Systematic Detection of Highways of Horizontal Gene Transfer

Mukul S. Bansal[1] Guy Banay[1] J. Peter Gogarten[2]
Ron Shamir[1]

[1]The Blavatnik School of Computer Science
Tel Aviv University, Israel
[2]Department of Molecular and Cell Biology
University of Connecticut, USA

What is Horizontal Gene Transfer?

Genetic material is typically transferred vertically from ancestor to descendant.

However, genetic material can also get transferred horizontally across species. This is called Horizontal Gene Transfer (HGT).

- Rare in eukaryotes, but rampant in prokaryotes.
- Bacterial conjugation (cell to cell contact), phages.
- 10% to 20% of genes in many bacteria.
- Important role in genome diversification and evolution. Responsible for drug resistance in bacteria.

What is Horizontal Gene Transfer?

Genetic material is typically transferred vertically from ancestor to descendant.
However, genetic material can also get transferred horizontally across species. This is called Horizontal Gene Transfer (HGT).

- Rare in eukaryotes, but rampant in prokaryotes.
- Bacterial conjugation (cell to cell contact), phages.
- 10% to 20% of genes in many bacteria.
- Important role in genome diversification and evolution. Responsible for drug resistance in bacteria.

What is Horizontal Gene Transfer?

Genetic material is typically transferred vertically from ancestor to descendant.

However, genetic material can also get transferred horizontally across species. This is called Horizontal Gene Transfer (HGT).

- Rare in eukaryotes, but rampant in prokaryotes.
- Bacterial conjugation (cell to cell contact), phages.
- 10% to 20% of genes in many bacteria.
- Important role in genome diversification and evolution. Responsible for drug resistance in bacteria.

What is Horizontal Gene Transfer?

Genetic material is typically transferred vertically from ancestor to descendant.

However, genetic material can also get transferred horizontally across species. This is called Horizontal Gene Transfer (HGT).

- Rare in eukaryotes, but rampant in prokaryotes.
- Bacterial conjugation (cell to cell contact), phages.
- 10% to 20% of genes in many bacteria.
- Important role in genome diversification and evolution. Responsible for drug resistance in bacteria.

What is Horizontal Gene Transfer?

Genetic material is typically transferred vertically from ancestor to descendant.
However, genetic material can also get transferred horizontally across species. This is called Horizontal Gene Transfer (HGT).

- Rare in eukaryotes, but rampant in prokaryotes.
- Bacterial conjugation (cell to cell contact), phages.
- 10% to 20% of genes in many bacteria.
- Important role in genome diversification and evolution. Responsible for drug resistance in bacteria.

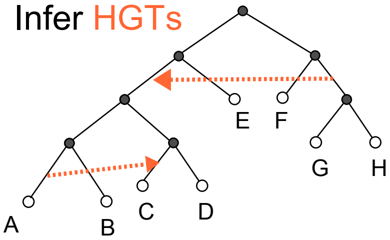
What is Horizontal Gene Transfer?

Genetic material is typically transferred vertically from ancestor to descendant.
However, genetic material can also get transferred horizontally across species. This is called Horizontal Gene Transfer (HGT).

- Rare in eukaryotes, but rampant in prokaryotes.
- Bacterial conjugation (cell to cell contact), phages.
- 10% to 20% of genes in many bacteria.
- Important role in genome diversification and evolution. Responsible for drug resistance in bacteria.

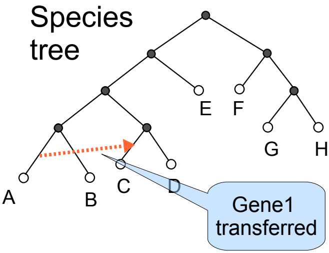
What is Horizontal Gene Transfer?

Basic Problem: Infer the HGT events in the evolutionary history of a given set of species.

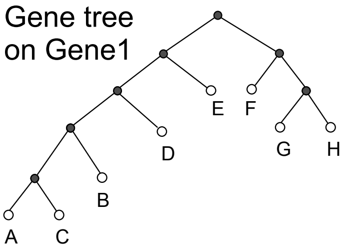


Inferring HGTs

- Use fact that evolutionary history of horizontally transferred genes does not agree with species phylogeny. For example,



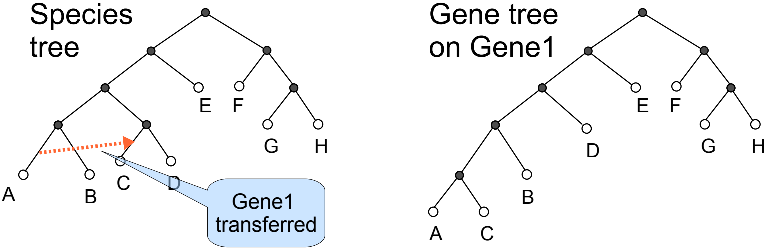
Gene1 transferred



HGT Inference Problem: Find the fewest HGT edges on a given species tree to explain a given gene tree. This problem is NP-hard [Hallett and Lagergren (2001), Bordewich and Semple (2005)].

Inferring HGTs

- Use fact that evolutionary history of horizontally transferred genes does not agree with species phylogeny. For example,



HGT Inference Problem: Find the fewest HGT edges on a given species tree to explain a given gene tree. This problem is NP-hard [Hallett and Lagergren (2001), Bordewich and Semple (2005)].

Highways of HGT

- Typically, only one or a few genes are transferred along any HGT edge.
- However, it has been observed that certain HGT edges are responsible for transferring many different genes [Beiko et al., PNAS 2005]. These edges are called Highways of HGT.

Inferring highways is important for understanding evolution of prokaryotes.

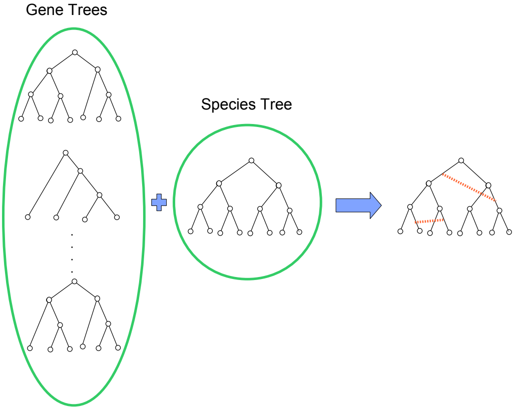
Highways of HGT

- Typically, only one or a few genes are transferred along any HGT edge.
- However, it has been observed that certain HGT edges are responsible for transferring many different genes [Beiko et al., PNAS 2005]. These edges are called Highways of HGT.

Inferring highways is important for understanding evolution of prokaryotes.

The Highway Problem

The Highway Problem: Given a set of (rooted or unrooted) gene trees $\mathcal{G}$ and the corresponding (rooted) species tree S, infer the highways of gene transfer on S.



Existing Approach for the Highway Problem

Existing approach:

- Considers each gene tree one at a time.
- Heuristically infers the HGT edges on the species tree, for that gene tree.
- Combines these individual solutions to paint an overall picture.

But...

- Solving the HGT inference problem is NP-hard.
- Many optimal solutions for each gene tree.
- The model itself may be inadequate for correctly inferring the full history of HGTs for a gene.

Our approach tackles the highway problem directly and avoids these pitfalls.

Existing Approach for the Highway Problem

Existing approach:

- Considers each gene tree one at a time.
- Heuristically infers the HGT edges on the species tree, for that gene tree.
- Combines these individual solutions to paint an overall picture.

But...

- Solving the HGT inference problem is NP-hard.
- Many optimal solutions for each gene tree.
- The model itself may be inadequate for correctly inferring the full history of HGTs for a gene.

Our approach tackles the highway problem directly and avoids these pitfalls.

Existing Approach for the Highway Problem

Existing approach:

- Considers each gene tree one at a time.
- Heuristically infers the HGT edges on the species tree, for that gene tree.
- Combines these individual solutions to paint an overall picture.

But...

- Solving the HGT inference problem is NP-hard.
- Many optimal solutions for each gene tree.
- The model itself may be inadequate for correctly inferring the full history of HGTs for a gene.

Our approach tackles the highway problem directly and avoids these pitfalls.

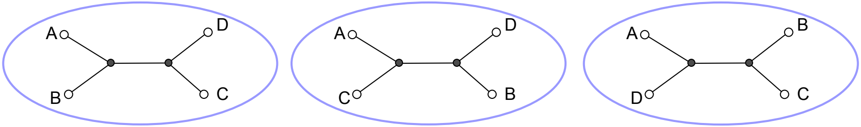
Our Approach for the Highway Problem

- Decompose each gene tree into its constituent set of quartet trees.
- Combine all the quartet trees into a single weighted set.
- Identify the quartet trees that are inconsistent with given species tree.
- Horizontal edges that can explain the most (normalized) inconsistent quartets are the proposed highways.

Our Approach for the Highway Problem

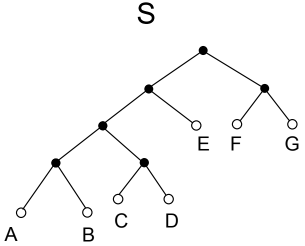
Based on quartets.

A quartet is a set of four species. Given a tree, each quartet, say $\{A, B, C, D\}$, induces one of these three topologies in the tree.



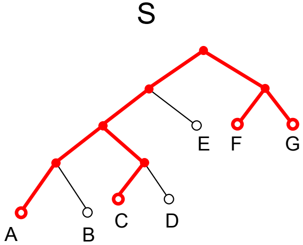
Our Approach for the Highway Problem

For example, given S and quartet $\{A, C, F, G\}$.



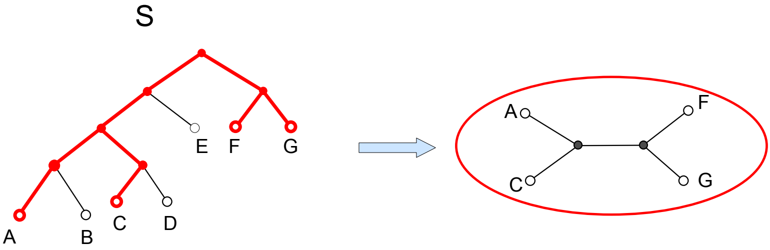
Our Approach for the Highway Problem

For example, given S and quartet $\{A, C, F, G\}$.



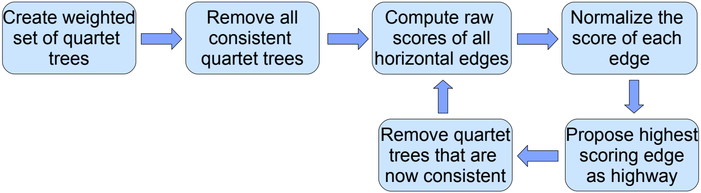
Our Approach for the Highway Problem

For example, given S and quartet $\{A, C, F, G\}$.



Our Approach for the Highway Problem

Our algorithm proceeds iteratively, proposing one highway per iteration.


Create weighted set of quartet trees

Remove all consistent quartet trees

Compute raw scores of all horizontal edges

Normalize the score of each edge

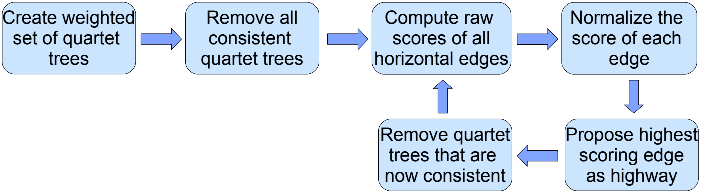
Propose highest scoring edge as highway

Remove quartet trees that are now consistent

- Terminates either after predefined number of steps or based on the distribution of edge scores.

Our Approach for the Highway Problem

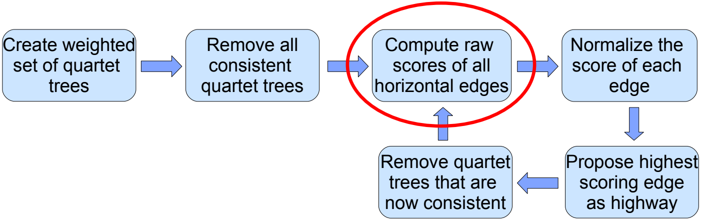
Our algorithm proceeds iteratively, proposing one highway per iteration.


Create weighted set of quartet trees
Remove all consistent quartet trees
Compute raw scores of all horizontal edges
Normalize the score of each edge
Propose highest scoring edge as highway
Remove quartet trees that are now consistent

- Terminates either after predefined number of steps or based on the distribution of edge scores.

Our Approach for the Highway Problem

Our algorithm proceeds iteratively, proposing one highway per iteration.


Create weighted set of quartet trees
Remove all consistent quartet trees
Compute raw scores of all horizontal edges
Normalize the score of each edge
Propose highest scoring edge as highway
Remove quartet trees that are now consistent

- Terminates either after predefined number of steps or based on the distribution of edge scores.

Main Computational Problem

Highway scoring problem: Given a weighted set of inconsistent quartet trees and a rooted species tree on n taxa, find the (raw) score of every horizontal edge.

What is the score?

The (raw) score of any horizontal edge is the total weight of all inconsistent quartet trees that can be explained by an HGT event along that horizontal edge.

- Can be solved naïvely in $O(n^6)$ time: Evaluate each of the $O(n^2)$ horizontal edges, on the $O(n^4)$ quartet trees.
- We developed an algorithm that computes all scores in $O(n^4)$ time. This is optimal.

Main Computational Problem

Highway scoring problem: Given a weighted set of inconsistent quartet trees and a rooted species tree on n taxa, find the (raw) score of every horizontal edge.

What is the score?

The (raw) score of any horizontal edge is the total weight of all inconsistent quartet trees that can be explained by an HGT event along that horizontal edge.

- Can be solved naïvely in $O(n^6)$ time: Evaluate each of the $O(n^2)$ horizontal edges, on the $O(n^4)$ quartet trees.
- We developed an algorithm that computes all scores in $O(n^4)$ time. This is optimal.

Main Computational Problem

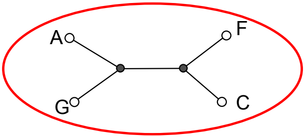
Highway scoring problem: Given a weighted set of inconsistent quartet trees and a rooted species tree on n taxa, find the (raw) score of every horizontal edge.

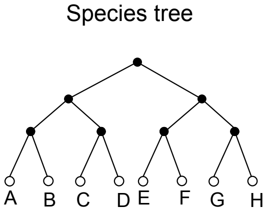
What is the score?

The (raw) score of any horizontal edge is the total weight of all inconsistent quartet trees that can be explained by an HGT event along that horizontal edge.

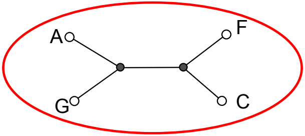
- Can be solved naïvely in $O(n^6)$ time: Evaluate each of the $O(n^2)$ horizontal edges, on the $O(n^4)$ quartet trees.
- We developed an algorithm that computes all scores in $O(n^4)$ time. This is optimal.

Crucial Structural property

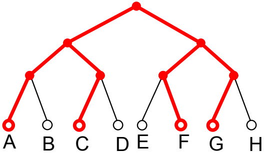




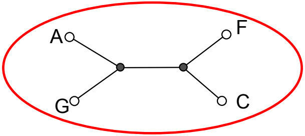
Crucial Structural property



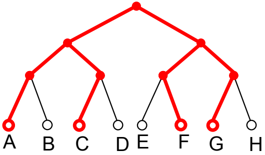
Species tree



Crucial Structural property

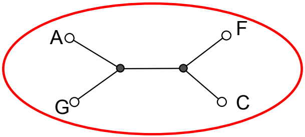


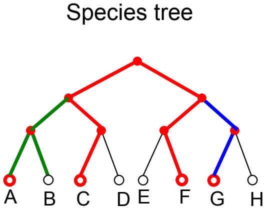
Species tree



1. Move G towards A.
2. Move A towards G.
3. Move C towards F.
4. Move F towards C.

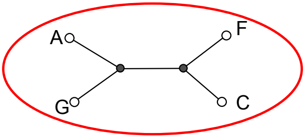
Crucial Structural property

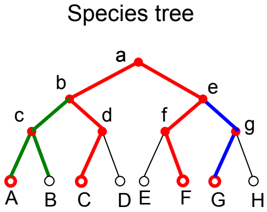




1. Move G towards A.
2. Move A towards G.
3. Move C towards F.
4. Move F towards C.

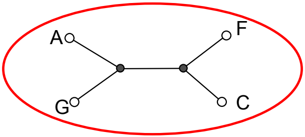
Crucial Structural property



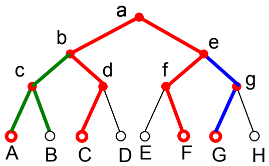


1. Move G towards $A \Rightarrow \langle c, e \rightarrow G \rangle$.
2. Move A towards G.
3. Move C towards F.
4. Move F towards C.

Crucial Structural property

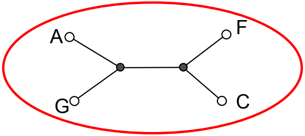


Species tree

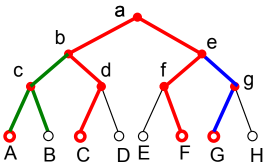


1. Move G towards $A \Rightarrow \langle c, e \rightarrow G \rangle$.
2. Move A towards $G \Rightarrow \langle g, b \rightarrow A \rangle$.
3. Move C towards $F \Rightarrow \langle f, b \rightarrow C \rangle$.
4. Move F towards $C \Rightarrow \langle d, e \rightarrow F \rangle$.

Crucial Structural property



Species tree



1. Move G towards $A \Rightarrow \langle c, e \rightarrow G \rangle$.
2. Move A towards $G \Rightarrow \langle g, b \rightarrow A \rangle$.
3. Move C towards $F \Rightarrow \langle f, b \rightarrow C \rangle$.
4. Move F towards $C \Rightarrow \langle d, e \rightarrow F \rangle$.

Depending on how the quartet is embedded in the species tree, we can always get something similar.

Our Algorithm

We make use of this characterization to decorate the species tree with the subtree-path pairs for all the quartet trees.

We then use a dynamic programming algorithm to compute the score of each horizontal edge.

Our Algorithm

- Decorating the species tree, for each quartet tree, takes $O(1)$ time.
- The dynamic programming algorithm takes $O(n^2)+$ $O(\text{Number of quartet trees})$ time.
- Total time Complexity: $= O(n^2)+$ $O(\text{Number of quartet trees})$, i.e., $O(n^4)$.

Normalization

Different HGTs can explain different numbers of inconsistent quartets.

Normalization of edge scores: Divide raw score of a horizontal edge by the maximum number of inconsistent quartets that could be explained by an HGT event along that edge.

We developed an $O(n^2)$-time algorithm to normalize all the raw cores.

Normalization

Different HGTs can explain different numbers of inconsistent quartets.

Normalization of edge scores: Divide raw score of a horizontal edge by the maximum number of inconsistent quartets that could be explained by an HGT event along that edge.

We developed an $O(n^2)$-time algorithm to normalize all the raw cores.

Normalization

Different HGTs can explain different numbers of inconsistent quartets.

Normalization of edge scores: Divide raw score of a horizontal edge by the maximum number of inconsistent quartets that could be explained by an HGT event along that edge.

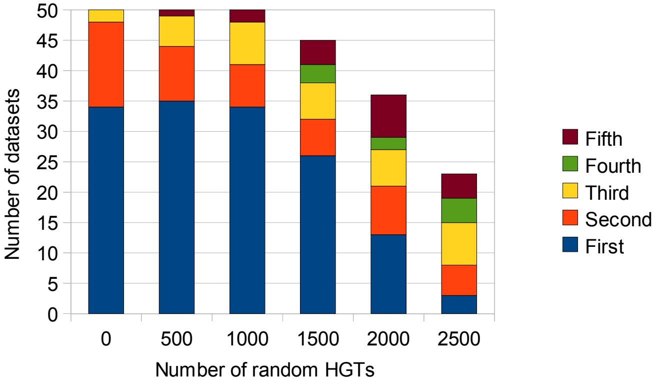
We developed an $O(n^2)$-time algorithm to normalize all the raw cores.

Simulation Study

Measure the effect of noise on highway inference:

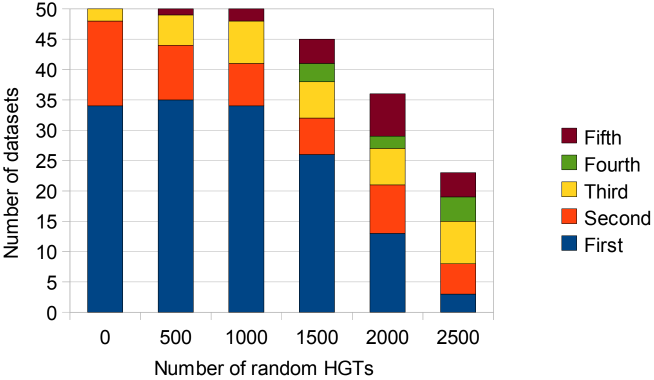
- Simulated data sets of 50 taxa, 1000 gene trees.
- Randomly implanted highway affecting 10% of genes.
- Noise levels: 0, 500, 1000, 1500, 2000, and 2500 HGTs, with each HGT event affecting a gene chosen at random with replacement.
- 50 data sets for each noise level.

Simulation Study



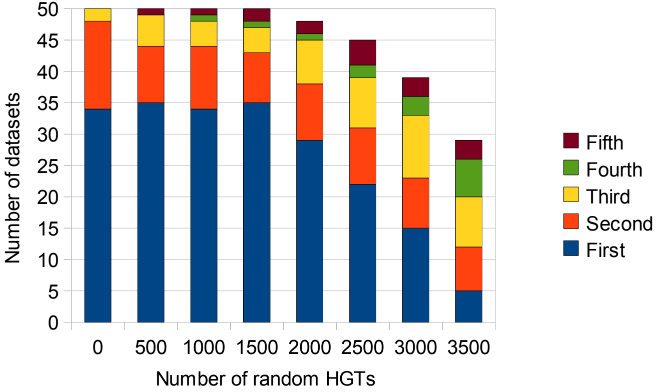
Average ranks of the implanted highways: 1.36, 1.46, 1.58, 2.56, 5.26, and 19.20 respectively (out of over 4000 candidates).

Simulation Study



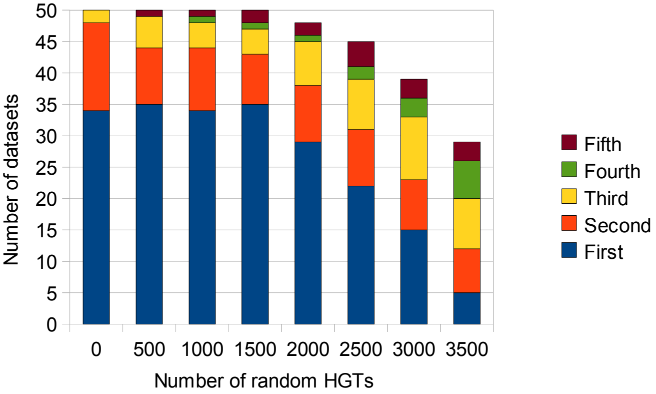
Average ranks of the implanted highways: 1.36, 1.46, 1.58, 2.56, 5.26, and 19.20 respectively (out of over 4000 candidates).

Simulation Study



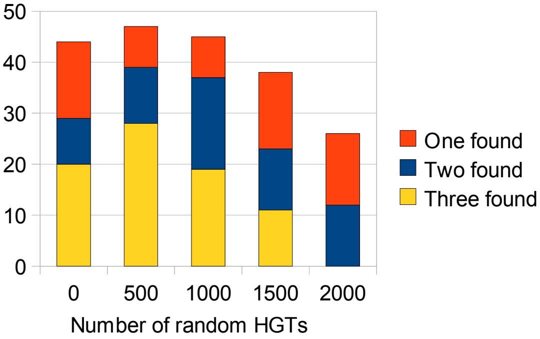
Average ranks of the implanted highways: 1.36, 1.46, 1.50, 1.54, 1.90, 2.60, 4.08, and 9.24 respectively.

Simulation Study



Average ranks of the implanted highways: 1.36, 1.46, 1.50, 1.54, 1.90, 2.60, 4.08, and 9.24 respectively.

Simulation Study



Running time

To analyze any dataset we:

1. Decompose the gene trees to generate the weighted set of quartets: $O(t \cdot n^4)$ time
2. Solve the highway scoring problem: $O(n^4)$ time

The bottleneck is thus the quartet decomposition step.

Still, we can quickly analyze fairly large datasets:

- 50 taxa, 1000 gene trees: 30 seconds,
- 100 taxa, 1000 gene trees: 15 minutes,
- 200 taxa, 1000 gene trees: 5 hours.

Running time

To analyze any dataset we:

1. Decompose the gene trees to generate the weighted set of quartets: $O(t \cdot n^4)$ time
2. Solve the highway scoring problem: $O(n^4)$ time

The bottleneck is thus the quartet decomposition step.

Still, we can quickly analyze fairly large datasets:

- 50 taxa, 1000 gene trees: 30 seconds,
- 100 taxa, 1000 gene trees: 15 minutes,
- 200 taxa, 1000 gene trees: 5 hours.

Running time

To analyze any dataset we:

1. Decompose the gene trees to generate the weighted set of quartets: $O(t \cdot n^4)$ time
2. Solve the highway scoring problem: $O(n^4)$ time

The bottleneck is thus the quartet decomposition step.

Still, we can quickly analyze fairly large datasets:

- 50 taxa, 1000 gene trees: 30 seconds,
- 100 taxa, 1000 gene trees: 15 minutes,
- 200 taxa, 1000 gene trees: 5 hours.

Shortcomings of This Approach

Two main shortcomings:

1. We lose information about individual gene trees. So, even though any gene is transferred in only one of the directions along a highway, our normalization is unable to take that into account.

   E.g., Consider some highway $\{x_1, y_1\}$ that transfers 100 genes. Let $x_1 \rightarrow y_1$ give 10 inconsistent quartets and $y_1 \rightarrow x_1$ give 30 inconsistent quartets.
   All transfers in $x_1 \rightarrow y_1$ direction imply total normalized score of $10 \times 100/NF$, while all if transfers are in the other direction then score is $30 \times 100/NF$.

2. If a gene tree has too many missing leaves, it cannot be used in the analysis.

Shortcomings of This Approach

Two main shortcomings:

1. We lose information about individual gene trees. So, even though any gene is transferred in only one of the directions along a highway, our normalization is unable to take that into account.

   E.g., Consider some highway $\{x_1, y_1\}$ that transfers 100 genes. Let $x_1 \rightarrow y_1$ give 10 inconsistent quartets and $y_1 \rightarrow x_1$ give 30 inconsistent quartets.
   All transfers in $x_1 \rightarrow y_1$ direction imply total normalized score of $10 \times 100/NF$, while all if transfers are in the other direction then score is $30 \times 100/NF$.
2. If a gene tree has too many missing leaves, it cannot be used in the analysis.

Shortcomings of This Approach

Two main shortcomings:

1. We lose information about individual gene trees. So, even though any gene is transferred in only one of the directions along a highway, our normalization is unable to take that into account.

   E.g., Consider some highway $\{x_1, y_1\}$ that transfers 100 genes. Let $x_1 \rightarrow y_1$ give 10 inconsistent quartets and $y_1 \rightarrow x_1$ give 30 inconsistent quartets.
   All transfers in $x_1 \rightarrow y_1$ direction imply total normalized score of $10 \times 100/NF$, while all if transfers are in the other direction then score is $30 \times 100/NF$.
2. If a gene tree has too many missing leaves, it cannot be used in the analysis.

Our New Method

Same basic idea, but look at the quartet decomposition of each gene tree separately and normalize scores at gene tree level.

Addresses both the problems mentioned in previous slide, and improves performance drastically.

Our New method

1: For each input gene tree G_i, for $1 \leq i \leq k$,

  1(a): Decompose G_i into its constituent set of quartet trees, denoted Φ_i.

  1(b): Remove from Φ_i all those quartet trees that are consistent with S (or that can be explained by a previously inferred highway).

  1(c): For each horizontal edge $\{u, v\} \in H(S)$, compute the normalized score $NS(\{u, v\}, \Phi_i)$.

2: For each horizontal edge $\{u, v\} \in H(S)$, compute its final score, denoted $score\{u, v\}$, to be $\sum_{i=1}^{k} NS(\{u, v\}, \Phi_i)$.

3: Select the highest scoring horizontal edge as a highway.

Our New method

The normalized score, w.r.t any Φ_i, is computed in essentially the same way as before; except that we can now also take the direction into account.

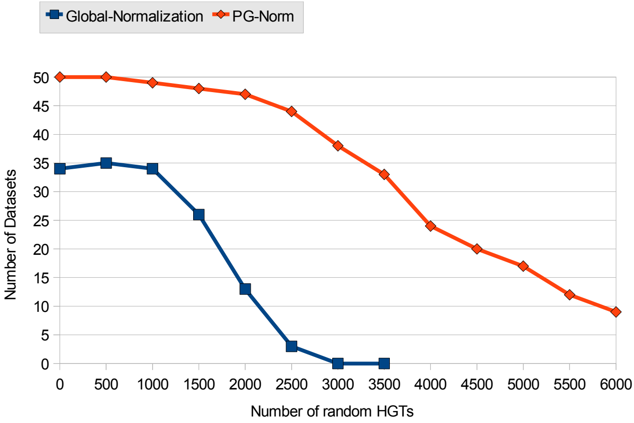
So we compute normalized scores of the directed HGT events and assign $NS(\{u, v\}, \Phi_i) = \max\{NS((u, v), \Phi_i), NS((v, u), \Phi_i)\}$.

Our New method

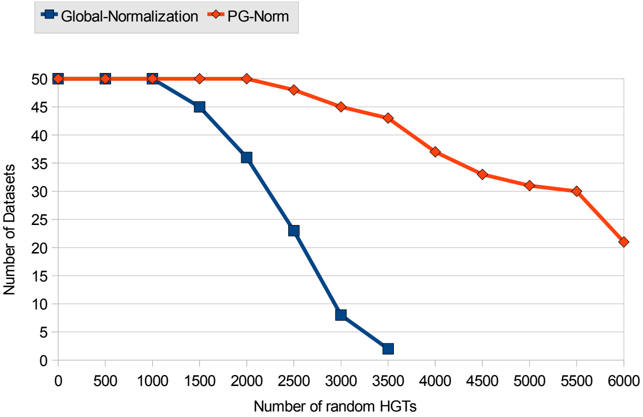
The normalized score, w.r.t any Φ_i, is computed in essentially the same way as before; except that we can now also take the direction into account.

So we compute normalized scores of the directed HGT events and assign $NS(\{u, v\}, \Phi_i) = \max\{NS((u, v), \Phi_i), NS((v, u), \Phi_i)\}$.

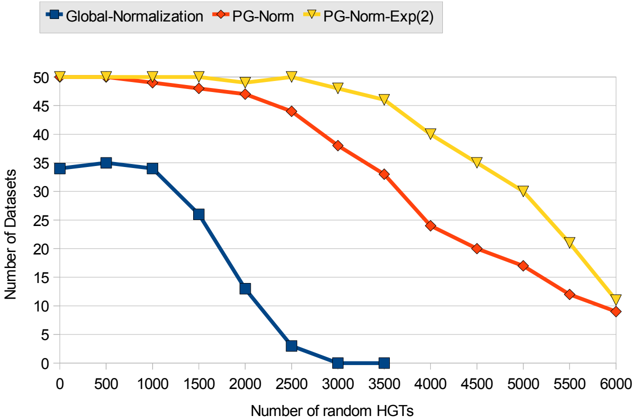
Comparative Results



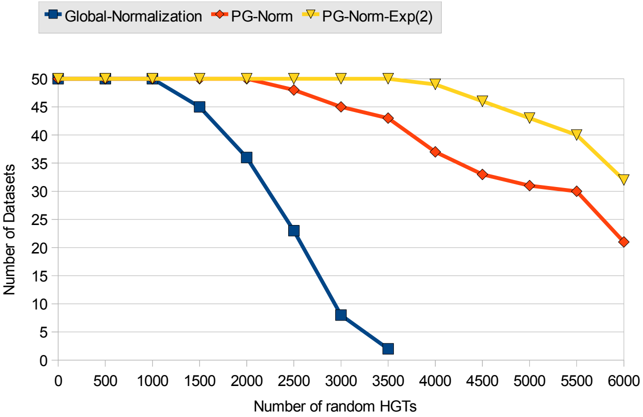
Comparative Results



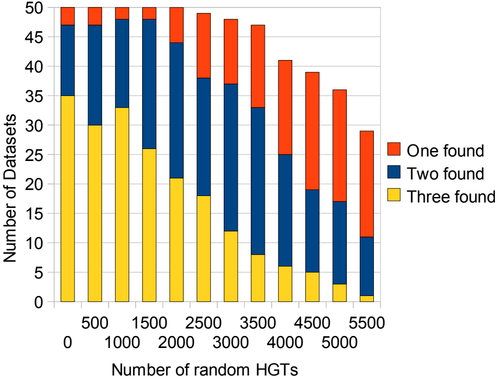
Comparative Results



Comparative Results



Comparative Results



Time Complexity

Essentially the same as the previous method: $O(t \cdot n^4)$

Slower by approximately a constant factor of 16.

Biological Data Analysis

- Dataset of 144 prokaryotic species and over 22000 gene families.
- Each gene family represented by 100 gene trees.
- Each iteration takes less than 6 hours!

Conclusion

Our new approach:

- Looks at combined data from all gene trees to detect fingerprints of highways.
- Bypasses need to infer individual HGT events.
- Highly accurate at even high rates of HGT or noise.
- Capable of efficiently handling even very large datasets.
- Can deal cleanly with uncertainty in gene trees.

Thank You!

Questions!