Conditioned Genome Reconstruction: How to Avoid Choosing the Conditioning Genome

Matthew Spencer*[,1,2], David Bryant[3,4], Edward Susko[1]

1 *Department of Mathematics and Statistics, Dalhousie University, Halifax, Nova Scotia, B3H 3J5, Canada.*

2 *Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, Nova Scotia, B3H 4H7, Canada.*

3 *Department of Mathematics, University of Auckland, Private Bag 92019, Auckland, New Zealand.*

4 *McGill Centre for Bioinformatics, McGill University, 3775 University Street, Duff Medical Building, Montreal, Quebec, H3A 2B4, Canada.*

* Author for correspondence. Present address: School of Biological Sciences, University of Liverpool, Liverpool L69 7ZB, UK. Email m.spencer@liverpool.ac.uk, Phone +44 (0) 151 795 4399.

Abstract

Genome phylogenies can be inferred from data on the presence and absence of genes across taxa. Logdet distances may be a good method, because they allow expected genome size to vary across the tree. Recently, Lake and Rivera proposed conditioned genome reconstruction (calculation of logdet distances using only those genes present in a conditioning genome) to deal with unobservable genes that are absent from every taxon of interest. We prove that their method can consistently estimate the topology for almost any choice of conditioning genome. Nevertheless, the choice of conditioning genome is important for small samples. For real bacterial genome data, different choices of conditioning genome can result in strong bootstrap support for different tree topologies. To overcome this problem, we developed supertree methods that combine information from all choices of conditioning genome. One of these methods, based on the BIONJ algorithm, performs well on simulated data and may have applications to other supertree problems. However, an analysis of 40 bacterial genomes using this method supports an incorrect clade of parasites. This is a common feature of model-based gene content methods, and is due to parallel gene loss.

Keywords: BIONJ, conditioned genome reconstruction, consistency, gene content, logdet, supertrees

Variation in gene content makes it difficult to estimate organismal phylogeny from nucleotide or amino acid sequences. Within a lineage, genes are often gained (for example, by lateral transfer) and lost (by deletion). Both lateral transfer (e.g. Doolittle et al., 2003) and differential loss of paralogous genes (e.g. Martin and Burg, 2002) can mislead sequence-based phylogenetic estimation. Instead of viewing gains and losses of genes as problems for sequence-based phylogenetics, we could estimate phylogenies from patterns of presence and absence of gene families.

A gene family is the set of all genes belonging to a group of repeated sequences,

derived sufficiently recently from a common ancestor (Graur and Li, 2000, p. 264). In practice, the level of similarity required for membership of a gene family is somewhat arbitrary, and this may have important consequences for analysis of such data (Hughes et al., 2005). For brevity, we usually refer to gene presence/absence in this paper, with presence meaning that at least one member of a given gene family is present in a genome. The data used to estimate phylogenies are presence/absence states for some defined set of gene families: the choice of this set is a central theme in this paper. The set of families present in a genome is viewed as dynamic: there is a continuous turnover of genes even when the total number of genes in the genome remains roughly constant (Snel et al., 2002). We assume that each gene family is independent. This cannot be true because genes located close together on a genome may be transferred or deleted together, and because there are functional relationships between gene families. However, violation of this assumption is not critical, because we focus on the marginal behaviour of an individual family. Similar assumptions are made and violated in models of sequence evolution, especially for RNA, and the consequences are not too severe (Tillier and Collins, 1995). We also assume that lineages evolve independently once they have diverged from their common ancestor. This implies that the loss of a family from one lineage does not affect the probability that it can be gained in another lineage. This cannot be strictly true because a family must exist in at least one lineage at a given time in order to be transferred to another lineage. However, if the sampled lineages represent a small fraction of those actually existing (as is certainly the case), gain or loss in one lineage may have almost no effect on gain or loss in another.

Most early analyses used either ad-hoc distances or parsimony (e.g. Fitz-Gibbon and House, 1999; Tekaia et al., 1999; Snel et al., 1999; Montague and Hutchinson III, 2000; Wolf et al., 2001, 2002). Because these methods did not take account of the relative rates of gene gain and loss, they are likely to be unreliable. Distance estimates from Markov models for the evolution of gene content (e.g. Gu, 2000; Gu and Zhang, 2004; Huson and Steel, 2004; Zhang and Gu, 2004; Spencer et al., 2006) do take account of these relative rates. However, the substantial variation in genome size across taxa suggests that there

may be different rates of gene gain and loss in different parts of the tree (Lake and Rivera, 2004). The distance measure used in the SHOT program (Korbel et al., 2002; Dutilh et al., 2004) attempts to deal with variation in genome size and seems to perform well in practice, although it is only an approximate evolutionary distance. Another and more rigorous approach is to use logdet (also known as paralinear) distances (Lake, 1994; Lockhart et al., 1994), which do not assume the same Markov model on every edge. Two recent studies have applied logdet distances to gene content data (Lake and Rivera, 2004; Rivera and Lake, 2004).

There is an important difference between gene content data and sequence data. An unknown number of genes exist or once existed but are not found in any sequenced genome (Lake and Rivera, 2004). We will therefore underestimate the frequency of genes that are absent from both members of a pair of taxa, and will get the wrong estimated pairwise distances. Numerical examples (available on request) show that this can lead to inconsistency in topology estimation. There is no such problem with sequence data. Lake and Rivera (2004) suggested calculating conditioned logdet distances, using only those genes present in a conditioning genome. The conditioning genome is an additional taxon which will not be included in the estimated tree topology (for example, node c is a conditioning genome for the four taxa w, x, y and z in figure 1). Rivera and Lake (2004) used trees based on conditioned logdet distances to make inferences about the relationships among the three kingdoms of life. In this paper, we use theory, simulations and real data to evaluate and improve the performance of conditioned logdet methods.

Outline

This paper is structured as follows:

– We first show that under some suitable assumptions, conditioned logdet distances are tree-additive and non-negative, and thus permit consistent reconstruction of phylogenies, for almost any choice of conditioning genome. This appears to support the idea that 'the choice of the conditioning genome should not significantly affect

the outcome of the analysis' (Rivera and Lake, 2004, Supplementary Data).

– We then show that for realistic sample sizes, the choice of conditioning genome can have substantial effects. Theoretical considerations suggest that a large conditioning genome, far from any of the taxa of interest, should be a good choice. We show that this is true for simulated data.

– We evaluate the effects of the choice of conditioning genome on the bootstrap support for alternative topologies in real bacterial genome data. Although the size and location of the conditioning genome are possible predictors of performance, it would be better to avoid choosing a single conditioning genome.

– We then examine supertree methods by which we can combine data from all choices of conditioning genome into a single estimate of topology. One of these methods, based on a modification of the BIONJ algorithm (Gascuel, 1997a), performs well on simulated data for 5-taxon trees. However, simpler methods do better in simulations on a 40-taxon tree.

– We apply the modified BIONJ algorithm to a real 40-taxon bacterial data set, and discuss the problems that may be caused by parallel gene loss in parasites.

Markov models for gene content and conditioned logdet distances

We assume as above that gene families are independent, and that lineages evolve independently. We further assume that the evolution of gene family states can be described by a two-state Markov model in continuous time, not necessarily time-homogeneous. We assume that the same rates of gain and loss apply to all gene families in a given genome at a point in time. This is unlikely to be true, and in the Discussion we highlight the possible consequences of violating this assumption.

Let X_x be the presence/absence state at node x, with value 0 indicating absence of a gene and 1 indicating presence. Let the probabilities of gene absence and presence at a node x be $\Pr\{X_x = 0\} = \pi_0^{(x)}$ and $\Pr\{X_x = 1\} = \pi_1^{(x)}$. Let $\mathbf{\Pi}^{(x)}$ be the diagonal matrix of

state probabilities at node x. Let $\mathbf{F}^{(wx)}$ be the matrix of pattern probabilities, whose ijth entry $f_{ij}^{(wx)}$ is the probability of observing state i in w and state j in x. The standard (unconditional) logdet distance between w and x is

$$d_{wx} = -1/2 \log \det[\mathbf{\Pi}^{(w)}]^{-1/2} \mathbf{F}^{(wx)} [\mathbf{\Pi}^{(x)}]^{-1/2} \tag{1}$$

The initial constant 1/2 applies to two-state models. More generally, scaling by the reciprocal of the number of states means that the unconditional logdet distance can be interpreted as the expected number of substitutions for stationary models with equal frequencies of each state (Lockhart et al., 1994).

The conditioned logdet distance between w and x, calculated for genes present in a conditioning genome, is

$$d'_{wx} = -1/2 \log \det[\mathbf{\Pi}'^{(w)}]^{-1/2} \mathbf{F}'^{(wx)} [\mathbf{\Pi}'^{(x)}]^{-1/2} \tag{2}$$

Here, $\mathbf{F}'^{(wx)}$, $\mathbf{\Pi}'^{(w)}$ and $\mathbf{\Pi}'^{(x)}$ are defined as above, except that they include only those genes present in the conditioning genome.

When we use a conditioning genome, we have a nonrandom sample of genes, whose composition depends on the location of the conditioning genome. The conditional Markov model will generally be nonstationary even if the unconditional model was stationary. This means that the estimated distance between a pair of taxa depends on the length of the path leading to the conditioning genome (figure 2a). Because the model is conditional, it is not obvious that the arguments used to justify it in Lake and Rivera (2004) are applicable. Here we show that the model can indeed be justified, for almost all choices of conditioning genome.

Theorem 1 *If gene families evolve independently of each other, and independently along the edges of a tree according to a two-state continuous-time Markov model, not necessarily time homogeneous, but with the same gain and loss rates for all genes, then conditioned logdet distances are tree-additive and non-negative, so long as the conditioning genome is*

not one of the taxa included in the tree or an internal node.

Proof Appendix 1.

Given these properties, we can obtain a consistent estimate of tree topology (Chang and Hartigan, 1991).

Small Samples and the Choice of Conditioning Genome

We showed in Appendix 1 that conditioned logdet distances are tree-additive for almost any choice of conditioning genome. Nevertheless, for small sample sizes, the choice of conditioning genome may be important. If the conditioning genome is at or very close to an internal or terminal node on the tree, the conditional probability of gene absence at this node will be zero or almost zero. It is unlikely that we will be able to calculate logdet distances in such cases, and if we can calculate them, they will have large variance. An approximate sampling variance for the logdet distance between taxa w and x is

$$\sigma^2(\hat{d}_{wx}) \approx \frac{1}{4s^2n}\sum_i \left[\frac{1+4\hat{\pi}_i^{(w)}}{\hat{\pi}_i^{(x)}} + \frac{1}{\hat{\pi}_i^{(w)}}\left((4\sum_j [\hat{\mathbf{P}}^{-1}]_{ji}^2 \hat{p}_{ij}) - 3\right) - 2\sum_j \frac{\hat{p}_{ij}}{\hat{\pi}_j^{(x)}}\right] - \frac{1}{sn} \qquad (3)$$

where $\hat{\mathbf{P}}$ is a sample estimate of the stochastic matrix whose ijth entry $\hat{p}_{ij}$ is the probability that taxon x has state j, given that taxon w has state i, $\hat{\pi}_i^{(w)}$ is the sample estimate of the probability that taxon w has state i, s is the number of states (in this case 2) and n is the number of observations (in this case, genes). Here, all these estimates are conditional, calculated from the sample of genes that are present in the conditioning genome. Equation 3 was derived by the delta method, using a similar approach to Barry and Hartigan (1987).

Equation 3 contains the reciprocals of the state probabilities. Therefore, for the same reason that the logdet distance gets large as one or more state probabilities at a node approach zero, the sampling variance gets large (figure 2b). The consistency result implies that bias and variance will both get small as the number of genes increases, but both bias and variance could be relatively large with small samples. We may therefore estimate the wrong topology in some situations. If we want to minimize these problems, a large

conditioning genome is clearly a good idea. The arguments above also suggest that a conditioning genome far from the taxa of interest might perform better.

To illustrate this, we simulated gene presence/absence data on the topology of Figure 1, with a stationary homogeneous continuous-time two-state Markov model. Throughout the tree, genes were gained at an instantaneous rate of 0.625 and lost at an instantaneous rate of 2.5 per unit time. This gives a stationary probability 0.2 of gene presence. We varied the edge t_k leading to the conditioning genome and the edge t_u separating internal nodes u and k, while keeping the sum of t_u and t_v constant. This means that the unconditioned distances between taxa other than c are constant. In each replicate, logdet distances were calculated from 5000 genes. This is about twice the number of genes in the real data analyzed by Rivera and Lake (2004). For the conditioned case, this means that we had to simulate until 5000 genes were present in the conditioning genome. This choice of parameter values might be realistic. For example, the COG database (Tatusov et al., 2003) contains 4873 gene families, with an average of 1328 families present per bacterial genome, and an unknown number of families absent from all genomes.

For the parameters we used, unconditioned logdet distances could be calculated in every replicate, and we almost always recovered the correct topology for taxa w, x, y and z using least squares (mean 98%, minimum 96% over parameter combinations). The outcome is unaffected by the location of the conditioning genome, since it has no effect on distances between the taxa included in the tree. If we conditioned on presence of genes in taxon c, sampling variability meant that we were often unable to calculate logdet distances. This was particularly likely if the conditioning genome was on a short edge (figure 3a, small values of t_k). For cases in which logdet distances could be calculated, the correct tree topology was less likely to be recovered when the conditioning genome was on a short edge and connected close to internal node v (figure 3b, small values of t_k and large values of t_u). This is the pattern we would expect if high sampling variance reduces accuracy. Overall, the true topology was recovered less often (mean 87%, minimum 63% where conditioned logdet distances could be calculated, mean 67%, minimum 2% over all replicates) than in the unconditioned case, even though we used the same number of genes to calculate

distances in each case.

Effects of the Conditioning Genome for Real Data

To examine the effects of choice of conditioning genome for real data, we analyzed four four-taxon subsets of the 50 bacteria in the COG database (Tatusov et al., 2003), downloaded 13 May 2004 from `ftp://ftp.ncbi.nih.gov/pub/COG/`. The version we used contains data on the number of members of 4873 gene families in each genome. For all taxa, we scored the presence or absence of at least one member of each gene family. The mean number of gene families was 1328 (minimum 362, maximum 2243, standard deviation 562). For each four-taxon subset, we then used each of the 46 other bacterial genomes in turn as a conditioning genome, selecting only those gene families that were present in the conditioning genome. We then bootstrap resampled the gene families present in the conditioning genome 1000 times, and recorded the number of times each of the three possible unrooted tree topologies was estimated by unweighted least-squares without constraints on edge length. For comparison, we calculated least-squares trees using SHOT distances (Korbel et al., 2002) on the same data. The SHOT distance between two taxa w and x is

$$dS_{wx} = -\log\left(n_{PP}\frac{\sqrt{a^2+b^2}}{ab\sqrt{2}}\right) \tag{4}$$

where n_{PP} is the number of gene families present in both taxa, and a and b are the numbers of gene families present in w and x respectively. SHOT distances seem to give good results in practice, although they are not based on an explicit model of genome evolution. We used the SHOT results and other biological information to decide on the probable true topology in each case.

For subset a (Figure 4a), the dominant topology was (*Synechocystis* sp., *Mycoplasma genitalium*) | (*Escherichia coli* K12, *Mesorhizobium loti*), with mean 68% bootstrap support over all conditioning genomes, and maximum support 99%, when conditioning on *Treponema pallidum*. This is probably the correct topology, and was supported by the SHOT method. The mean bootstrap support for the true topology was

not overwhelming, probably because the internal edge is short. However, both of the other topologies had substantial support from some conditioning genomes. For example, pairing *Synechocystis* with *E. coli* (mean support 20%) received 79% bootstrap support when conditioning on *Vibrio cholerae*, and pairing *Synechocystis* with *M. loti* (mean support 12%) received 68% support when conditioning on *Mycoplasma pneumoniae*.

For subset b (Figure 4b), the topology (*Bacillus subtilis*, *Bacillus halodurans*) | (*Haemophilus influenzae*, *Pasteurella multocida*) was supported by a mean of 99.98% of bootstrap replicates over all conditioning genomes. No conditioning genome gave less than 99% support for this topology. The two *Bacillus* species have the same presence/absence state for 89% of gene families, and are almost certainly sister taxa.

For subset c (Figure 4c), only two of the three topologies had high frequencies. The dominant topology was (*Aquifex aeolicus*, *Yersinia pestis*) | (*Buchnera* sp. APS, *Ureaplasma urealyticum*) (mean support 65%, maximum support 100%, when conditioning on *Nostoc* sp. PCC 7120). The second topology, (*A. aeolicus* with *Buchnera*), had mean support 34%, and maximum support 99% (when conditioning on *Mycoplasma pulmonis*). No conditioning genome gave more than 4% support for the third topology. Two other model-based gene content methods, Gu and Zhang's extended gene content distance (Gu and Zhang, 2004) and a more complex Markov model (M. Spencer, E. Susko and A. J. Roger, unpublished) also supported the dominant topology. Nevertheless, all three methods are probably wrong. *Buchnera* and *U. urealyticum* are both members of a group of parasites and endosymbionts with reduced genomes. The loss of a common set of genes that are not required for an intracellular lifestyle misleads most gene content methods (Wolf et al., 2001). The SHOT method (Korbel et al., 2002), which treats the shared absence of a gene family as uninformative, grouped *A. aeolicus* with *U. urealyticum*. A whole-genome method based on average BLAST scores (Gophna et al., 2005), also ignores shared absence of genes and groups *A. aeolicus* with *U. urealyticum*.

For subset d (Figure 4d), the dominant topology was (*Corynebacterium glutamicum*, *Lactococcus lactis*) | (*Salmonella typhimurium* LT2, *Campylobacter jejuni*) (mean 77% support, maximum 100% support with conditioning genomes *Aquifex aeolicus* and

Mesorhizobium loti). The SHOT method also supported this topology, in which the proteobacteria *S. typhimurium* and *C. jejuni* are together. Pairing *C. glutamicum* with *S. typhimurium* had mean 12% and maximum 67% support (conditioning on *Pasteurella multocida*). Pairing *C. glutamicum* with *C. jejuni* had mean 11% and maximum 72% support (conditioning on *Mycoplasma pneumoniae*). As in the first case, any of the three topologies could receive strong bootstrap support from some choices of conditioning genome.

We suggested above that a good conditioning genome should be far from the taxa of interest and contain many gene families. To test this, we did logistic regression analyses (Agresti, 2002, chapter 5) of the four-taxon data sets. We used the pairwise SHOT distance from the conditioning genome to the closest taxon of interest ($dS_{\min}$) and the number of gene families in the conditioning genome to predict bootstrap support for the dominant topology. $dS_{\min}$ is a surrogate for the unknown true evolutionary distance, which may introduce some extra variability. However, we obtained similar results using distances estimated from a stationary Markov model of gene family size (M. Spencer, A.J. Roger and E. Susko, unpublished). We used a binomial logistic model with overdispersion, implemented in R Version 2.0.1 (R Development Core Team, 2004). We did not analyze dataset b, because the dominant topology was almost always the only one found. Table 1 summarizes the results. In all cases, the bootstrap support for the dominant topology was significantly higher for conditioning genomes further from the closest taxon of interest. In all cases except dataset a, bootstrap support for the dominant topology was also significantly higher for larger conditioning genomes. The P values are probably underestimates for two reasons. First, the residual deviance is very large. Second, we treated each conditioning genome as an independent point, ignoring the phylogenetic relationships among them. For dataset c, we analyzed support for the dominant topology, which is probably wrong. Conditions which we would expect to improve performance (a larger conditioning genome, further from the taxa of interest) instead increased support for the wrong topology. This is because the data on shared absence of genes are positively misleading for conditioned logdet distances among parasites. Figure 5 shows the

corresponding univariate relationships (these are easier to visualize than the full model, but the parameter estimates are similar). Overall, $dS_{\min}$ seems to be a more reliable predictor of support for the dominant topology than conditioning genome size. In particular, there are many large conditioning genomes in dataset c for which there is low support for the dominant topology (Figure 5e), even though the overall relationship is positive. In all cases, there is a lot of unexplained variability. In summary, the data provide some support for our suggestion that a large conditioning genome, far from the taxa of interest, might be a good choice. Nevertheless, it may be difficult to choose a good conditioning genome in practice, because there is a lot of unexplained variability, and we do not know the true location of each genome.

Supertree Methods

In this section, we develop ways to avoid choosing a single conditioning genome. We focus on methods that we can show to be consistent. Intuitively, consistency is important because our estimates should approach the true values for very large amounts of data (Silvey, 1975). However, there are other desirable properties, such as simplicity and good performance on small samples. We evaluate small-sample performance for several methods using simulations. However, we have not included every possible method. For example, we do not evaluate standard supertree methods such as Matrix Representation with Parsimony (MRP: Baum and Ragan, 2004) or Average Consensus Supertrees (ACS: Lapointe and Cucumel, 1997). Our methods have some theoretical properties that suggest they should perform well, but there is plenty of scope for empirical evaluations of other methods.

We first show that simply averaging distance matrices over all choices of conditioning genome (the ACS approach) may not give a tree-additive distance matrix. This suggests the need for more sophisticated methods. We describe two: a least-squares approach; and an agglomerative supertree method based on BIONJ (Gascuel, 1997a). We evaluate the performance of these methods and others using 5-taxon and 40-taxon simulations. Lastly, we use the agglomerative supertree method to produce a tree for 40

real bacterial genomes.

Average Distance Matrices are not Generally Tree-Additive

For a set of m taxa, there are m possible choices of conditioning genome and $m-2$ estimates of each pairwise distance. An obvious idea is to average the estimates of each pairwise distance. We might hope that because the distance matrix for any choice of conditioning genome is tree-additive, the average distance matrix will also be tree additive. Unfortunately, this is not generally so.

First, the expectation of the conditioned logdet distance between a pair of taxa depends on the choice of conditioning genome. To understand why this is, consider Figures 1 and 2a. Figure 2a shows that the expectation of the conditioned logdet distance between a pair of taxa (e.g. w, x in Figure 1) depends on the conditioning genome distance (the distance from the conditioning genome to the path connecting the two taxa). Conditioning on taxon c in Figure 1, the conditioning genome distance is $t_v + t_k$. Conditioning on taxon z, the conditioning genome distance is $t_v + t_u + t_z$, and the conditioned logdet distance between w and x has a different expectation. Furthermore, because the conditioning genome distance is different for different pairs, distance matrices from different choices of conditioning genome are not the same up to a scalar. Consider the pairs (w, x) and (y, z) in Figure 1, with conditioning genome c. The conditioning genome distance $t_k + t_v$ to (w, x) is not necessarily the same as the conditioning genome distance $t_k + t_u$ to (y, z). Thus, even if $t_w + t_x = t_y + t_z$, the expectation of the conditioned logdet distance between w and x may be different from the expectation of the conditioned logdet distance between y and z.

Second, we show that when different distance matrices are not the same up to a scalar, the average of a set of tree additive distance matrices is not necessarily tree-additive. Let $\tau^{(k)*} \in \tau^*$ be the true subtree for all taxa other than k. Consider the kth distance matrix $\mathbf{D}^{(k)}$, which is additive on $\tau^{(k)*}$ and has x, yth element $d_{xy}^{(k)}$, the conditioned distance between x and y for conditioning genome k. The kth row and column are missing, but we can extend $\mathbf{D}^{(k)}$ to a matrix $\mathbf{C}^{(k)}$ with no missing elements by subdividing the

appropriate edge on $\tau^{(k)*}$. In other words, we attach taxon k at its correct position in τ^*, setting the unknown terminal edge length to zero. Any arbitrary positive edge length would also work here, because we will choose weights below so that distances involving this edge are ignored. $\mathbf{C}^{(k)}$ is then additive on τ^*. Now consider the weighted sum of squares

$$f(p) = \sum_k \sum_{xy} (\mathbf{C}_{xy}^{(k)} - p_{xy})^2 w_{xy}^{(k)} \tag{5}$$

where $f(p)$ is a function of the estimated overall edge lengths p_{xy}, and w_{xy} are weights:

$$w_{xy}^{(k)} = \begin{cases} 0, & k = x, y \\ 1, & \text{otherwise} \end{cases} \tag{6}$$

The p_{xy} that minimize $f(p)$ are

$$p_{xy} = \left[\sum_k w_{xy}^{(k)} \mathbf{C}_{xy}^{(k)} \right] / \left[\sum_k w_{xy}^{(k)} \right] \tag{7}$$

which are simple weighted averages of the $\mathbf{C}_{xy}^{(k)}$. This is equivalent to averaging over the original $\mathbf{D}_{xy}^{(k)}$ for which $d_{xy}^{(k)}$ was defined. All the $\mathbf{C}^{(k)}$ are in the same subspace C, because they are additive on τ^*. If the weights did not depend on the choice of x, y, then p would be a linear combination of the $\mathbf{C}^{(k)}$, which would be in C and additive on τ^*. However, for our case, the weights are different for different x, y (for fixed k, the weights are given by Equation 6), and in general p will not be in C. Averaging each distance over those matrices for which it is defined is equivalent to the Average Consensus Supertree method (Lapointe and Cucumel, 1997), for the special case where the kth row and column are missing from the kth distance matrix. Lapointe and Cucumel (1997) noted that the distance matrix obtained by ACS may not be additive even if all the input distance matrices are. Such cases can be constructed for all topologies with more than five taxa (details on request), and do occur for conditioned logdet distances (for example, the expected distances in the simulations described in *Simulations* below). This means that ACS will perform poorly for

some choices of edge lengths. The inability to recover the correct topology with increasing amounts of data also suggests poor estimation with small samples. However, the lack of tree-additivity at some edge length settings does not imply this will be the case for all edge length settings. For some trees, ACS might perform well, a conjecture worthy of further investigation.

Summing over Subtrees

An unrooted full topology τ for a set of $m > 4$ taxa can be decomposed into a unique set of $(m-1)$-taxon unrooted subtree topologies $\tau^{(k)} \in \tau$ formed by deleting each conditioning genome $k = 1 \ldots m$ and the resulting internal node of degree 2. For example, the full topology $(((w,x),c),(y,z))$ of the tree in figure 1 has the subtrees: $(w,x)|(y,z)$; $(c,x)|(y,z)$; $(c,w)|(y,z)$; $(w,x)|(c,z)$; and $(w,x)|(c,y)$. None of the other 14 unrooted topologies for these five taxa has the same set of subtrees. We showed in Appendix 1 that given sufficient data we can obtain a tree-additive distance matrix for any choice of conditioning genome, so we can consistently estimate the topology of every subtree. Thus we can consistently estimate the topology but not the edge lengths of the full tree by estimating all the subtrees. With $m \leq 4$, $\tau^{(k)}$ is the same for all three possible unrooted topologies, so we cannot identify the full topology from the subtrees.

Suppose that we select the true subtree topology $\tau^{(k)*}$ using

$$\tau^{(k)*} = \arg\min_{\tau^{(k)}} f(\tau^{(k)}, \mathbf{D}^{(k)}) \tag{8}$$

where $f(\tau^{(k)}, \mathbf{D}^{(k)})$ is some objective function (such as the sum of squares) of the subtree topology and the pairwise distances $\mathbf{D}^{(k)}$ for all taxa other than conditioning genome k. We minimize the objective function separately over all possible $(m-1)$-taxon topologies. Then we can select the full tree topology τ^* that has the minimum sum of objective

functions over all its subtrees:

$$\tau^* = \arg\min_{\tau} \sum_k f(\tau^{(k)}, \mathbf{D}^{(k)}), \tau^{(k)} \in \tau \tag{9}$$

With the limiting logdet distances, the consistency result implies that the sum of squares is zero on each subtree topology formed by deleting one conditioning genome from the true full topology, and is greater than zero on any other subtree topology (assuming that there are no edges of length zero). The sum of these sums of squares over conditioning genomes is zero on the true full topology, and on no other full topology. In principle, τ runs over all possible tree topologies (we do this in the 5-taxon simulations below). In practice, a heuristic would be necessary for all but the smallest numbers of taxa.

The consistency result implies that we could apply a general purpose supertree method, such as quartet supertrees (Piaggio-Talice et al., 2004) or MRP (Baum and Ragan, 2004) to the partial trees constructed from each conditioning genome. These methods have the disadvantage that the partial trees are constructed independently of one another, so information from one choice of conditioned genome only indirectly informs the inference of trees for the other conditioned genomes. The alternative we present here is a method that uses information from all choices of conditioned genome directly in the construction of a global tree. This suggests that our method should perform better. It will be interesting to see whether this is actually the case.

BIONJ for Conditioned Genome Reconstruction

Another approach is to adapt BIONJ (Gascuel, 1997a), an improved version of neighbor-joining. BIONJ is a good choice because it is simple, consistent and fast. We obtain an overall topology for all taxa, but not an overall set of edge lengths, from a set of $m > 4$ taxa. We first describe the original BIONJ algorithm, then explain how it can be adapted. Agglomerative methods such as BIONJ are easy to extend to supertree cases. We would have to develop a completely different approach for optimality-based methods such

as minimum evolution.

The Original BIONJ Algorithm.– At each step of the original BIONJ algorithm, we choose a pair of taxa i and j so as to minimize the criterion

$$Q_{ij} = (r-2)d_{ij} - S_i - S_j \tag{10}$$

where r is the number of taxa remaining, $S_i = \sum_{h=1}^{r} d_{ih}$, and d_{ij} is the distance between taxa i and j. This criterion is guaranteed to select a true pair given sufficient data (Gascuel, 1997b; Bryant, 2005). We then replace i and j with their common ancestor u, and estimate edge lengths from u to i and j using:

$$d_{iu} = \frac{1}{2}\left(d_{ij} + \frac{S_i - S_j}{r-2}\right) \tag{11}$$

BIONJ now chooses a parameter λ so as to minimize the sum of the variances of the new distances:

$$\lambda = \frac{1}{2} + \frac{1}{2(r-2)v_{ij}} \sum_{h=1, h\neq i,j}^{r} (v_{jh} - v_{ih}) \tag{12}$$

where v_{ij} is the variance of d_{ij}, and $0 \leq \lambda \leq 1$. We use d_{ij} as an initial estimate of v_{ij} (up to a constant which we do not need to know). This is reasonable for many evolutionary models (Gascuel, 1997a), including conditioned logdet distances, provided the distances are not too large.

We then calculate the variances v_{ch} of the new distances from each taxon h to the centre c of the cluster $\{i, j\}$, where c is defined such that $d_{ch} = \lambda d_{ih} + (1-\lambda)d_{jh}$:

$$v_{ch} = \lambda v_{ih} + (1-\lambda)v_{jh} - \lambda(1-\lambda)v_{ij} \tag{13}$$

and the new distances

$$d_{uh} = \lambda d_{ih} + (1-\lambda)d_{jh} - \lambda d_{iu} - (1-\lambda)d_{ju} \tag{14}$$

We then decrease r by one and iterate until only three nodes remain.

Adapting BIONJ to obtain a supertree from conditioned logdet distances.– Our modified BIONJ algorithm consists of three steps, that are applied iteratively:

1. Identify a candidate pair to aggregate from each distance matrix
2. Select the best such pair over all distance matrices, and aggregate the subtrees containing this pair in all distance matrices
3. Update the distance matrices

Identifying a candidate pair.– Given a set of m distance matrices, each from a different choice of conditioning genome k, we can modify the BIONJ criterion (equation 10) to choose a candidate pair of taxa ($i^{(k)} \neq k, j^{(k)} \neq k$) to combine from each matrix. Because we have no information about distances involving the conditioning genome, we replace S_i by

$$S_i^{(k)} = \sum_{h=1, h\neq k}^{r^{(k)}} d_{ih}^{(k)} \tag{15}$$

where $d_{ih}^{(k)}$ is the distance between i and h with conditioning genome k. We calculate $S_j^{(k)}$ similarly. The number of taxa remaining under conditioning genome k at any given iteration is $r^{(k)}$. For m taxa, the initial value of $r^{(k)}$ is $m-1$ for all k, because the kth taxon is missing from the distance matrix. At later aggregation steps, $r^{(k)}$ will either be $r-1$ (if taxon k has not yet been aggregated) or r (if it has). For all matrices in which $r^{(k)} > 3$, we obtain the candidate pair $(i^{(k)}, j^{(k)})$ that minimizes

$$Q_{ij}^{(k)} = (r^{(k)} - 2)d_{ij}^{(k)} - S_i^{(k)} - S_j^{(k)} \tag{16}$$

Assume the distances are tree-additive; this is the case for the limiting log-det distance. Then the choice $(i^{(k)}, j^{(k)})$ is consistent for tree-additive distances (Gascuel, 1997a,b), so any such pair will be a true pair in the kth matrix. However, the location of taxon k cannot

be inferred from the kth matrix. We therefore need to check whether the true pair in the overall topology is $(i^{(k)}, j^{(k)})$, $(i^{(k)}, k)$ or $(k, j^{(k)})$, using information from the other matrices.

If taxon k has already been aggregated, then the subtree in which it is contained is represented in the kth distance matrix. Therefore, in this case $(i^{(k)}, j^{(k)})$ are a true pair in the overall topology. If taxon k has not already been aggregated, the true pair may be $(i^{(k)}, j^{(k)})$, $(i^{(k)}, k)$ or $(k, j^{(k)})$. With tree-additive distances, one of these must be a pair of neighbours in the overall topology, if we have sufficient data. We can establish which is the true pair by examining all matrices k' in which $i^{(k)}$, $j^{(k)}$ and k are all present (i.e. $k' \neq i^{(k)}$, $j^{(k)}, k$). We treat $i^{(k)}$, $j^{(k)}$ and k as fixed. In the following three equations, the dependence on k is implicit. We show in Appendix 2 that if a given taxon i has a neighbour, the criterion Q_{iv} for fixed i will be minimized over all other taxa v at that neighbour j. For fixed b, let

$$R_{a|b}^{(k')} = \begin{cases} 1 & Q_{ab}^{(k')} \text{minimal among } a = i^{(k)}, j^{(k)}, k, \text{ with } a \neq b \\ 0 & \text{otherwise} \end{cases} \tag{17}$$

Also let

$$R_{ab} = \sum_{k'} (R_{a|b}^{(k')} + R_{b|a}^{(k')}) w^{(k')} \tag{18}$$

where $w^{(k')}$ is some suitable positive weight, discussed later. Two of $i^{(k)}$, $j^{(k)}$ and k are true neighbours; without loss of generality, assume it is $i^{(k)}$ and $j^{(k)}$. Then for large samples, $R_{i^{(k)}|j^{(k)}}^{(k')} = 1$ and $R_{j^{(k)}|i^{(k)}}^{(k')} = 1$ for all k'. Thus $R_{i^{(k)}j^{(k)}} = 2(m-3)\sum_{k'} w^{(k')}$. For large samples, we also have $R_{k|i^{(k)}}^{(k')} = 0$ for all k', because k is not the true neighbour of $i^{(k)}$. It is possible that $R_{i^{(k)}|k}^{(k')} = 1$ for large samples, but only if k does not have a true neighbour. Thus $R_{i^{(k)}k} \leq (m-3)\sum_{k'} w^{(k')}$. By symmetry, $R_{j^{(k)}k} \leq (m-3)\sum_{k'} w^{(k')}$.

With estimated distances we therefore choose the pair

$$(a', b') = \arg\max_{ab} R_{ab} \quad a, b = (i^{(k)}, j^{(k)}, k), a \neq b \tag{19}$$

We then set the indicator function $I_{a,b}^{(k)} = 1$ if $a = a', b = b'$ and 0 otherwise.

Selecting the best pair overall.– Overall, we will choose the pair (i, j), and aggregate the subtrees containing i and j in all distance matrices where $i \neq k, j \neq k$, using

$$(i, j) = \arg\max_{x,y} \sum_k I_{xy}^{(k)} w^{(k)} \tag{20}$$

We showed above that any candidate pair is a true pair given a large sample size, so any choice of weights $w^{(k)}$ will be consistent. However, a good choice of weight will improve performance on small samples. We consider two choices for the weights.

First, we could choose the pair that was selected as a candidate pair from the largest number of distance matrices with $r^{(k)} > 3$. This means $w^{(k)} = 1$ (the vote-counting method). We will break ties at random:

$$(i, j) = \arg\max_{x,y} \left[\sum_k I_{xy}^{(k)} + U(0, 1/2) \right] \tag{21}$$

The uniform random number $U(0, 1/2)$ breaks ties without changing the ranking of untied pairs.

This method does not account for differences in reliability between distance matrices. Weighting each pair by the inverse of the sum of all variances and covariances between distances would account for such differences. The variances of the distances are inversely proportional to the size of the conditioning genome (Equation 3), and increase as the distances increase (Figure 2). In BIONJ, we assume that variances are proportional to distances, and covariances are proportional to shared path lengths:

$$\mathrm{cov}(d_{ab}^{(k)}, d_{xy}^{(k)}) = \begin{cases} \frac{1}{2}(v_{ab}^{(k)} + v_{bx}^{(k)} - v_{ax}^{(k)}) & b = y \\ 0 & \text{otherwise} \end{cases} \tag{22}$$

(Gascuel, 1997a). The shared path length is zero unless two pairs of taxa have a member in common, because we are estimating covariances on a star tree for those nodes that have not yet been aggregated. Under this assumption, the row of the variance-covariance matrix corresponding to $d_{ab}^{(k)}$ sums to $(r^{(k)} - 1)v_{ab}^{(k)}$. This gives the inverse-variance method for

choosing a pair:

$$w^{(k)} = \frac{n^{(k)}}{(r^{(k)} - 1) \sum_{ab} v_{ab}^{(k)}} \tag{23}$$

where $n^{(k)}$ is the number of genes in the kth conditioning genome and the variances are summed over all taxa that have not yet been aggregated. The term $(r^{(k)} - 1)$ in the denominator appears because the number of taxa not yet aggregated may not be the same for every conditioning genome.

Updating the distance matrices.– At each iteration, we decrease $r^{(k)}$ by one unless $i = k$ or $j = k$. Where either i or j is the conditioning genome k, we do not make any aggregation, the distance matrix remains unchanged until the next iteration, and we do not decrease $r^{(k)}$. We calculate distances $d_{iu}^{(k)}$ to the new node under each conditioning genome k as in equation 11, using $d_{ij}^{(k)}$, $S_i^{(k)}$, $S_j^{(k)}$, and $r^{(k)}$. We estimate the parameter $\lambda^{(k)}$ for distances with the kth conditioning genome using:

$$\lambda^{(k)} = \frac{1}{2} + \frac{1}{2(r^{(k)} - 2)v_{ij}^{(k)}} \sum_{h=1, h \neq i,j,k}^{r^{(k)}} (v_{jh}^{(k)} - v_{ih}^{(k)}) \tag{24}$$

and $0 \leq \lambda^{(k)} \leq 1$. We are excluding variances $v_{ih}^{(k)}$ where $h = k$ because we have no information on them. We substitute $r^{(k)}$, $\lambda^{(k)}$, distances $d^{(k)}$ and variances $v^{(k)}$ into equations 13 and 14, to get distances $d_{uh}^{(k)}$ and variances $v_{ch}^{(k)}$ for each conditioning genome k, for all nodes $h \neq k$.

In some distance matrices, i and/or j may already have been aggregated into different subtrees at an earlier iteration. In such cases, we aggregate the subtrees containing i and j. For example, suppose that we initially have taxa w, x, y, z, c, and that at the first iteration we chose to aggregate (y, z). These taxa will not be aggregated in distance matrices ($k = y$ and $k = z$) where one of the pair is the conditioning genome. At the second iteration, suppose that we chose to aggregate y and w, based on votes from the distance matrices $k = y$ and $k = z$. In the other distance matrices, we will aggregate $((y, z), w)$. Once a pair has been aggregated, it is replaced by a new common ancestor node

in every distance matrix where both members are present. This means that no distance matrix can vote for a pair that has already been aggregated into the same subtree.

Once $r^{(k)} \leq 3$ for any k, this distance matrix makes no further contribution to the choice of topology. The algorithm terminates when $r = 3$. We then have an overall topology in the global list of taxa. When every taxon has been aggregated, $r^{(k)} = r$ in every matrix. This means that at every step with $r \geq 4$, there will be at least one matrix k with $r(k) \geq 4$, so that we will always be able to completely resolve the topology.

The edge length estimates are specific to each distance matrix, and the edge length to the conditioning genome is missing in each case. We therefore cannot get edge lengths on the overall tree.

The algorithm is outlined in Figure 6. The code for the modified version of BIONJ is available at `http://www.liv.ac.uk/~matts/`.

5-taxon simulations

We used simulated 5-taxon data sets to evaluate the performance of the summing-over-subtrees method and the modified BIONJ algorithm. We compared these methods with the original conditioned logdet method, using BIONJ on distances from each single conditioning genome. If we do not combine information from subtrees, we can only claim a correct result that is independent of the choice of conditioning genome if all subtrees are correct. We therefore scored a correct result from the original method only if it recovered the right 4-taxon subtree for every choice of conditioning genome. We also recorded the number of times we correctly recovered the 5-taxon topology using SHOT distances (Korbel et al., 2002) and BIONJ. SHOT distances are not tree-additive evolutionary distances and are not based on an explicit model of genome evolution. Nevertheless, they include a correction for variation in genome size among taxa, and seem to perform well in practice.

We simulated data on the tree shown in Figure 7, with different expected genome sizes in different parts of the tree. All the internal edges and the terminal edges leading to

taxa c, x and y were the same length (0.1: the black edges in Figure 7). The two terminal edges t_w and t_z leading to taxa w and z (gray edges in Figure 7) were varied together between 0.1 and 1. All the black edges in Figure 7 had the rate matrix parameters $q_{01}^{(1)} = 0.2$, $q_{10}^{(1)} = 0.8$. The gray edges had $q_{10}^{(2)} = 0.8$, and $q_{01}^{(2)}$ from 0.1 to 1. We rooted the tree as shown in the figure, and set the probabilities of gene absence and presence at the root to their stationary values, 0.8 and 0.2. We simulated 5000 genes, so the expected number of genes present at the root was 1000. The longer the gray edges, the more strongly the distributions of absence and presence at w and z diverged from those of the other taxa. We ran 1000 replicates of each set of conditions, but excluded any replicates in which the conditioned logdet distances could not be calculated for every conditioning genome. On average, conditioned logdet distances could be calculated for 805 replicates per set of conditions.

Averaging over all conditions, the summing-over-subtrees method performed worst, giving the correct topology in only 45% of replicates (Figure 8c). Even though we proved that this method is consistent, it may require very large numbers of data in order to do well, so we do not discuss it further. BIONJ on SHOT distances recovered the true topology in 61% of cases (Figure 8a: the average success rate and overall pattern of performance for SHOT were unchanged if replicates where conditioned logdet distances could not be calculated where included). Separate BIONJ on conditioned logdet distances with each conditioning genome in turn gave all 4-taxon subtrees correct in 64% of cases (Figure 8b). The modified BIONJ algorithm with vote-counting gave the correct topology in 78% of cases (Figure 8d). Modified BIONJ with inverse-variance weighting was by far the best method, giving the correct topology in 92% of cases (Figure 8e).

When $q_{01}^{(2)}$ is large and the edges t_w and t_z are long, taxa w and z will tend to have much larger genomes than the other taxa. All methods did worse in these conditions (Figure 8, top right-hand corner of each panel), but the modified BIONJ method with inverse variance weighting was least affected. The SHOT method did very well when t_w and t_z were short and $q_{01}^{(2)}$ was small, but very badly when t_w and t_z were long and $q_{01}^{(2)}$ was large. In the extreme case $q_{01}^{(2)} = 1$, $t_w = t_z = 1$, SHOT distances gave the incorrect tree

$(x, ((z, w), y), c)$ from all replicates (240 where conditioned logdet distances could be calculated, and 760 where they could not). BIONJ with SHOT distances gave the same incorrect tree even when distances were calculated from the exact pattern probabilities, and is therefore inconsistent for these parameters. In SHOT, both the correction for multiple changes (which is important with long t_w and t_z) and the correction for variation in genome size (which is important with large $q_{01}^{(2)}$) are approximate, so we expect to find cases where they do not work. This is an example of long-branch attraction with a mis-specified model (Susko et al., 2004). In contrast, all the methods based on conditioned logdet distances gave the correct tree when distances were calculated from the exact pattern probabilities. This supports the claim that these methods are consistent, although there may sometimes be strong small-sample effects.

The difference between vote-counting and inverse-variance weighting in modified BIONJ (Figure 8d and e) is interesting. Both methods are consistent, but vote-counting performs much less well on small samples because it does not take account of the reliability of each distance matrix. For example, when $q_{01}^{(2)} = 1$, $t_w = t_z = 1$, genomes w and z are both large and far from internal nodes. This will tend to increase the reliability of distance matrices conditioned on presence of genes in w or z relative to those from other taxa (Figure 3). Weighting by the reliability of evidence often improves the performance of other supertree methods (Ronquist, 1996; Bininda-Emonds and Sanderson, 2001).

40-taxon simulations

We also evaluated the performance of several different methods on simulated data on a 40-taxon tree. We used the set of 40 bacterial taxa from the COG database for which all conditioned logdet distances were real. We estimated the maximum likelihood tree topology for 16s rRNA sequences (from the Ribosomal Database Project II, release 9, `http://rdp.cme.msu.edu/`, downloaded 3 May 2006) using PHYML Online (Guindon et al., 2005, http://atgc.lirmm.fr/phyml/, accessed 18 May 2006) with discrete gamma rate variation, four gamma categories and a general time-reversible model. There is little

correlation between the edge lengths on the 16s tree and the numbers of gene insertions, duplications and deletions in bacterial genomes (Hao and Golding, 2004). We therefore generated uniform random edge lengths on the interval $[0.005, 0.015]$. For each of 1000 replicate sets of edge lengths, we simulated presence-absence data for 5000 gene families. We used Seq-Gen version 1.3.2 (Rambaut and Grassly, 1997) to generate nucleotide data under the GTR model (parameters -f 0.4 0.1 0.4 0.1 -r 1 0 1 1 0 1), then recoded as presence/absence data ($A, G \to 0$, $C, T \to 1$). This gives a stationary probability of 0.2 for gene presence. For each dataset, we estimated a tree topology using several methods: BIONJ with naive logdet (conditioning on the set of genes present in at least one taxon), BIONJ with SHOT, modified BIONJ with inverse variance weighting, modified BIONJ with vote counting, and the original conditioned logdet method (as above, recording a success only if the correct subtree was obtained for every choice of conditioning genome). Conditioned logdet distances could be calculated for every choice of conditioning genome in 686 replicates. Naive logdet and SHOT gave the correct topology in all replicates (including those where conditioned logdet distances could not be calculated for every choice of conditioning genome). We know that both methods can be inconsistent, but here they performed perfectly. Modified BIONJ gave the correct result in 99% of cases where conditioned logdet distances could be calculated for every conditioning genome, with either inverse-variance weighting or vote counting. Although we proved that this method is consistent, it did not do as well as the simpler methods in this case, perhaps because of high variances (Rosenberg and Kumar, 2003). The lack of difference between inverse-variance weighting and vote-counting may be because ties are less frequent with large numbers of taxa. The original conditioned logdet method gave the right topology for all choices of conditioning genome in only 17% of cases.

We also simulated a heterogeneous case, with smaller equilibrium genome sizes in the parasites (two *Rickettsias*, *Mycoplasma*, *Ureaplasma*, *Borrelia*, and *Treponema*). On the edges leading to these taxa, we changed the frequency parameters to give a stationary probability of 0.03 for gene presence (-f 0.485 0.015 0.485 0.015). Conditioned logdet distances could be calculated for every choice of conditioning genome in only 113 out of

1000 replicates. Again, naive logdet and SHOT performed well, with the correct topology in 99.9% and 97.9% of cases respectively (100% and 96.5% of replicates where all conditioned logdet distances could be calculated). Modified BIONJ with inverse variance weighting gave the correct topology in 93% of cases where all conditioned logdet distances could be calculated. With vote-counting, the correct topology was obtained in 92% of cases. Original conditioned logdet did not give the right topology for all choices of conditioning genome in any replicate. The degree of heterogeneity in these simulations was modest. The parasite genomes still contained around 800 genes in most cases, compared to 1000 in the other taxa. With longer edges leading to the parasites or increased gene loss rate, conditioned logdet distances could only be calculated in a very small proportion of replicates.

From these results, it is clear that we can do much better than the original conditioned logdet method. Naive logdet and SHOT performed very well in these cases, but we know both can be inconsistent. Modified BIONJ did not do quite as well, even though it is consistent for this case. However, in real parasite genomes, only some genes may show increased loss rates, in which case modified BIONJ may not be consistent.

Modified BIONJ Applied to Bacterial Genomes

We applied the modified BIONJ algorithm to the real bacterial genome data analyzed above. We selected the subset of 40 taxa for which all pairwise distances were real. We then generated 1000 bootstrap pseudosamples for these taxa from the gene families in the COG database. We generated these by taking samples of size n with replacement from the set of n gene families present in at least one of the 40 taxa. For each bootstrap replicate, we calculated conditioned logdet distances using each conditioning genome in turn. If any distances were not real, we discarded the replicate and sampled again. We used the modified BIONJ algorithm with inverse-variance weighting to generate a tree for each bootstrap replicate, and obtained a majority consensus tree using PHYLIP CONSENSE (Felsenstein, 2005). We did not attempt to deal with rate variation among

gene families. Methods such as pattern filtering (Lake, 1998; Rivera and Lake, 2004) may be useful here, but are not yet well-developed for gene content data.

Figure 9 shows the results. The most important feature is that the six parasite genomes (two *Rickettsias*, *Mycoplasma*, *Ureaplasma*, *Borrelia*, and *Treponema*) form a clade with 97% bootstrap support. This is almost certainly incorrect. As Figure 4c suggested, conditioned logdet distances do not deal correctly with parallel gene loss in unrelated taxa. There are several other disagreements with commonly-accepted relationships. Three taxa (*Xylella fastidiosa* and *Pseudomonas aeruginosa* from the γ-proteobacteria, and *Ralstonia solanacearum* from the β-proteobacteria) are not placed with the other members of their major groups. The relative positions of *Vibrio cholerae* and the (*Haemophilus influenzae*,*Pasteurella multocida*) pair are reversed in a tree based on the concatenation of 205 proteins (Lerat et al., 2003). A sister relationship between *Sinorhizobium meliloti* and *Mesorhizobium loti* is not supported by a synthesis of 34 gene trees (Bapteste et al., 2005). Placing *Aquifex aeolicus* with the ϵ-proteobacteria and *Thermotoga maritima* with the firmicutes is also controversial, but the usual placement of these thermophiles basal to the other bacteria (e.g. Bocchetta et al., 2000) may also be an artefact (Brochier and Philippe, 2002). Other than these discrepancies, the major groups of bacteria appear as expected in the tree and have strong internal bootstrap support (weakest for the gram positives). Many edges have 100% bootstrap support. This does not mean these edges are certainly correct. Instead, it means that stochastic errors are negligible, while systematic errors may still be important (Phillips et al., 2004; Jeffroy et al., 2006), as they almost certainly are for the parasites.

Discussion

We showed that we can obtain a consistent estimate of tree topology from conditioned logdet distances, for almost any choice of conditioning genome. Nevertheless, sampling variance means that the choice of conditioning genome will be important for real data. In our analyses of real bacterial genomes, different choices of conditioning genome

resulted in strong bootstrap support for different tree topologies.

Supertree methods that combine information from all choices of conditioning genome are preferable. In five-taxon simulations, our method based on a modified BIONJ algorithm with inverse-variance weighting performed much better than the other approaches we tried. There are some possibilities for further improvement. BIONJ assumes that the distance matrix elements have variances proportional to their magnitudes. We could use the variance estimate in Equation 3 instead. However, calculating the covariances between distances would be more complicated, and we expect the gain in performance would be small. In 40-taxon simulations, our method did not do as well as simpler approaches such as SHOT and naive logdet distances. This may be because of higher variances, but there is scope for much more work in this area.

Our method may have other applications. For example, the average consensus supertree method (Lapointe and Cucumel, 1997) attempts to obtain a supertree topology from a matrix of average path lengths on a set of trees. The average matrix is not generally tree-additive, so this method may be inconsistent. Our modified BIONJ algorithm is consistent, and although we only applied it to the case where row and column i are missing from the ith distance matrix, it might be adaptable to other cases of missing distances.

Conditioned genome reconstruction does not correctly place the parasites in a bacterial genome tree based on the COG database. It is possible that other ways of defining orthologs would give different results, because there is strong evidence that phylogenies based on gene content are sensitive to ortholog definitions (Hughes et al., 2005). It will therefore be valuable to test conditioned genome reconstruction on other databases. However, it is also possible that this is an artefact arising from varying rates of gene loss. Snel et al. (2005) draw an interesting parallel between parallel gene loss in parasites and the kind of heterotachy studied by Kolaczkowski and Thornton (2004) for DNA sequence models. In both cases, there is a subset of observations (sites or genes) showing the same change in the rate of evolution on widely-separated edges. If this is not accounted for in a model of evolution, the wrong topology may be inferred. Kolaczkowski and Thornton (2004) suggested that parsimony might be less vulnerable than maximum

likelihood to the effects of heterotachy, but this is highly controversial (Gadagkar and Kumar, 2005; Gaucher and Miyamoto, 2005; Lockhart et al., 2005; Philippe et al., 2005; Spencer et al., 2005). Snel et al. (2005) similarly suggested that ad-hoc corrections for genome size such as SHOT might be preferable to model-based methods for gene content data. SHOT is not misled by parallel gene loss, because it ignores shared absences. However, we showed that variation in genome size can cause SHOT to be inconsistent, in cases where conditioned logdet distances are consistent and perform well. Thus no current method is invulnerable to artefacts caused by changing evolutionary processes across the tree. If the parasite group is not simply an artefact of the COG database, the best solution might be to develop mixture models for gene content: such models work well for heterotachy in DNA sequences (Spencer et al., 2005) and have been shown to be identifiable in some cases (Allman and Rhodes, 2006).

The interpretation of a genome phylogeny is less clear than the interpretation of a gene phylogeny. Lake and Rivera (2004) suggested that 'any method that can properly model genomic evolution will be invariant to the confounding effects of HGT [horizontal gene transfer].' However, such a method would have to properly account for the sources of lateral transfers. All of our results are based on the assumption of independent evolution along edges. If lateral transfers are common between taxa that share the same habitat, this assumption may be violated, and the tree estimated by a gene content method may not reflect the tree of vertical descent. Such cases may or may not be common, but if they occur one might not reliably estimate the tree that best represents vertical transmission. We could think of the tree as representing the dominant pathways of transmission, whether vertical or horizontal (Wolf et al., 2002). Alternatively, we could pursue phylogenetic methods such as Neighbor-Net (Bryant and Moulton, 2004) that do not impose a tree-like model on non-tree-like data. The interpretation of trees obtained from conditioned genome reconstruction has been criticized (Bapteste and Walsh, 2005). Nevertheless, the use of conditioned logdet distances is a useful step towards more rigorous analyses of gene content data. The major challenges facing such analyses are parallel gene loss and non-random lateral gene transfer.

Acknowledgments

This work was funded by the Genome Atlantic/Genome Canada Prokaryotic Genome Evolution and Diversity Project. E.S. is supported by the Natural Sciences and Engineering Council of Canada. We are grateful to Eric Bapteste, Robert Charlebois, Ford Doolittle, Uri Gophna, James Lake, Andrew Roger, Berend Snel, Olga Zhaxybayeva and the Statistical Evolutionary Bioinformatics group at Dalhousie for comments and suggestions. The manuscript was greatly improved by constructive criticism from Cécile Ané, Olivier Gascuel, and two anonymous referees.

References

Agresti, A. 2002. Categorical Data Analysis. Second ed. John Wiley and Sons, Hoboken.

Allman, E. S. and J. A. Rhodes. 2006. The identifiability of tree topology for phylogenetic models, including covarion and mixture models. J. Comput. Biol. 13:1101–1113.

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

Bapteste, E. and D. A. Walsh. 2005. Does the 'Ring of Life' ring true? Trends Microbiol. 13:256–261.

Barry, D. and J. A. Hartigan. 1987. Asynchronous distance between homologous DNA sequences. Biometrics 43:261–276.

Baum, B. R. and M. A. Ragan. 2004. The mrp method. Pages 17–34 *in* Phylogenetic supertrees: combining information to reveal the tree of life (O. R. P. Bininda-Emonds, ed.). Kluwer Academic Publishers, Dordrecht.

Bininda-Emonds, O. R. P. and M. J. Sanderson. 2001. Assessment of the accuracy of matrix representation with parsimony analysis supertree construction. Syst. Biol. 50:565–579.

Bocchetta, M., S. Gribaldo, A. Sanangelantoni, and P. Cammarano. 2000. Phylogenetic depth of the bacterial genera Aquifex and Thermotoga inferred from analysis of ribosomal protein, elongation factor, and RNA polymerase subunit sequences. J. Mol. Evol. 50:366–380.

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

Bryant, D. 2005. On the uniqueness of the selection criterion in neighbor-joining. J. Classif. 22:3–15.

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

Chang, J. T. and J. A. Hartigan. 1991. Reconstruction of evolutionary trees from pairwise distributions on current species. Pages 254–257 *in* Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface (E. M. Keramidas, ed.).

Doolittle, W. F., Y. Boucher, C. L. Nesbø, C. J. Douady, J. O. Andersson, and A. J. Roger. 2003. How big is the iceberg of which organellar genes in nuclear genomes are but the tip? Phil. Trans. Roy. Soc. Lond. B Biol. Sci. 358:39–58.

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

Felsenstein, J. 2005. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

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

Gadagkar, S. R. and S. Kumar. 2005. Maximum likelihood outperforms maximum parsimony even when evolutionary rates are heterotachous. Mol. Biol. Evol. 22:2139–2141.

Gascuel, O. 1997a. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol. Biol. Evol. 14:685–695.

Gascuel, O. 1997b. Concerning the NJ algorithm and its unweighted version, UNJ. Pages 149–170 *in* Mathematical Hierarchies and Biology (B. Mirkin, F. R. McMorris, F. S. Roberts, and A. Rzhetsky, eds.) DIMACS Series in Discrete Mathematics and Theoretical Computer Science. American Mathematical Society, Providence.

Gaucher, E. A. and M. M. Miyamoto. 2005. A call for likelihood phylogenetics even when the process of sequence evolution is heterogeneous. Mol. Phylogenet. Evol. 37:928–931.

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

Graur, D. and W.-H. Li. 2000. Fundamentals of Molecular Evolution. second ed. Sinauer, Massachusetts.

Gu, X. 2000. A simple evolutionary model for genome phylogeny based on gene content. Pages 515–523 *in* Comparative Genomics (D. Sankoff and J. H. Nadeau, eds.). Kluwer Academic Publishers, Dordrecht.

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

Guindon, S., L. F., P. Duroux, and O. Gascuel. 2005. Phyml online – a web server for fast maximum likelihood-based phylogenetic inference. Nucleic Acids Res. 33:W557–W559.

Hao, W. and G. B. Golding. 2004. Patterns of bacterial gene movement. Mol. Biol. Evol. 21:1294–1307.

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

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

Jeffroy, O., H. Brinkmann, F. Delsuc, and H. Philippe. 2006. Phylogenomics: the beginning of incongruence? Trends Genet. 22:225–231.

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

Korbel, J. O., B. Snel, M. A. Huynen, and P. Bork. 2002. SHOT: a web server for the construction of genome phylogenies. Trends Genet. 18:158–162.

Lake, J. A. 1994. Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances. Proc. Natl. Acad. Sci. Unit. States Am. 91:1455–1459.

Lake, J. A. 1998. Optimally recovering rate variation information from genomes and sequences: pattern filtering. Mol. Biol. Evol. 15:1224–1231.

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

Lapointe, F.-J. and G. Cucumel. 1997. The average consensus procedure: combination of weighted trees containing identical or overlapping sets of taxa. Syst. Biol. 46:306–312.

Lerat, E., V. Daubin, and N. A. Moran. 2003. From gene trees to organismal phylogeny in prokaryotes: the case of the γ-proteobacteria. PLoS Biology 1:101–109.

Lockhart, P., N. P., B. G. Milligan, J. Riden, A. Rambaut, and T. Larkum. 2005. Heterotachy and tree building: a case study with plastids and eubacteria. Mol. Biol. Evol. 23:40–45.

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

Martin, A. P. and T. M. Burg. 2002. Perils of paralogy: using HSP70 genes for inferring organismal phylogenies. Syst. Biol. 51:570–587.

Montague, M. G. and C. A. Hutchinson III. 2000. Gene content phylogeny of herpesviruses. Proc. Natl. Acad. Sci. Unit. States Am. 97:5334–5339.

Norris, J. R. 1997. Markov Chains. Cambridge University Press, Cambridge, England.

Philippe, H., Y. Zhou, H. Brinkmann, N. Rodrigue, and F. Delsuc. 2005. Heterotachy and long-branch attraction in phylogenetics. BMC Evol. Biol. 5:50.

Phillips, M. J., F. Delsuc, and D. Penny. 2004. Genome-scale phylogeny and the detection of systematic biases. Mol. Biol. Evol. 21:1455–1458.

Piaggio-Talice, R., J. G. Burleigh, and O. Eulenstein. 2004. Quartet supertrees. Pages 173–191 *in* Phylogenetic supertrees: combining information to reveal the tree of life (O. R. P. Bininda-Emonds, ed.). Kluwer Academic Publishers, Dordrecht.

R Development Core Team. 2004. R: A language and environment for statistical computing. R Foundation for Statistical Computing Vienna, Austria.

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.

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

Ronquist, F. 1996. Matrix representation of trees, redundancy, and weighting. Syst. Biol. 45:247–253.

Rosenberg, M. S. and S. Kumar. 2003. Heterogeneity of nucleotide frequencies among evolutionary lineages and phylogenetic inference. Mol. Biol. Evol. 20:610–621.

Silvey, S. D. 1975. Statistical Inference. Chapman and Hall, London.

Singer, B. and S. Spilerman. 1976. The representation of social processes by markov models. Am. J. Sociol. 82:1–54.

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

Snel, B., P. Bork, and M. A. Huynen. 2002. Genomes in flux: the evolution of archaeal and proteobacterial gene content. Genome Res. 12:17–25.

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

Spencer, M., E. Susko, and A. J. Roger. 2005. Likelihood, parsimony, and heterogeneous evolution. Mol. Biol. Evol. 22:1161–1164.

Spencer, M., E. Susko, and A. J. Roger. 2006. Modelling prokaryote gene content. Evol. Bioinformatics Online 2:165–186.

Susko, E., Y. Inagaki, and A. J. Roger. 2004. On inconsistency of the neighbor-joining, least squares, and minimum evolution estimation when substitution processes are incorrectly modeled. Mol. Biol. Evol. 21:1629–1642.

Tatusov, R. L., N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V. Koonin, D. M. Krylov, R. Mazumder, S. L. Mekhedov, A. V. Sverdlov, S. Vasudevan, Y. I. Wolf, J. J. Yin, and D. A. Natale. 2003. The cog database: an updated version includes eukaryotes. BMC Bioinformatics 4:41.

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

Tillier, E. R. M. and R. A. Collins. 1995. Neighbor joining and maximum likelihood with rna sequences: addressing the interdependence of sites. Mol. Biol. Evol. 12:7–15.

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

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

Zhang, H. and X. Gu. 2004. Maximum likelihood for genome phylogeny on gene content. Stat. Appl. Genet. Mol. Biol. 3:article 31.

Appendix 1: Tree-Additivity and Non-Negativity of Conditioned Logdet

The Unconditioned Case

We first briefly review a proof that the logdet distance between two taxa w and x is tree-additive and non-negative if all genes are observable. For simplicity, we assume that the path from w to x includes one internal node v and two edges with separate Markov processes, but the generalization to many internal edges with different Markov processes is simple.

The unconditional logdet distance between w and x is defined in Equation 1. Let $\mathbf{P}^{(wx)}$ be a stochastic matrix with ijth entry $p_{ij}^{(wx)}$, the probability that taxon x has state j, given that taxon w has state i. The pattern probabilities satisfy

$$\begin{aligned} \mathbf{F}^{(wx)} &= \mathbf{\Pi}^{(w)}\mathbf{P}^{(wx)} \\ &= \mathbf{\Pi}^{(w)}\mathbf{P}^{(wv)}\mathbf{P}^{(vx)} \end{aligned} \tag{25}$$

If v is the common ancestor of w and x, then $\mathbf{P}^{(wv)}$ is a time-reversed process. Provided the forward process has an irreducible transition matrix, this time-reversed process will also be a Markov process, although it will not necessarily be the same as the forward process (Norris, 1997, pp. 47-48).

Pre- and post-multiplying by the $-1/2$ powers of the state probabilities gives

$$\begin{aligned} [\mathbf{\Pi}^{(w)}]^{-1/2}\mathbf{F}^{(wx)}[\mathbf{\Pi}^{(x)}]^{-1/2} &= [\mathbf{\Pi}^{(w)}]^{-1/2}\mathbf{\Pi}^{(w)}\mathbf{P}^{(wx)}[\mathbf{\Pi}^{(x)}]^{-1/2} \\ &= [\mathbf{\Pi}^{(w)}]^{-1/2}\mathbf{\Pi}^{(w)}\mathbf{P}^{(wv)}[\mathbf{\Pi}^{(v)}]^{-1}\mathbf{\Pi}^{(v)}\mathbf{P}^{(vx)}[\mathbf{\Pi}^{(x)}]^{-1/2} \\ &= [\mathbf{\Pi}^{(w)}]^{-1/2}\mathbf{F}^{(wv)}[\mathbf{\Pi}^{(v)}]^{-1/2}[\mathbf{\Pi}^{(v)}]^{-1/2}\mathbf{F}^{(vx)}[\mathbf{\Pi}^{(x)}]^{-1/2} \end{aligned} \tag{26}$$

Note that $\log\det(\mathbf{AB}) = \log\det\mathbf{A} + \log\det\mathbf{B}$. Then the last line of equation 26 shows that $d_{wx} = d_{wv} + d_{vx}$. Thus the standard logdet distances are tree-additive.

To prove that the distance between a pair of adjacent nodes w and v is

non-negative, we first write

$$\begin{aligned}[\mathbf{\Pi}^{(w)}]^{-1/2}\mathbf{F}^{(wv)}[\mathbf{\Pi}^{(v)}]^{-1/2} &= [\mathbf{\Pi}^{(w)}]^{-1/2}\mathbf{\Pi}^{(w)}\mathbf{P}^{(wv)}[\mathbf{\Pi}^{(v)}]^{-1/2} \\ &= [\mathbf{\Pi}^{(w)}]^{1/2}\mathbf{P}^{(wv)}[\mathbf{\Pi}^{(v)}]^{-1/2} \\ &= [\mathbf{\Pi}^{(w)}]^{-1/2}[\mathbf{P}^{(vw)}]^{\mathrm{T}}[\mathbf{\Pi}^{(v)}]^{1/2}\end{aligned} \tag{27}$$

where $^{\mathrm{T}}$ indicates transpose. Then

$$\begin{aligned}d_{wv} &= \frac{1}{4}\{-\log\det[\mathbf{\Pi}^{(w)}]^{1/2}\mathbf{P}^{(wv)}[\mathbf{\Pi}^{(v)}]^{-1/2} - \log\det[\mathbf{\Pi}^{(w)}]^{-1/2}[\mathbf{P}^{(vw)}]^{\mathrm{T}}[\mathbf{\Pi}^{(v)}]^{1/2}\} \\ &= \frac{1}{4}\{-\log\det\mathbf{P}^{(wv)} - \log\det\mathbf{P}^{(vw)}\}\end{aligned} \tag{28}$$

We know that $\log\det\mathbf{P} = \mathrm{tr}(\log\mathbf{P})$, where tr is the trace (the sum of the diagonal elements) and log is the matrix logarithm. Let $\mathbf{Q}^{(wv)} = 1/t_{wv}\log\mathbf{P}^{(wv)}$ be the instantaneous rate matrix along the edge of length $t_{wv} = t_{vw} \geq 0$ from w to v. We then have

$$d_{wv} = \frac{1}{4}\{-\mathrm{tr}(\mathbf{Q}^{(wv)}t_{wv}) - \mathrm{tr}(\mathbf{Q}^{(vw)}t_{vw})\} \tag{29}$$

For an instantaneous rate matrix, -1 times the trace is the sum of the rates of leaving each state. This is positive unless every state is absorbing, so the logdet distance is nonnegative. Also, if $t_{wv} = 0$, the logdet distance is zero.

The Conditioned Case

The logdet distance between w and x, calculated for genes present in a conditioning genome, is defined in Equation 2 and can be rewritten

$$d'_{wx} = -1/2\log\det[\mathbf{\Pi}'^{(w)}]^{1/2}\mathbf{P}'^{(wx)}[\mathbf{\Pi}'^{(x)}]^{-1/2} \tag{30}$$

$\mathbf{P}'^{(wx)}$ is defined as in the unconditioned case, except that it includes only those genes present in the conditioning genome.

There are three different rooted trees for a pair of taxa w and x and a conditioning

genome c (figure 10). However, cases ii and iii differ only in the labelling of the taxa of interest, so we do not need to consider case iii separately. We now show that conditioned logdet distances are tree-additive and non-negative for cases i and ii, so long as the $\mathbf{\Pi}'$ and $\mathbf{F}'$ matrices are not singular.

Consider case i in figure 10. Connecting the conditioning genome c to the path from w to x separates the edge wx into two new edges wv and vx, meeting at a new internal node v. $\mathbf{F}'^{(wx)}$ has ijth entry

$$\begin{aligned} f_{ij}'^{(wx)} &= \Pr\{X_x = j, X_w = i | X_c = 1\} \\ &= \Pr\{X_x = j | X_w = i, X_c = 1\} \Pr\{X_w = i | X_c = 1\} \end{aligned} \tag{31}$$

Each of the terms is conditional on genes being present in the conditioning genome. The first term in equation 31 can be written as

$$\begin{aligned} \Pr\{X_x = j | X_w = i, X_c = 1\} &= \sum_k \Pr\{X_x = j | X_v = k, X_c = 1\} \Pr\{X_v = k | X_w = i, X_c = 1\} \\ &= \sum_k p_{ik}'^{(wv)} p_{kj}'^{(vx)} \\ &= \left[\mathbf{P}'^{(wv)} \mathbf{P}'^{(vx)}\right]_{ij} \end{aligned} \tag{32}$$

Here, $p_{ik}'^{(wv)}$ is the ikth conditional transition probability in $\mathbf{P}'^{(wv)}$.

The second term in equation 31 can be written as

$$\begin{aligned} \Pr\{X_w = i | X_c = 1\} &= \sum_k \Pr\{X_w = i | X_v = k\} \Pr\{X_v = k | X_c = 1\} \\ &= \sum_k \Pr\{X_w = i | X_v = k\} \sum_l \Pr\{X_v = k | X_r = l\} \Pr\{X_r = l | X_c = 1\} \end{aligned} \tag{33}$$

Let $\pi_l'^{(r)} = \Pr\{X_r = l | X_c = 1\}$ be the probability of state l at node r conditional on

state 1 in the conditioning genome. Then

$$\begin{aligned}\sum_l \Pr\{X_v = k|X_r = l\} \Pr\{X_r = l|X_c = 1\} &= \sum_l p_{lk}^{(rv)} \pi_l'^{(r)} \\ &= \pi_k'^{(v)}\end{aligned} \tag{34}$$

We can then rewrite equation 33 as

$$\begin{aligned}\Pr\{X_w = i|X_c = 1\} &= \sum_k \Pr\{X_w = i|X_v = k\} \pi_k'^{(v)} \\ &= \sum_k p_{ki}^{(vw)} \pi_k'^{(v)} \\ &= \pi_i'^{(w)}\end{aligned} \tag{35}$$

Finally, we can substitute equations 32 and 35 back into equation 31 to get

$$f_{ij}'^{(wx)} = \pi_i'^{(w)} \left[\mathbf{P}'^{(wv)} \mathbf{P}'^{(vx)}\right]_{ij} \tag{36}$$

The proof that conditioned logdet distances are tree-additive is then exactly as in the unconditional case.

The proof that conditioned logdet distances are non-negative is a little different. In the unconditional case, we used the properties of the instantaneous rate matrix $\mathbf{Q}$. We have not shown that there is an instantaneous rate matrix $\mathbf{Q}'$ corresponding to a conditional transition probability matrix $\mathbf{P}'$.

When going away from the conditioning genome, e.g. $v \to w$ in case i:

$$\begin{aligned}\Pr\{X_w = i|X_v = k, X_c = 1\} &= \Pr\{X_w = i|X_v = k\} \Pr\{X_v = k|X_c = 1\} \\ &= p_{ki}^{(vw)} \pi_k'^{(v)}\end{aligned} \tag{37}$$

Thus, going away from the conditioning genome, we have the same $\mathbf{P}$ matrix as in the unconditional case, and can use the same proof.

Going towards the conditioning genome, e.g. $v \leftarrow w$ in case i:

$$\begin{aligned}
[\mathbf{P}'^{(wv)}]^{\mathrm{T}}\mathbf{\Pi}'^{(w)} &= \mathbf{\Pi}'^{(v)}\mathbf{P}^{(vw)} \\
\det([\mathbf{P}'^{(wv)}]^{\mathrm{T}}\mathbf{\Pi}'^{(w)}) &= \det(\mathbf{\Pi}'^{(v)}\mathbf{P}^{(vw)}) \\
\det[\mathbf{P}'^{(wv)}]^{\mathrm{T}} \det \mathbf{\Pi}'^{(w)} &= \det \mathbf{\Pi}'^{(v)} \det \mathbf{P}^{(vw)} \\
\det \mathbf{P}'^{(wv)} \det \mathbf{\Pi}'^{(w)} &= \det \mathbf{\Pi}'^{(v)} \det \mathbf{P}^{(vw)} \\
\det \mathbf{P}'^{(wv)} &= \frac{\det \mathbf{\Pi}'^{(v)}}{\det \mathbf{\Pi}'^{(w)}} \det \mathbf{P}^{(vw)}
\end{aligned} \tag{38}$$

For any stochastic matrix $\mathbf{M}$, we know that

$$0 \leq |\det \mathbf{M}| \leq 1 \tag{39}$$

(because the determinant is the product of the eigenvalues, and none of these has absolute value greater than 1). This means that $\det \mathbf{P}'^{(wv)} \leq 1$. For $\mathbf{\Pi}'^{(v)}$ and $\mathbf{\Pi}'^{(w)}$, we know that the determinants are non-negative (because they are the products of the probabilities of states 0 and 1). This fact, together with equation 38 and inequality 39, gives us

$$0 \leq \det \mathbf{P}^{(vw)} \leq 1 \leftrightarrow 0 \leq \det \mathbf{P}'^{(wv)} \leq 1 \tag{40}$$

If there was a continuous-time Markov process along the edge in the unconditional case, then we know that $\log \det \mathbf{P}^{(vw)} \leq 0$, with equality only when the expected number of substitutions is zero (as shown above in the unconditional case). This establishes that the LHS of relation 40 is true. The conditioned logdet distance (equation 30) is non-negative if both the $\mathbf{\Pi}'$ matrices and the $\mathbf{P}'$ matrix have determinants between zero and one, which we have just shown. This also establishes that there is an instantaneous rate matrix $\mathbf{Q}'$ corresponding to $\mathbf{P}'$ for the two-state case (reviewed in Singer and Spilerman, 1976, pp. 9-10).

Case ii differs from case i in that the path from w to x includes r as well as v, but

the proof is similar. The first term in equation 31 becomes

$$\begin{aligned}
&\Pr\{X_x = j | X_w = i, X_c = 1\} = \\
&\sum_l \Pr\{X_x = j | X_r = l, X_c = 1\} \sum_k \Pr\{X_r = l | X_v = k, X_c = 1\} \Pr\{X_v = k | X_w = i, X_c = 1\} \\
&\qquad = \sum_l \sum_k p'^{(wv)}_{ik} p'^{(vr)}_{kl} p'^{(rx)}_{lj} \\
&\qquad = \left[\mathbf{P}'^{(wv)} \mathbf{P}'^{(vr)} \mathbf{P}'^{(rx)}\right]_{ij}
\end{aligned} \tag{41}$$

For case ii, let $\pi'^{(v)}_k = \Pr\{X_v = k | X_c = 1\}$. The second term in equation 31 then becomes

$$\begin{aligned}
\Pr\{X_w = i | X_c = 1\} &= \sum_k \Pr\{X_w = i | X_v = k\} \Pr\{X_v = k | X_c = 1\} \\
&= \sum_k p^{(vw)}_{ki} \pi'^{(v)}_k \\
&= \pi'^{(w)}_i
\end{aligned} \tag{42}$$

so we finish with

$$f'^{(wx)}_{ij} = \pi'^{(w)}_i \left[\mathbf{P}'^{(wv)} \mathbf{P}'^{(vr)} \mathbf{P}'^{(rx)}\right]_{ij} \tag{43}$$

Proving tree-additivity and non-negativity is then the same as in case i, but with two internal nodes instead of one.

Appendix 2: If i has a Neighbour, Minimizing Q_{iv} will Find that Neighbour

We want to show that if a fixed taxon i has a neighbour, minimizing the criterion Q_{iv} over all other taxa v will identify that neighbour, for tree-additive distances on a bifurcating tree with positive internal edge lengths. We have dropped the superscripts indicating the conditioning genome for clarity. Let

$$\begin{aligned} g(v) &= d_{iv} - (S_i + S_v)/(r-2) \\ &= Q_{iv}/(r-2) \end{aligned} \tag{44}$$

We therefore need to show that if i has neighbour j, then $g(v)$ is minimized at j. Let u be the internal node that is the parent of i and j. Let s be the internal node adjacent to u with $d_{us} > 0$, and let l be some terminal node other than i, j.

$$\begin{aligned} S_j - S_l &= \sum_v d_{jv} - d_{lv} \\ &= (d_{ij} - d_{il}) + (d_{jj} - d_{jl}) + (d_{jl} - d_{ll}) + \sum_{v \neq i,j,l} d_{jv} - d_{lv} \\ &= (d_{ij} - d_{il}) + \sum_{v \neq i,j,l} d_{jv} - d_{lv} \end{aligned} \tag{45}$$

For $v \neq i, j, l$, the path from j to v includes the nodes s and u. Using the triangle inequality,

$$\begin{aligned} d_{jv} - d_{lv} &= d_{ju} + d_{us} + d_{sv} - d_{lv} \\ &\geq d_{ju} + d_{us} + d_{sv} - (d_{ls} + d_{sv}) \\ &= d_{ju} + d_{us} - d_{ls} \end{aligned} \tag{46}$$

Since the path from l to i includes nodes s and u, $d_{li} = d_{ls} + d_{su} + d_{ui}$, or $d_{ls} = d_{li} - d_{su} - d_{ui}$. Substituting this into Equation 46,

$$\begin{aligned} d_{jv} - d_{lv} &\geq d_{ju} + d_{us} - d_{li} + d_{su} + d_{ui} \\ &= d_{ij} - d_{li} + 2d_{us} \\ &> d_{ij} - d_{il} \end{aligned} \tag{47}$$

Substituting in Equation 45,

$$\begin{aligned} S_j - S_l &> (d_{ij} - d_{il}) + (r-3)(d_{ij} - d_{il}) \\ &= (r-2)(d_{ij} - d_{il}) \end{aligned} \tag{48}$$

so that

$$\begin{aligned} g(l) - g(j) &= d_{il} - (S_i + S_l)/(r-2) - d_{ij} + (S_i + S_j)/(r-2) \\ &= (d_{il} - d_{ij}) + (S_j - S_l)/(r-2) \\ &> (d_{il} - d_{ij}) + (d_{ij} - d_{il}) \\ &= 0 \end{aligned} \tag{49}$$

Table 1: Logistic regression analyses of the datasets in Figure 4.

	Estimate[a]	Standard error	P[b]
Dataset a			
$dS_{\min}$[c]	5.10	1.65	3×10^{-3}
Conditioning genome size[d]	-6.93×10^{-5}	3.85×10^{-4}	0.86
Null deviance	20371		
Residual deviance	15934		
Dataset c			
$dS_{\min}$	15.68	2.13	4×10^{-9}
Conditioning genome size	1.69×10^{-3}	4.18×10^{-4}	2×10^{-4}
Null deviance	26068		
Residual deviance	8339		
Dataset d			
$dS_{\min}$	7.80	2.10	6×10^{-4}
Conditioning genome size	1.86×10^{-3}	4.35×10^{-4}	1×10^{-4}
Null deviance	18318		
Residual deviance	11920		

[a] The response is bootstrap support for the dominant topology. We assumed a binomial logistic regression model with overdispersion. Entries are parameter estimates on the logit scale. We did not analyze dataset b because there was so little variation in the dominant topology.

[b] In all cases, there were 45 null and 43 residual d.f.

[c] SHOT distance from the conditioning genome to the closest taxon of interest

[d] Number of gene families in conditioning genome

Figure 1: A four-taxon tree with a conditioning genome. The taxa of interest are w, x, y and z, and c is the conditioning genome. The node connecting c to the rest of the tree is k. Edge lengths are $t_k \ldots t_z$.

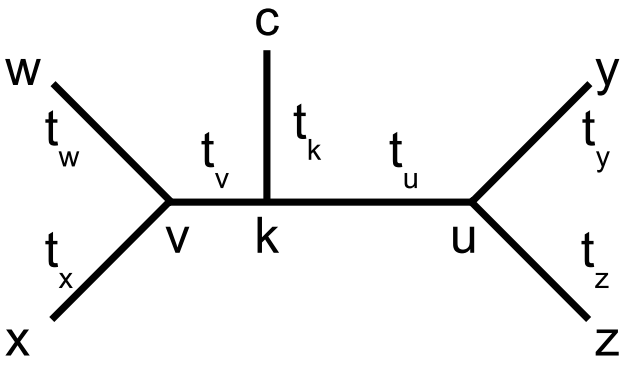

Figure 2: (a) Unconditioned (dashed line: equation 1) and conditioned (solid line: equation 2) logdet distances d'_{wx} between a pair of taxa w and x (see figure 1 for definitions of edge lengths). The conditioning genome c is a distance $t_c = t_k + t_v$ from the path separating the pair of taxa. (b) Approximate sampling variance $\sigma^2(\hat{d}'_{wx})$ (equation 3) for unconditioned (dashed line) and conditioned (solid line) logdet distances, assuming a sample of 1000 genes in each case. Other parameters: $t_w = 0.1$, $t_x = 0.2$; instantaneous rate of gene gain 0.5625; instantaneous rate of gene loss 4.5. Logdet distances were calculated from the expected pattern frequencies.

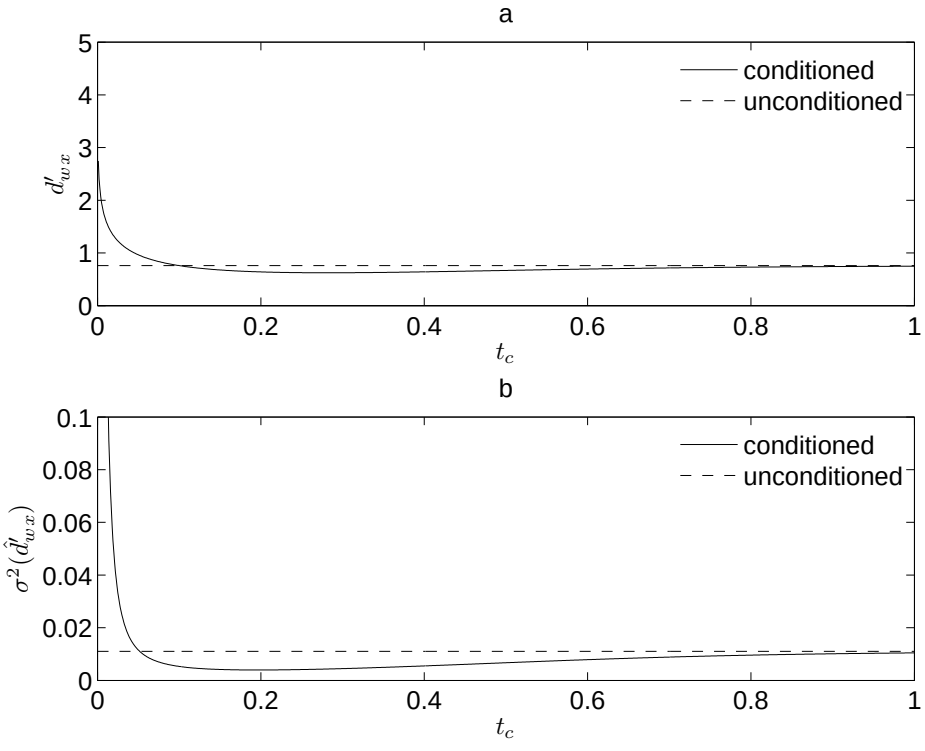

Figure 3: Recovery of a four-taxon subtree using conditioned logdet distances on simulated data, with a single conditioning genome: (a) proportion of replicates for which logdet distances could be calculated; (b) proportion of such replicates for which the correct tree topology (excluding the conditioning genome) was recovered by unweighted least squares with no constraints on edge lengths. In all cases, we simulated 1000 replicates, and calculated distances from 5000 genes. Parameters (see figure 1): $t_w = t_x = t_y = t_z = 0.1$; $t_u + t_v = 0.02$ (varying t_u between 0.01 and 0.02); t_k between 0 and 0.01; instantaneous rate of gene gain 0.625; instantaneous rate of gene loss 2.5.

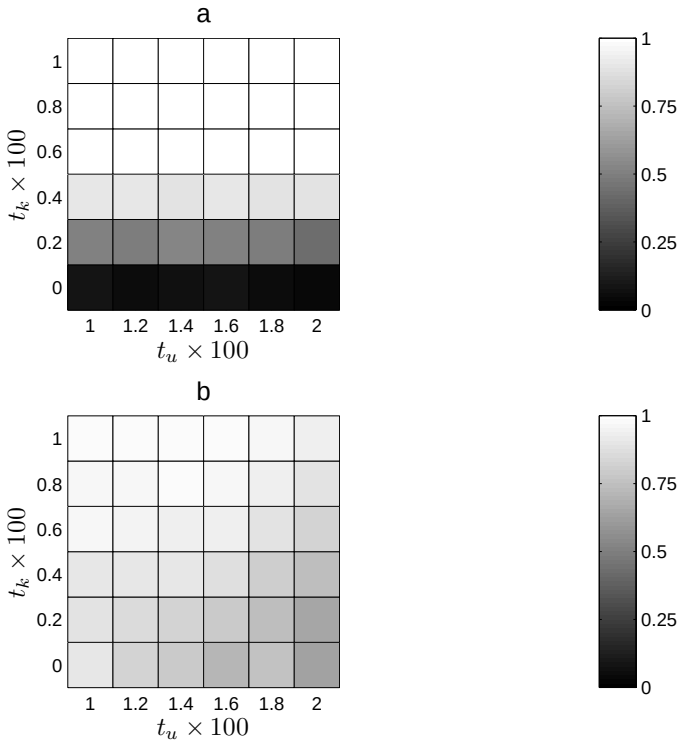

Figure 4: Bootstrap proportions of the three topologies wx, wy and wz inferred using conditioned logdet distances and unweighted least-squares for four-taxon subsets of a 50-taxon, 4873-gene-family bacterial genome database. Each point is from one choice of conditioning genome. The vertices are 100% support for one topology, and open circles at the vertices indicate what we think are the correct topologies. The four-taxon data sets were: (a) w=*Synechocystis* sp., x=*Escherichia coli* K12, y=*Mesorhizobium loti*, z=*Mycoplasma genitalium*; (b) w=*Bacillus subtilis*, x=*Bacillus halodurans*, y=*Haemophilus influenzae*, z=*Pasteurella multocida*; (c) w=*Aquifex aeolicus*, x=*Yersinia pestis*, y=*Buchnera* sp. APS, z=*Ureaplasma urealyticum*; (d) w=*Corynebacterium glutamicum*, x=*Lactococcus lactis*, y=*Salmonella typhimurium* LT2, z=*Campylobacter jejuni*. In each case, each of the remaining 46 bacterial taxa from the COG database (downloaded 13 May 2004 from ftp://ftp.ncbi.nih.gov/pub/COG/) was used as a conditioning genome, and 1000 bootstrap replicates were run after conditioning.

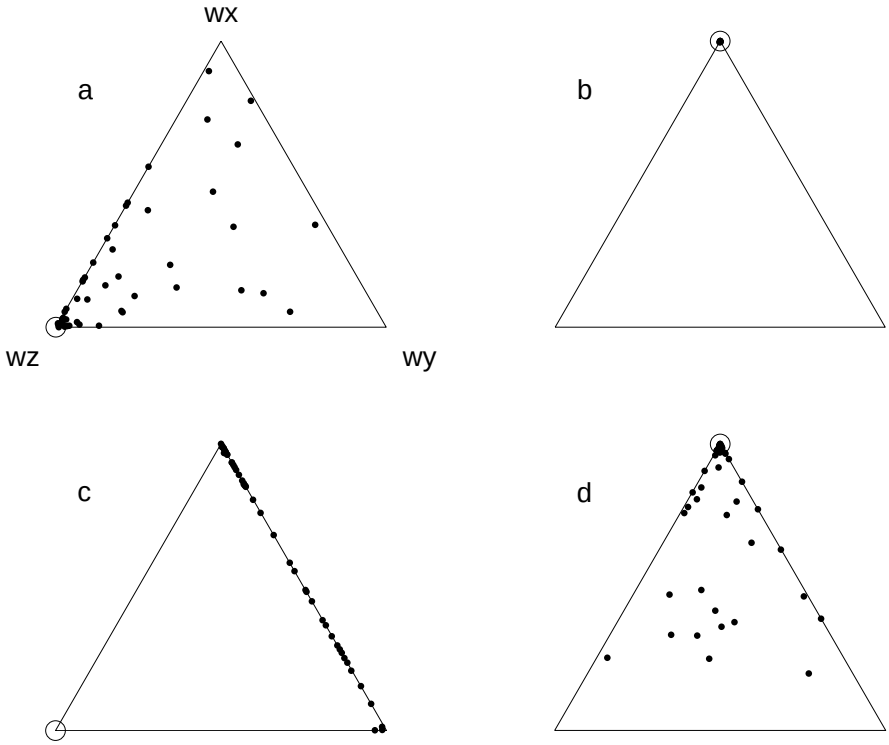

Figure 5: Logistic regression analyses of the data in Figure 4. Panels a and d are dataset a; b and e are dataset c; c and f are dataset d (we did not analyze dataset b). The upper three panels are relationships between $dS_{\min}$ (SHOT distance from the conditioning genome to the closest taxon of interest) and bootstrap support for the dominant topology. The lower three panels are relationships between size of the conditioning genome and bootstrap support for the dominant topology. Circles are bacterial conditioning genomes. The fitted lines are logistic regressions, as in Table 1 except that only one predictor variable is fitted at a time for clarity.

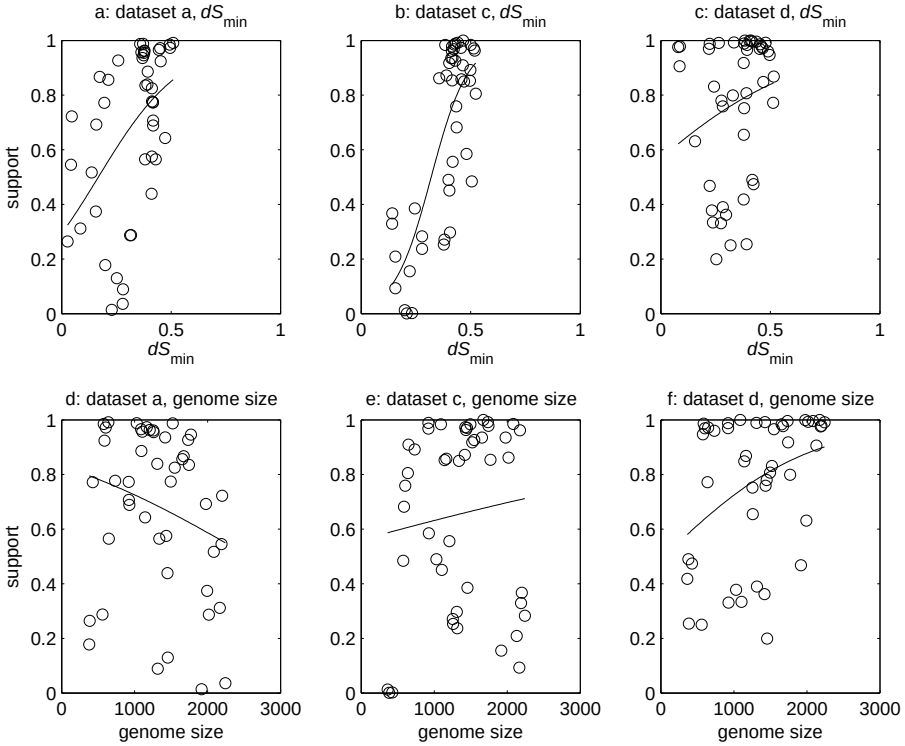

Figure 6: The modified BIONJ algorithm, adapted from Gascuel (1997a).

1. Input a set of m $m \times m$ distance matrices, each with a different conditioning genome k. Row and column k are missing from the kth distance matrix $\mathbf{D}^{(k)}$.
2. Initialize the global number of taxa $r \leftarrow m$ and the global list of taxa
3. Initialize the number of taxa in each distance matrix $r^{(k)} \leftarrow (r-1)$
4. Initialize each variance matrix $\mathbf{V}^{(k)} \leftarrow \mathbf{D}^{(k)}$
5. While the global number of taxa $r > 3$:
   (a) Compute the sums $S_i^{(k)}$, $S_j^{(k)}$ (equation 15) from each distance matrix $\mathbf{D}^{(k)}$ where $r^{(k)} > 3$
   (b) Choose a candidate pair of taxa $(i^{(k)}, j^{(k)})$ from each distance matrix $\mathbf{D}^{(k)}$ where $r^{(k)} > 3$, by minimizing equation 16
   (c) Determine the true pair out of $(i^{(k)}, j^{(k)}, k)$ using equations 17-19
   (d) Choose a pair (i, j) to aggregate in all distance matrices, by either the vote-counting method (equation 21) or the inverse-variance method (equation 23)
   (e) Aggregate the subtrees containing this pair in the global list of taxa
   (f) For all distance matrices $\mathbf{D}^{(k)}$ where $k \neq i, j$ and $r^{(k)} > 3$:
      i. Add a new internal node to which the subtrees containing i and j are connected, and compute new edge lengths (equation 11 with variables for the kth distance matrix)
      ii. Compute $\lambda^{(k)}$ (equation 24)
      iii. Update distances and variances using equations 13 and 14, and variables for the kth distance matrix
      iv. Reduce $r^{(k)}$ by 1
   (g) Reduce r by 1
6. Output the tree from the global list of taxa

Figure 7: Simulations for evaluating conditioned logdet distance methods. All the internal edges (black) had length 0.1 and rate parameters $q_{01}^{(1)} = 0.2$, $q_{10}^{(1)} = 0.8$. The edges t_w and t_z (gray) were varied together from 0.1 to 1. The rate parameters on these edges were $q_{01}^{(2)}$ from 0.1 to 1, $q_{10}^{(2)} = 0.8$.

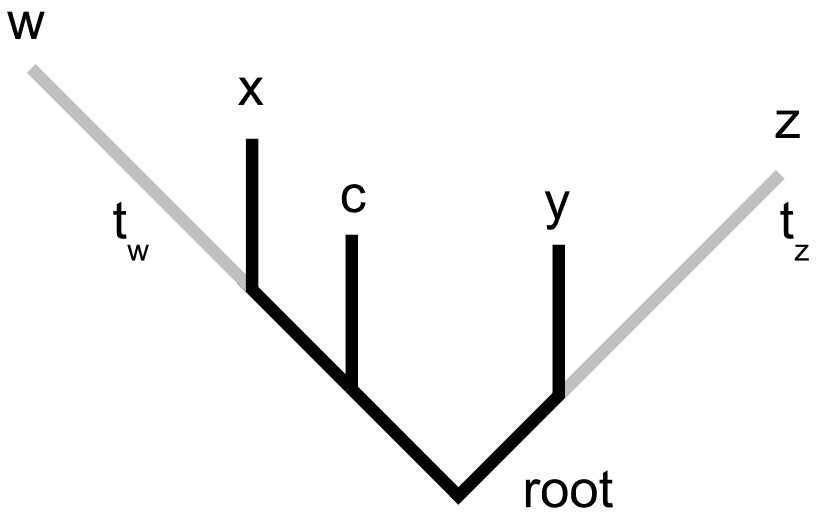

Figure 8: Results of simulations for evaluating conditioned logdet distance methods. In each panel, the horizontal axis is the lengths of the edges leading to taxa w and z, and the vertical axis is the rate of transition from absent to present on these edges. Methods were: (a) BIONJ on SHOT distances; (b) BIONJ on conditioned logdet distances from each conditioning genome separately (scored as correct if the correct subtree was recovered from all conditioning genomes); (c) summing the sum of squares over subtrees; (d) modified BIONJ, using vote-counting; (e) modified BIONJ, using inverse-variance weighting. The simulation setup is explained in Figure 7. Lighter values indicate recovery of the correct topology from a higher proportion of replicates.

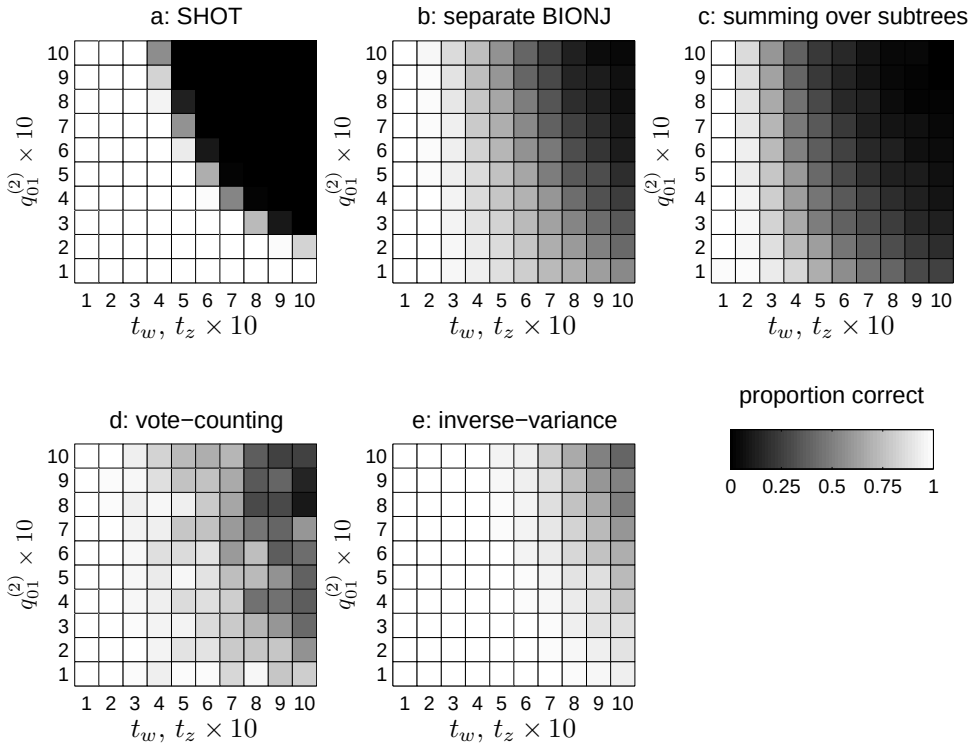

Figure 9: Unrooted bootstrap consensus phylogeny for 40 bacterial genomes from the COG database, estimated using conditioned logdet distances, modified BIONJ with inverse-variance weighting, and 1000 bootstrap replicates. Edge labels are the percentage of bootstrap replicates supporting the edge. Edges are not drawn to scale.

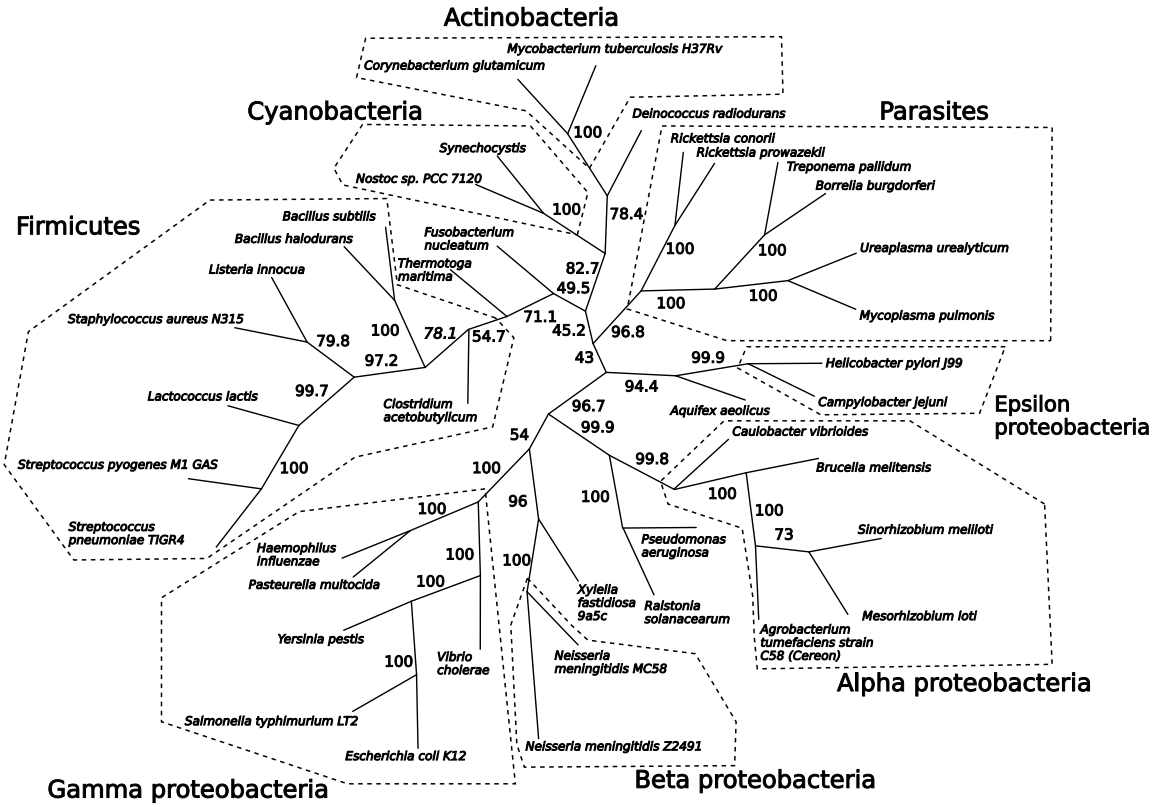

Figure 10: The three possible arrangements of two genomes w and x between which we want to calculate the distance, together with a conditioning genome c, the root r and an internal node v.