Gene Tree Parsimony for Incomplete Gene Trees

Md. Shamsuzzoha Bayzid and Tandy Warnow

Bangladesh University of Engineering and Technology

ILLINOIS

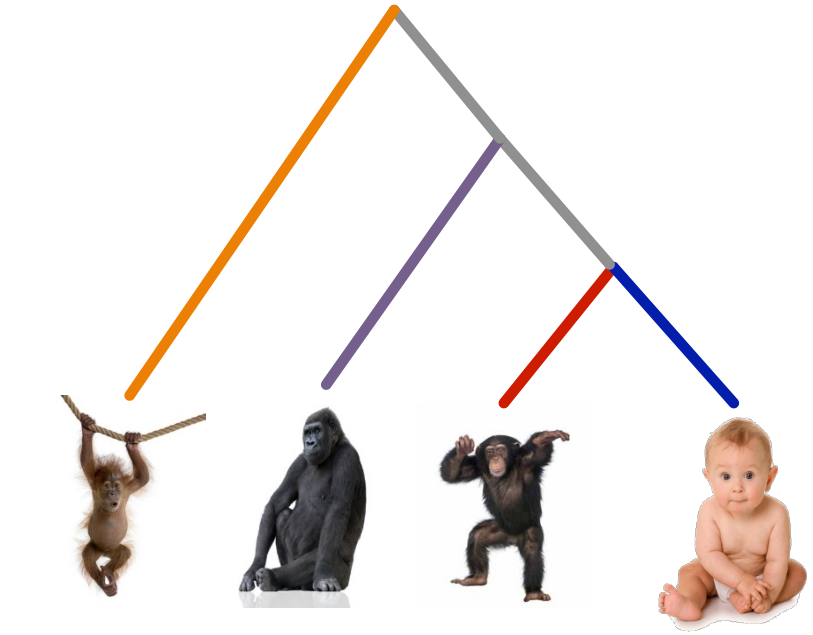
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Outline

- Background
  - Gene trees and species trees
  - Species tree estimation techniques
- GTP for Incomplete gene trees
  - Summary of our contributions
  - Descriptions of our algorithms
- Conclusion

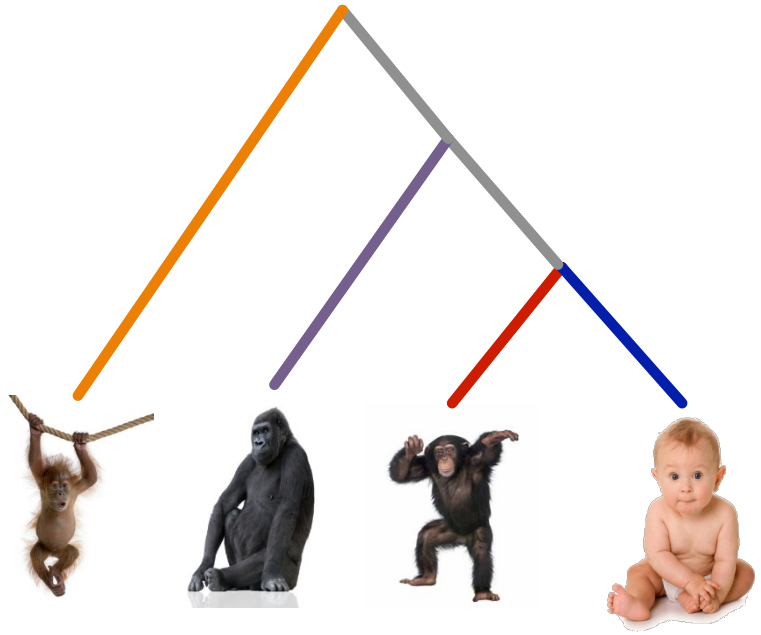
Species tree

- represents the evolutionary history of a group of organisms.

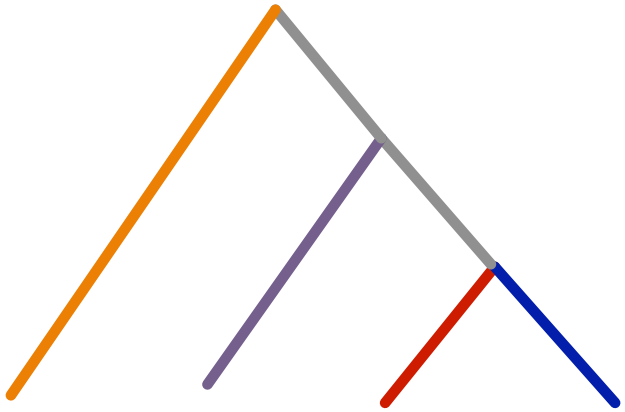


Gene trees and species tree

- Species tree – Pattern of branching of species lineages via speciation.
- Gene tree – A phylogenetic tree that depicts how a *single* gene has evolved in a group of related species.



Orangutan Gorilla Chimpanzee Human

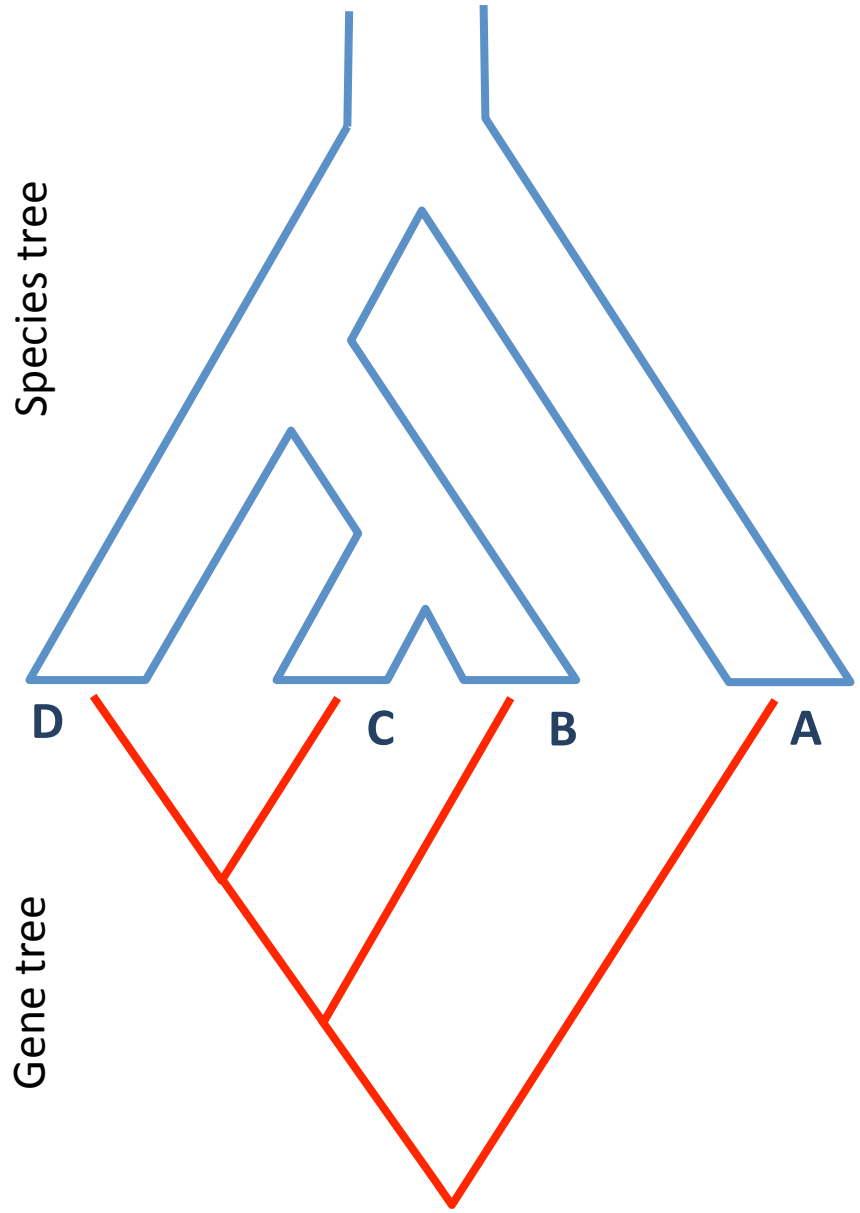


Hemoglobin

@

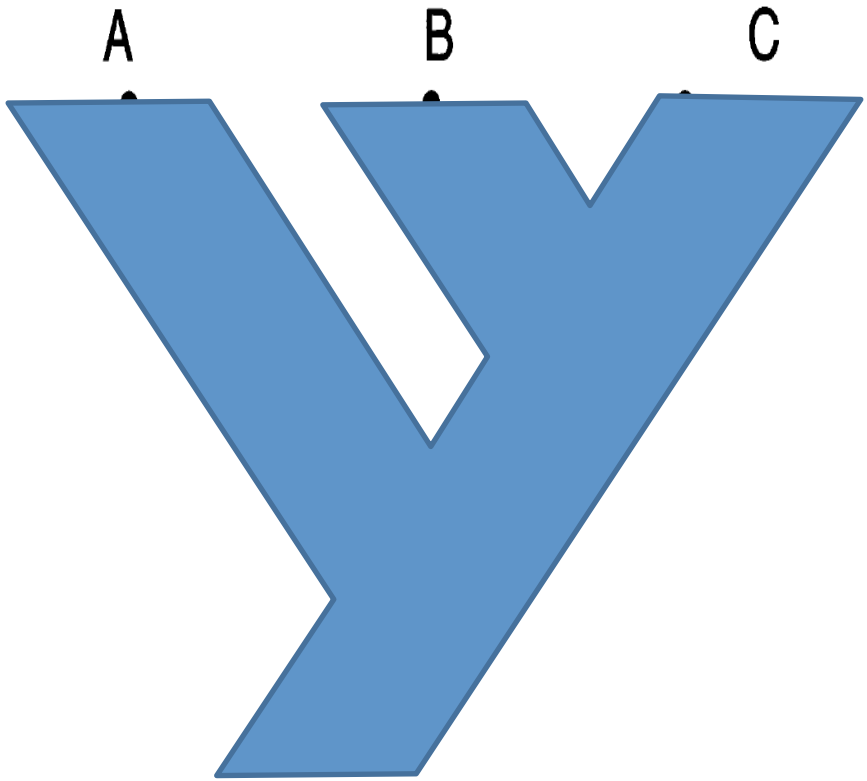
Orangutan Gorilla Chimpanzee Human

Discordance



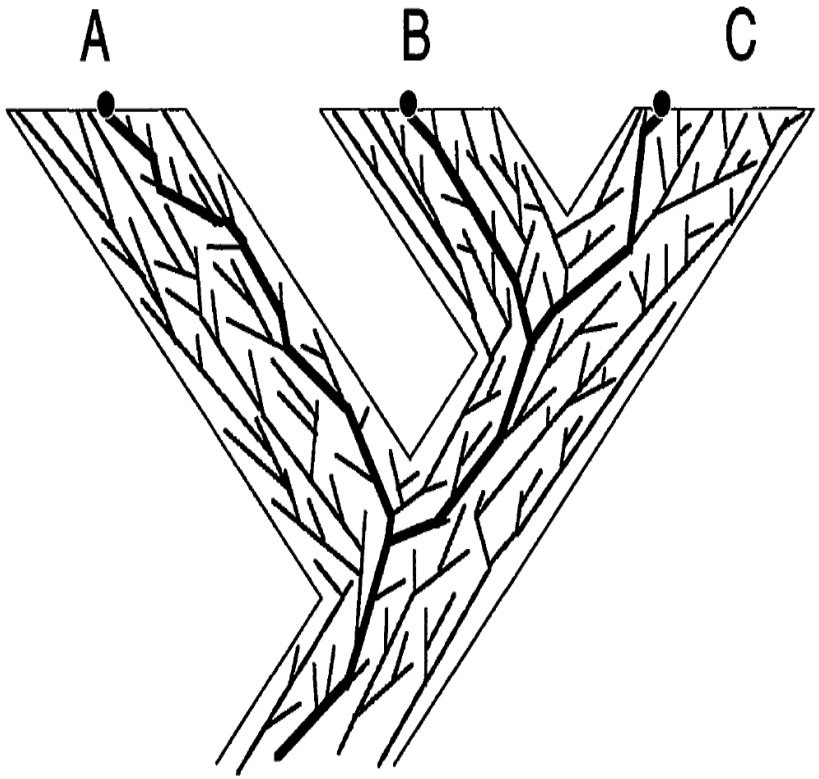
- Gene trees don't necessarily show the same branching pattern as their containing species tree

Gene trees in species tree



[Maddison, Syst.biol., 1997]

Gene trees in species tree



[Maddison, Syst.biol., 1997]

Causes of gene tree discordance

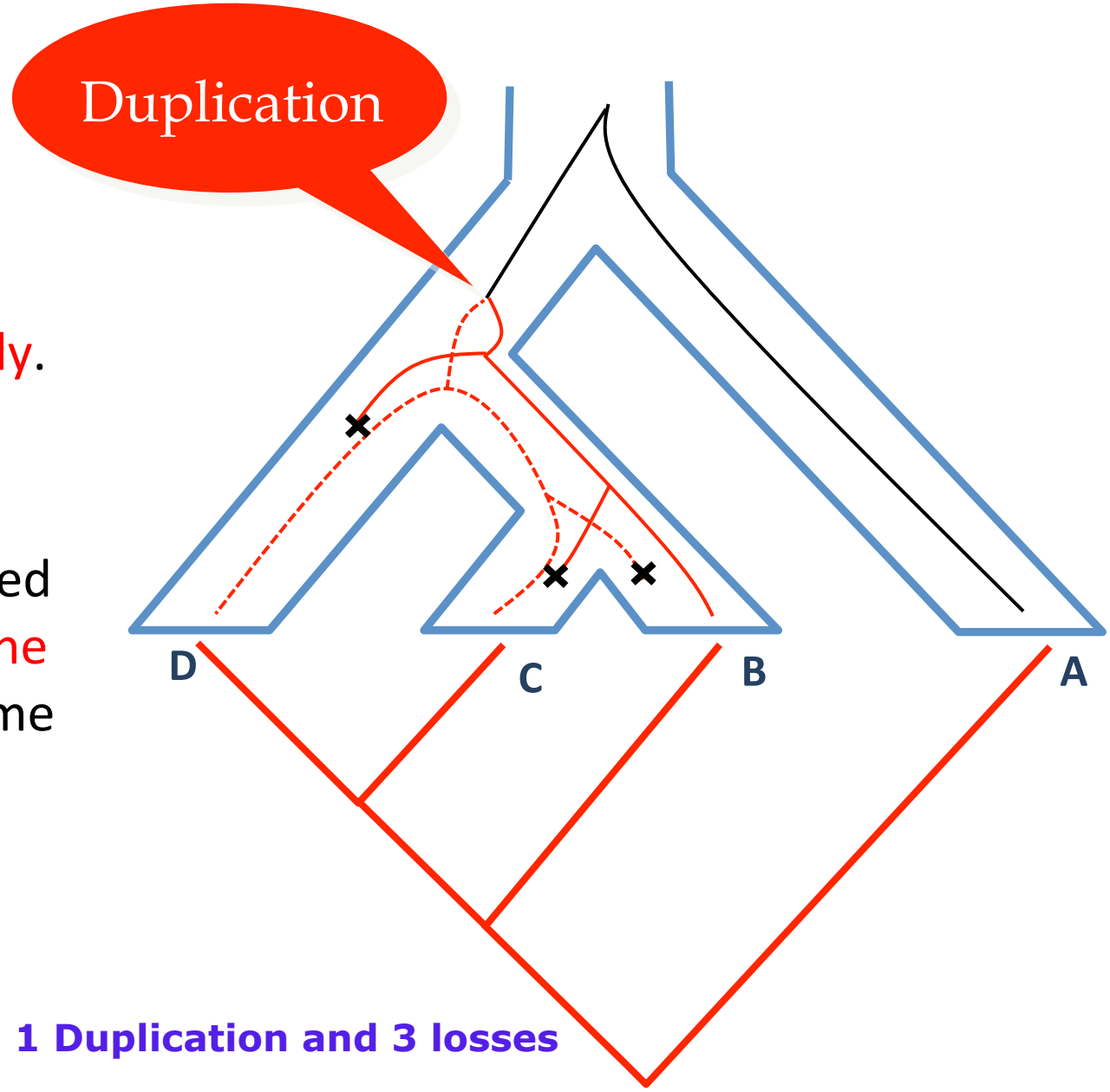
- Discord can arise from
  - Deep Coalescence (ILS = incomplete lineage sorting)
  - Gene Duplication/Loss (GDL)
  - Horizontal Gene Transfer (HGT) etc.

- Estimation error may also introduce discordance.

Gene Duplication/Loss

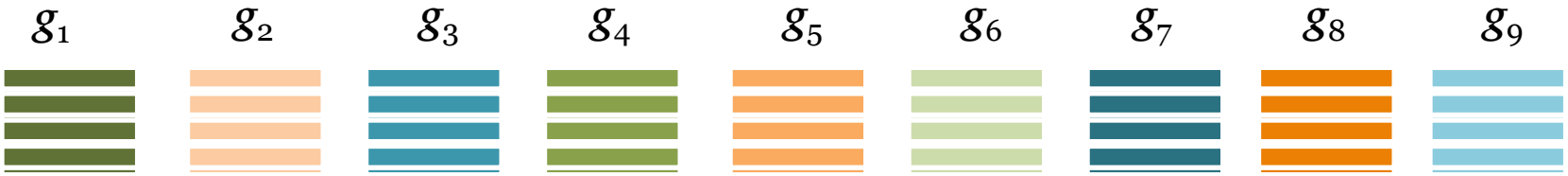
- A gene might get duplicated and both copies descend and evolve independently.

- Discordance can occur if some sampled copies come from one locus and others come from another locus



1 Duplication and 3 losses

Species tree estimation – concatenation?



Supergene alignment g^*

Sequence-based tree estimation method

Species Tree

Species Tree Estimation

Concatenation – standard approach, but: needs single copy of each species, and does not take gene tree heterogeneity into account

Species Tree Estimation

Concatenation – standard approach, but: needs single copy of each species, and does not take gene tree heterogeneity into account

Co-estimation of gene trees and species trees (e.g., PhylDog) – very powerful but slow

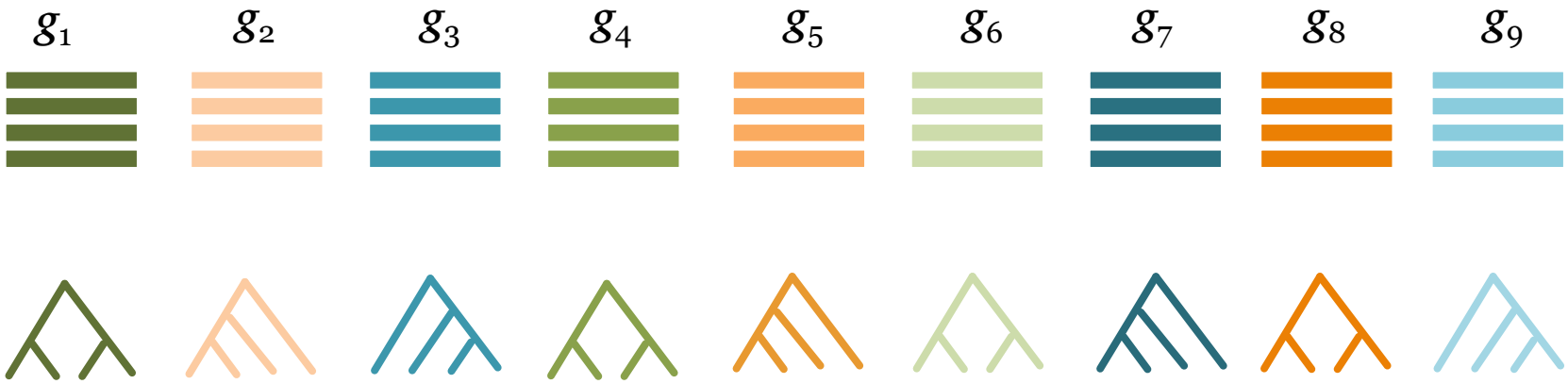
Species Tree Estimation

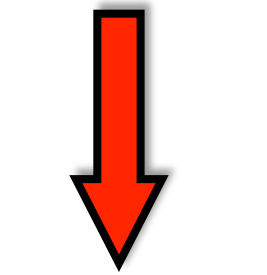
Concatenation – standard approach, but: needs single copy of each species, and does not take gene tree heterogeneity into account

Co-estimation of gene trees and species trees (e.g., PhylDog) – very powerful but slow

Summary methods (e.g., gene tree parsimony) – NP-hard optimization problems, but fast in practice

Species tree estimation: Summary methods



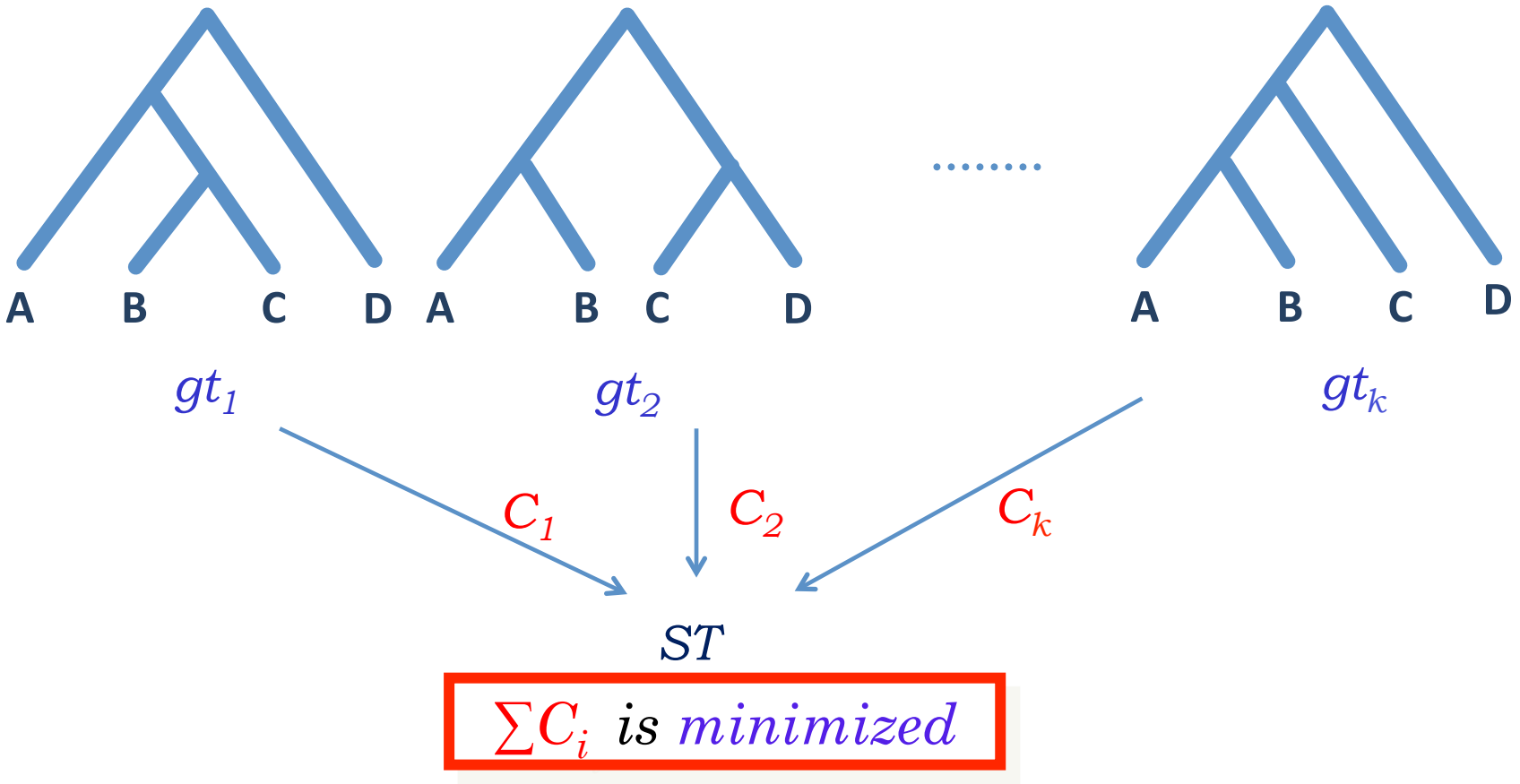


Species Tree

Gene Tree Parsimony (GTP, formulated by Guigo, first method by Rod Page), Supertree methods

GTP: Minimize Gene Duplication+Loss

- Input: A set of rooted binary gene trees (multi-copy)
- Output: A species tree *ST* that minimizes total number of duplications and losses



GTP: Minimize Gene Duplication+Loss

- Input: A set of rooted binary gene trees (multi-copy)
- Output: A species tree *ST* that minimizes total number of duplications and losses

Scoring a single species tree with respect to a set of gene trees is polynomial time

Finding a best species tree is NP-hard, but good heuristics exist:

iGTP (Chaudhary, Bansal, Wehe, Fernandez-Baca, and Eulenstein. BMC Bioinformatics 2010)

DupTree (Wehe, Bansal, Burleigh, and Eulenstein, Bioinformatics 2008)

Incomplete gene trees

- Incomplete gene tree: not all gene trees have individuals from all the species.

  - Sampling Error
    - The gene may be available in the species' genome, but it was not sampled when the gene tree was estimated
  - True biological gene loss
    - Gene birth/death

Summary of our contributions

- We prove that the standard calculation correctly computes losses when incompleteness is due to sampling

Summary of our contributions

- We prove that the standard calculation correctly computes losses when incompleteness is due to sampling

- We show by example that the standard calculation for losses in GTP can be incorrect when incompleteness is due to true biological loss

Summary of our contributions

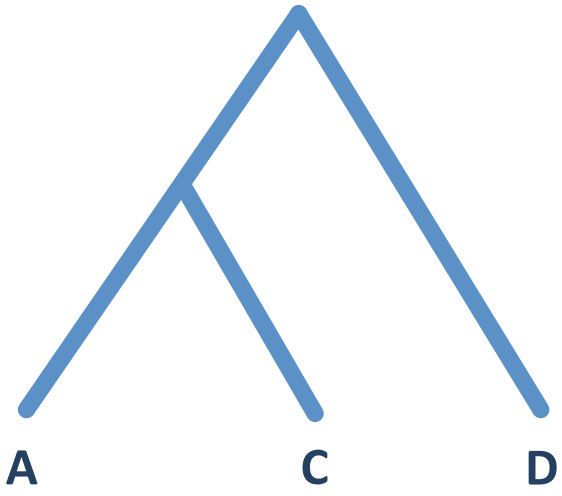
- We prove that the standard calculation correctly computes losses when incompleteness is due to sampling
- We show by example that the standard calculation for losses in GTP can be incorrect when incompleteness is due to true biological loss
- We show how to compute the number of losses implied by a gene tree and species tree, when incompleteness is due to true biological loss

Summary of our contributions

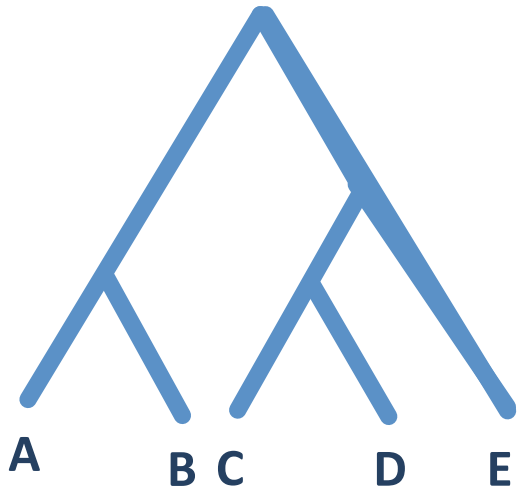
- We prove that the standard calculation correctly computes losses when incompleteness is due to sampling
- We show by example that the standard calculation for losses in GTP can be incorrect when incompleteness is due to true biological loss
- We show how to compute the number of losses implied by a gene tree and species tree, when incompleteness is due to true biological loss
- We formulate variants of the GTP problem (when gene tree incompleteness is due to true biological loss) as minimum weight maximum clique problems, and we show a dynamic programming algorithm to find the optimal species tree.

Reconciliation

- Given a gene tree *gt* and a species tree *ST*,
  - *the objective is to explain the differences in terms of gene duplication and loss*



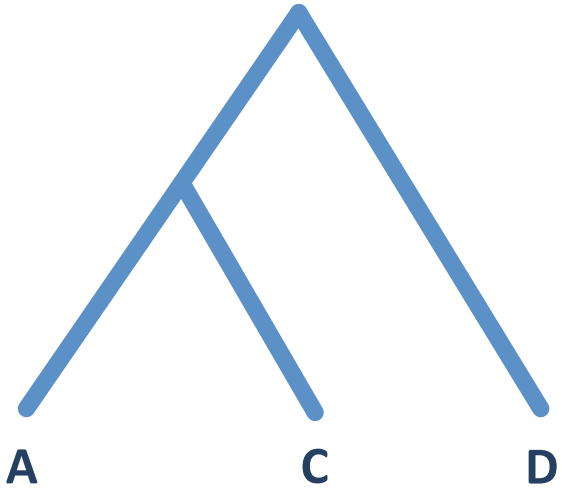
gt



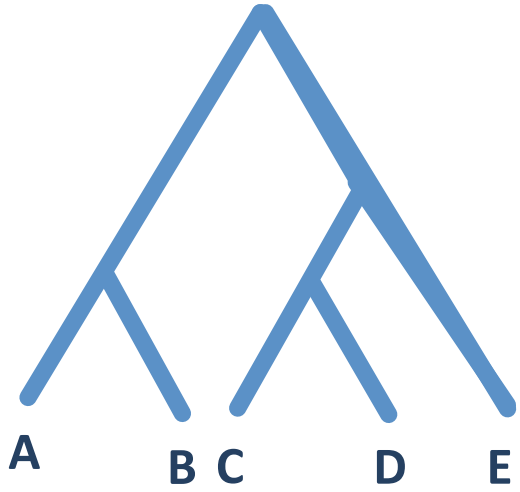
ST

Standard Reconciliation

- Step 1: Restrict the two trees to the same leafset
- Step 2: Map each internal node in the gene tree to MRCA in the species tree
- Step 3: Identify duplication nodes in gene tree
- Step 4: Calculate losses



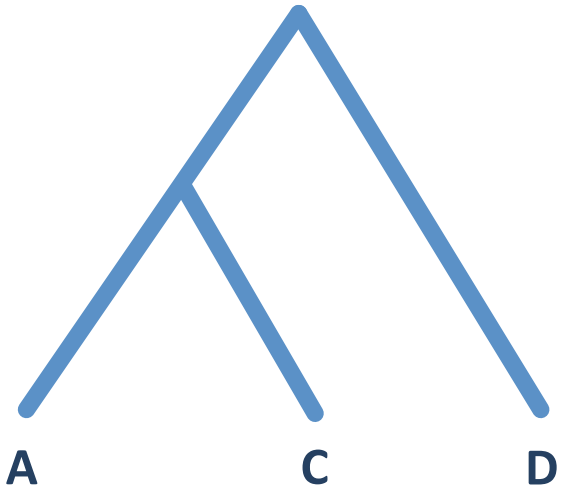
gt



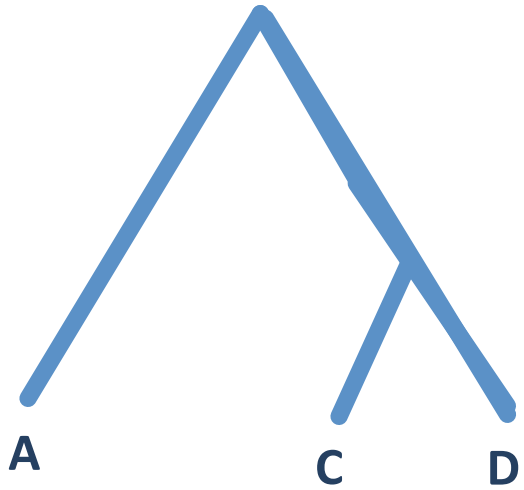
ST

Step 1: Restrict to the same leafset

- Step 1: Restrict to the same leafset
  - Given a gene tree *gt* and a species tree *ST*, *ST(gt)* is the *homeomorpic* subtree of *ST* induced by the leafset of *gt*.



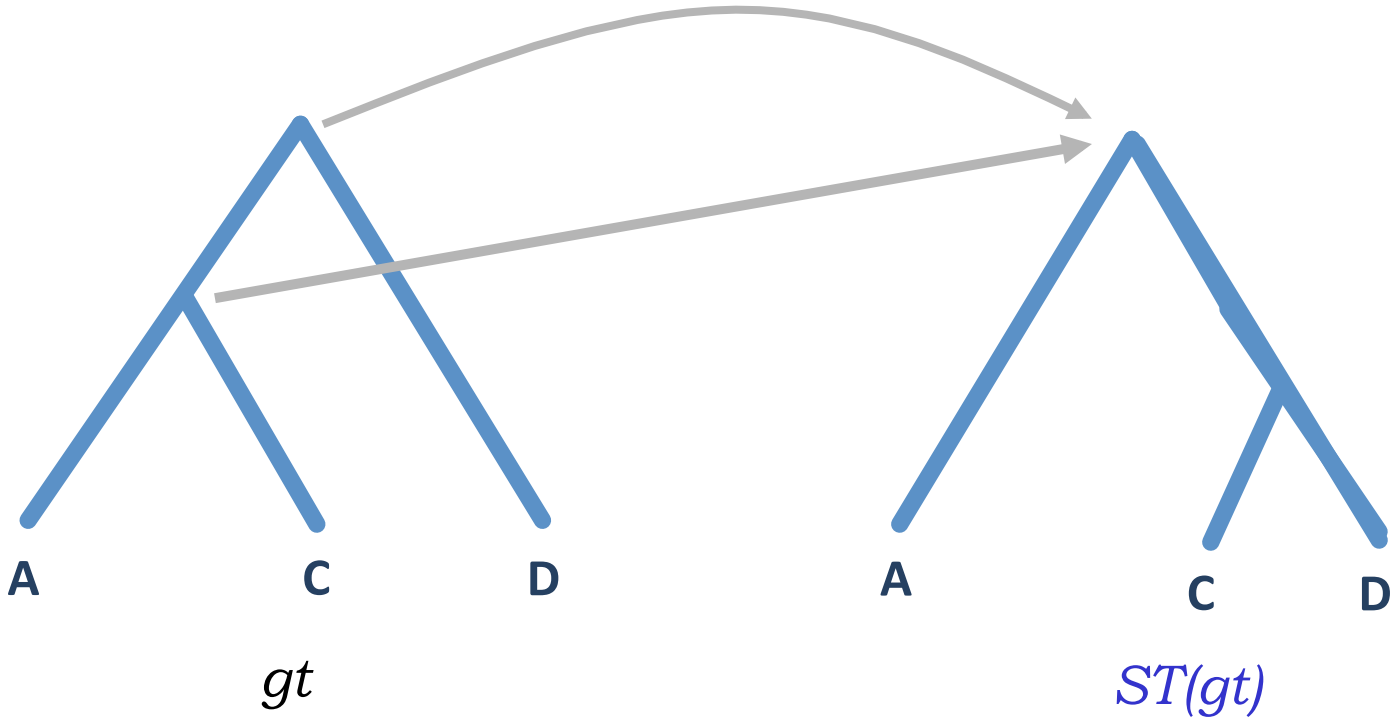
gt



ST(gt)

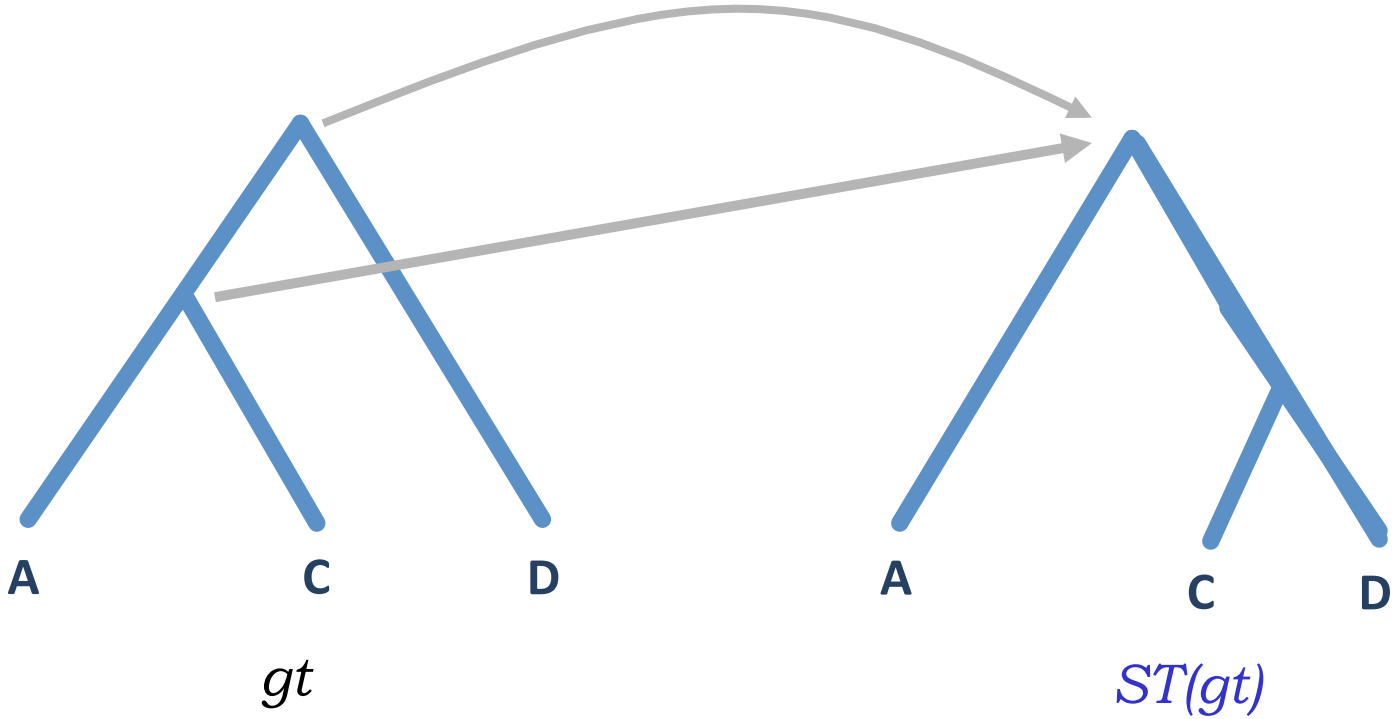
Step 2: Map nodes in gene tree to species tree

- The standard approach maps the internal nodes in gt to the nodes in ST(gt) using MRCA mapping, called “M”.



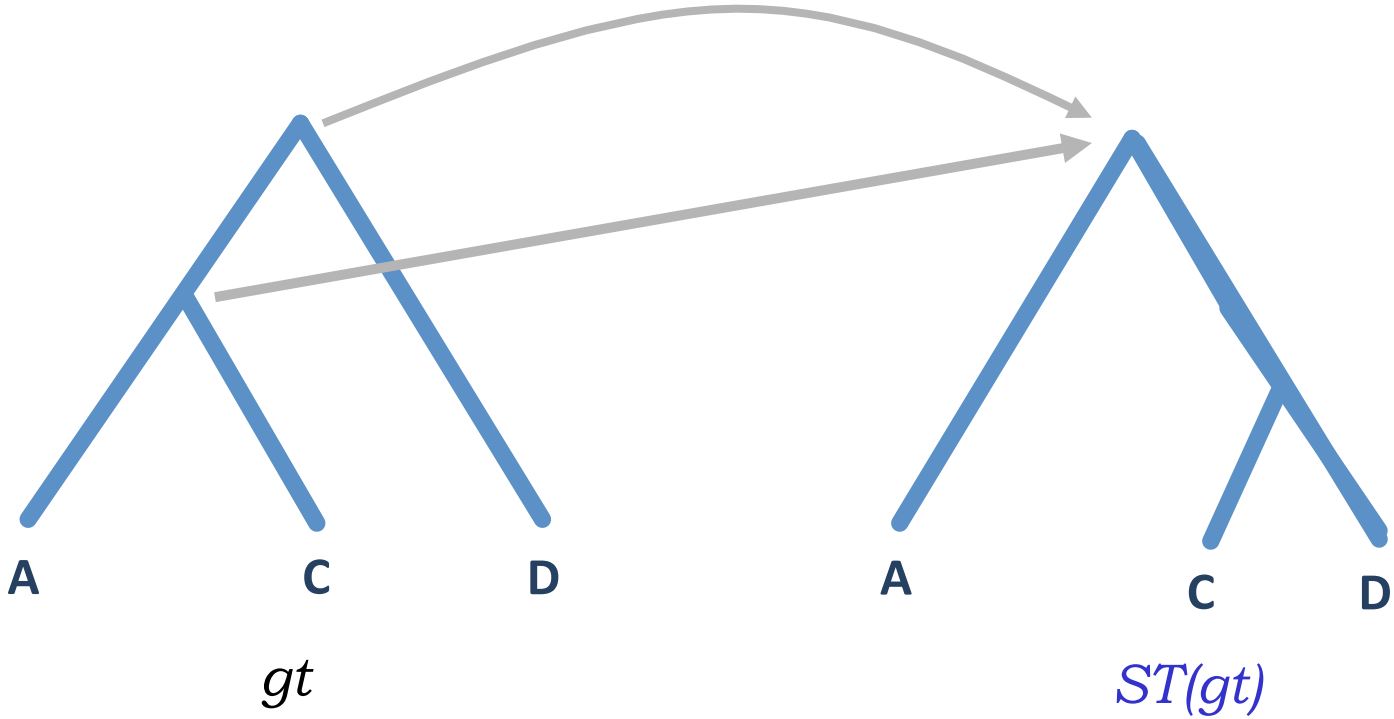
Step 3: Identify duplication nodes in gt

- Every node v in gt that has a child v' for which M(v)=M(v') is a duplication node (Guigo et al. 1996, Ma et al. 2000); all others are speciation nodes.



Step 4: Calculating losses

- Losses are associated to nodes in the gene tree.
- Each node u has two children l (left) and r (right)
- Calculation of losses depends on MRCA mapping of u, l, r



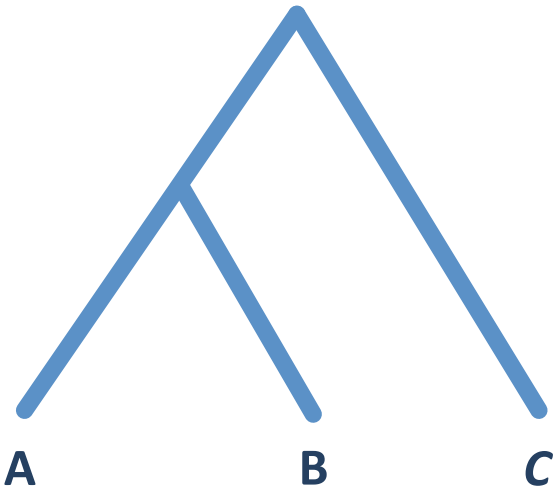
Step 4: Standard technique for calculating losses

- Let d(x,y) denote the number of vertices in the path between x and y. Then (by Ma et al. 2000, Gorecki 2004),

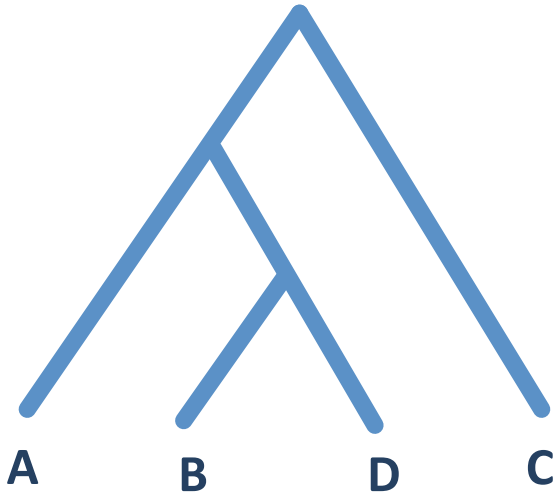
$$F(u,T) = \begin{cases} d(\mathcal{M}(r),\mathcal{M}(u)) + 1 & \text{if } \mathcal{M}(r) \neq \mathcal{M}(u) \ \& \ \mathcal{M}(l) = \mathcal{M}(u), \\ d(\mathcal{M}(l),\mathcal{M}(u)) + 1 & \text{if } \mathcal{M}(l) \neq \mathcal{M}(u) \ \& \ \mathcal{M}(r) = \mathcal{M}(u), \\ d(\mathcal{M}(r),\mathcal{M}(u)) & \\ +d(\mathcal{M}(l),\mathcal{M}(u)) & \text{if } \mathcal{M}(r) \neq \mathcal{M}(u) \ \& \ \mathcal{M}(l) \neq \mathcal{M}(u), \\ 0 & \text{if } \mathcal{M}(r) = \mathcal{M}(l) = \mathcal{M}(u). \end{cases}$$

$$L_{std}(gt, ST) = \sum_{u \in V_{int}(gt)} F(u, ST(gt))$$

What would the reconciliation cost be?



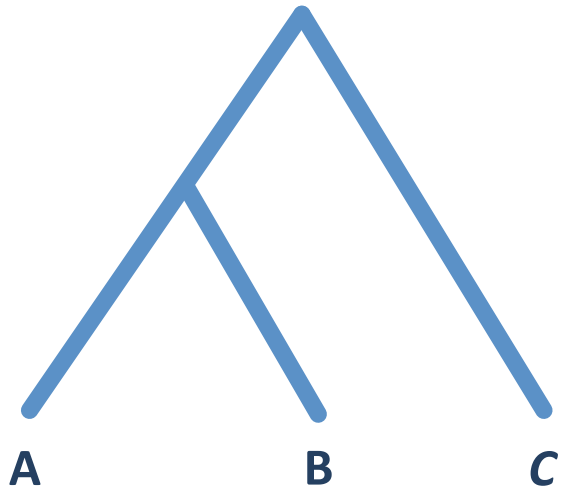
gt



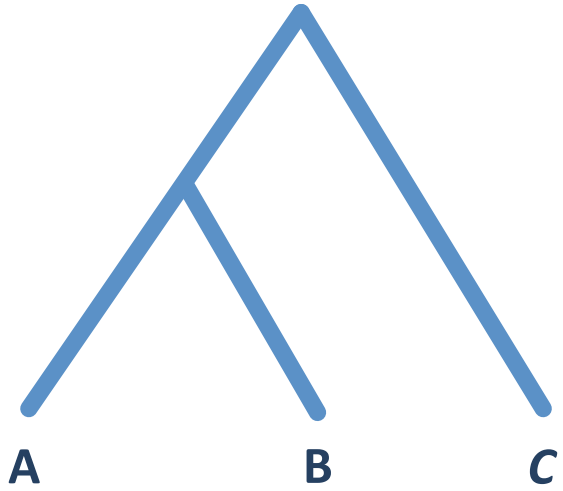
ST

Answer using standard formula: 0 losses!

- Standard formula by calculating the homeomorphic tree ST(gt) implies zero loss!

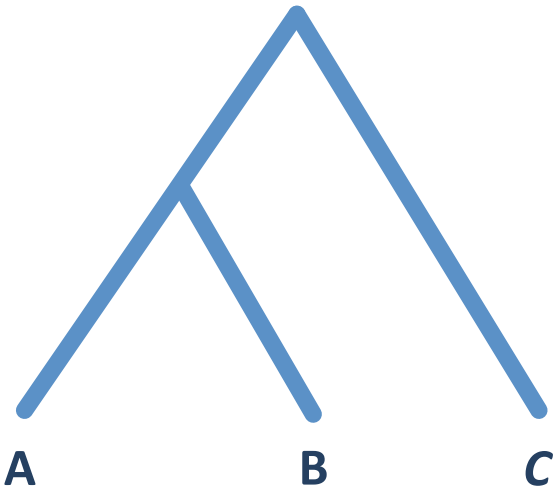


gt

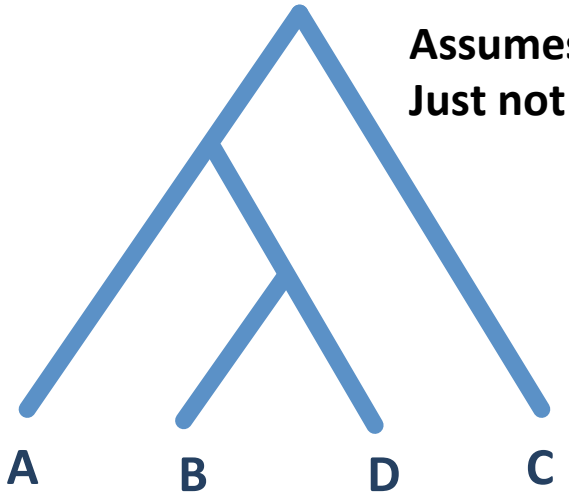


ST(gt)

Incompleteness due to sampling



gt



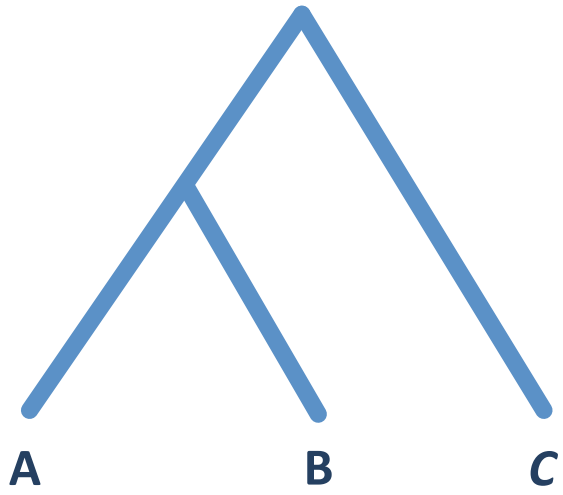
Assumes D was
Just not sampled.

ST

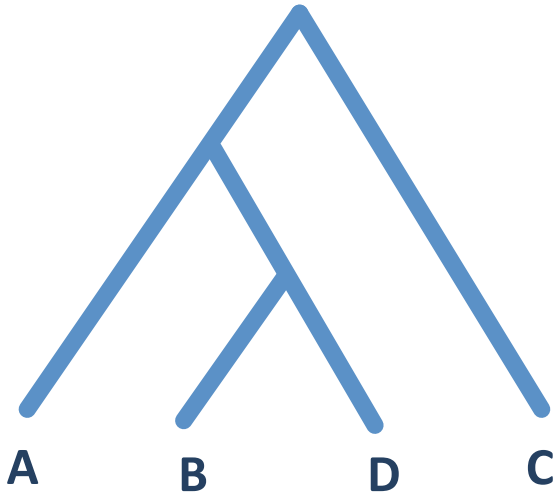
$L_{std}(gt, ST) = L_{samp}(gt, ST)$

▶ **Theorem 1.** *Given a binary rooted gene tree gt and a binary rooted species tree ST such that* $L(gt) \subseteq L(ST)$*, the MRCA mapping defines a reconciliation that minimizes the number of duplications, the number of losses, and hence also the total number of duplications and losses, where we treat losses as due to sampling. Furthermore,* $L_{std}(gt, ST) = L_{samp}(gt, ST)$*, which means the standard formula correctly computes the number of losses when we treat incompleteness as due to sampling.*

Incompleteness due to gene birth/death

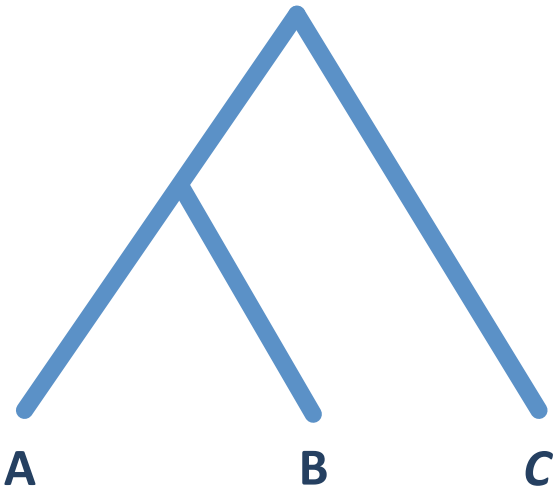


gt

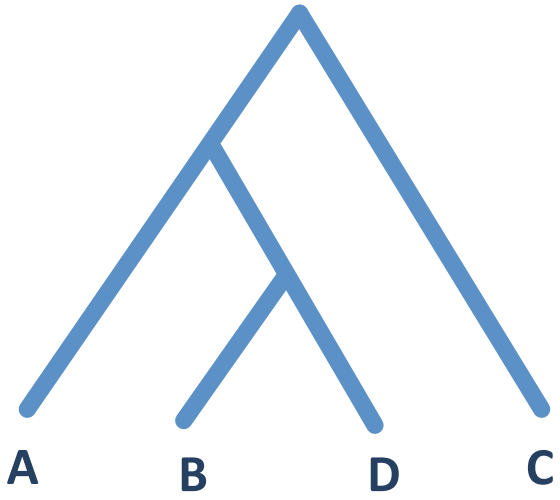


ST

What should the reconciliation cost be?

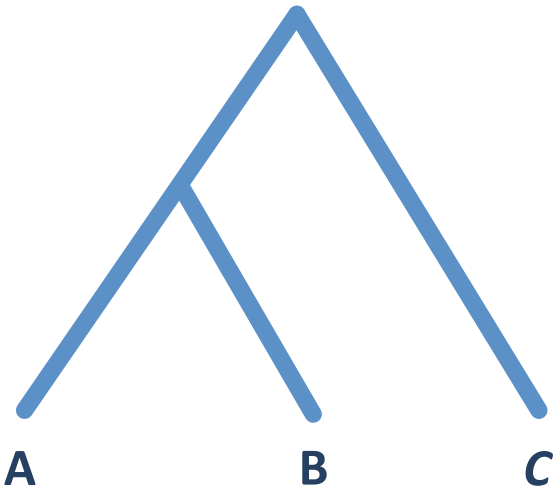


gt

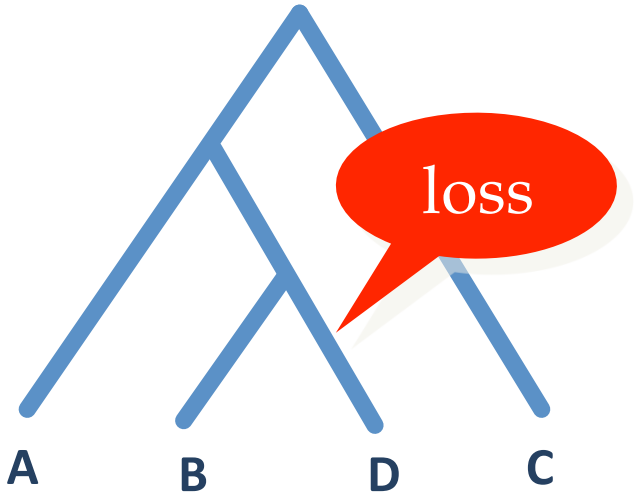


ST

What should the reconciliation cost be?



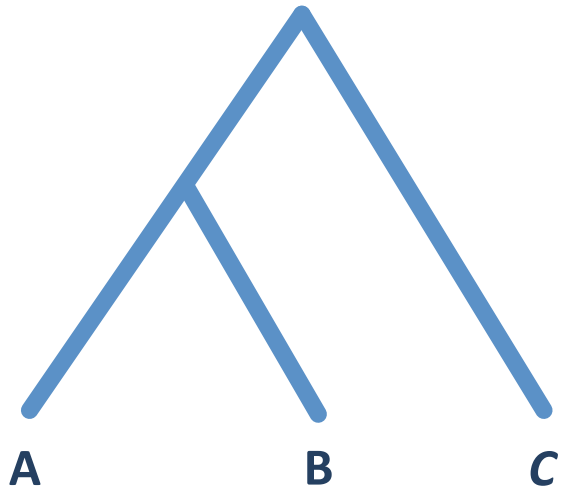
gt



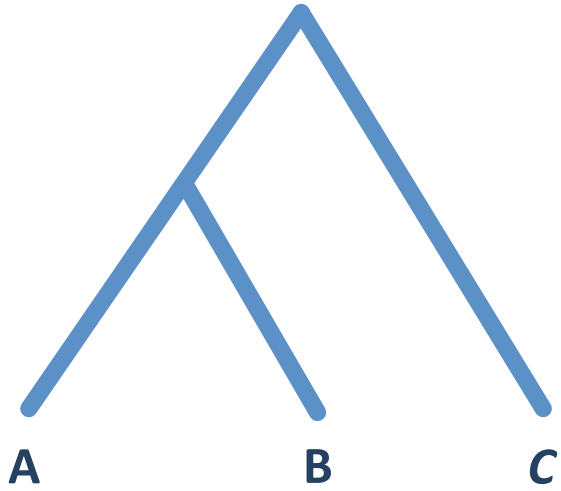
ST

What should the reconciliation cost be?

- Standard formula by calculating the homeomorphic tree ST(gt) implies zero loss!



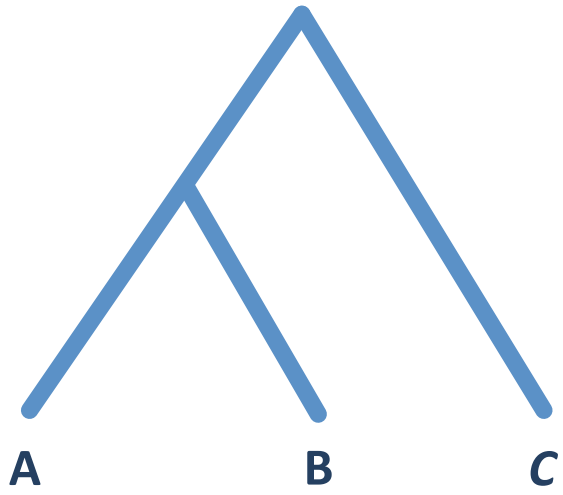
gt



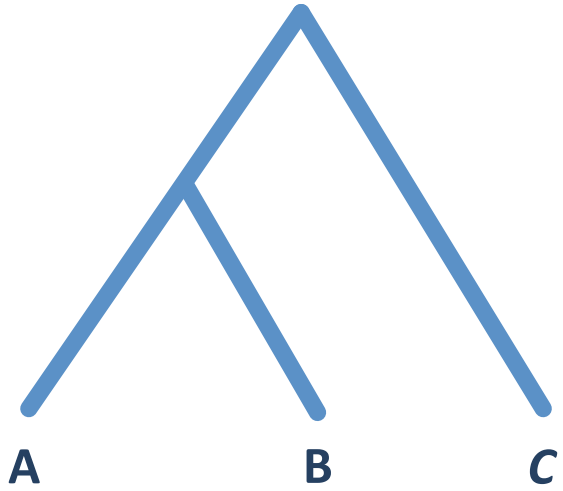
ST(gt)

Standard Formula doesn't work here

- Standard formula by calculating the homeomorphic tree ST(gt) implies zero loss!

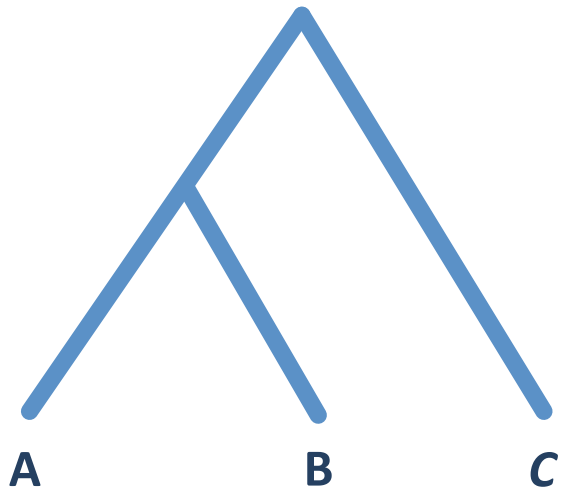


gt

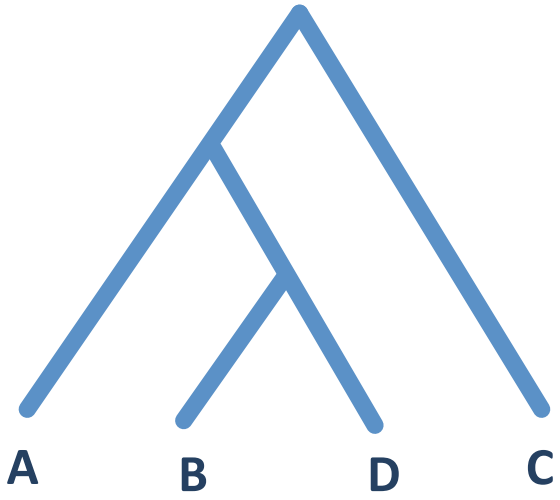


ST(gt)

Solution: Use ST instead of ST(gt)



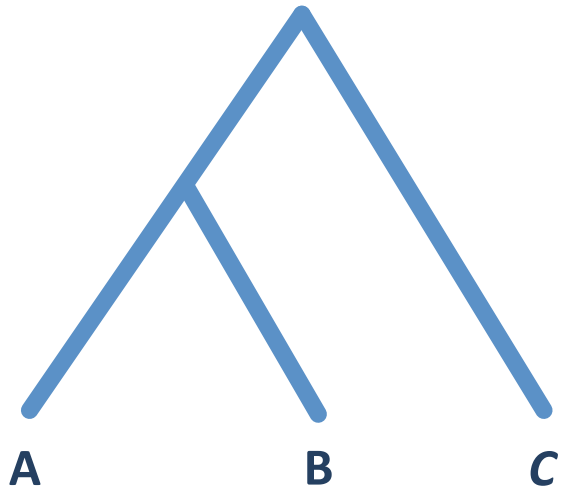
gt



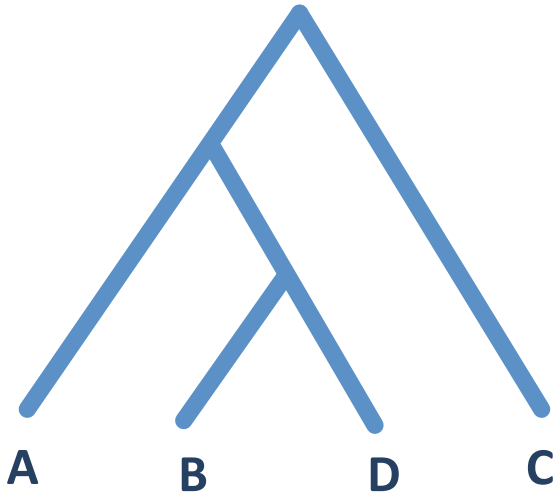
ST

Use ST instead of ST(gt) for reconciliation

No problem with calculating duplications



gt

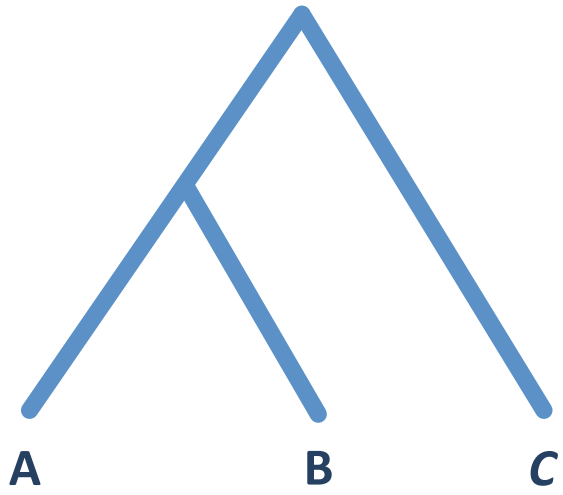


ST

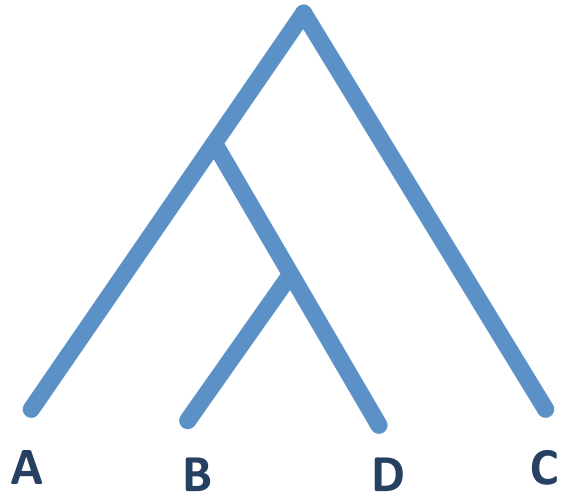
Use ST instead of ST(gt) for reconciliation

No problem with calculating duplications

Standard formula for losses with ST in place of ST(gt) works



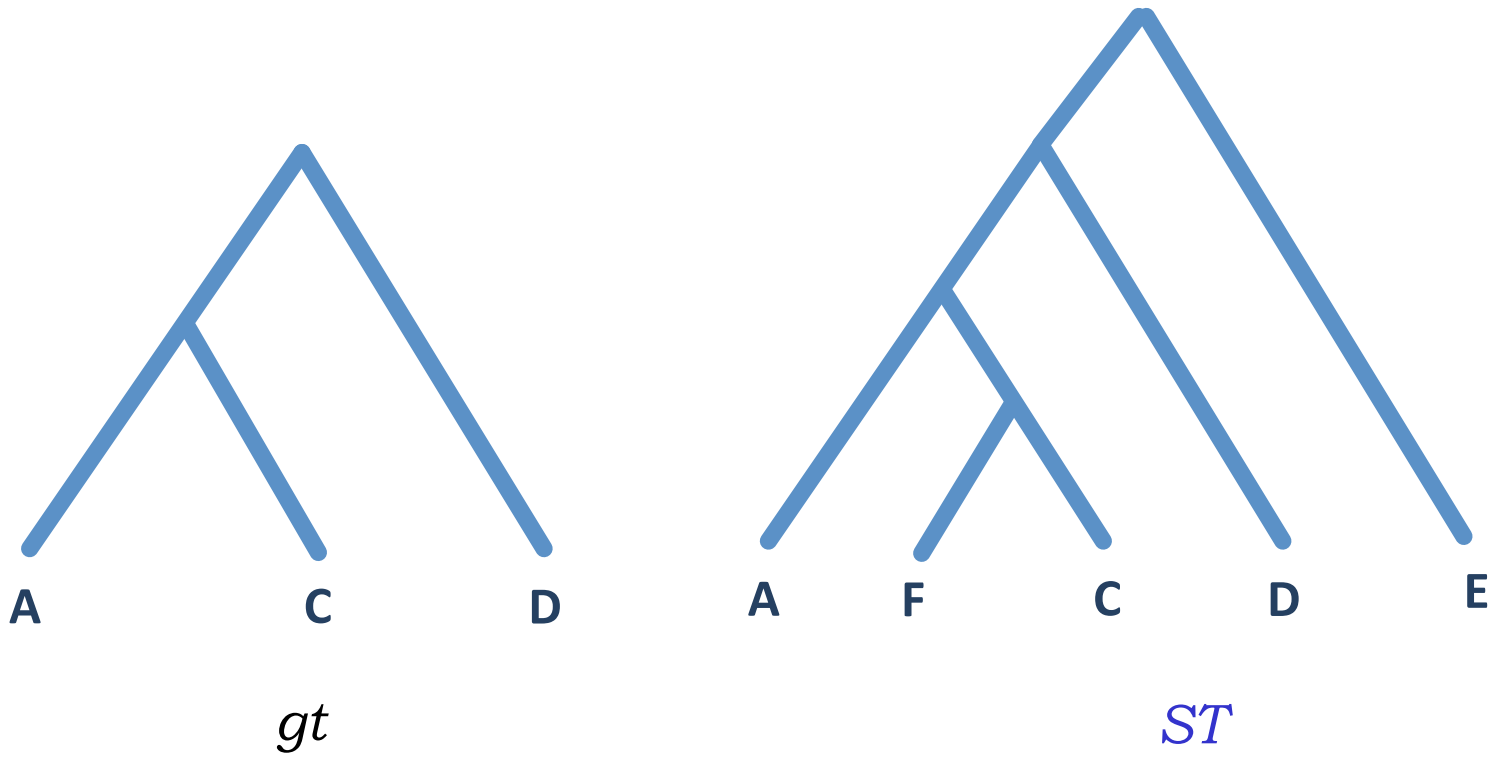
gt



ST

Losses due to gene birth/death

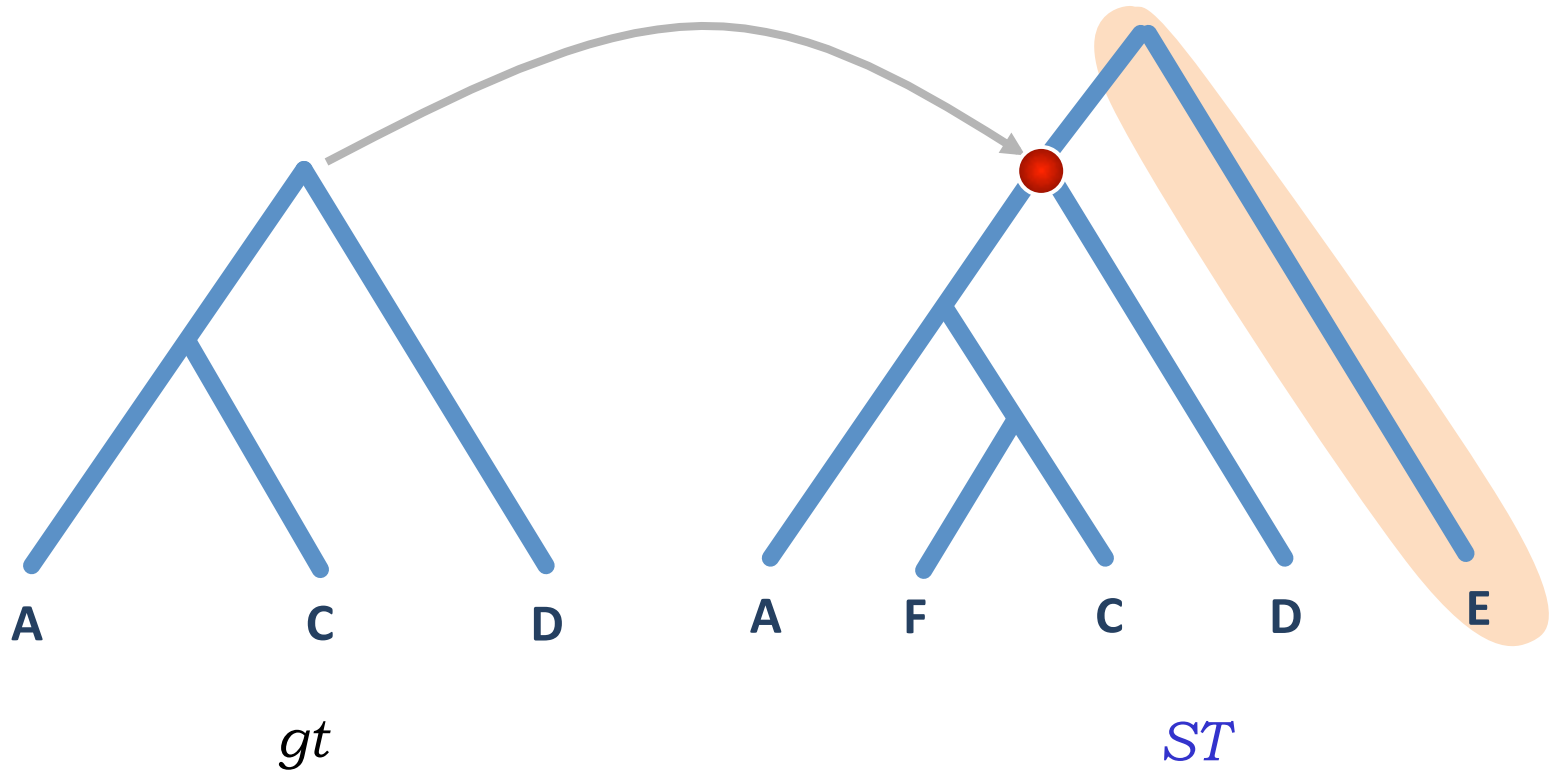
- Original species tree ST instead of the restriction ST(gt)



gt *ST*

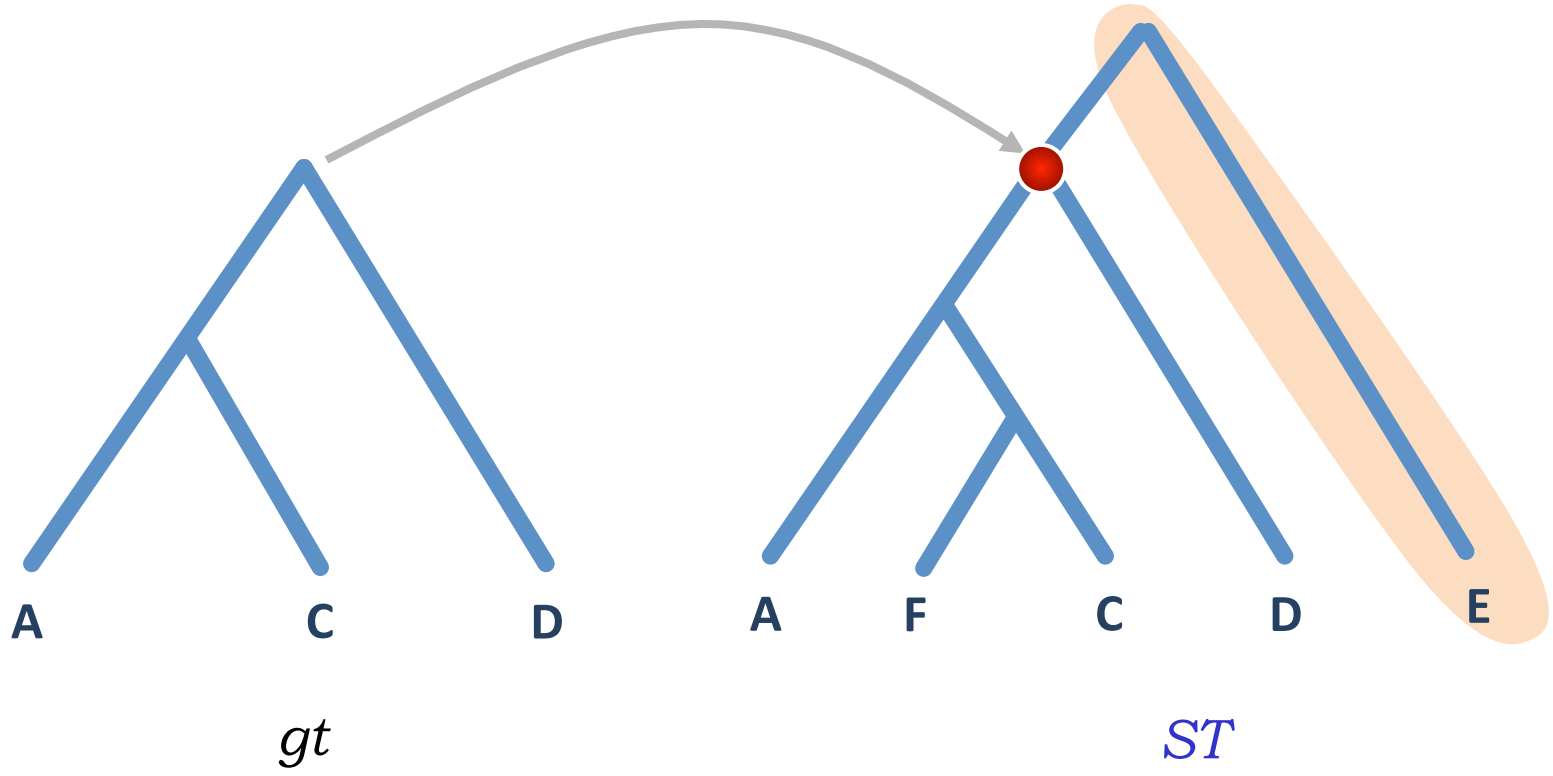
Losses due to gene birth/death

- Original species tree ST instead of the restriction ST(gt)
  - Not enough!



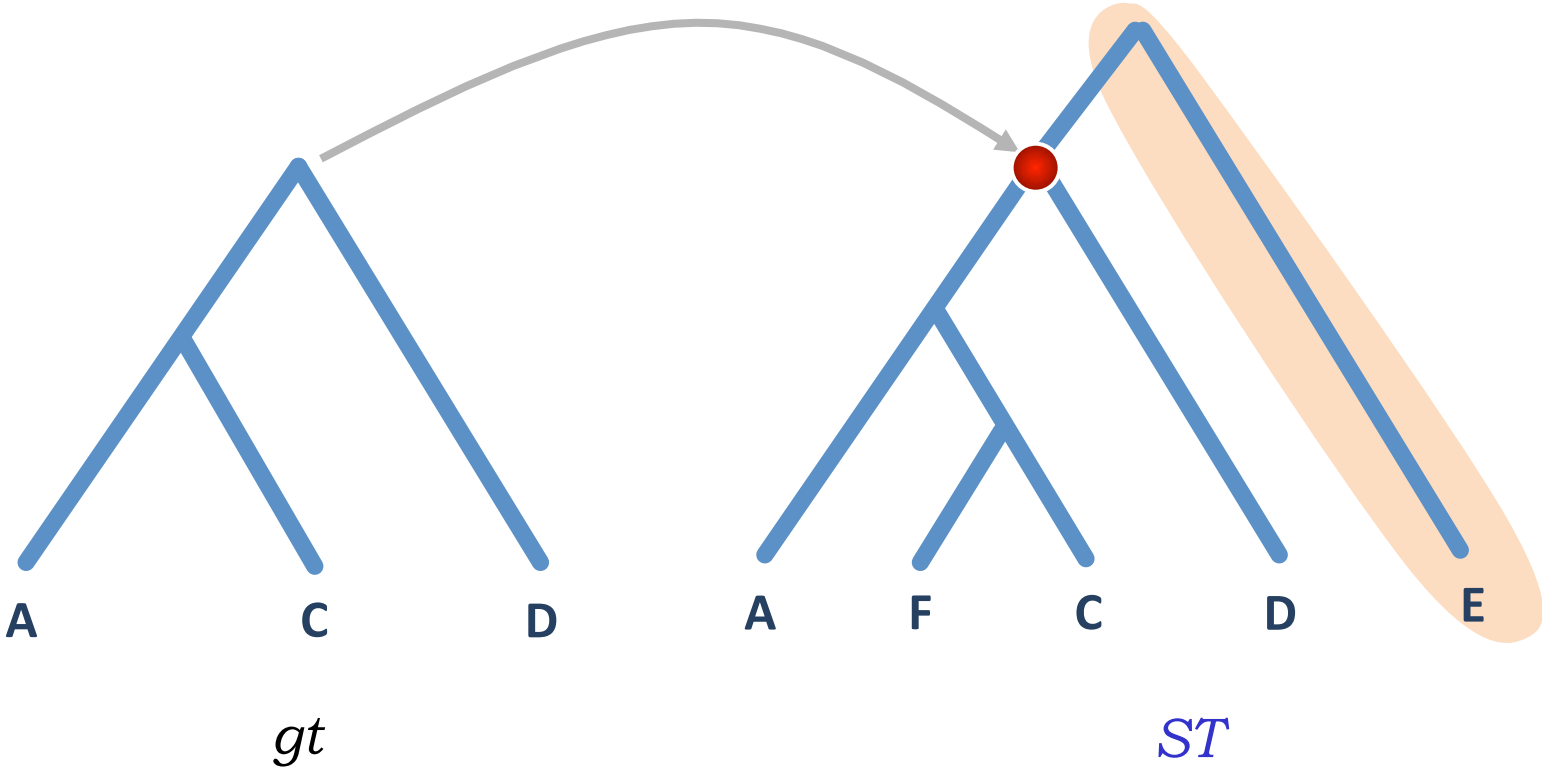
Losses due to gene birth/death

- Original species tree ST instead of the restriction ST(gt)
  - Not enough!
- Depends upon whether one assumes, *a priori*, that the gene is present in the root of the ST.



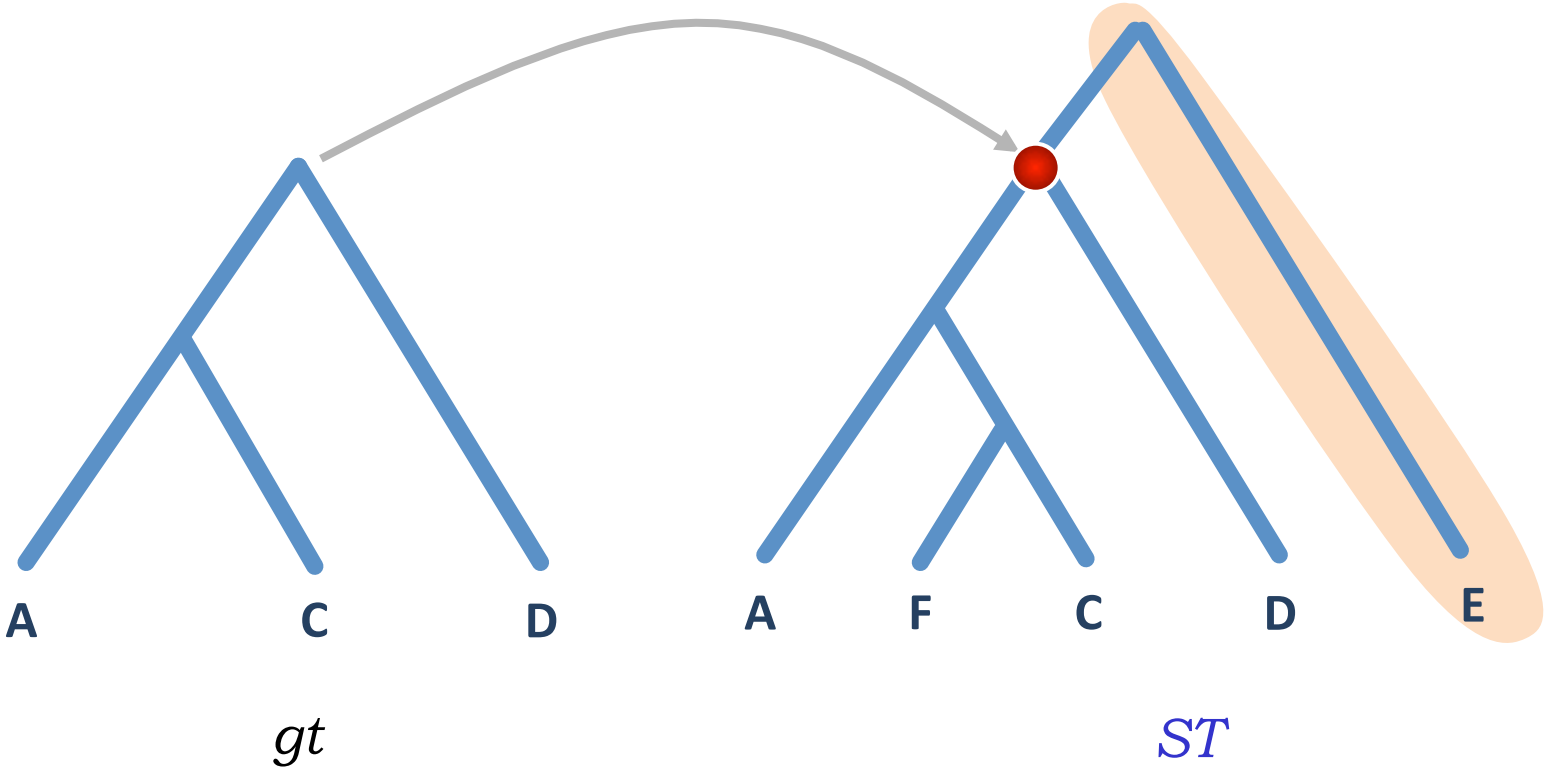
Losses due to gene birth/death

- Depends upon whether one assumes, *a priori*, that the gene is present in the root of the ST.
  - The gene was present in the r(ST)
    - Need to consider the maximal clades above M(r(gt))



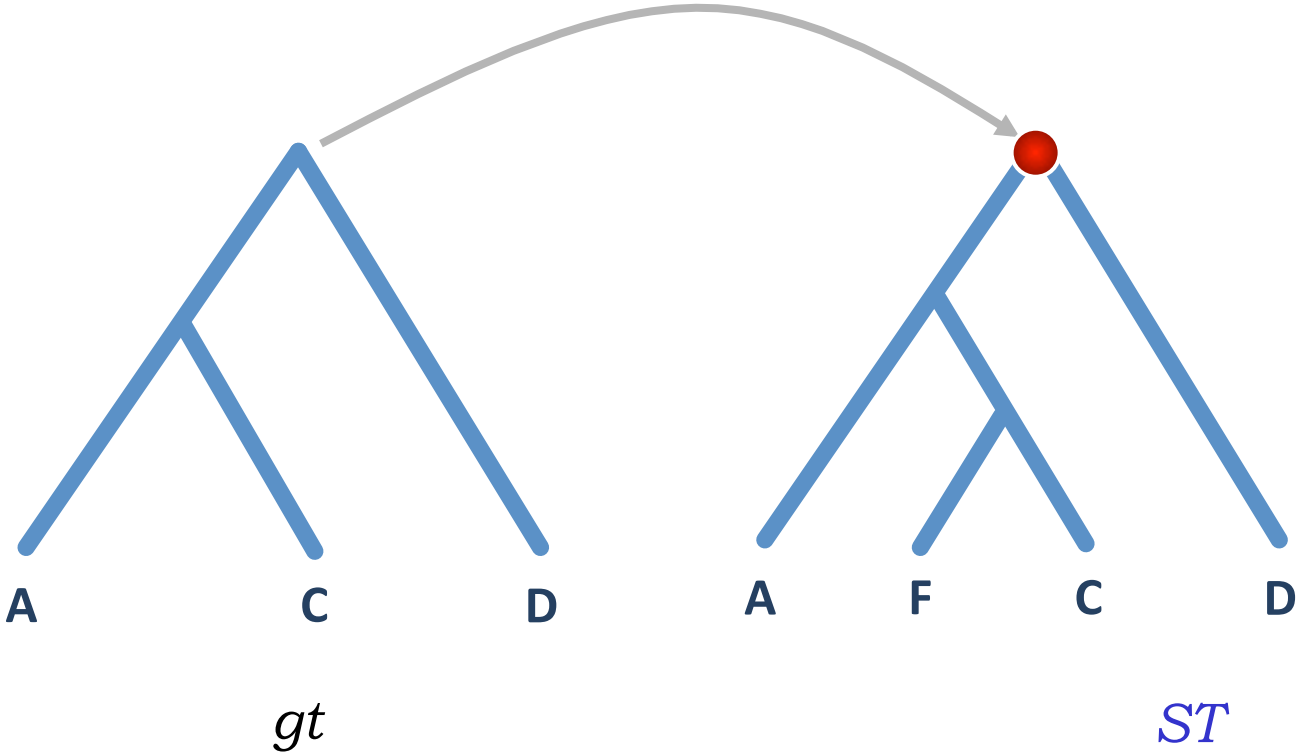
Losses due to gene birth/death

- Depends upon whether one assumes, *a priori*, that the gene is present in the root of the ST.
  - The gene was present in r(ST)
    - Need to consider the maximal clades above M(r(gt))
  - The gene was born in M(r(gt))



Losses due to gene birth/death

- Depends upon whether one assumes, *a priori*, that the gene is present in the root of the ST.
  - The gene was present in r(ST)
    - Need to consider the maximal clades above M(r(gt))
  - The gene was born in M(r(gt))
    - Standard formula with ST in place of ST(gt) works



Losses due to gene birth/death

- See Theorem 2 in the paper for mathematical proofs!

Species tree estimation

Input: Set of rooted binary gene trees, and costs for duplication and losses

Output: Rooted binary species tree ST, minimizing the total (weighted) duplication-loss cost (treating incompleteness as true biological loss)

Species tree estimation

Input: Set of rooted binary gene trees, and costs for duplication and losses

Output: Rooted binary species tree ST, minimizing the total (weighted) duplication-loss cost (treating incompleteness as true biological loss)

NP-hard! (also NP-hard if you treat incompleteness as due to sampling)

Species tree estimation

- **Our approach:** We extend the technique from Bayzid, Mirarab, and Warnow PSB 2011, which found optimal species trees for weighted duploss problem, treating incompleteness as sampling error.

Species tree estimation

Input: Set of rooted binary gene trees, and costs for duplication and losses

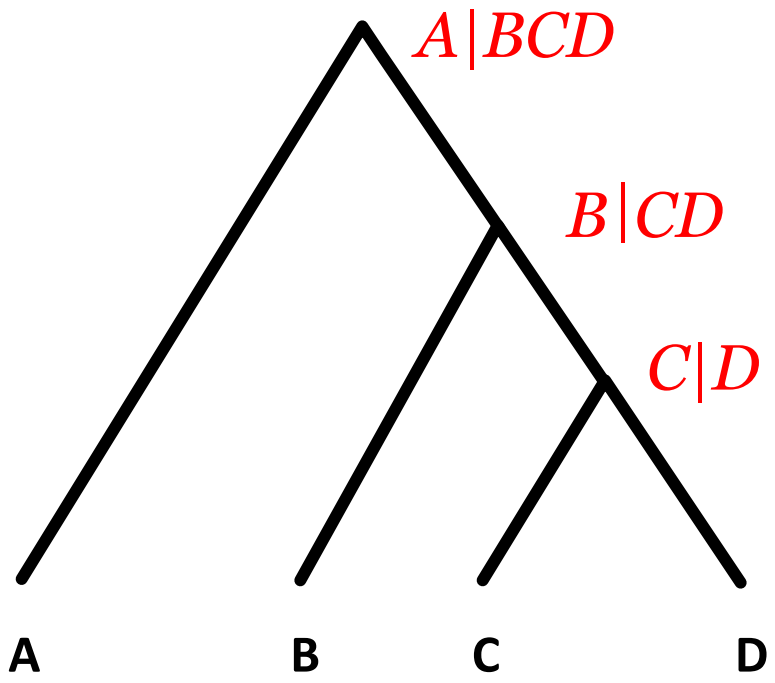
Output: Rooted binary species tree ST, minimizing the total (weighted) duplication-loss cost (treating incompleteness as true biological loss)

Constrained version:
Consider set *S* of "allowed subtree-bipartitions", and find species tree ST that draws its subtree-bipartitions from *S* and optimizes this weighted duploss score.

Terminology: Subtree-bipartition

- Subtree-bipartition
  - For an internal node u in a *binary-rooted* tree T,

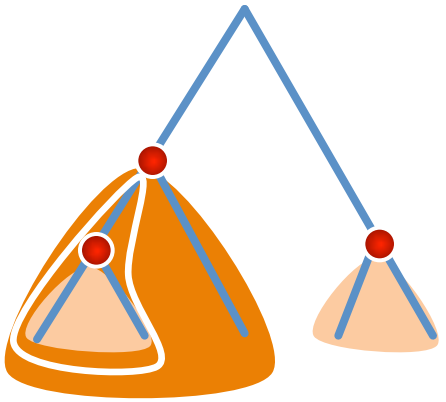
$$SBP(u) = cluster(T_L) \mid cluster(T_R)$$



Terminologies: Compatibility

- *Compatibility*
  - X|Y and P|Q are *compatible* if they can "co-exist" in a binary rooted tree.

Containment



Disjoint

Theorem: Two subtree-bipartitions are *compatible* if one *contains* the other, or they are *disjoint*

Species tree estimation

- Bayzid, Mirarab, and Warnow PSB 2011: approach for optimal species tree construction for weighted duploss problem, treating incompleteness as sampling error:

Species tree estimation

- Bayzid, Mirarab, and Warnow PSB 2011: approach for optimal species tree construction for weighted duploss problem, treating incompleteness as sampling error:
  - Computes "compatibility graph" CG (vertices correspond to subtree-bipartitions, edges correspond to pairs of compatible subtree-bipartitions)

Species tree estimation

- Bayzid, Mirarab, and Warnow PSB 2011: approach for optimal species tree construction for weighted duploss problem, treating incompleteness as sampling error:
  - Computes "compatibility graph" CG (vertices correspond to subtree-bipartitions, edges correspond to pairs of compatible subtree-bipartitions)
  - Assigns appropriate weights on the vertices of the compatibility graph

Species tree estimation

- Bayzid, Mirarab, and Warnow PSB 2011: approach for optimal species tree construction for weighted duploss problem, treating incompleteness as sampling error:
  - Computes "compatibility graph" CG (vertices correspond to subtree-bipartitions, edges correspond to pairs of compatible subtree-bipartitions)
  - Assigns appropriate weights on the vertices of the compatibility graph
  - Proves that the optimal species tree ST corresponds to the minimum weight maximum clique in CG.

Species tree estimation

- Bayzid, Mirarab, and Warnow PSB 2011: approach for optimal species tree construction for weighted duploss problem, treating incompleteness as sampling error:
  - Computes "compatibility graph" CG (vertices correspond to subtree-bipartitions, edges correspond to pairs of compatible subtree-bipartitions)
  - Assigns appropriate weights on the vertices of the compatibility graph
  - Proves that the optimal species tree ST corresponds to the minimum weight maximum clique in CG.
  - Presents efficient dynamic programming algorithm to find the optimal ST.
    - Exponential time algorithm for an exact solution
    - Polynomial time algorithm for a constrained version

Species tree estimation

- Bayzid, Mirarab, and Warnow PSB 2011: approach for optimal species tree construction for weighted duploss problem, treating incompleteness as sampling error:
  - Computes "compatibility graph" CG (vertices correspond to subtree-bipartitions, edges correspond to pairs of compatible subtree-bipartitions)
  - Assigns appropriate weights on the vertices of the compatibility graph
  - Proves that the optimal species tree ST corresponds to the minimum weight maximum clique in CG.
  - Presents efficient dynamic programming algorithm to find the optimal ST.
    - Exponential time algorithm for an exact solution
    - Polynomial time algorithm for a constrained version

Summary

» Investigated how different reasons for gene tree incompleteness affect the mathematical formulation of gene loss

- Sampling
- True Biological loss

» Presented mathematical formulation to model missing taxa due to true biological loss

» Proposed exact and heuristic algorithms to infer species trees from a set of incomplete gene trees by minimizing gene duplications and losses when the incompleteness is due to true biological loss

Md. S. Bayzid

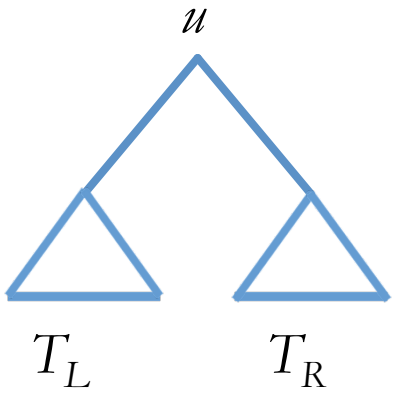
PhD received 2016
Now at BUET (Bangladesh University of Engineering and Technology)

http://cse.buet.ac.bd/faculty/facdetail.php?id=bayzid

Research supported by NSF 1062335 and Fulbright Fellowship

Dynamic Programming approach

- Minimum Weight Clique problem is NP-hard!
- DP-based approach would be more efficient.



$$weight(T) = weight(T_L) + weight(T_R) + weight(u)$$

- The DP algorithm will compute a rooted, binary tree T_A for every cluster A such that T_A maximizes the sum, over all gene trees t, of the number of subtree-bipartitions in t that are dominated by some subtree-bipartition in T_A. We will denote this total number by value(A).

Dynamic Programming Contd.

weight(X|Y) = #sbp in gene trees dominated by X|Y

value(A) = weight ($a_1|a_2$); if A =$\{a_1,a_2\}$
value(A) = 0; if A =$\{a_1\}$ (base case)

value(A) = min$\{$value(A_1) + value(A-A_1) + weight($A_1|A-A_1$)$\}$;

if |A| > 2 (recursive step)

($A_1|A-A_1$)

Global Optimal Solution - if we allow any subtree-bipartition on A

Constrained version - if ($A_1|A-A_1$) has to come from set *S*

Running Time

- Depends on the number of subtree-bipartitions and number n of species.
- Let S be the set of subtree-bipartitions.
  - $O(n|S|^2)$ for finding the domination relationships (for every pair).
  - value(A) can be computed in $O(|S|)$ time, since at worst we need to look at every subtree-bipartition in S.
  - Running time is $O(n|S|^2)$.