A Model of Horizontal Gene Transfer and the Bacterial Phylogeny Problem

Nicolas Galtier

Systematic Biology, Volume 56, Issue 4, August 2007, Pages 633–642
“Meaningfulness” of a tree of life, especially for prokaryotes

- There still exists a species tree-like structure even in the presence of lateral gene transfer.
- HGT would make a species tree so reticulated it doesn’t even make sense to think about a “tree of life.”

Phylogenetic Tree of Life

Phylogenetic tree based on Woese et al. rRNA analysis. The vertical line at bottom represents the last universal common ancestor (LUCA).

https://en.wikipedia.org/wiki/Carl_Woese
LGT/HGT Background

Def: movement of genetic material between two organisms

- **Transformation**
  - Uptake of exogenous material into a cell membrane from its surroundings
- **Conjugation**
  - Physical contact between two organisms results in a transfer of genetic material (bacterial sex)
- **Transduction**
  - Foreign genetic material is introduced into a cell by a virus

Main Research Questions

1. How much HGT is required to preclude any attempt at recovering a tree?
2. Given a dataset, is there a way to determine if HGT or other phylogenetic artifacts are causing the lack of resolution?
Approach

1. Develop a stochastic model of sequence evolution that incorporates HGT
2. Investigate the impact ability to reconstruct species tree by varying parameters such as:
   - tree length
   - deviation from the molecular clock
   - HGT rates on the ability to reconstruct a species tree
3. Investigate three real datasets
4. Look for consistent trends between simulated and real datasets to try to infer some features that might indicate the presence of HGT
Stochastic Model

**Input:** A clock-like species tree and set of parameters

**Output:** $m$ gene trees incorporating hgt and non-clocklike evolution

1. Incorporate substitution rate changes at the genome level
2. Generate $m$ gene trees by incorporating mean of tau HGT events per tree
3. Gene specific branch scaling
4. Round two of substitution rate changes at a gene and lineage specific level
5. Prune taxa from gene trees by sampling from binomial distribution
A closer look at random HGT

Subtree pruning regrafting

- Select a donor species (X) and a recipient species (Y). X and Y must both exist at the same point in time (contemporaneous).
- Prune Y and regraft at the mrca of X and it’s sibling (C).
- Not all HGT results in changes to topology.
Performance Measures

- **Reliability (R)** - percentage of correct internal branches in the estimated species tree (R=100% means perfect reconstruction)
- **Congruence (C)** - 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)
- **Average observed diameter (D)** - average of the longest leaf to leaf path of each reconstructed gene tree
- **Average branch length asymmetry (A)** - average ratio of longest root to leaf to shortest root to leaf path in each gene tree (A = 1 means gene trees are clock-like)
- **Average Node Depth (N)** - average of relative node depth of internal nodes where relative node depth is defined a d\_down / (d\_down + d \_up) where d\_down is the average distance from a node to its leaves and d\_up is the distance from the node to the root.

*Measures A & N require rooted gene trees, so the gene trees are arbitrarily rooted at the middle of the longest leaf-to-leaf pathway*
Phylogenetic Reconstruction

- **Gene trees** were estimated from sequence data using a maximum likelihood (ML) approach (PHYML) under the JTT model of sequence evolution.
- **Species tree** was estimated using MRP.
- **Branch lengths** were assigned to the species tree using ML on the concatenated dataset using PHYML.
Simulation Study

- 13 distinct conditions
- Replicated 50 times
- $m = 20$ genes
- $n = 40$ species

<table>
<thead>
<tr>
<th>Condition</th>
<th>$b_k$</th>
<th>$p$</th>
<th>$p'$</th>
<th>$r$</th>
<th>Reliability</th>
<th>Congruence</th>
<th>Diameter</th>
<th>Asymmetry</th>
<th>Node depth</th>
</tr>
</thead>
<tbody>
<tr>
<td>C0000</td>
<td>1-1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>97.3</td>
<td>90.5</td>
<td>0.86</td>
<td>1.52</td>
<td>0.29</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<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>
</tr>
<tr>
<td>C0001</td>
<td>1-2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>96.5</td>
<td>90.2</td>
<td>1.14</td>
<td>1.53</td>
<td>0.30</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<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>
</tr>
<tr>
<td>C0002</td>
<td>2-3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>96.6</td>
<td>90.0</td>
<td>1.59</td>
<td>1.54</td>
<td>0.32</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<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>
</tr>
<tr>
<td>C0010</td>
<td>1-1</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>96.0</td>
<td>88.2</td>
<td>0.91</td>
<td>2.20</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<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>
</tr>
<tr>
<td>C0020</td>
<td>1-1</td>
<td>16</td>
<td>0</td>
<td>0</td>
<td>95.3</td>
<td>87.4</td>
<td>1.05</td>
<td>2.73</td>
<td>0.36</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<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>
</tr>
<tr>
<td>C0100</td>
<td>1-1</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>97.1</td>
<td>88.4</td>
<td>0.94</td>
<td>2.23</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<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>
</tr>
<tr>
<td>C0200</td>
<td>1-1</td>
<td>0</td>
<td>16</td>
<td>0</td>
<td>96.5</td>
<td>86.7</td>
<td>1.00</td>
<td>2.71</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<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>
</tr>
<tr>
<td>C0220</td>
<td>2-3</td>
<td>16</td>
<td>16</td>
<td>0</td>
<td>93.5</td>
<td>84.9</td>
<td>1.94</td>
<td>2.90</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<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>
</tr>
<tr>
<td>C1000</td>
<td>1-1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>96.0</td>
<td>82.6</td>
<td>0.83</td>
<td>1.52</td>
<td>0.30</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<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>
</tr>
<tr>
<td>C2000</td>
<td>1-1</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>96.2</td>
<td>71.6</td>
<td>0.86</td>
<td>1.53</td>
<td>0.32</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<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>
</tr>
<tr>
<td>C5000</td>
<td>1-1</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>93.9</td>
<td>59.8</td>
<td>0.86</td>
<td>1.33</td>
<td>0.35</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<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>
</tr>
<tr>
<td>C6000</td>
<td>1-1</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td>86.8</td>
<td>45.8</td>
<td>0.87</td>
<td>1.54</td>
<td>0.40</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<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>
</tr>
<tr>
<td>C2222</td>
<td>2-3</td>
<td>16</td>
<td>16</td>
<td>12</td>
<td>80.9</td>
<td>37.9</td>
<td>1.87</td>
<td>2.95</td>
<td>0.40</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<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>
</tr>
</tbody>
</table>

Table 1. Properties of multigene data sets simulated under various conditions (95% confidence intervals are given within brackets). $b_k$ is the average diameter of true gene trees (maximal distance between any two leaves, expressed in per site average number of substitutions); $p$ and $p'$ are the average numbers of genome and gene-specific events of rate change, respectively; $r$ is the average number of horizontal gene transfer events; reliability measures the success in recovering the true species tree (percentage of correct internal branches in the estimates species tree); congruence measures the homogeneity between estimated gene trees (percentage of internal branches of the estimated gene trees shared by the estimated species tree); diameter is the average diameter of estimated gene trees (expressed in per site average number of substitutions); asymmetry is the average ratio of longest to shortest root-to-leaf distance in estimated gene trees; node depth measures the star-likeness of the estimated gene trees (see text).
Real Data

- HGT Events Not Suspected
  - MITO (but is known to be a misleading dataset)
  - EUCARYA
- HGT Events Suspected
  - BACTERIA

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>
Main Research Question 1

1. How much HGT is required to preclude any attempt at recovering a tree?
   ○ Phylogeny inference methods like MRP are fairly insensitive to moderate amounts of HGT (under a random model)
   ○ Even under the simulation conditions with high rates of HGT, the reliability of the species tree was still good
Main Research Question 2

2. Given a dataset, is there a way to determine if HGT or other phylogenetic artifacts are causing the lack of resolution?

- Congruence alone is not conclusive
- **Correlation between sequence length and gene tree congruence indicates presence of the HGT in simulations**
  - A significant correlation implies the dataset is immune from HGT but the reverse is not necessarily true (low correlation does not indicate the presence of HGT)
Model Limitations

- HGT events are equiprobable and do not account for ecological proximity or phylogenetic distance.
- Only accounts for HGT within a group (Bacterial genes cannot come from Archaea in this model).
- Constant rate of HGT for all genes but some genes may be more heavily transferred than others.
- Does not account for non-random HGT, like highways, where two edges in the phylogeny undergo an extremely high number of HGT events.
Conclusions

- Takes a key first step in investigating the possibility of deriving a useful species tree under HGT
- Provides some very rough measures to determine if HGT should be suspected in a dataset but use with caution!
- How do the simplifying assumptions about random HGT transfer impact the conclusions?
*Reanalyze the impact on species tree estimation with more plausible biological models of HGT*

- **Ecological Proximity & Phylogenetic Distance**

*EvolSimulator* - specializes in gene duplication & loss and HGT simulation

Questions?
**Bounded-rates model** The following assumption was introduced in [DR10] and is related to a common assumption in the mathematical phylogenetics literature.

**Definition 4 (Bounded-rates model)** Let $0 < \rho_\lambda < 1$ and $0 < \rho_\tau < 1$ be constants. Let further $0 < \bar{\sigma} < +\infty$ be a constant and $0 < \bar{\lambda} < +\infty$ be a value possibly depending on $n^+$. Under the Bounded-rates model, we consider the set of phylogenies $T_s = (V_s, E_s, L_s; r, \tau)$ with $n^+ > 0$ extant leaves and $n^- \geq 0$ extinct leaves and extant phylogeny $T_s^+ = (V_s^+, E_s^+, L_s^+; r, \tau^+)$ such that the following conditions are satisfied:

$$\lambda \equiv \rho_\lambda \bar{\lambda} \leq \lambda(e) \leq \bar{\lambda}, \quad \forall e \in E_s,$$

and

$$\tau \equiv \rho_\tau \bar{\tau} \leq \tau^+(e^+) \leq \bar{\tau}, \quad \forall e^+ \in E_s^+.$$
Key Findings from Simulations

- With greater departure from the molecular clock, asymmetry increases resulting in slight decreases in reliability and congruence
- Gene specific rate changes impact congruence more than reliability
- As HGT rate increases, congruences drops substantially but does not have as significant an impact on reliability
- Average node depth increases as HGT rate increases, implying that the estimated species trees have a more star-like shape
Further Analysis

- Relationship between sequence length and gene tree congruence
- When matching simulation conditions to the real datasets in terms of genes and species, sequence length had higher correlation with gene tree congruence
- Oddly, EUCARYA behaves more like a set with HGT than the BACTERIA dataset

<table>
<thead>
<tr>
<th>Data set</th>
<th>nb Genes</th>
<th>nb Species</th>
<th>$r^2$</th>
<th>Sim. c_{9000}</th>
<th>Sim. c_{0222}</th>
<th>Sim. c_{0000}</th>
</tr>
</thead>
<tbody>
<tr>
<td>MITO</td>
<td>9</td>
<td>50</td>
<td>0.657**</td>
<td>0.330 91%</td>
<td>0.239 95%</td>
<td>0.080 99%</td>
</tr>
<tr>
<td>BACTERIA</td>
<td>36</td>
<td>30</td>
<td>0.222**</td>
<td>0.173 57%</td>
<td>0.130 85%</td>
<td>0.028 99%</td>
</tr>
<tr>
<td>EUCARYA</td>
<td>59</td>
<td>49</td>
<td>0.040NS</td>
<td>0.191 4%</td>
<td>0.142 14%</td>
<td>0.025 64%</td>
</tr>
</tbody>
</table>

*Squared correlation coefficient between sequence length and gene tree congruence to the estimated species tree for real data sets; **, significant at the 1% level; NS, not significant.
*Median squared correlation coefficient in 100 simulated data sets.
*Proportion of simulated data sets for which the correlation coefficient was lower than the observed one.