A Model of Horizontal Gene Transfer and the Bacterial Phylogeny Problem

NICOLAS GALTIER

Institut des Sciences de l’Evolution (UM2-CNRS), Université Montpellier 2, Place E. Bataillon, CC64, 34095 Montpellier, France; E-mail: galtier@univ-montp2.fr

Abstract.—How much horizontal gene transfer (HGT) between species influences bacterial phylogenomics is a controversial issue. This debate, however, lacks any quantitative assessment of the impact of HGT on phylogenies and of the ability of tree-building methods to cope with such events. I introduce a Markov model of genome evolution with HGT, accounting for the constraints on time—an HGT event can only occur between concomitantly living species. This model is used to simulate multigene sequence data sets with or without HGT. The consequences of HGT on phylogenomic inference are analyzed and compared to other well-known phylogenetic artefacts. It is found that supertree methods are quite robust to HGT, keeping high levels of performance even when gene trees are largely incongruent with each other. Gene tree incongruence per se is not indicative of HGT. HGT, however, removes the (otherwise observed) positive relationship between sequence length and gene tree congruence to the estimated species tree. Surprisingly, when applied to a bacterial and a eukaryotic multigene data set, this criterion rejects the HGT hypothesis for the former, but not the latter data set. [Bacteria; eucarya; horizontal gene transfer; Markov models; phylogenomics; tree congruence.]

The occurrence of horizontal gene transfers (HGTs) between species is an important challenge to molecular phylogeny: if HGT events were frequent enough, the notion of a unique organismal phylogeny would be essentially meaningless. Current literature suggests that HGTs are much more frequent in prokaryotes than in eukaryotes, so that the debate about their importance in phylogenomics has naturally focused on the former domain. Responding to arguments about the irrelevance of bacterial phylogeny (Doolittle, 1999; Gogarten et al., 2002; Baptiste et al., 2005), several authors have claimed that a “core” of genes more or less immune to HGT could serve to reconstruct the bacterial species tree (Brochier et al., 2002; Daubin et al., 2003; Kurland et al., 2003), a statement that was illustrated by providing a well-resolved, consistent phylogeny of Gamma proteobacteria (Lerat et al., 2003). Even this restricted data set, however, could be polluted by HGTs (Baptiste et al., 2004; Susko et al., 2006), despite the relatively recent origin of this clade, as compared to the divergence between bacterial phyla.

One reason for suspecting that HGT might obscure the bacterial phylogeny is the fact that the bacterial tree is still largely unresolved despite the large amount of data available. It is striking to notice that the global picture we have of this group is not very different from the one provided by Woese 30 years ago from oligonucleotide catalogues (i.e., small fragments of ribosomal RNA, Woese and Fox, 1977). The molecular revolution has added much to our apprehension of microbial diversity, e.g., by allowing the characterization of uncultivable species, but neither full ribosomal RNA sequences (Woese, 1987), nor protein data sets (e.g., Lloyd and Sharp, 1993; Galtier and Gouy, 1994; Bustard and Gupta, 1997), nor even full genome comparisons (Wolf et al., 2001; Brown et al., 2002; Daubin et al., 2002) could resolve the branching orders between the major bacterial phyla. That such an increase in the size of the data set has not improved the resolution power suggests that the divergence between bacterial phyla could be a hard-polytomy, in which distinct genes have had distinct phylogenetic histories.

Alternatively, the relative lack of success of bacterial phylogenomics might be explained by the very old divergences considered and the resulting reduction in signal/noise ratio due to saturation; i.e., the accumulation of numerous substitutions at the same site. It might be the case that a core of genes sharing a common species tree actually exists, but that the available data and methods do not allow us to recover it. Departure from the molecular clock, i.e., the existence of rapidly evolving and slowly evolving lineages, is an additional factor potentially generating inconsistent reconstructed gene trees through the well known long-branch attraction effect. Finally, the bacterial phylogenetic problem could be intrinsically difficult if the major phyla had evolved through a rapid radiation, leaving little time for phylum-specific synapomorphies to appear.

Distinguishing between these various hypotheses would be an important advance. If we knew that a core of bacterial genes not (or marginally) influenced by HGT actually existed, then our first goal should be to identify these genes and analyze them carefully. If, however, HGT was a prominent evolutionary force having affected virtually every bacterial gene, then we should accept the idea of considering several, perhaps many, bacterial phylogenies, and try and explore the evolutionary pathways that have lead to present-day bacterial genomes through genetic exchanges, perhaps thanks to the use of phylogenetic networks (e.g., Huson and Bryant 2006)—the two approaches are not incompatible.

This largely discussed topic raises two practical questions that I plan to address in this study: (i) How much HGT is required to preclude any attempt of building a species tree? (ii) Is there a way to determine, for a given data set, whether HGT or phylogenetic artefacts are the cause of the lack of resolution? Exploring these issues obviously requires us to characterize, both quantitatively and qualitatively, the influence of HGT on phylogenetic reconstructions. Remarkably, such an assessment is currently lacking, despite the huge amount of literature on HGT. I introduce for this purpose a stochastic model of...
Materials and Methods

Simulation Procedure

A model of multigene sequence evolution incorporating HGT and non-clock-like evolution is presented. Parameters include: clock-like, rooted species tree \( T \) (\( n \) leaves) with relative branch lengths \( \lambda_i (1 < i < 2n - 2) \), average number of genomic rate change events \( \rho \), number of gene trees \( m \), average number of HGT events per gene tree \( \tau \), gene tree diameters \( \delta_k (1 < k \leq m) \), average number of gene-specific rate change events \( \rho' \), gamma distribution of rates across lineages (shape \( \alpha_L \)), gamma distribution of rate across sites (shape \( \alpha_S \)), substitution matrix \( M \), gene length \( l_k (1 < k < m) \), and sampling effort \( p_{sk} (1 < k \leq m, 0 < p_{sk} < 1) \). The simplest way to introduce the details of the model is probably to describe the various steps of the simulation procedure, given a set of parameters (Fig. 1).

![Simulation Procedure Diagram](image-url)

**FIGURE 1.** Simulation procedure. Starting from species tree \( T \) (top), \( m = 3 \) gene trees are generated by introducing genomic rate change events (step 1), horizontal gene transfer (HGT) events (step 2, moving lineages bold-faced), scaling branch lengths (step 3), introducing gene-specific rate change events (step 4), and taxon sampling (step 5). The next step is the simulation of sequence evolution along gene trees (not shown). Details of step 2 are given in Figure 2. Details of steps 1 and 4 are given in Figure 3. Parameters relevant to each step are given.
Species tree $T$ is taken as input. This tree is rooted and clock-like, which means that all root-to-leaf pathways have equal lengths. Genomic departure from the molecular clock is simulated by inserting a Poisson-distributed number (mean $\rho$) of events of substitution rate change; i.e., acceleration or slow-down. The variance of rates across lineages is controlled by parameter $\alpha_i$. Details about how relative rates are assigned to branches are given below. These lineage-specific rates represent genomic averages; they will influence the evolution of all genes. $m$ Gene trees are then generated from $T$ by incorporating HGTs, each gene tree undergoing a Poisson-distributed number of HGT events of mean $r$. Details about how HGTs are simulated are given below. For every gene tree, all branch lengths are multiplied by a scaling factor so that the diameter of the tree (sum of branch lengths of the longest leaf-to-leaf pathway) is equal to $4(1 < k < m)$.

Parameters $\delta_i$ therefore control for the average level of sequence divergence in a gene-specific way. Then a second round of substitution rate change events is simulated (mean number of events $\rho'$) separately for every gene, adding some gene-specific, lineage-specific variation of substitution rate. The next step of the simulation process is a random pruning of taxa: for each gene tree, a binomial $B(n, p^i)$ number of species is retained, the other ones being removed, thus introducing missing data. Finally, for each gene tree a sequence data set of length $\gamma_k$ is simulated in a standard way, using substitution matrix $M$, assuming independent sites and stationary evolution. An as-shaped Gamma distribution of relative rates across sites is assumed.

**Modeling HGT**

Tree $T$ is intended to represent species evolution, so that its branch lengths are proportional to time and the depths of its nodes to speciation dates. During the evolution of a gene along $T$, HGT events are assumed to follow a Poisson process of rate $r/L$, where $L$ is the sum of branch lengths in $T$. This is equivalent to saying that a Poisson-distributed number of rate-change events with mean $r$ are randomly placed on $T$. Simulating an HGT event means cutting a subtree and pasting it at some other place in the tree (Fig. 1, step 1; Fig. 2). What moves is the subtree underlying the location of the HGT event (recipient lineage). The place where it moves (donor lineage) is randomly drawn in the portion of $T$ older than the time of the HGT event. This reflects the fact that the donor species must have been living at the time of the HGT event (so that the target lineage cannot be younger than the moving one) but might be extinct now, or not sampled (so that the branching point traces back to its common ancestor with any sampled species; Fig. 2), as noted by Maddison (1997). Such a topological move is called subtree pruning regrafting in the tree-searching phylogenetic literature. Setting $r = 0$ means no HGT, which involves all gene trees being identical to each other and to the species tree.

**Modeling Substitution Rate Change**

Departure from the molecular clock is achieved by running a discrete model of substitution rate evolution. Rate-changing events are assumed to follow a Poisson process of rate $\rho/L$ (or $\rho'/L$), where $L$ is the sum of branch lengths. This is equivalent to saying that a Poisson-distributed number of rate-change events with mean $\rho$ (or $\rho'$) are randomly placed on the tree. The ancestral relative substitution rate is assumed to be equal to one. When a rate-change event occurs, the current relative substitution rate is replaced by a new relative substitution rate drawn from a gamma distribution of mean 1 and shape parameter $\alpha_i$. The rate-change process therefore divides the tree in parts, and assigns a relative rate to every part (Fig. 3). This process recalls Galtier’s (2001) site-specific rate change model, although in the present study relative
FIGURE 3. Modeling departure from the molecular clock. (Top) A gene tree undergoes a Poisson distributed number of events of substitution rate changes, represented by dots. After each rate change event the current relative rate is replaced by a new relative rate drawn from a Gamma distribution. In this example, three events of rate change occurred, cutting the tree in four parts. In the topmost part of the tree evolution proceeds at ancestral relative rate \( r_0 \), whereas each of the other three parts has its specific rate \( r_x \). The length of each segment of the tree is multiplied by its relative rate before simulating sequence evolution. Assuming \( r_0 = 1 \), \( r_x = 0.5 \), \( r_2 = 2.5 \), \( r_3 = 2 \), for instance, would yield the non-clock-like tree displayed at the bottom. This is applied twice during the simulation process: at step 1, when shared by all genes, and at step 4, in a gene-specific way.

Rates are shared by all sites, and the clock-relaxed model of Huelsenbeck et al. (2000), although in the latter study the successive relative rates were not independent.

The insertion of rate change events occurs twice during the simulation process; i.e., at step 1 (genomic rate change events, rate \( \rho \)) and at step 4 (gene-specific rate change events, rate \( \rho' \); Fig. 1). The net substitution rate of a given lineage for a given gene is the product of the genomic and gene-specific rates. The model therefore accounts for global genomics trends, but allows each gene to follow its own history. Sequence evolution is then simulated according to these relative rates, which is typically achieved by first multiplying branch lengths by their relative rate, and then running a constant-rate Markov process of sequence evolution. Clock-like evolution can be assumed by setting \( \rho = \rho' = 0 \) (no event of rate change) or \( \alpha_L = \infty \) (no variance between rates).

A C++ program implementing the whole simulation process was developed using the BIO++ library (Dutheil et al., 2006). It is available from http://162.38.181.25/HGT.

**Phylogenetic Reconstruction**

Simulated multigene data sets were analyzed as follows. First, a phylogenetic tree was built for every gene using the maximum-likelihood PHYML program (Guindon and Gascuel, 2003) and the JTT model of sequence evolution (Jones et al., 1992). This model is different from the model used to simulate sequences in that a constant distribution of rates across sites is assumed. This was done to save computing time. The species tree was estimated using the supertree MRP algorithm (Ragan, 1992), which minimizes the conflict between internal branches from distinct trees. Branch lengths were assigned to the estimated species tree by maximum likelihood fitting to the concatenated data set. This was achieved using program PHYML again, which accounts for missing data.

**Tree-Based Statistics**

Simulated data sets were characterized by various statistics relative to tree congruence, shape, and size.

1. Reliability \( R \) was defined as the percentage of correct internal branches in the estimated species tree; i.e., internal branches shared by the species tree. \( R = 100\% \) means that the true species tree was successfully reconstructed.

2. Congruence \( C \) was defined as the percentage of internal branches in reconstructed gene trees involving no conflict with the estimated species tree. \( C = 100\% \) means that all reconstructed gene trees are perfectly congruent with each other, irrespective of the species tree.

3. Average observed diameter \( D \) was calculated by taking the average of the longest leaf-to-leaf pathway (sum of branch lengths) in reconstructed gene trees.

4. Average branch length asymmetry \( A \) was defined as the average ratio, over reconstructed gene trees, of longest root-to-leaf pathway to shortest root-to-leaf pathway. \( A = 1 \) means that all reconstructed gene trees appear clock-like.

5. Average node depth \( N \), finally, was defined as the average of internal node relative depths in the estimated species tree. The relative depth of node \( nd \) is calculated as the \( d_{down}/(d_{down} + d_{up}) \) ratio, where \( d_{down} \) is the average distance (sum of branch lengths) from \( nd \) to its underlying leaves, and \( d_{up} \) the distance from root to
to \( nd \). The relative depth is 0 for leaves, 1 for root; average depth \( N \) measures whether a given tree is star-like; i.e., has short internal branches.

Note that calculating statistics \( A \) and \( N \) requires knowledge of the precise location of the root, which is typically not available, especially when real data are analysed. I arbitrarily rooted each reconstructed gene tree at the middle of the longest leaf-to-leaf pathway, hence providing a conservative measure of branch length asymmetry \( A \).

**Real Data Sets**

Three existing multigene data sets were analyzed. The MITO data set consisted of nine mammalian mitochondrial proteins. I first randomly drew 50 species out of the >150 mammals for which the full mitochondrial genome is available (Jameson et al., 2003). This was done to make the data set closer to the simulated ones and decrease the running time. The nine genes longer than 200 amino acids were selected. Missing data were then introduced by removing a random proportion (0% to 50%) of taxa from each alignment. Mammalian mitochondrial sequences are known to be largely misleading from a phylogenetic point of view (Springer et al., 2001) but no HGT events are suspected for this genome, although arguments for mitochondrial HGT have been made in other groups (Bergthorsson et al., 2003; Alvarez et al., 2006). The EUCARYA data set consisted of 59 proteins sampled in various animal and fungi taxa. Starting from the 146 proteins analyzed in Philippe et al. (2005), I selected the alignments for which a subset of at least 20 taxa sharing at least 200 complete sites (no gap, no X) could be sought. The total number of taxa was 49, varying from 20 to 43 between proteins. Again, no HGT events are suspected for this data set. The BACTERIA data set, finally, was built in a similar way from data in Daubin et al. (2002), again requiring 20 taxa and 200 sites without gap. It was made of 36 proteins encompassing 30 bacterial species, between which the occurrence of HGT events is suspected.

**RESULTS**

**Simulations**

Multigene sequence data sets were generated following the above-described model under 13 distinct conditions, each condition being replicated 50 times. Each data set was made of \( m = 20 \) genes covering a total of \( n = 40 \) species. Species tree were randomly generated using PHYLY-O-GEN (http://evolve.zoo.ox.ac.uk/software/PhyloGen/main.html). Rate of HGT \( \tau \), genomic and gene-specific rates of rate-change events \( \rho \) and \( \rho' \), and gene tree diameters \( \delta_t \) varied between conditions (Table 1, and see below). The shape parameter of the gamma distribution of rates across lineages was set to

<table>
<thead>
<tr>
<th>Condition</th>
<th>( \delta_t )</th>
<th>( \rho )</th>
<th>( \rho' )</th>
<th>( \tau )</th>
<th>Reliability</th>
<th>Congruence</th>
<th>Diameter</th>
<th>Asymmetry</th>
<th>Node depth</th>
</tr>
</thead>
<tbody>
<tr>
<td>c_{000}</td>
<td>1-1 0 0 0 0</td>
<td>97.3</td>
<td>90.5</td>
<td>0.86</td>
<td>1.52</td>
<td>0.29</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[91.9, 1.0]</td>
<td>[84.4, 95.1]</td>
<td>[0.78, 0.92]</td>
<td>[1.41, 1.65]</td>
<td>[0.19, 0.39]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c_{001}</td>
<td>1-2 0 0 0 0</td>
<td>96.5</td>
<td>90.2</td>
<td>1.14</td>
<td>1.53</td>
<td>0.30</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[91.9, 1.0]</td>
<td>[84.3, 95.0]</td>
<td>[1.01, 1.25]</td>
<td>[1.39, 1.67]</td>
<td>[0.20, 0.40]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c_{002}</td>
<td>2-3 0 0 0 0</td>
<td>96.6</td>
<td>90.0</td>
<td>1.59</td>
<td>1.54</td>
<td>0.32</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[89.2, 1.0]</td>
<td>[84.2, 95.2]</td>
<td>[1.46, 1.73]</td>
<td>[1.38, 1.73]</td>
<td>[0.21, 0.41]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c_{010}</td>
<td>1-1 0 0 0 0</td>
<td>96.0</td>
<td>88.2</td>
<td>0.91</td>
<td>2.20</td>
<td>0.27</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[86.5, 1.0]</td>
<td>[73.8, 95.4]</td>
<td>[0.66, 1.33]</td>
<td>[1.45, 4.59]</td>
<td>[0.17, 0.40]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c_{020}</td>
<td>1-1 16 0 0 0</td>
<td>95.3</td>
<td>87.4</td>
<td>1.05</td>
<td>2.73</td>
<td>0.02</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[83.8, 1.0]</td>
<td>[67.9, 94.7]</td>
<td>[0.73, 1.68]</td>
<td>[1.58, 8.70]</td>
<td>[0.16, 0.37]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c_{090}</td>
<td>1-1 0 0 0 0</td>
<td>97.1</td>
<td>88.4</td>
<td>0.94</td>
<td>2.23</td>
<td>0.28</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[91.9, 1.0]</td>
<td>[81.0, 93.6]</td>
<td>[0.81, 1.03]</td>
<td>[1.79, 3.00]</td>
<td>[0.19, 0.38]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c_{200}</td>
<td>1-1 0 0 0 0</td>
<td>96.5</td>
<td>86.7</td>
<td>2.67</td>
<td>2.71</td>
<td>0.28</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[91.9, 1.0]</td>
<td>[78.8, 92.7]</td>
<td>[0.87, 1.13]</td>
<td>[2.07, 4.30]</td>
<td>[0.18, 0.38]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c_{222}</td>
<td>1-1 0 0 1 0</td>
<td>93.5</td>
<td>84.9</td>
<td>1.94</td>
<td>2.90</td>
<td>0.28</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[81.1, 1.0]</td>
<td>[74.9, 91.2]</td>
<td>[1.44, 2.52]</td>
<td>[2.15, 4.76]</td>
<td>[0.17, 0.40]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c_{000}</td>
<td>1-1 0 0 0 3</td>
<td>96.0</td>
<td>82.6</td>
<td>0.83</td>
<td>1.52</td>
<td>0.30</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[89.2, 1.0]</td>
<td>[76.4, 88.5]</td>
<td>[0.76, 0.89]</td>
<td>[1.40, 1.67]</td>
<td>[0.21, 0.40]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c_{000}</td>
<td>1-1 0 0 0 0</td>
<td>96.2</td>
<td>71.6</td>
<td>0.86</td>
<td>1.53</td>
<td>0.32</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[89.2, 1.0]</td>
<td>[64.4, 79.3]</td>
<td>[0.82, 0.92]</td>
<td>[1.40, 1.65]</td>
<td>[0.23, 0.41]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c_{000}</td>
<td>1-1 0 0 0 6</td>
<td>93.9</td>
<td>59.8</td>
<td>1.86</td>
<td>1.53</td>
<td>0.35</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[86.5, 1.0]</td>
<td>[53.4, 65.8]</td>
<td>[0.79, 0.91]</td>
<td>[1.42, 1.67]</td>
<td>[0.26, 0.44]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c_{010}</td>
<td>1-1 0 0 12 8</td>
<td>86.8</td>
<td>45.8</td>
<td>0.87</td>
<td>1.54</td>
<td>0.40</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[75.7, 97.3]</td>
<td>[37.3, 54.7]</td>
<td>[0.80, 0.92]</td>
<td>[1.44, 1.68]</td>
<td>[0.30, 0.50]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c_{222}</td>
<td>1-1 16 0 12 12</td>
<td>80.9</td>
<td>37.9</td>
<td>1.87</td>
<td>2.95</td>
<td>0.40</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[62.2, 91.9]</td>
<td>[27.6, 54.2]</td>
<td>[1.38, 2.66]</td>
<td>[2.25, 4.04]</td>
<td>[0.29, 0.50]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The reference condition, called \( c_{0000} \), was one without HGT (\( \tau = 0 \)), without departure from the molecular clock (\( \rho = \rho' = 0 \)), and with a moderate amount of divergence between taxa (\( \beta_k = 1 \) for every gene tree). This condition was chosen to make the phylogenetic inference problem an easy one. Then the problem was made harder in four distinct ways, independently from each other. Under conditions \( c_{0001} \) and \( c_{0002} \), a higher diameter is assumed for gene trees. Under conditions \( c_{0010} \) and \( c_{0020} \), moderate to strong genomic departure from the molecular clock hypothesis is introduced. Under conditions \( c_{0100} \) and \( c_{0020} \), moderate to strong gene-specific departure from the molecular clock hypothesis is assumed. Condition \( c_{0222} \) combines the three effects: high diameter, genomic and gene-specific rate variation across lineages. Conditions \( c_{1000}, c_{2000}, c_{0000}, \) and \( c_{0000} \), finally, assume a low to strong amount of HGT, and condition \( c_{4222} \) combines all four effects.

Under reference condition \( c_{0000} \), as expected, the congruence between gene trees was high (average \( C: 90.5\% \)), and estimated species trees were very close to true trees (average \( R: 97.3\% \), Table 1). Increasing the diameter of gene trees had little effect on the results, despite the fact that actual diameters were largely underestimated (Table 1). When departure from the molecular clock was introduced, the observed tree asymmetry was substantially increased. Congruence and reliability were slightly decreased, as compared to the reference condition. Gene-specific events of substitution rate change affected congruence between gene trees more than reliability: although individual gene trees are reconstructed less accurately, the average phylogenetic signal is unchanged. Genomic events of rate changes, in contrast, increased the error rate slightly: when genomes as a whole evolve at different rates, many genes can agree in supporting wrong branching orders. Combining a high diameter and substantial genomic and gene-specific departure from the clock (\( c_{0222} \)) yielded a reduced congruence and reliability. The error rate \( 1 - R \), however, was still quite low: only 6.5\% of the reconstructed internal branches were wrong, on average, despite the seemingly difficult conditions. It should be noted that the variance between replicates was higher when whole-genome rate variation was introduced: some simulated multigene data sets resulted in a reliability as low as 75\%. This is consistent with the empirical observation that rapidly evolving genomes are one of the most difficult problems in phylogenomics (Philippe et al. 2005).

When HGT was introduced in the simulation process, the congruence between gene trees dropped dramatically. A single average event of HGT per gene tree (\( c_{1000} \)) affected congruence more seriously than the combination of difficulties modelled in \( c_{0222} \). Remarkably, this substantial drop of congruence between gene trees did not result in a strong decrease of reliability. When an average of six HGT events per gene tree were assumed, so that 40\% of the internal branches of gene trees were in conflict with the estimated species tree, the performance of the supertree method was still good (\( R = 93.9\% \)), and comparable to condition \( c_{0222} \), for which congruence was as high as 85\%. HGT also modified the tree shape. The average relative depth of internal nodes in the estimated species tree was significantly increased when substantial amounts of HGT were simulated, whereas this statistic was essentially unaffected by variations of the tree diameter or of lineage-specific substitution rate. This suggests that supertree methods tend to produce star-like trees when the gene trees to be reconciled are incongruent. Under condition \( c_{4222} \), finally, the effects of rate changes and HGT are combined and lead to a further decrease in congruence and reliability.

**Real Data**

The same statistics were calculated for three real data sets, of course excepting reliability, because the true species tree is unknown (Table 2). The congruence between gene trees was much lower for all three real data sets than for data sets simulated without HGT. This was true even for data sets supposedly unaffected by HGT. Many assumptions of the substitution model used for building trees (e.g., homogeneity, stationarity, independent sites, constant rate in time for each site) are not met by real data sets, resulting in reconstruction errors. The signal/noise ratio is much higher in simulated data, although I constrained sequence length to be higher than 200 for real data sets, but only 100 in simulations. A consequence is that absolute values of incongruence can probably not be compared between real and simulated data sets. Observing an average congruence to estimated species tree below 70\% is not sufficient to conclude that HGT has occurred, despite the fact that data sets simulated without HGT in this study never reached such a low congruence value.

The relatively small MITO data set illustrates this point. Here, phylogenetic incongruence is probably

---

**Table 2.** Real data analysis (maximal and minimal values are given within brackets).

<table>
<thead>
<tr>
<th>Data set</th>
<th>nb Genes</th>
<th>nb Species</th>
<th>Congruence</th>
<th>Diameter</th>
<th>Asymmetry</th>
<th>Node depth</th>
</tr>
</thead>
<tbody>
<tr>
<td>MITO</td>
<td>9</td>
<td>50</td>
<td>55.5</td>
<td>0.67</td>
<td>3.14</td>
<td>0.47</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>[34.8, 77.3]</td>
<td>[0.22, 1.17]</td>
<td>[1.93, 6.66]</td>
<td></td>
</tr>
<tr>
<td>EUCARYA</td>
<td>59</td>
<td>49</td>
<td>67.7</td>
<td>0.97</td>
<td>2.76</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>[47.4, 89.5]</td>
<td>[0.30, 2.17]</td>
<td>[1.31, 6.16]</td>
<td></td>
</tr>
<tr>
<td>BACTERIA</td>
<td>36</td>
<td>30</td>
<td>61.4</td>
<td>1.55</td>
<td>2.42</td>
<td>0.47</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>[29.4, 90.0]</td>
<td>[0.83, 2.28]</td>
<td>[1.49, 3.91]</td>
<td></td>
</tr>
</tbody>
</table>

---

\( \alpha_L = 1 \). The JTT model of amino-acid sequence evolution was assumed, with a gamma distribution of rates across sites of shape parameter \( \alpha_S = 0.5 \). Sequence length randomly varied between 100 and 500, and sampling effort between 50\% and 100\%, across genes.
largely explained by extensive variations of substitution rates across lineages—the average observed asymmetry is 3.14; i.e., higher than in the most extreme simulation condition \(c_{222}\). Consequently, the estimated species tree (not shown) was totally inconsistent with the now well-resolved mammalian phylogeny (Springer et al., 2004), many mammalian orders or supraorders being split as polyphyletic groups. The BACTERIA versus EUCARYA comparison is more directly relevant to the issue of the importance of HGT in deep phylogenies. Congruence was higher in EUCARYA than in BACTERIA. This is consistent with the idea, but does not demonstrate, that the BACTERIA data set could be affected by HGT. I then compiled, for the three data sets, the values of statistics \(C\), \(D\), and \(A\) for every gene tree and analyzed their relationship. Gene tree congruence to the estimated species tree, \(C\), was not, or very weakly, correlated to gene tree diameter or asymmetry, although a negative relationship could have been expected. This was true for all three data sets. Congruence, however, was positively and significantly correlated to sequence length for the MITO \((r^2 = 0.657, P < 0.01; \text{Fig. 4a})\) and BACTERIA \((r^2 = 0.222, P < 0.01; \text{Fig. 4b})\) data sets, but not the EUCARYA data set \((r^2 = 0.040; \text{Fig. 4c})\). Such a relationship is not obviously expected since the MRP supertree method puts the same weight on every gene tree, irrespective of sequence length.

To assess further the significance of this result, I applied the same analysis to data sets simulated under the \(c_{0000}\) (basic, no HGT), \(c_{222}\) (difficult, no HGT), and \(c_{2000}\) (HGT) conditions. Here the sizes of the simulated data sets (number of species, number of genes, sampling effort) were equated to the real ones, and 100 replicates were performed. Without HGT, a positive relationship was frequently found between sequence length and gene tree congruence to estimated species, but this was rarely the case with HGT (Table 3). This result makes sense. When all genes share a common species tree, gene trees supported by long sequences tend to be closer to the true tree, because longer sequences carry more signal. When gene trees differ from each other, even long sequences can support a tree different from the species tree; what controls in the first place the congruence to the species tree is how many HGT events have occurred, and how much they have perturbed the phylogeny. Quite unexpectedly, therefore, the EUCARYA data set behaved like data sets simulated with HGT, and the BACTERIA one behaved like data sets simulated without HGT, as far as the congruence/sequence length relationship was concerned. BACTERIA-like simulations, in particular, led to a rejection of the \(c_{2000}\) HGT model for the BACTERIA data set (Table 3): only 1 of the 100 simulated data sets showed a correlation coefficient higher than the one measured from the real data set.

**DISCUSSION**

**A HGT Model**

In this study, a stochastic model of HGT in molecular phylogeny is introduced. This model makes a number of simplifying assumptions. HGT events are assumed to be equally probable between any pair of lineage, irrespective of phylogenetic and ecological proximity. The model does not account for HGT events whose donor species is an outgroup to the sampled taxa. HGT from Archaea to Bacteria (e.g., Calteau et al., 2005), for instance, can have affected the BACTERIA data set, but such events were not simulated. Simulations, finally, assume a constant rate of HGT for all genes, whereas in nature some genes are heavily transferred, and other genes little or not transferred (Lerat et al., 2005).

The model, however, accounts for the constraint on time: the donor and recipient species must be contemporaneous, which means that the target location of moving
lineages must be closer to the root than their initial location—the model therefore takes a rooted species tree as parameter. Meeting this constraint on time has some importance. It means that old HGT events will perturb the phylogeny less strongly than recent ones. When the ancestor of the moving subtree (Y in Fig. 2) is close to the root, the tree resulting from the move cannot be very different from the initial one. Recent HGT events, in contrast, can move lineages far away from their initial locations and generate strong phylogenetic conflicts. Suchard (2005) and Jin et al. (2006) used models in which HGT events occur through SPR moves, failing to account for the timing constraint. It is unclear whether, and how much, this approximation biases the Bayesian and maximum-likelihood inference procedures proposed by Suchard and Jin et al., respectively.

It should be noted that the uniform sampling of the destination of moving subtrees used in this study is one (natural but arbitrary) option out of many. A model incorporating the rates of speciation and extinction and the sampling effort could help providing a more rigorous definition of the subtree-moving process during HGT. To check the robustness of the results to the procedure used here, I reconduted the simulations using a distinct, somewhat extreme HGT model in which subtrees are only moving towards contemporaneous lineages—this assumes no extinction and exhaustive sampling. Qualitatively identical results were obtained (not shown).

**Lessons from Simulations**

Simulation were designed to address several general questions about HGT: How much HGT is required to remove the phylogenetic signal? How does HGT compare to other phylogenetic artefacts with this respect? Is there a way to determine whether a given multigene data set is affected by HGT? I now examine these various issues in the light of the simulation results.

The first question got a relatively clear answer: surprisingly, good levels of performance were kept even when a substantial amount of HGT was introduced. The congruence between gene trees, furthermore, is not a good indicator of the reliability of the estimated species tree: conditions $c_{0222}$ and $c_{1000}$ yielded similar levels of performance, although the average congruence to the estimated species tree was much lower in the latter. This seems to be good news for phylogenomic projects in Bacteria: there is apparently some hope to recover the bacterial species tree even from genes (moderately) affected by HGT. It should be noted, however, that the assumption of equiprobable donor and recipient species pairs used in this study is probably the ideal case. Here HGT occurs randomly, so that the heterogeneity between gene trees is increased, but the “average” gene tree is unchanged. The performances would probably be worse if trends existed; e.g., if HGT occurred more frequently between specific taxa. Such a directed HGT process would presumably bias species tree estimates, leading to the artefactual grouping of taxa exchanging genetic material (Zhaxybayeva et al., 2004). It is currently unclear whether HGT have occurred randomly, or more frequently between specific lineages, during bacterial evolution (Beiko et al., 2005).

Comparing the impact of HGT to that of other factors, such as tree diameter and rate variation across lineages, was another goal of this study. Simulations suggest that HGT results in a much stronger decrease of congruence between gene trees, as compared to other artefacts. Real data, however, demonstrated the irrelevance of absolute congruence measures: low levels of congruence between gene trees can be reached even in the (almost certain) absence of HGT. This discrepancy between real and simulated data presumably comes from the fact that real data, but not simulated data, depart from many assumptions of the models used to reconstruct trees. This is one of the reasons why no attempt to fit the model to data was made in this study. Because the “basic” level of incongruence between trees is obviously underestimated when standard models of sequence evolution are used, an inference model allowing HGT would presumably overestimate the rate of HGT, and detect significant HGT as soon as congruence between gene trees is lower than ~80%, although we know this can occur without HGT. This is a very general problem, applying to any attempt of quantifying the rate/amount of HGT from the analysis of incongruence between gene trees (Daubin and Ochman, 2004; Beiko et al., 2005; MacLeod et al., 2005). Ge et al. (2005) addressed this problem by combining two measures of congruence, namely the size of the maximum agreement subtree and the Robinson-Foulds distance, with the goal of detecting gene trees showing one (or a couple of) strong incongruity(ies) with the consensus, but otherwise consistent phylogenetic signal.
Our simulation study, finally, could serve the purpose of identifying features that might distinguish data generated with or without HGT. In addition to a decrease of congruence between gene trees, simulations suggest that HGT tends to make gene trees star-like: the average node depth increased when r increased. These trends, however, can hardly be used to decide whether a given data set has or has not undergone HGT. Again, real data sets can yield star-like estimated species trees and incongruent gene trees even in the absence of HGT. More interesting is the discovery that multigene data sets simulated without HGT frequently show a positive relationship between sequence length and gene tree congruence to the estimated species tree, whereas this is usually not the case with HGT, in which case even well-resolved gene trees can be incongruent with the species tree. Although not fully discriminative, this criterion can help distinguishing between the two alternative models, as I now discuss for the three real data sets analyzed in this study.

Incongruence Patterns in Bacteria and Eucarya

The EUCARYA data set showed a higher level of congruence between gene trees than the BACTERIA one. This appears consistent with our prior about the relative importance of HGT in the two domains. This conclusion, however, is challenged by the gene-by-gene pattern of congruence. Surprisingly, a significant correlation between sequence length and gene tree congruence to the estimated species tree was detected in Bacteria, but not in Eucarya. This suggests that the Daubin et al. (2002) BACTERIA data set is little affected by HGT. When an average of three HGT events per gene tree were introduced in the BACTERIA-like simulations (condition c2000), the relationship between sequence length and congruence essentially vanished, and only 1 simulated data set out of 100 showed a higher correlation than in the real BACTERIA data set (Table 3). Such a moderate amount of HGT apparently does not perturb much the phylogenetic inference: the reliability of the MRP method under condition c2000 was hardly distinguishable from the reference condition c0000 (Table 1). This supports again the conclusion that HGT is probably not the main problem we face in bacterial phylogenomics. The correlation between sequence length and gene tree congruence was lower than in most data sets simulated under condition c0000 or c022 (no HGT) but was typical for data sets simulated under condition c2000 (with HGT; see Table 3).

This result, of course, does not demonstrate that metazoan and fungi genomes have undergone HGT. In real data, sequence length is probably less strongly correlated to the amount of phylogenetic information carried by each gene than in simulated data. It should also be noted that the species tree reconstructed in this study is nonoptimal. When the tree recovered by the detailed analysis of Philippe et al. (2005; Fig. 4) was taken as the estimated species tree, the correlation between sequence length and gene tree accuracy was increased (r² = 0.11). The EU-
CARYA/BACTERIA discrepancy, however, asks for a deeper investigation of the gene trees/species tree relationship in the Philippe et al. (2005) data set. Gene and genome duplications are frequent in Eucarya. Like HGT, hidden paralogy can lead to incongruent gene trees, and presumably to a removal of the relationship between sequence length and congruence.

The correlation between sequence length and gene tree congruence to the estimated species tree is proposed as a new potential criterion for detecting HGT in phylogenomics. Its empirical relevance obviously requires further assessment. This study suggests that even a moderate amount of HGT essentially removes any influence of gene length on gene tree accuracy. So detecting a significant correlation should be taken as solid evidence that the considered data set is largely immune from HGT. The reciprocal statement is less clear, however. Not observing a correlation between sequence length and gene-tree congruence to the supertree can probably happen for many reasons, as discussed above; one should not claim for the occurrence of HGT based on this sole evidence.

CONCLUSION

Thanks to the use of a new stochastic model, this study suggests that supertree methods are robust enough to the occurrence of HGT, at least when HGT events occur randomly. A new criterion having some power to distinguish between HGT and other confounding factors in phylogenomics is proposed, namely the correlation between sequence length and gene tree congruence to the estimated species tree. Further assessment of the empirical relevance of this criterion is needed. Additional perspectives include extending the approach to supermatrix methods and taking into account the statistical support for internal branches.

ACKNOWLEDGEMENTS

The author thanks H. Philippe and V. Daubin for sharing data sets, and N. Lartillot, Y. Song, and M. Steel for their thoughtful comments. This work was supported by the Centre National de la Recherche Scientifique and Action Concertée Incitative "Informatique, Mathématique et Informatique..."
et Physique pour la Biologie” MODEL.PHYLO. This is publication ISEM 2007-042.

REFERENCES


First submitted: 11 October 2006; revisions returned 29 December 2006; final acceptance 1 May 2007

Associate Editor: Mike Steel