Recovering the Treelike Trend of Evolution Despite Extensive Lateral Genetic Transfer: A Probabilistic Analysis

Sebastien Roch, Sagi Snir

presentation by Minghao

April 1, 2020
The problem

- Species tree from gene trees generated with LGT
The problem

- Species tree from gene trees generated with LGT
- Lateral genetic transfer
  - Referred to as Horizontal Genetic Transfer (HGT) in the book and other places
The problem

- Species tree from gene trees generated with LGT
- Lateral genetic transfer
  - Referred to as Horizontal Genetic Transfer (HGT) in the book and other places
  - Gene may transfer horizontally, between distantly related organisms. Common in prokaryotes.
The problem

- Species tree from gene trees generated with LGT
- Lateral genetic transfer
  - Referred to as *Horizontal Genetic Transfer* (HGT) in the book and other places
  - Gene may transfer horizontally, between distantly related organisms. Common in prokaryotes.
  - **Stochastic** HGT - attempt model gene transfer as a stochastic process
Start from our usual species tree, we generate a gene tree:

- Fix a rate of transfer $\lambda$, subsequently fixing the average total number of HGT events.
- Determine locations of transfer (recipients) on the species tree by a continuous Poisson process starting from the root and going down locations determined independently with a certain density. Locations can be in the middle of some edge.
- Select a transfer donor uniformly randomly from the locations that are contemporaneous (same branch length from the root) with the recipient within a certain phylogenetic distance $R$ (so somewhat related) of the recipient.
- Move the subtree under the recipient to the donor (SPR move).
Start from our usual species tree, we generate a gene tree:

- Fix a rate of transfer $\lambda$, subsequently fixing the average total number of HGT events.
Stochastic HGT model

Start from our usual species tree, we generate a gene tree:

- Fix a rate of transfer $\lambda$, subsequently fixing the average total number of HGT events.
- Determine locations of transfer *(recipients)* on the species tree by a continuous Poisson process.
Stochastic HGT model

Start from our usual species tree, we generate a gene tree:

- Fix a rate of transfer $\lambda$, subsequently fixing the average total number of HGT events.
- Determine locations of transfer (recipients) on the species tree by a continuous Poisson process:
  - start from the root and goes down
Start from our usual species tree, we generate a gene tree:

- Fix a rate of transfer $\lambda$, subsequently fixing the average total number of HGT events.
- Determine locations of transfer (recipients) on the species tree by a continuous Poisson process
  - start from the root and goes down
  - locations determined independently with a certain density
Stochastic HGT model

Start from our usual species tree, we generate a gene tree:

- Fix a rate of transfer $\lambda$, subsequently fixing the average total number of HGT events.
- Determine locations of transfer (recipients) on the species tree by a continuous Poisson process
  - start from the root and goes down
  - locations determined independently with a certain density
  - locations can be in the middle of some edge
Stochastic HGT model

Start from our usual species tree, we generate a gene tree:

- Fix a rate of transfer $\lambda$, subsequently fixing the average total number of HGT events.
- Determine locations of transfer (recipients) on the species tree by a continuous Poisson process:
  - start from the root and goes down
  - locations determined independently with a certain density
  - locations can be in the middle of some edge
Start from our usual species tree, we generate a gene tree:

- Fix a rate of transfer $\lambda$, subsequently fixing the average total number of HGT events.
- Determine locations of transfer (recipients) on the species tree by a continuous Poisson process
  - start from the root and goes down
  - locations determined independently with a certain density
  - locations can be in the middle of some edge
- Select a transfer donor uniformly randomly from the locations that are:
Start from our usual species tree, we generate a gene tree:

- Fix a rate of transfer $\lambda$, subsequently fixing the average total number of HGT events.
- Determine locations of transfer (recipients) on the species tree by a continuous Poisson process
  - start from the root and goes down
  - locations determined independently with a certain density
  - locations can be in the middle of some edge
- Select a transfer donor uniformly randomly from the locations that are:
  - contemporaneous (same branch length from the root) with the recipient
Stochastic HGT model

Start from our usual species tree, we generate a gene tree:

- Fix a rate of transfer $\lambda$, subsequently fixing the average total number of HGT events.
- Determine locations of transfer (recipients) on the species tree by a continuous Poisson process
  - start from the root and goes down
  - locations determined independently with a certain density
  - locations can be in the middle of some edge
- Select a transfer donor uniformly randomly from the locations that are:
  - contemporaneous (same branch length from the root) with the recipient
  - within a certain phylogenetic distance $R$ (so somewhat related) of the recipient
Stochastic HGT model

Start from our usual species tree, we generate a gene tree:

- Fix a rate of transfer $\lambda$, subsequently fixing the average total number of HGT events.
- Determine locations of transfer (recipients) on the species tree by a continuous Poisson process
  - start from the root and goes down
  - locations determined independently with a certain density
  - locations can be in the middle of some edge
- Select a transfer donor uniformly randomly from the locations that are:
  - contemporaneous (same branch length from the root) with the recipient
  - within a certain phylogenetic distance $R$ (so somewhat related) of the recipient
- Move the subtree under the recipient to the donor (SPR move).
Preferential HGT

A way of modeling the non-randomness of HGT events

- When selecting the donor, we can set a distance $R$ such that only donors within the distance $R$ can be selected.
A way of modeling the non-randomness of HGT events

- When selecting the donor, we can set a distance $R$ such that only donors within the distance $R$ can be selected.
- Distance is the sum of branch lengths, or phylogenetic distances from the recipient/donor to their closest common ancestor.

Preferential HGT
A way of modeling the non-randomness of HGT events

- When selecting the donor, we can set a distance $R$ such that only donors within the distance $R$ can be selected
- Distance is the sum of branch lengths, or phylogenetic distances from the recipient/donor to their closest common ancestor
- We are allowed to set $R = \infty$ to make HGT happen more randomly (which makes our estimation harder)
Stochastic HGT model

The model properties:

- HGT
Stochastic HGT model

The model properties:

- HGT

- Taxon sampling: for each gene tree, only a subset of taxa remains.
Questions:

- How high the transfer rate can be so that the species tree is still identifiable?
Questions:

- How high the transfer rate can be so that the species tree is still identifiable?
- How to reconstruct the species tree if it is identifiable?
Theorem (Recoverability)

Under the stochastic HGT model, with \( \Omega(\log n) \) number of gene trees, an average of \( O(n/\log n) \) total number of HGT events, and preferential distance \( R = \infty \), it is possible to reconstruct the species tree with high probability.

They gave a statistically consistent algorithm:

- The algorithm: for every 4 taxa, take the majority of all induced quartets on the gene trees.
Theorem (Recoverability)

Under the stochastic HGT model, with $\Omega(\log n)$ number of gene trees, an average of $O(n/\log n)$ total number of HGT events, and preferential distance $R = \infty$, it is possible to reconstruct the species tree with high probability.

They gave a statistically consistent algorithm:

- The algorithm: for every 4 taxa, take the majority of all induced quartets on the gene trees.
- The majority quartet will be the correct quartet with high probability.
Theorem (Recoverability)

Under the stochastic HGT model, with $\Omega(\log n)$ number of gene trees, an average of $O(n/\log n)$ total number of HGT events, and preferential distance $R = \infty$, it is possible to reconstruct the species tree with high probability.

They gave a statistically consistent algorithm:

- The algorithm: for every 4 taxa, take the majority of all induced quartets on the gene trees.
- The majority quartet will be the correct quartet with high probability.
- We can then reconstruct the tree with the quartets (using other algorithms).
Results

Theorem (Recoverability)

Under the stochastic HGT model, with $\Omega(\log n)$ number of gene trees, an average of $O(n / \log n)$ total number of HGT events, and preferential distance $R = \infty$, it is possible to reconstruct the species tree with high probability.

They gave a statistically consistent algorithm:

- The algorithm: for every 4 taxa, take the majority of all induced quartets on the gene trees.
- The majority quartet will be the correct quartet with high probability.
- We can then reconstruct the tree with the quartets (using other algorithms).
- More of a statistical property than an algorithm.
Theorem (Recoverability)

Under the stochastic HGT model, with $\Omega(\log n)$ number of gene trees, an average of $O(n/\log n)$ total number of HGT events, and preferential distance $R = \infty$, it is possible to reconstruct the species tree with high probability.

They gave a statistically consistent algorithm:

- The algorithm: for every 4 taxa, take the majority of all induced quartets on the gene trees.
- The majority quartet will be the correct quartet with high probability.
- We can then reconstruct the tree with the quartets (using other algorithms).
- More of a statistical property than an algorithm.

In Daskalakis & Roch ’16, the rate is improved to constant, so a total of $O(n)$ number of HGT events can happen and we can still recover the tree.
Implication: any method that solves (or statistically consistently approximates) **maximum quartet consistency** will also work under the stochastic HGT model.
Implication: any method that solves (or statistically consistently approximates) maximum quartet consistency will also work under the stochastic HGT model.

- ASTRAL, Quartet MaxCut, etc.
Theorem (Unrecoverability)

Under the stochastic HGT model, with \( \Omega(\log n) \) number of gene trees and an average of \( \Omega(n \log \log n) \) total number of HGT events, it is not possible to reconstruct the species tree in general.
Theorem (Unrecoverability)

*Under the stochastic HGT model, with $\Omega(\log n)$ number of gene trees and an average of $\Omega(n \log \log n)$ total number of HGT events, it is not possible to reconstruct the species tree in general.*

Proves that two topologically different species trees can generate the same gene tree with high probability, so we can’t determine the underlying species tree.
Motivation: some times in the phylogeny, there will be a high number of HGT events happening on some particular edges.

- A HGT highway is a pair of edges that shares contemporaneous locations.

Result: the same algorithm reconstruct the species tree, as long as the fraction of genes trees affected by the highways are small. Specifically, suppose there are $h$ highways, we need the fraction of genes trees affected by each highway to be smaller than $\frac{1}{2^h}$. 
HGT Highways

Motivation: some times in the phylogeny, there will be a high number of HGT events happening on some particular edges.

- A **HGT highway** is a pair of edges that shares contemporaneous locations.
- A highway will affect a subset of the gene trees.
HGT Highways

Motivation: some times in the phylogeny, there will be a high number of HGT events happening on some particular edges.

- **A HGT highway** is a pair of edges that shares contemporaneous locations.
- A highway will affect a subset of the gene trees.
- If a highway affects a gene tree, a deterministic HGT event will happen on that pair of edges, in addition to the regular random HGTs.
HGT Highways

Motivation: some times in the phylogeny, there will be a high number of HGT events happening on some particular edges.

- A **HGT highway** is a pair of edges that shares contemporaneous locations.
- A highway will affect a subset of the gene trees.
- If a highway affects a gene tree, a deterministic HGT event will happen on that pair of edges, in addition to the regular random HGTs.
- This models some non-random aspects of HGT
HGT Highways

Motivation: some times in the phylogeny, there will be a high number of HGT events happening on some particular edges.

- A **HGT highway** is a pair of edges that shares contemporaneous locations.
- A highway will affect a subset of the gene trees.
- If a highway affects a gene tree, a deterministic HGT event will happen on that pair of edges, in addition to the regular random HGTs.
- This models some non-random aspects of HGT

Result: the same algorithm reconstruct the species tree, as long as the fraction of genes trees affected by the highways are small.
HGT Highways

Motivation: some times in the phylogeny, there will be a high number of HGT events happening on some particular edges.

- A **HGT highway** is a pair of edges that shares contemporaneous locations.
- A highway will affect a subset of the gene trees.
- If a highway affects a gene tree, a deterministic HGT event will happen on that pair of edges, in addition to the regular random HGTs.
- This models some non-random aspects of HGT

Result: the same algorithm reconstruct the species tree, as long as the fraction of genes trees affected by the highways are small. Specifically, suppose there are $h$ highways, we need the fraction of genes trees affected by each highway to be smaller than $1/(2h)$. 
Contributions: give many of the first theoretical results regarding the stochastic HGT model
Contributions: give many of the first theoretical results regarding the stochastic HGT model

- Under the model with rate $O(1/\log n)$, quartet majority is statistically consistent.
Conclusions

Contributions: give many of the first theoretical results regarding the stochastic HGT model

- Under the model with rate $O(1/\log n)$, quartet majority is statistically consistent.
  - later improved to constant rate by Daskalakis & Roch ’16

Limitations:
- Result is unavoidably highly model-dependent
- Models HGT as random events, can use more biological realism
- Even with highways, the model is still quite random
Conclusions

Contributions: give many of the first theoretical results regarding the stochastic HGT model

- Under the model with rate $O(1/\log n)$, quartet majority is statistically consistent.
  - later improved to constant rate by Daskalakis & Roch '16
  - implies that methods based on quartet consistency retains their statistical consistency under this model

Limitations:
- Result is unavoidably highly model-dependent
- Models HGT as random events, can use more biological realism
- Even with highways, the model is still quite random
Contributions: give many of the first theoretical results regarding the stochastic HGT model

- Under the model with rate $O(1/\log n)$, quartet majority is statistically consistent.
  - later improved to constant rate by Daskalakis & Roch ’16
  - implies that methods based on quartet consistency retains their statistical consistency under this model

- Generalize the model to include highways. The result still holds if highways only affect a small fraction of genes.
Conclusions

Contributions: give many of the first theoretical results regarding the stochastic HGT model

- Under the model with rate $O(1/\log n)$, quartet majority is statistically consistent.
  - later improved to constant rate by Daskalakis & Roch ’16
  - implies that methods based on quartet consistency retains their statistical consistency under this model

- Generalize the model to include highways. The result still holds if highways only affect a small fraction of genes.

- The species tree is not identifiable with rate $\Omega(\log \log n)$. 
Conclusions

Contributions: give many of the first theoretical results regarding the stochastic HGT model

- Under the model with rate $O(1/\log n)$, quartet majority is statistically consistent.
  - later improved to constant rate by Daskalakis & Roch ’16
  - implies that methods based on quartet consistency retains their statistical consistency under this model
- Generalize the model to include highways. The result still holds if highways only affect a small fraction of genes.
- The species tree is not identifiable with rate $\Omega(\log \log n)$.

Limitations:
- Result is unavoidably highly model-dependent
Conclusions

Contributions: give many of the first theoretical results regarding the stochastic HGT model

- Under the model with rate $O(1/\log n)$, quartet majority is statistically consistent.
  - later improved to constant rate by Daskalakis & Roch ’16
  - implies that methods based on quartet consistency retains their statistical consistency under this model
- Generalize the model to include highways. The result still holds if highways only affect a small fraction of genes.
- The species tree is not identifiable with rate $\Omega(\log \log n)$.

Limitations:

- Result is unavoidably highly model-dependent
- Models HGT as random events, can use more biological realism
Conclusions

Contributions: give many of the first theoretical results regarding the stochastic HGT model

- Under the model with rate $O(1/\log n)$, quartet majority is statistically consistent.
  - later improved to constant rate by Daskalakis & Roch ’16
  - implies that methods based on quartet consistency retains their statistical consistency under this model

- Generalize the model to include highways. The result still holds if highways only affect a small fraction of genes.

- The species tree is not identifiable with rate $\Omega(\log \log n)$.

Limitations:

- Result is unavoidably highly model-dependent
- Models HGT as random events, can use more biological realism
- Even with highways, the model is still quite random
Questions?

References:

- S Roch, S Snir. (2013) Recovering the treelike trend of evolution despite extensive lateral genetic transfer: a probabilistic analysis
- C Daskalakis, S Roch. (2016) Species trees from gene trees despite a high rate of lateral genetic transfer: A tight bound.