Genome-scale phylogenetics through gene tree parsimony

Mukul S. Bansal

Computer Science & Engineering
University of Connecticut, Storrs, CT, USA

Gene Trees and Species Trees

A species tree is an evolutionary tree depicting the evolutionary history of a set of species.

A gene tree is an evolutionary tree depicting the evolutionary history of a gene family from some set of species.

- In general, one expects that the evolution of genes should mimic the evolution of the species themselves. But this is frequently not true.
- Gene trees built on different genes, taken from the same set of species, are often incongruent with one another and with the species tree.

Gene Trees and Species Trees

A species tree is an evolutionary tree depicting the evolutionary history of a set of species.

A gene tree is an evolutionary tree depicting the evolutionary history of a gene family from some set of species.

- In general, one expects that the evolution of genes should mimic the evolution of the species themselves. But this is frequently not true.
- Gene trees built on different genes, taken from the same set of species, are often incongruent with one another and with the species tree.

Gene Trees and Species Trees

A species tree is an evolutionary tree depicting the evolutionary history of a set of species.

A gene tree is an evolutionary tree depicting the evolutionary history of a gene family from some set of species.

- In general, one expects that the evolution of genes should mimic the evolution of the species themselves. But this is frequently not true.
- Gene trees built on different genes, taken from the same set of species, are often incongruent with one another and with the species tree.

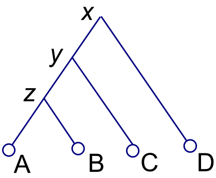
Gene Trees and Species Trees

This incongruence can be explained by several evolutionary phenomena:

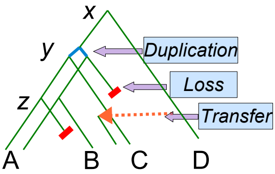
- Gene duplication and loss
- Horizontal gene transfer
- Incomplete lineage sorting
- Hybridization
- Gene conversion

Example: Gene Tree Evolution through D-T-L

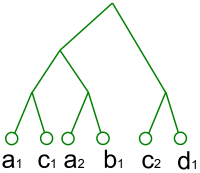
Species Tree *S*



Evolution of Gene Family *G*



Gene Tree on *G*



Impact on Species Tree Inference

This discordance among gene tree topologies confounds species tree inference!

Possible solutions:

- Use a single, or a few, "well-behaved" gene families.
  - Such genes may not exist or may not provide enough resolution.
- Phylogenomics or multi-locus phylogenetics.
  1. Concatenated analysis.
     - Can use only single-copy genes or orthogroups. Averages over discordant phylogenetic signals.
  2. Supertree analysis.
     - Can use only single-copy genes or orthogroups. Not "biology-aware".
  3. Coalescence-based methods.
     - Can use only single-copy genes or orthogroups. Often require universal genes. Assume discordance is due to ILS.
  4. Gene Tree Parsimony.
     - Can use all gene families. Generally assumes discordance is due to gene duplication and loss. Fairly robust to ILS.

Phylogenomics Through Gene Tree Parsimony

Given a gene tree and a species tree, the smallest number of evolutionary events that can explain their discordance is called their reconciliation cost.

Gene Tree Parsimony (GTP): Find a species tree that minimizes the reconciliation cost.

- Used mostly in the context of gene duplication and loss.
- The only way to deal cleanly with multi-copy gene families or MUL trees.
- Can construct truly genome-scale phylogenies.

Phylogenomics Through Gene Tree Parsimony

Given a gene tree and a species tree, the smallest number of evolutionary events that can explain their discordance is called their reconciliation cost.

Gene Tree Parsimony (GTP): Find a species tree that minimizes the reconciliation cost.

- Used mostly in the context of gene duplication and loss.
- The only way to deal cleanly with multi-copy gene families or MUL trees.
- Can construct truly genome-scale phylogenies.

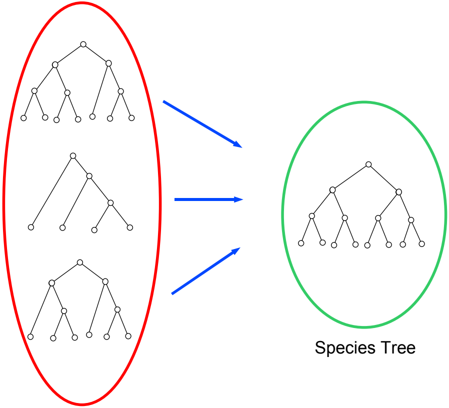
Phylogenomics Through Gene Tree Parsimony

Given a gene tree and a species tree, the smallest number of evolutionary events that can explain their discordance is called their reconciliation cost.

Gene Tree Parsimony (GTP): Find a species tree that minimizes the reconciliation cost.

- Used mostly in the context of gene duplication and loss.
- The only way to deal cleanly with multi-copy gene families or MUL trees.
- Can construct truly genome-scale phylogenies.

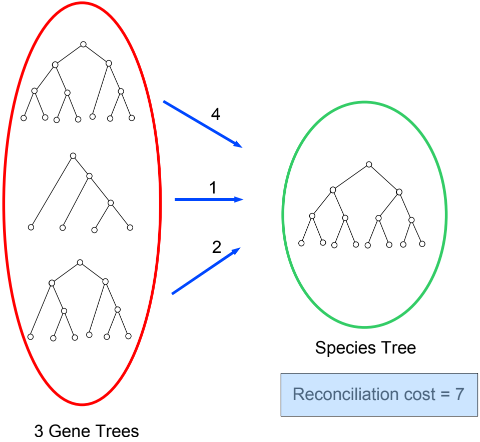
GTP under the Duplication-Loss Model


3 Gene Trees

Species Tree

* Goodman et al. (1979); Page (1994); Guigó et al. (1996); ...

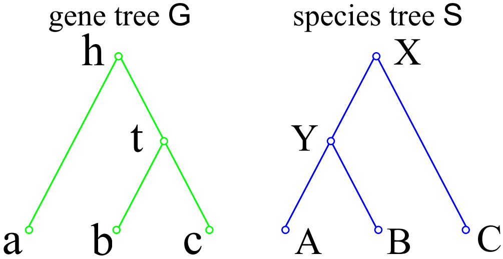
GTP under the Duplication-Loss Model



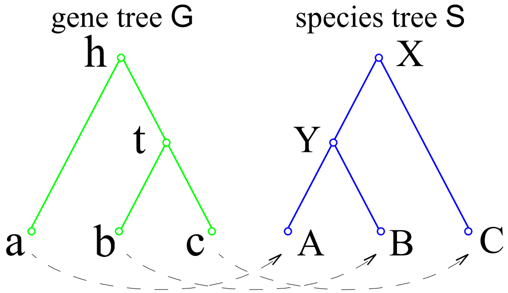
Reconciliation cost = 7

* Goodman et al. (1979); Page (1994); Guigó et al. (1996); ...

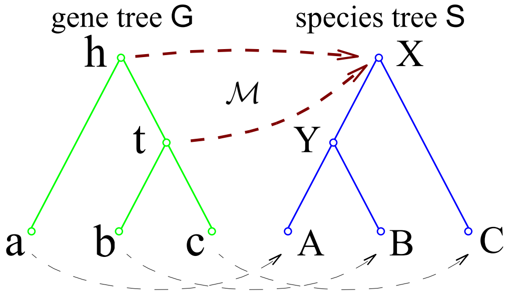
Computing the Reconciliation Cost



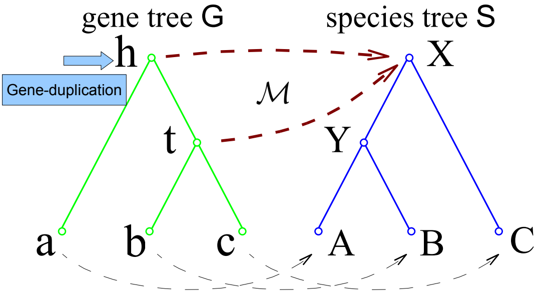
Computing the Reconciliation Cost



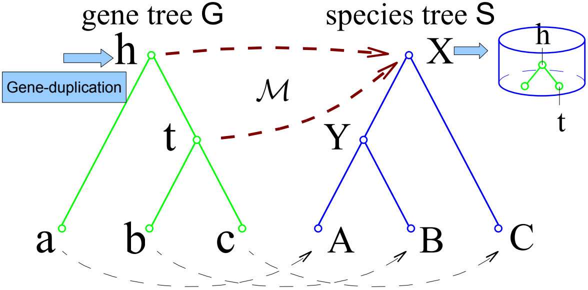
Computing the Reconciliation Cost



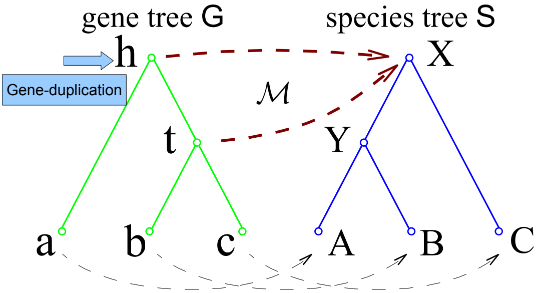
Computing the Reconciliation Cost



Computing the Reconciliation Cost

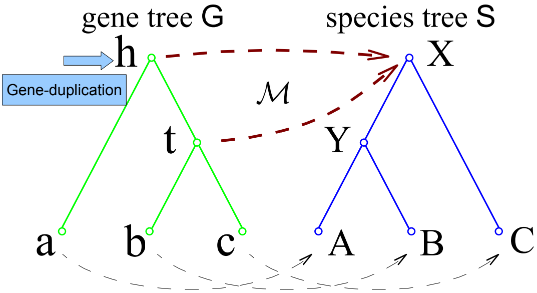


Computing the Reconciliation Cost



Duplication cost = 1

Computing the Reconciliation Cost



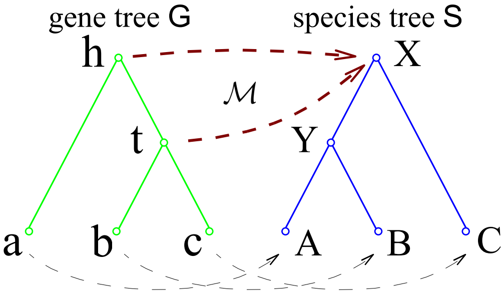
Gene-duplication

Duplication cost = 1

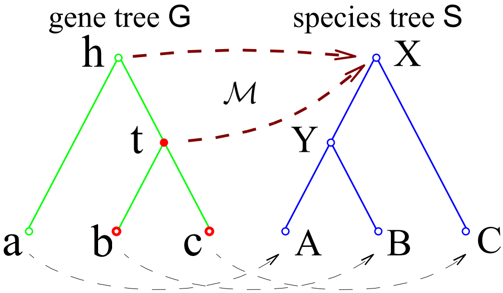
$Time\,Complexity : O\,(n)$

Complexity result by: Zhang (1997)

Computing the Reconciliation Cost

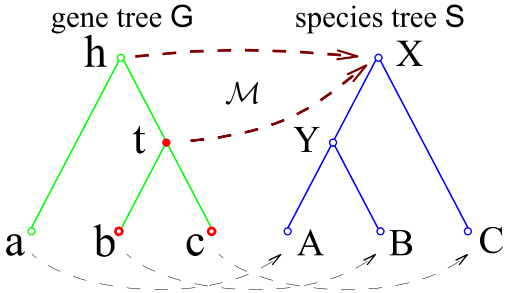


Computing the Reconciliation Cost



$$Loss(x) = \begin{cases} 0, \text{ if } \mathcal{M}(x) = \mathcal{M}(x') = \mathcal{M}(x'') \\ |d(\mathcal{M}(x), \mathcal{M}(x')) - 1| + |d(\mathcal{M}(x), \mathcal{M}(x'')) - 1|, \text{ otherwise} \end{cases}$$

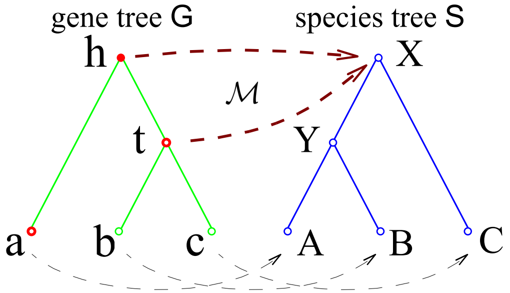
Computing the Reconciliation Cost



loss(t) = 1

$$Loss(x) = \begin{cases} 0, \text{ if } \mathcal{M}(x) = \mathcal{M}(x') = \mathcal{M}(x'') \\ |d(\mathcal{M}(x), \mathcal{M}(x')) - 1| + |d(\mathcal{M}(x), \mathcal{M}(x'')) - 1|, \text{ otherwise} \end{cases}$$

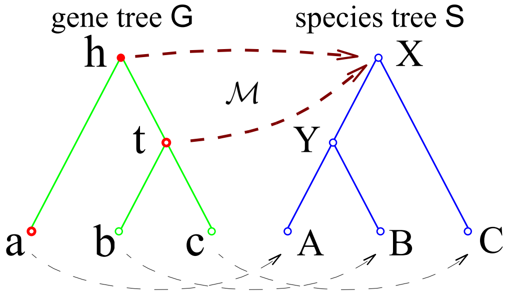
Computing the Reconciliation Cost



loss(t) = 1

$$Loss(x) = \begin{cases} 0, \text{ if } \mathcal{M}(x) = \mathcal{M}(x') = \mathcal{M}(x'') \\ |d(\mathcal{M}(x), \mathcal{M}(x')) - 1| + |d(\mathcal{M}(x), \mathcal{M}(x'')) - 1|, \text{ otherwise} \end{cases}$$

Computing the Reconciliation Cost

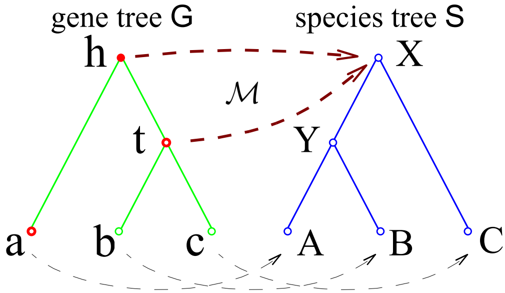


loss(t) = 1

loss(h) = 2

$$Loss(x) = \begin{cases} 0, \text{ if } \mathcal{M}(x) = \mathcal{M}(x') = \mathcal{M}(x'') \\ |d(\mathcal{M}(x), \mathcal{M}(x')) - 1| + |d(\mathcal{M}(x), \mathcal{M}(x'')) - 1|, \text{ otherwise} \end{cases}$$

Computing the Reconciliation Cost

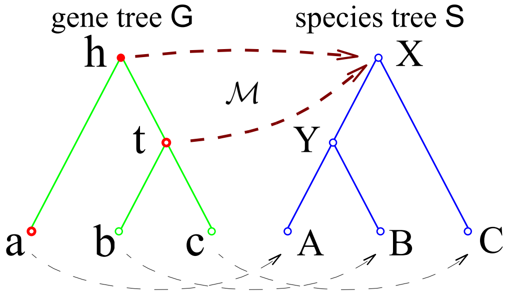


Loss Cost = 3

loss(t) = 1

loss(h) = 2

Computing the Reconciliation Cost



Loss Cost = 3

loss(t) = 1

loss(h) = 2

$Time\,Complexity : O\,(n)$

Duplication-Loss GTP Problem

Given

A set of gene trees.

Find

A species tree with minimum reconciliation cost.

Duplication-Loss GTP Problem

Given

A set of gene trees.

Find

A species tree with minimum reconciliation cost.

DupTree, DupLoss, and iGTP

Most popular software packages for GTP:

1. DupTree: Only counts gene duplications
   - Better if gene families could be incomplete (e.g., due to incomplete or patchy genome sequencing)
2. DupLoss: Counts both gene duplications and losses
   - Better if gene families are relatively complete.
3. iGTP: Graphical interface around DupTree and DupLoss (and DeepC)

GTP under Duplication-Loss is computationally hard (Bin Ma et al., 1998).

- DupTree and DupLoss use local search heuristics.

DupTree, DupLoss, and iGTP

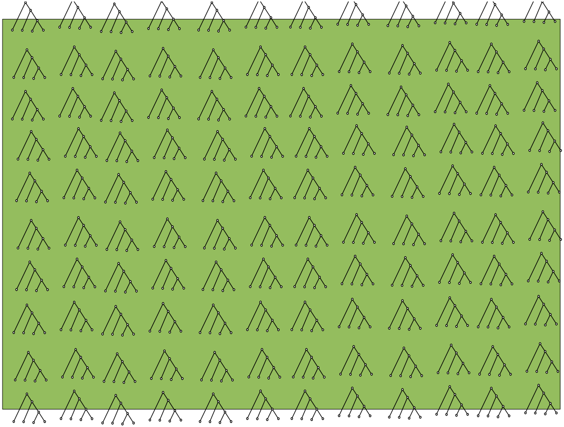
Most popular software packages for GTP:

1. DupTree: Only counts gene duplications
   - Better if gene families could be incomplete (e.g., due to incomplete or patchy genome sequencing)
2. DupLoss: Counts both gene duplications and losses
   - Better if gene families are relatively complete.
3. iGTP: Graphical interface around DupTree and DupLoss (and DeepC)

GTP under Duplication-Loss is computationally hard (Bin Ma et al., 1998).

- DupTree and DupLoss use local search heuristics.

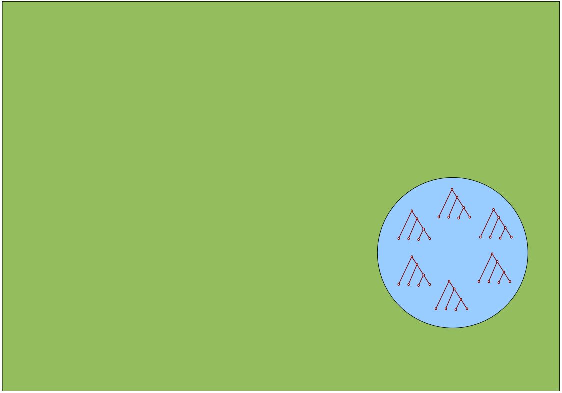
Local Search Heuristics



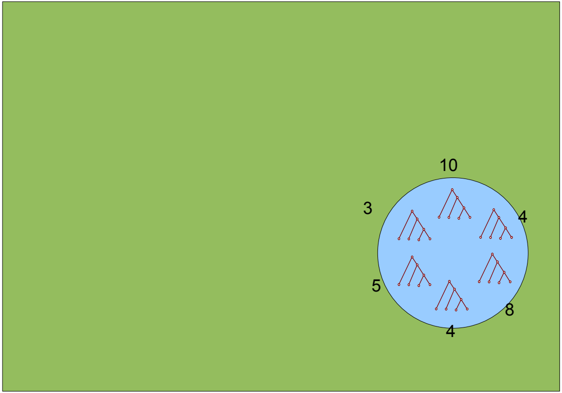
Local Search Heuristics

Local Search Heuristics

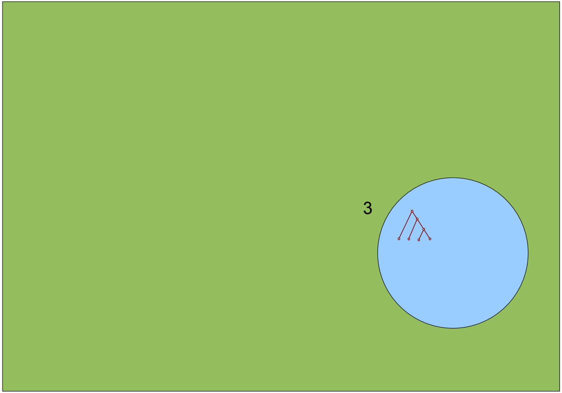
Local Search Heuristics



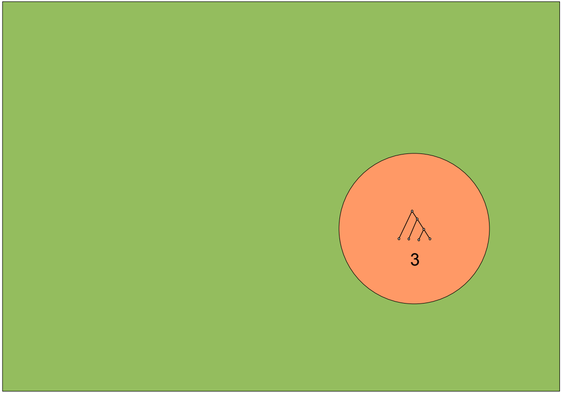
Local Search Heuristics



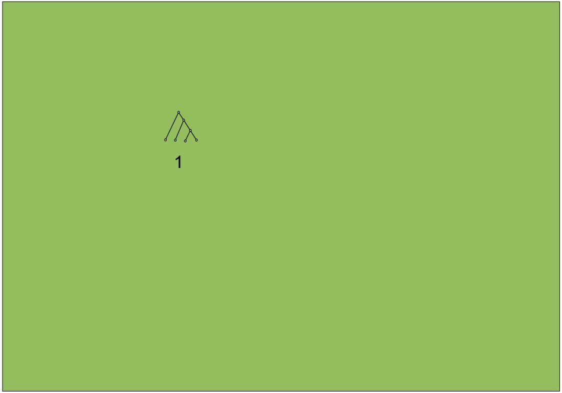
Local Search Heuristics



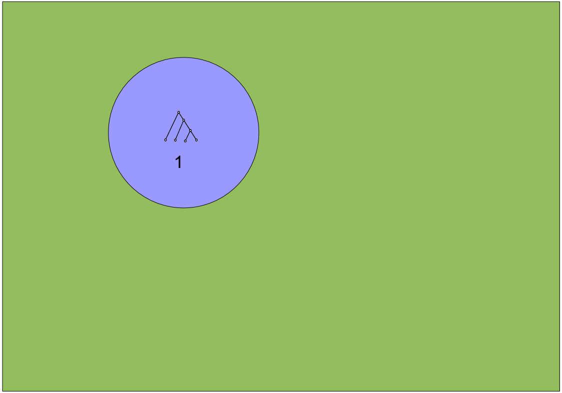
Local Search Heuristics



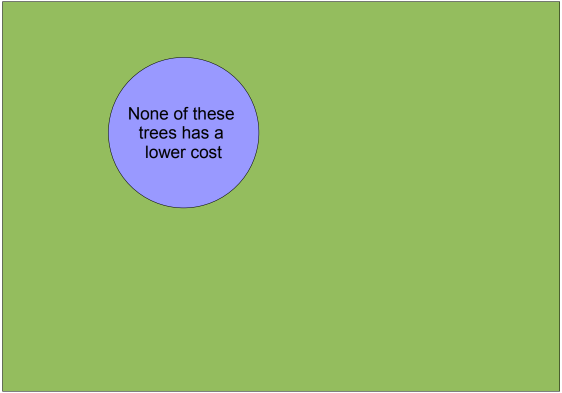
Local Search Heuristics



Local Search Heuristics

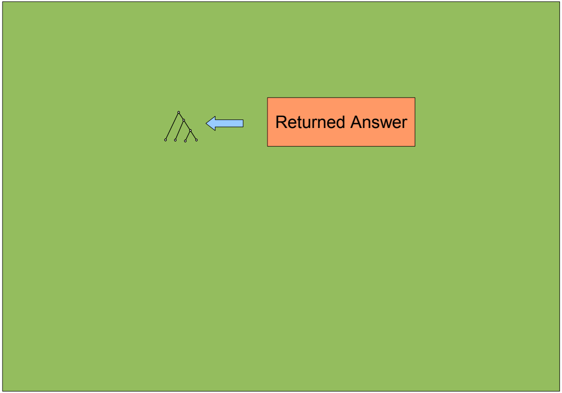


Local Search Heuristics



None of these trees has a lower cost

Local Search Heuristics



Returned Answer

DupTree/DupLoss/iGTP in practice

- Scalable to thousands of gene trees from many hundreds of taxa (tested up to 2000 taxa)
- Runtime in seconds for smaller datasets to a few days for larger ones.
- Used for multiple ground breaking applications:
  - 136 species plant phylogeny using 18,896 gene trees (Burleigh et al., SB 2011)
  - Melon genome sequence analysis (Garcia-Mas et al., PNAS 2012)
  - Rooting Eukaryotic Tree of Life (Katz et al., SB 2012)
  - Sugar beet genome sequence analysis (Dohm et al., Nature 2013)
  - King cobra genome and venom evolution (Vonk et al., PNAS 2013)
  - Crocodilian evolution (Green et al., Science 2014)

Using DupTree/DupLoss/iGTP

To build a genome-scale phylogeny for some set of species:

1. Cluster all gene from these species into gene families or homologous groups. Or use existing databases (NCBI Homologene, Ensembl Compara).
2. Align sequences and construct gene tree for each gene family (e.g., using RAxML, FastTree)
3. Root gene trees (known root or best guess, e.g., midpoint rooting).
4. Parse gene trees to replace each leaf label with name of species it came from.
5. Prepare single file containing all newick formatted gene trees.
6. Use DupTree/DupLoss/iGTP to construct species tree.

Best practices

- Filter out really bad gene trees.
  - E.g., gene trees with less then 30% average bootstrap support.
- Use “unrooted” gene tree option if not sure about gene tree roots.
  - Append [&U] tag before each gene tree.
- Run DupTree/DupLoss/iGTP multiple times and choose best result(s).
  - Generally about 10 runs should be sufficient.
  - Check if converging to almost the same score and tree. item Can take strict consensus of distinct optimal trees.
- Estimate support for species tree using gene tree bootstrapping.
  - Repeat analysis with different samples of gene tree bootstrap replicates and measure branch support across different runs.

Coming soon ...

- Improved search heuristics.
- Automation of some best practices.
- Dealing with uncertainty in gene trees.
- More inclusive reconciliation models for broader applicability.

Practice Exercise

Workshop page:

`https://compbio.engr.uconn.edu/software/desertworkshop/`

- These slides.
- Toy data set.
- Links to DupTree and iGTP download pages.
- Executables for DupLoss.

DupTree Commands

1.
   ```
   ./duptree --fast -i vertebrates.newick
   ```

2.
   ```
   ./duptree --fast -i vertebratesUNR.newick
   ```

3.
   ```
   ./duptree --fast --nogenetree -i vertebrates.newick -o outputfile.out
   ```

Repeat with different seeds:

1.
   ```
   ./duptree --fast --nogenetree --seed 1 -i vertebrates.newick -o outputfile1.out
   ```

2.
   ```
   ./duptree --fast --nogenetree --seed 2 -i vertebrates.newick -o outputfile2.out
   ```

DupLoss Commands

1.
   ```
   ./DupLoss --fast -i vertebrates.newick
   ```
2.
   ```
   ./DupLoss --fast -i vertebratesUNR.newick
   ```
3.
   ```
   ./DupLoss --fast --nogenetree -i vertebrates.newick -o outputfile.out
   ```

Repeat with different seeds:

1.
   ```
   ./DupLoss --fast --nogenetree --seed 1 -i vertebrates.newick -o outputfile1.out
   ```
2.
   ```
   ./DupLoss --fast --nogenetree --seed 2 -i vertebrates.newick -o outputfile2.out
   ```

Using iGTP

- Run multi-replicate analysis on the same sets of input trees.
- Visualize species trees (using Mike's tree display software Paloverde).