Deciphering Reticulate Evolution Using Phylogenetic Reconciliation

Mukul S. Bansal

Department of Computer Science and Engineering,
University of Connecticut,
USA

The Phylogenetic Networks Workshop
July 27, 2015

Supertrees or Supernetworks?

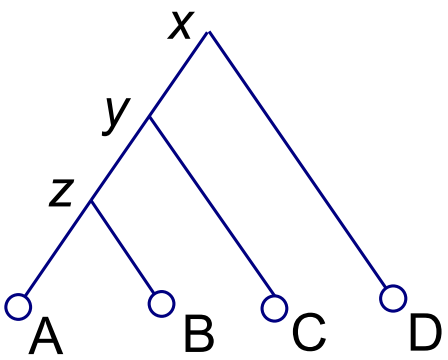
Gene Family Evolution

Problem

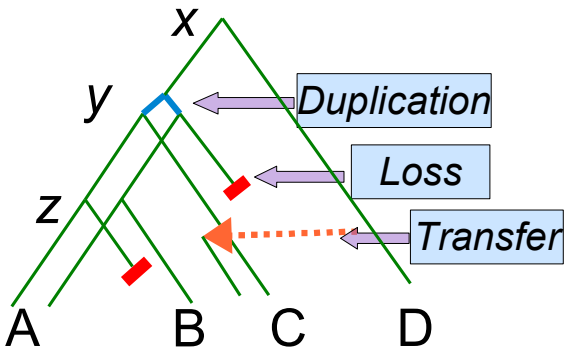
How did any given gene family evolve?

- Gene families evolve inside species trees.
- Affected by evolutionary events such as gene duplication, horizontal gene transfer, and gene loss.

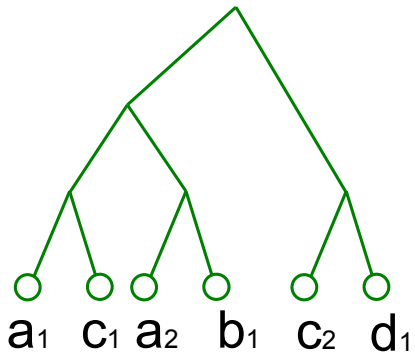
Species Tree *S*



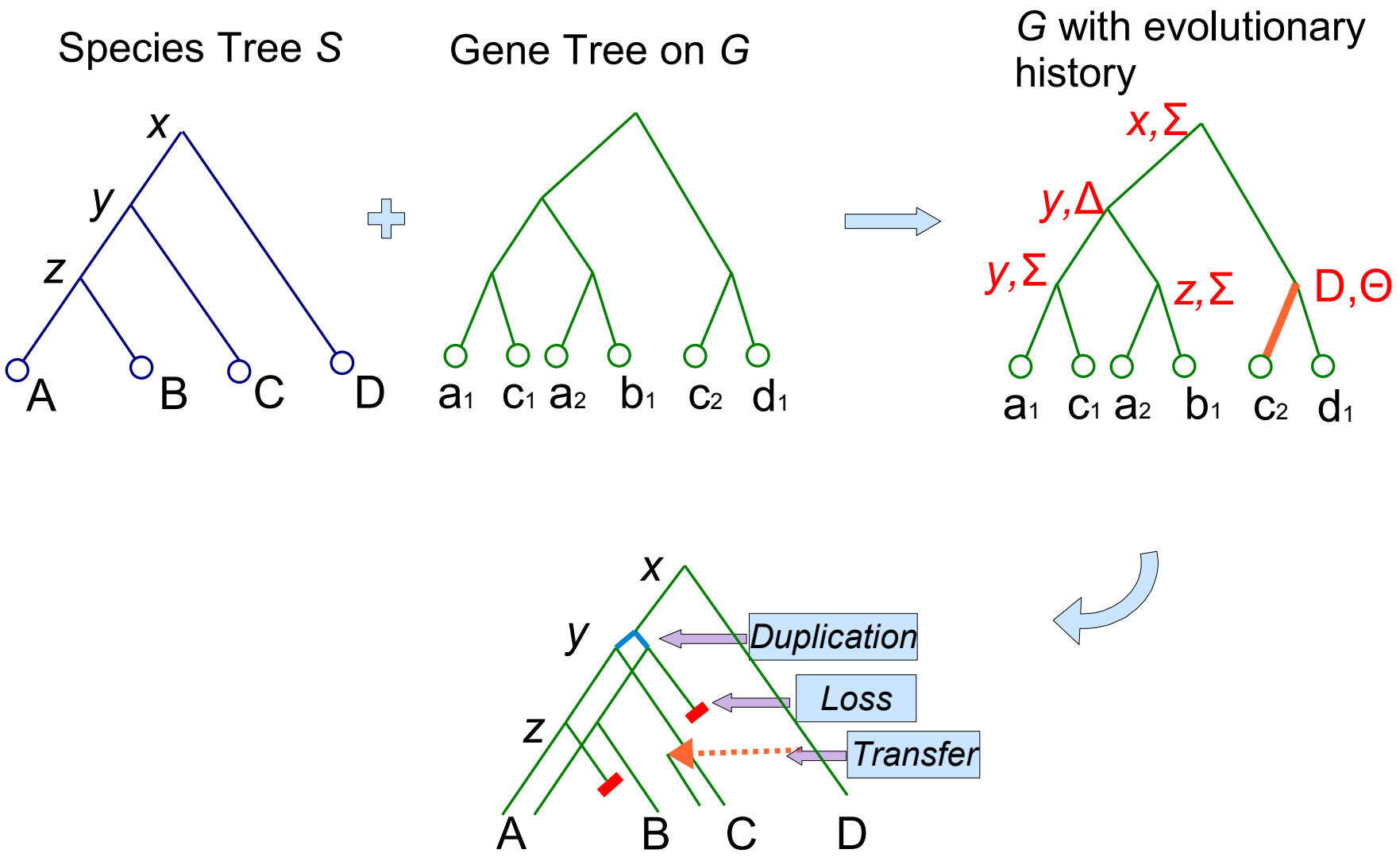
Evolution of Gene Family *G*



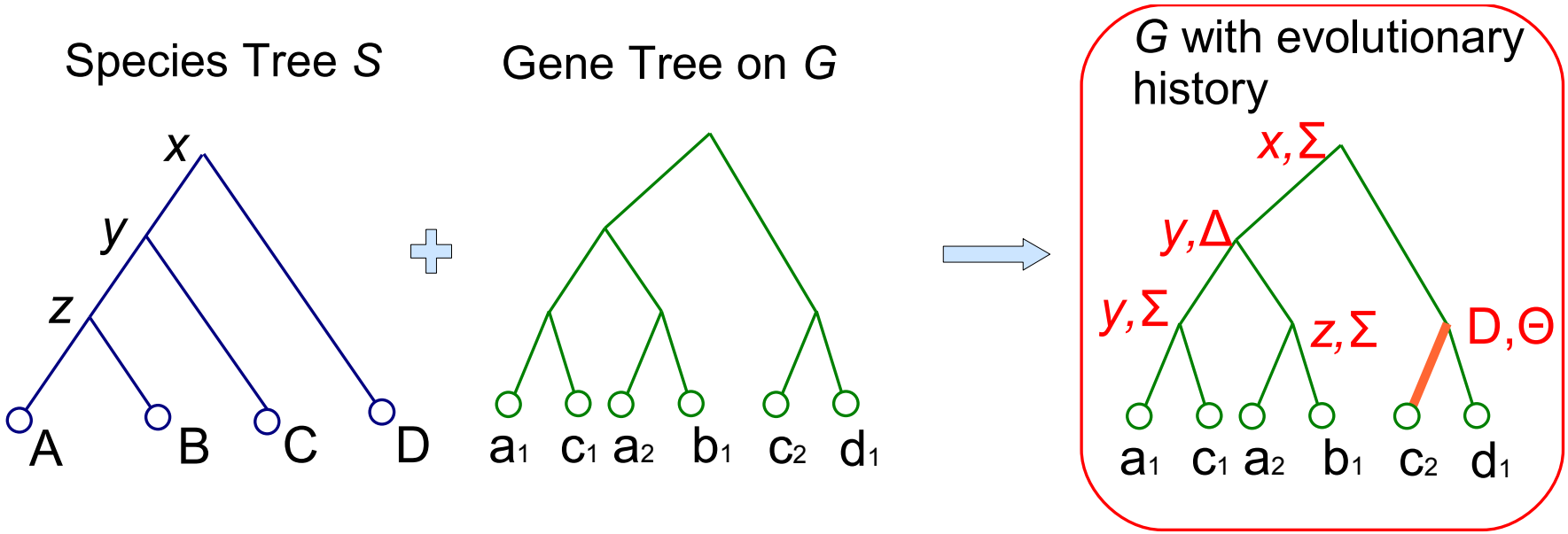
Gene Tree on *G*



Definition: DTL Reconciliation



Definition: DTL Reconciliation



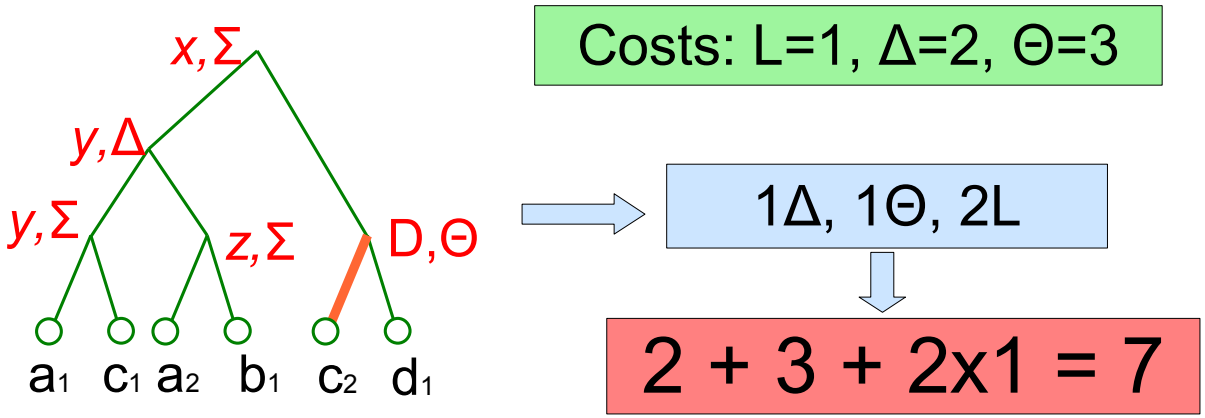
Input: A gene tree for that gene family, and a trusted rooted species tree.

Output: An evolutionary history of that gene family showing horizontal gene transfers, gene duplications, losses, and speciation events.

DTL Reconciliation Problem Formulation

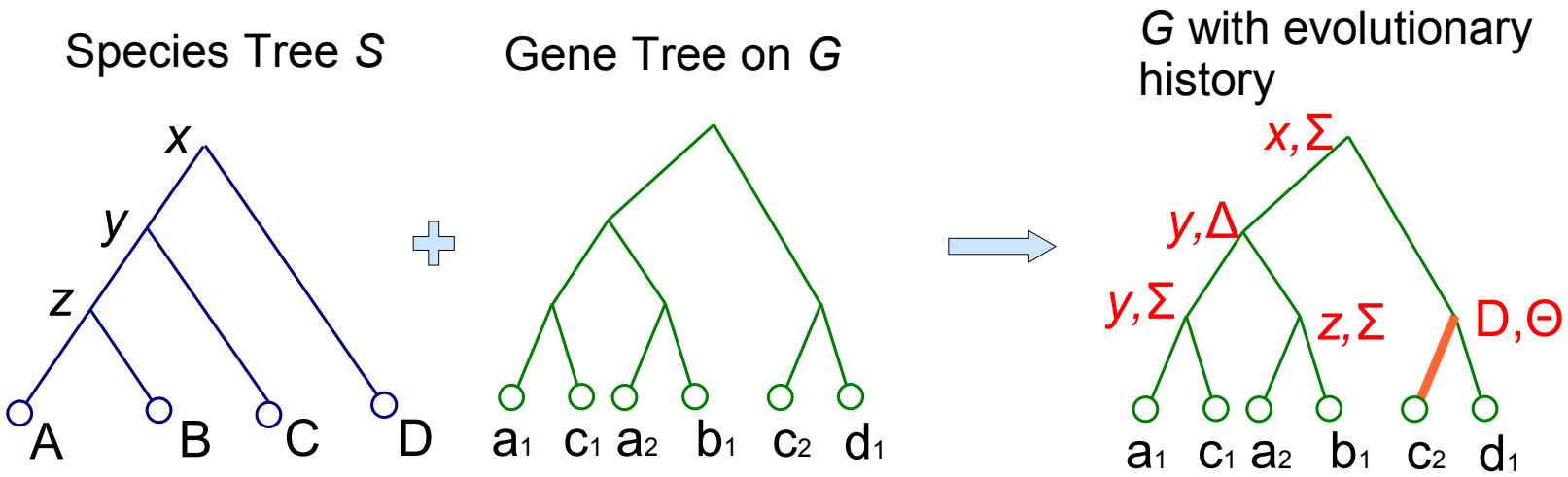
Parsimony formulation:

- Costs are assigned to duplications, transfers, and losses.
- Goal: Find the reconciliation that minimizes the total cost.
- Easy to compute cost for a given reconciliation.



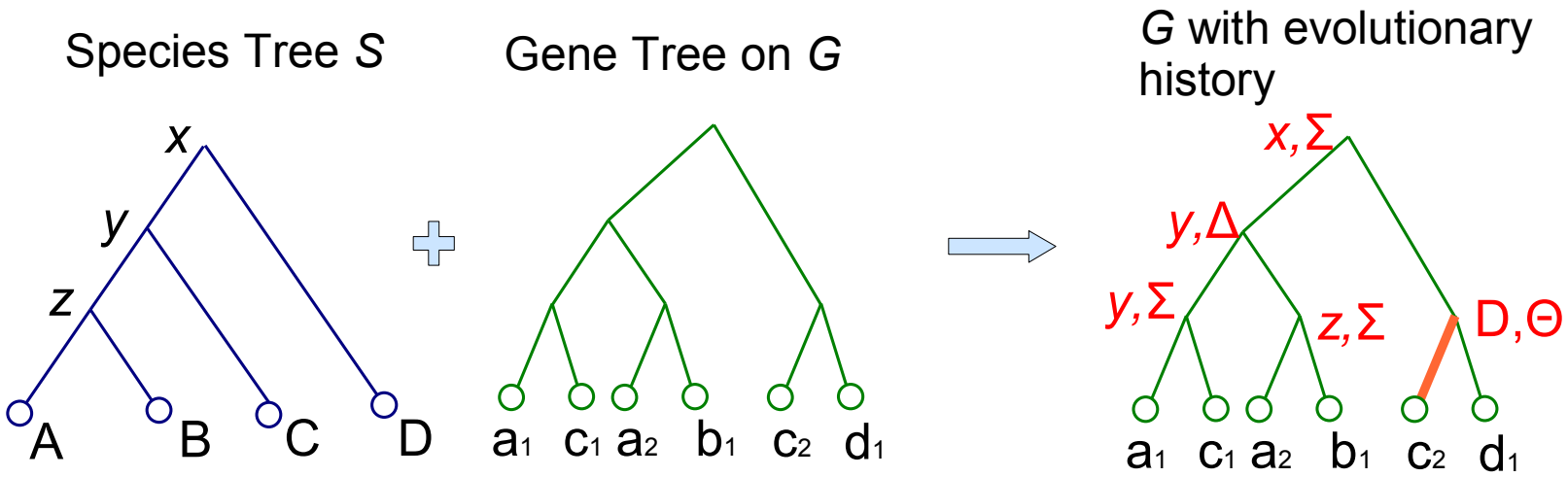
Different reconciliation could have different cost.

Applications of DTL Reconciliation



- Understanding how gene families evolve.
- Dating gene birth.
- Inferring orthologs/paralogs/xenologs.
- Gene tree error-correction.
- Whole genome species tree construction.
- Constructing species phylogenetic networks.

Applications of DTL Reconciliation



- Understanding how gene families evolve.
- Dating gene birth.
- Inferring orthologs/paralogs/xenologs.
- Gene tree error-correction.
- Whole genome species tree construction.
- Constructing species phylogenetic networks.

Background: Computing Optimal DTL Reconciliations

For undated species tree:

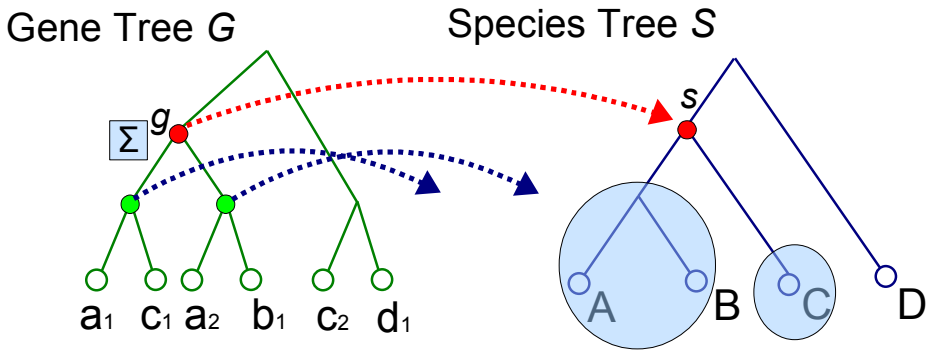
- Best time-consistent reconciliation: NP-hard (Tofigh et al., 2011; Ovadia et al., 2011)
- If time-consistency not enforced: $O(mn)$-time algorithm (Bansal et al., 2012)

For dated species trees:

- Best time-consistent reconciliation: $O(mn^2)$-time algorithm (Doyon et al., 2010)

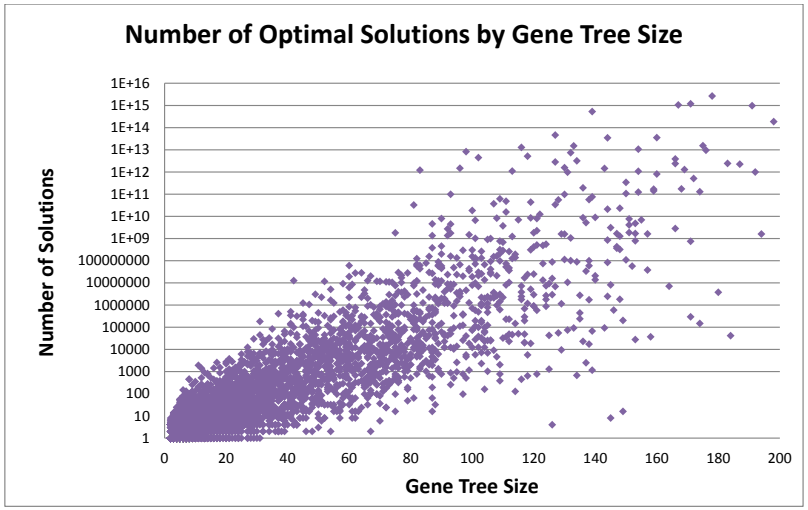
DTLI model:

- Considers ILS at unresolved species tree nodes (Stolzer et al., 2012)



Background: Handling Multiple Optima

- There can be many optimal DTL reconciliations.

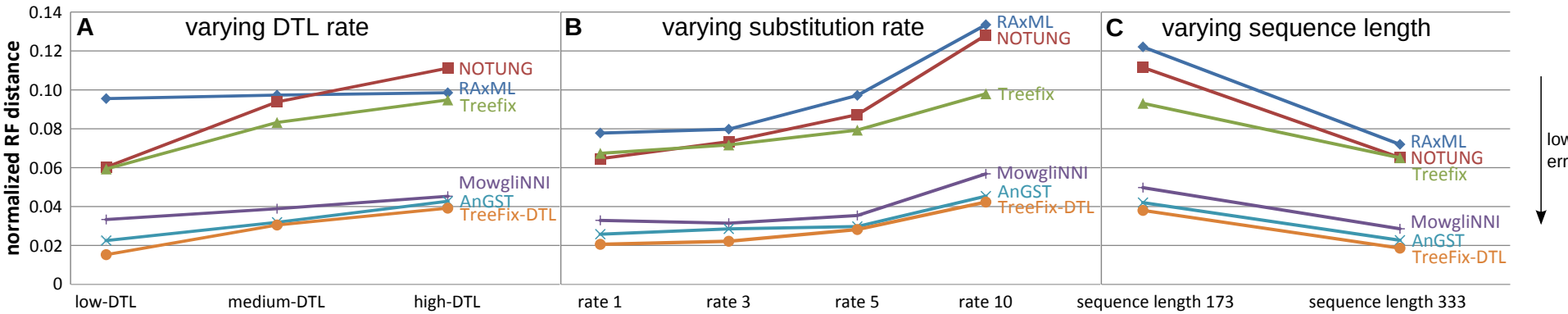


- Enumeration: Exponential in input size (Tofigh. et al., 2011; Chen et al., 2012)
- Uniform random sampling and aggregation: $O(mn^2)$-time (Bansal et al., 2013)
- Compact representation in reconciliation graph: $O(mn^3)$-time (Scornavacca et al., 2013)
- Median reconciliation (Nguyen et al., 2013)

Background: Gene Tree Error-Correction

Gene tree construction is highly error-prone. Garbage in, garbage out.

- First-attempts: Mowgli-NNI (Nguyen et al., 2012), AnGST (David and Alm, 2011)
- New methods: TreeFix-DTL (Bansal et al., 2015), ALE (szollosi et al. 2013)), TERA (Scornavacca et al., 2015))

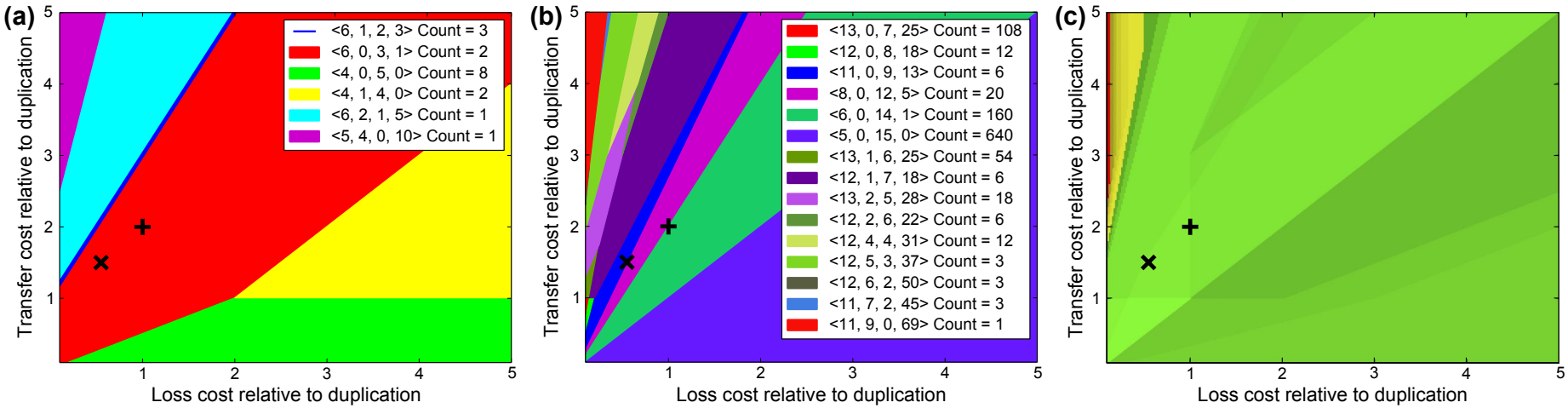


Background: Event Cost Assignment

Fundamental questions:

- What are the "best" event costs to use?
- How do reconciliations vary as we change event costs?

Algorithm to partition event cost space into equivalence regions based on pareto-optimality of event counts: $O(m^5 n \log m)$-time (Libeskind-Hadas et al., 2014)



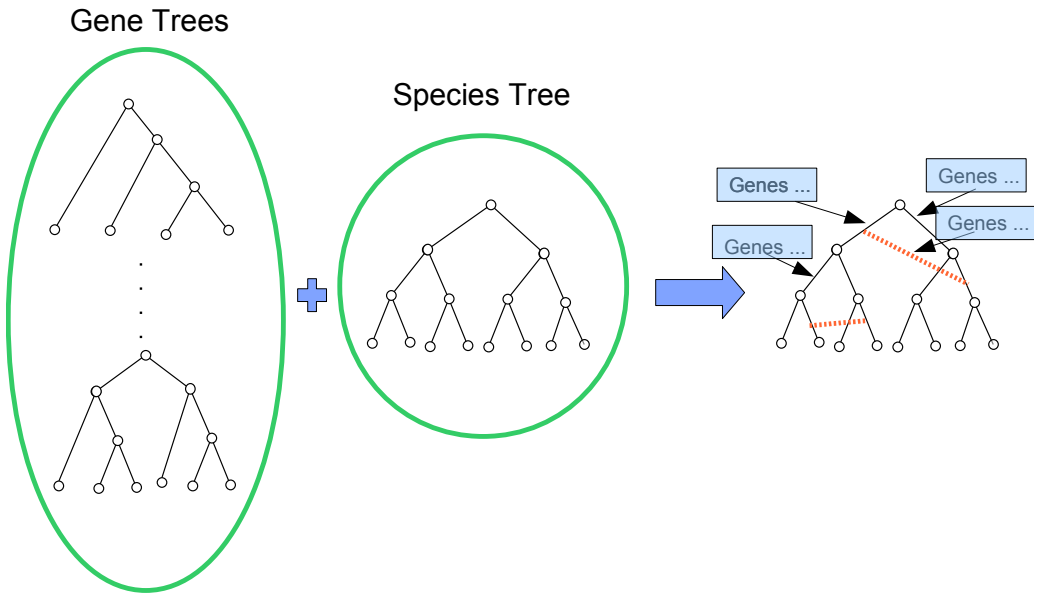
Constructing Phylogenetic Networks

- Microbial phylogenetic networks are distinct from hybridization networks.

Problem formulation

Input: A collection of gene families (sequence alignments) and a reference species tree.

Output: Species tree augmented with horizontal edges (representing transfer events), and labels for each vertical and horizontal edge specifying the genes that traveled through it.



Constructing Phylogenetic Networks

Advantages of using a reference species tree:

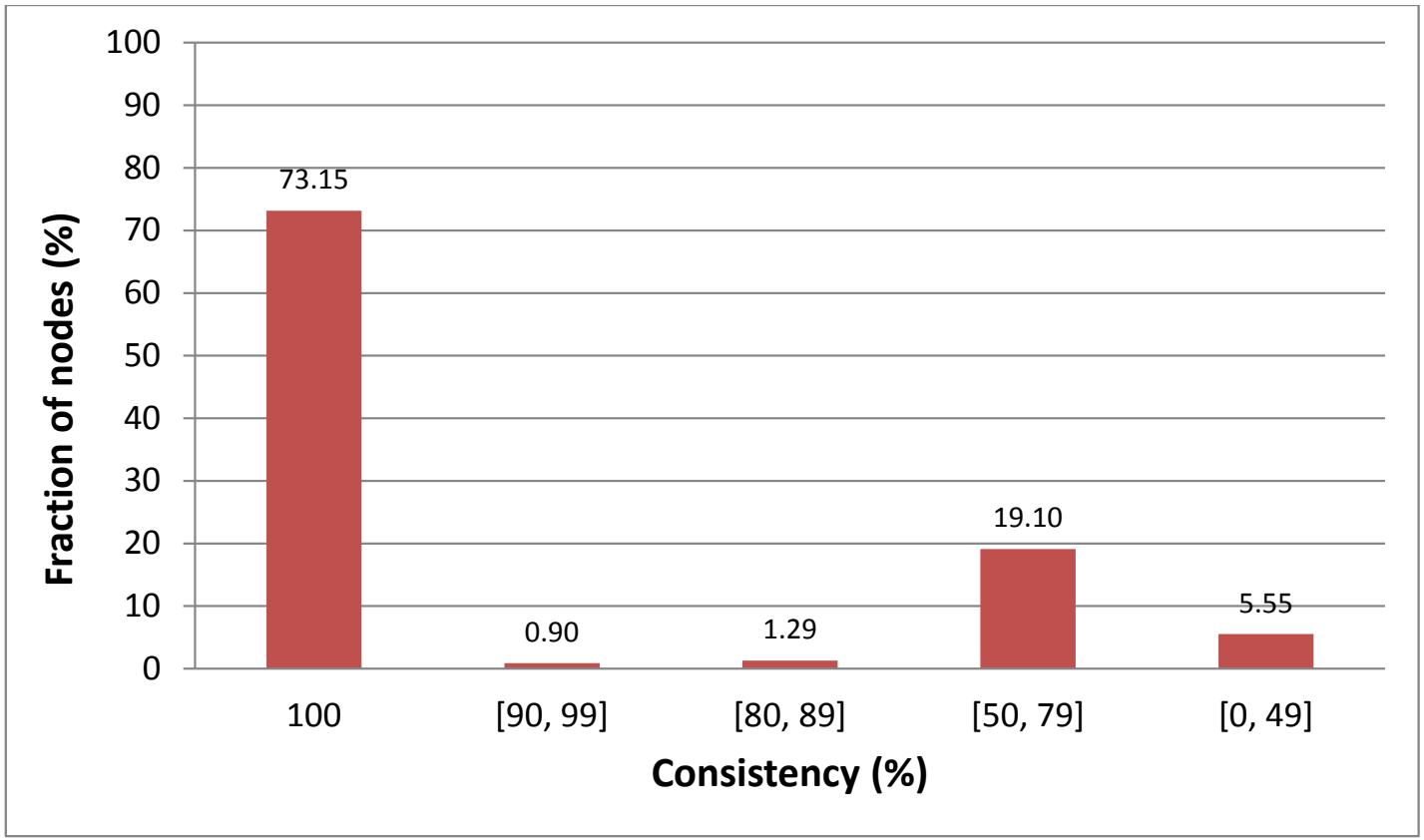
- Improved complexity and scalability.
- Inferred network does not depend on reference tree.
- Customizable and easily interpretable network view.

Basic implementation:

1. Infer gene trees.
2. Use DTL reconciliation to reconcile individual gene trees with reference tree.
3. Aggregate transfers inferred for gene trees onto species tree; e.g., NOTUNG.

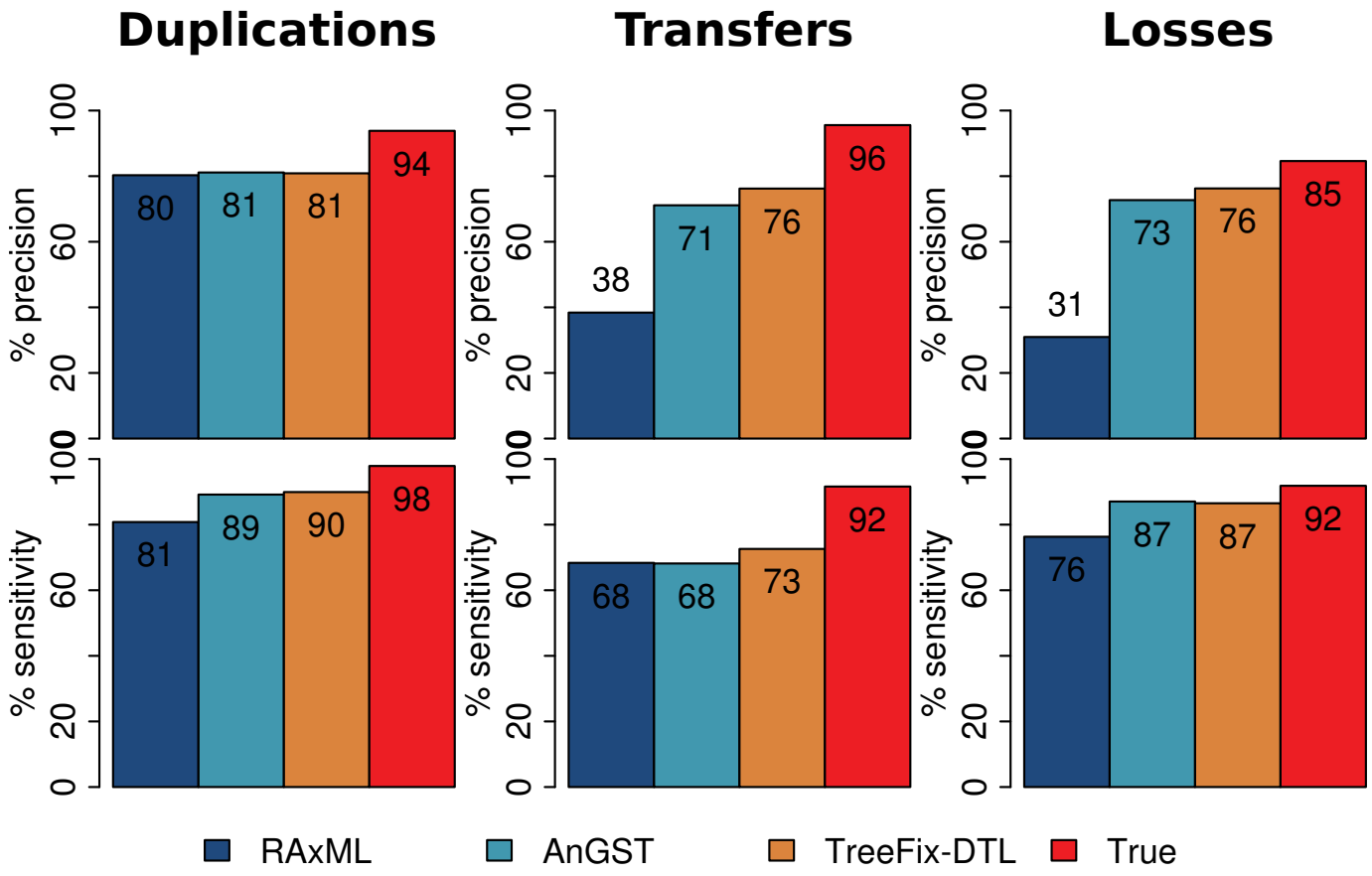
Challenges due to reconciliation uncertainty

- Multiple optima, gene tree error, and event cost assignment confound reconciliation accuracy.



Challenges due to reconciliation uncertainty

- Multiple optima, gene tree error, and event cost assignment confound reconciliation accuracy.



Possible Solution

Proposal: Distinguish between highly-supported and weakly-supported events across multiple optima, multiple event costs, and gene tree topologies.

- Easy to do for multiple optima and event costs based on developed algorithms.
- No such algorithms for gene tree error.

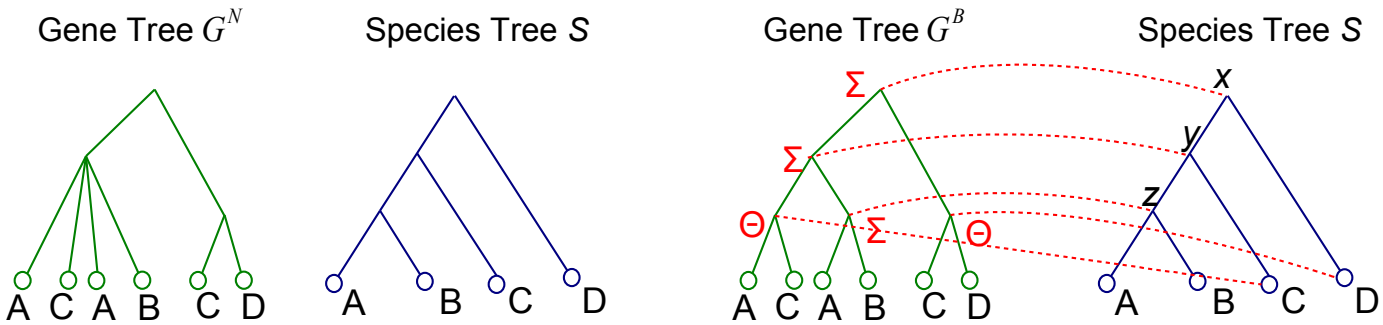
Goal: Develop algorithms to generate alternative gene tree topologies and study variability of reconciliation across them.

Optimal Gene Tree Resolution (OGTR)

Input: A non-binary gene tree G^N, a species tree S, and event costs.

Output: Find a binary resolution G^B of G^N such that, the most parsimonious DTL reconciliation of G^B and S has smallest reconciliation cost.

- Provides rigorous framework for uniformly sampling from *all* candidate gene trees.
- Non-binary gene tree obtained by collapsing weak edges.
- Fundamental question for DTL reconciliation.



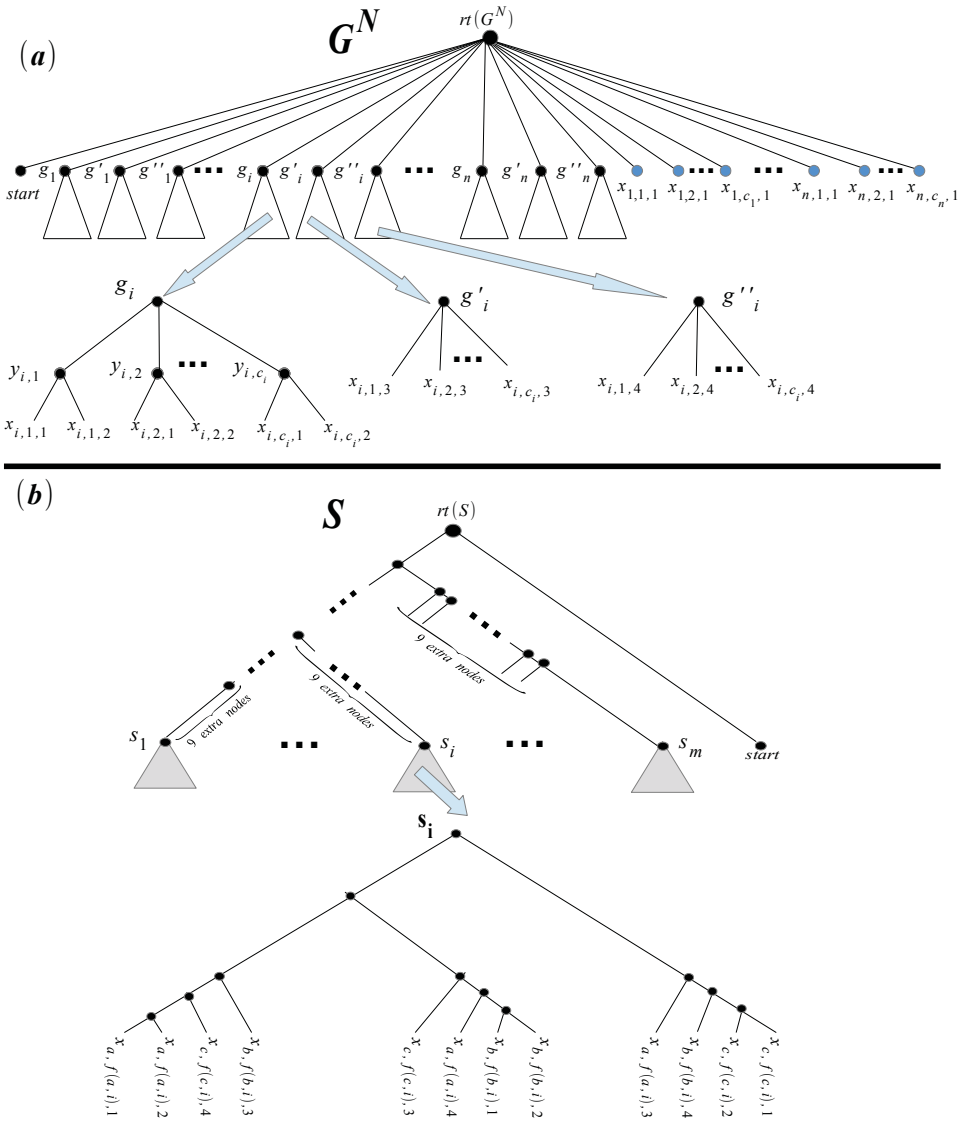
OGTR is NP-hard

- OGTR is NP-hard for both undated and dated species trees (Kordi and Bansal, 2015).
- Surprising since problem is linear-time solvable under duplication-loss model (Zheng and Zhang, 2014).

Key Ideas:

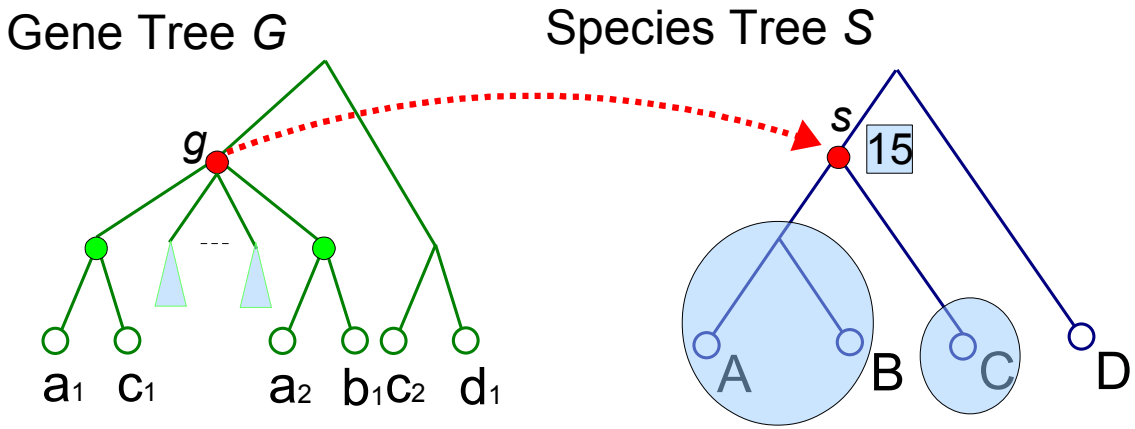
- Reduction from minimum 3 set cover problem.
- Subtrees in species tree correspond to sets
- Subtrees in unresolved gene tree correspond to elements of the universe.
- Structure of gadget forces root of each subtree (element) to map to a "set" in which that element occurs.
- Using more sets for the mapping results in higher reconciliation cost.

OGTR is NP-Hard

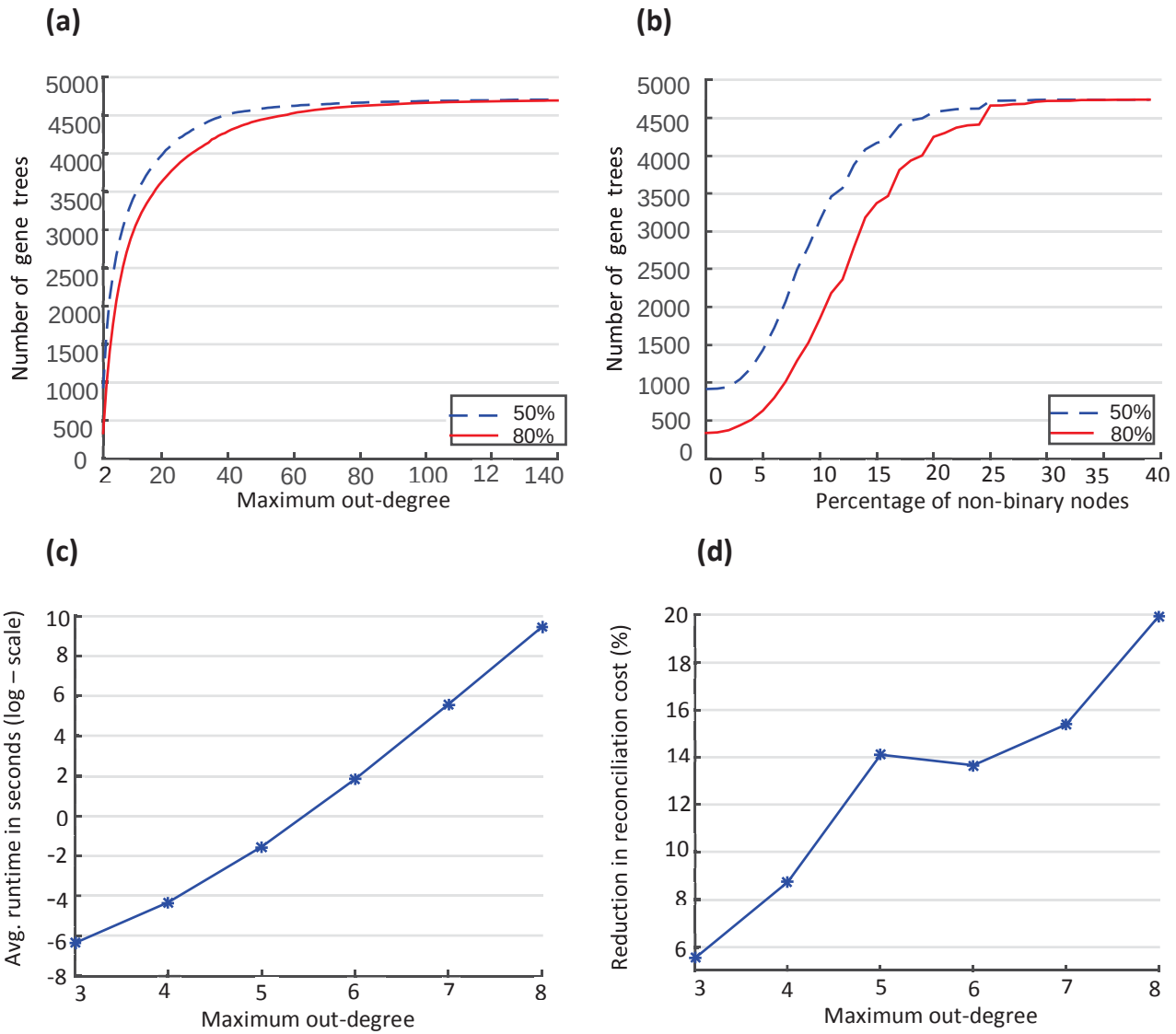


A Fixed-Parameter Algorithm for OGTR

- DP algorithms for binary gene trees can be extended to work with non-binary gene trees.
- Consider all possible resolutions at each non-binary gene tree node and fill DP-table for subproblem with best score over all resolutions.
- Gives $O(2^k \cdot k! \cdot ln + mn)$ and $O(2^k \cdot k! \cdot ln + mn^2)$-time algorithms for dated and undated species trees, respectively



A Fixed-Parameter Algorithm for OGTR



Uniform Sampling of Optimal Resolutions

- DP framework allows for sampling optimal resolutions uniformly at random.
- Samples can be combined with event cost and multiple optima analyses to assess robustness of individual events and mappings.
- Enables identification of highly- and weakly-supported events and mappings for network construction.

Future Directions: Assigning Weakly-Supported Events

1. Leverage high-support events to improve assignment of all other events:
   - Use highly-supported events to infer highways of transfers.
   - Improve assignment of other events based on inferred highways.
2. Use gene tree and species tree branch lengths to choose between alternative scenarios for low-support events.

Thank You!

Questions!