Gene tree inference under gene duplication, transfer and loss with

GeneRax
Resolving phylogenetic conflicts
About me

- Benoit Morel (benoit.morel@h-its.org)
- PhD student (advisor: Alexandros Stamatakis)
- Working on phylogenetics tools development
Phylogenetics, species and gene trees

Gene tree correction and reconciliation problem

Our solution

GeneRax
Resolving phylogenetic conflicts
Definitions

In the scope of this talk:

**Species**: group of organisms that share a pool of genes and that are able to interbreed

**Gene**: a DNA or protein sequence that belongs to a species

**Gene family (or homologous genes)**: set of genes that share a common ancestor.
(Phylogenetic) species tree

Represent the evolutionary history of species
Gene tree

- Represent the evolutionary history of homologous genes
- High correlation with the species tree
Gene trees evolve along a species tree.
Gene trees evolve along a species tree

Ancestral gene
Gene trees evolve along a species tree.
Gene trees evolve along a species tree.

Loss event
Gene trees evolve along a species tree
Gene trees evolve along a species tree
Gene trees evolve along a species tree
Gene trees evolve along a species tree.
The reconciliation scenario corresponds to a gene tree.
The UndatedDTL model

A gene lineage present in a given species lineage can either:

- Duplicate with probability $p_D$
- Speciate with probability $p_S$
- Get lost with probability $p_L$
- Get transferred to another species with probability $p_T$
Probability of a scenario

\[ P(\text{scenario}) = P_s^6 \times P_I \times P_d \times (P_t / \#\text{species}) \]
The reconciliation likelihood

- $P(G, S) = \text{probability of observing a gene tree } G \ \text{given a rooted species tree } S$

- Sum over all possible DTLS scenarios that would generate $G$

![Diagram showing a gene tree $G$ with nodes a1, a2, c, and b.](attachment:image.png)
Gene sequences evolve along a gene tree

Speciation, duplication or HGT
The phylogenetic likelihood

Sequence evolution models: JC, GTR, LG etc.

\[ P(A,G) = \text{probability of observing an alignment } A \text{ given a gene tree } G \]

(Implemented in RaxML, IQTree, PhyML)

Speciation, duplication or HGT
The whole process

The gene tree evolves along the species tree

Speciation

The sequences evolve along the gene tree

Speciation, duplication or HGT

AGGAGCTTC

AGGAGCTTC

AGGAGCTTC

AGGAGCTTC

A

T

G

C

The gene tree evolves along the species tree

The sequences evolve along the gene tree
Gene tree and species tree are different.
Gene tree and species tree are different.
Gene tree and species tree are different

- Tree inference error
- DTL events
- (Others: Incomplete lineage sorting, hybridization etc.)
Gene tree and species tree are different

- Tree inference error
  - Need to be fixed! (Gene tree correction)

- DTL events
  - Need to be explained! (Gene tree reconciliation)
Resolving the conflicts to recover the truth

- **Gene tree inference:** infer a gene tree from the sequences

- **Gene tree correction:** correct the gene tree topology using the species tree

- **Gene tree – species tree reconciliation:** explain how the gene tree evolved within the species tree
Gene tree inference (from the sequences)

Maximize $P(A|G)$

Standard software: RAxML, IQTree, PhyML, (FastTree)
Gene tree inference (from the sequences)

Limitations:

- Innaccurate (lack of signal)
- Does not infer DTL events
Gene tree correction and reconciliation with a species tree

Rooted species tree

MSA (DNA or protein)

Unrooted gene tree estimation

Reconciled rooted gene tree
Related work

- Parsimony methods:
  - Notung, EcceTERA, Treerecs, Ranger-DTL etc.

- Bayesian methods:
  - ALE
Related work: parsimony

- Infer a gene tree from the sequences
- Identify parts of the tree we do not trust
- Rearrange with parsimonious reconciliation

A gene tree with support values

Low support values are in red
Related work: parsimony

Tools: Notung, EcceTERA, Ranger-DTL etc.

→ the correction step is very fast
→ BUT inferring support values is slow
→ how to pick the support value threshold??
Related work: bayesian

- Infer a gene tree distribution from the sequences (MrBayes, ExaBayes etc.)

- Sample gene trees from this distribution under the UndatedDTL model
Related work: bayesian

Tool: ALE

→ More accurate than parsimony methods

→ BUT inferring the gene tree distribution is very slow
Maximum likelihood

Parsimony → not accurate
   → expansive precomputations

Bayesian → very expansive precomputations

What about trying maximum likelihood?
Our solution: maximum likelihood method accounting for both sequences and species tree

→ more accurate than parsimony

→ does not need any slow precomputation step
GeneRax

Maximizes the joint likelihood:

$$L(G|S) = P(A|G) \cdot P(G|S)$$

Phylogenetic likelihood

Reconciliation likelihood

Speciation, duplication or HGT
Tree search heuristic

• Start from an initial gene tree

• Apply small changes to the tree and keep the ones that improve the likelihood

• Stop when we cannot find any better tree
Result: the gene tree that is the “best” compromise between the signal from the sequences and the signal from the species tree.
Reconciliation with GeneRax

Once we inferred the best gene tree, we can either return:

- the maximum likelihood scenario
- a stochastic sample of plausible scenarios
Displaying reconciliations

Output in RecPhyloXML

Can be viewed with RecPhyloVisu

http://phylariane.univ-lyon1.fr/recphyloxml/recphylovisu
GeneRax is suitable for large analyses

- Process single or multiple gene families in parallel with MPI
  - can also be deployed on a cluster

- Checkpoint system (a run can be restarted after an interruption)
Results (accuracy)
Results (accuracy)
Results (runtime)

![Bar graph showing runtime comparison for various tools]

- RAXML-NG
- Notung
- Treerecs
- EcceTERA
- ALE
- GeneRax-Raxml
- GeneRax-Random

The y-axis represents time in seconds, ranging from $10^3$ to $10^7$. Each bar is divided into two sections, one for the tool and one for pre-processing.
GeneRax strengths

- fast and accurate
- mathematically satisfying
- suitable for large analyses
- simplifies pipelines
- actively maintained
- available on github and bioconda
Future work on GeneRax

- Infer the species tree from gene trees
- Provide post-analysis tools
- Ask the users what they would like to have!
Thanks!

Contact: benoit.morel@h-its.org

GeneRax on Github: https://github.com/BenoitMorel/GeneRax/

GeneRax on BioConda: https://anaconda.org/bioconda/generax