analysis. Curr. Opin. Genet. Dev. **10**:606–611.

Higgins, D. G., J. D. Thompson, and T. J. Gibson. 1996. Using CLUSTAL for multiple sequence alignments. Methods Enzymol. **266**:383–402.

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.

Kishino, H., and M. Hasegawa. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. J. Mol. Evol. **2**:170–179.

Moreira, D., and H. Philippe. 2000. Molecular phylogeny: pitfalls and progress. Int. Microbiol. **3**:9–16.

Nesbo, C. L., Y. Boucher, and W. F. Doolittle. 2001. Defining the core of nontransferable prokaryotic genes: the euryarchaeal core. J. Mol. Evol. **53**:340–350.

Nieselt-Struwe, K., and A. von Haeseler. 2001. Quartet-mapping, a generalization of the likelihood-mapping procedure. Mol. Biol. Evol. **18**:1204–1219.

Ochman, H., J. G. Lawrence, and E. A. Groisman. 2000. Lateral gene transfer and the nature of bacterial innovation. Nature **405**:299–305.

Philippe, H., and E. Douzery. 1994. The pitfalls of molecular phylogeny based on four species as illustrated by the Cetacea/Artiodactyla relationships. J. Mam. Evol. **2**:133–152.

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

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

———. 1997. Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. Proc. Natl. Acad. Sci. USA **94**:6815–6819.

Zhaxybayeva, O., and J. P. Gogarten. 2002. Bootstrap, Bayesian probability and maximum likelihood mapping: exploring new tools for comparative genome analyses. BMC Genomics **3**:4.

Zwickl, D. J., and D. M. Hillis. 2002. Increased taxon sampling greatly reduces phylogenetic error. Syst Biol. **51**:588–598.

Hervé Phillipe, Associate Editor

Accepted August 8, 2003