Quintet Rooting: Rooting Species Trees under the Multi-Species Coalescent Model

Yasamin Tabatabaee, Kowshika Sarker, Tandy Warnow

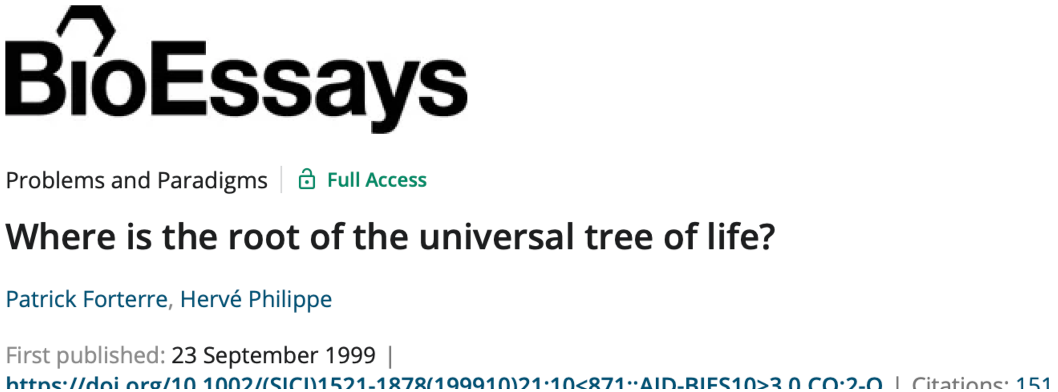
University of Illinois at Urbana-Champaign

Intelligent Systems for Molecular Biology (ISMB)

July 2022

Why Rooting Species Trees?

- Multiple applications throughout biology
- Understanding
  - Adaptation
  - Biodiversity
  - Phylogeography
  - Co-Evolution
- Most species tree estimation methods don't produce rooted trees



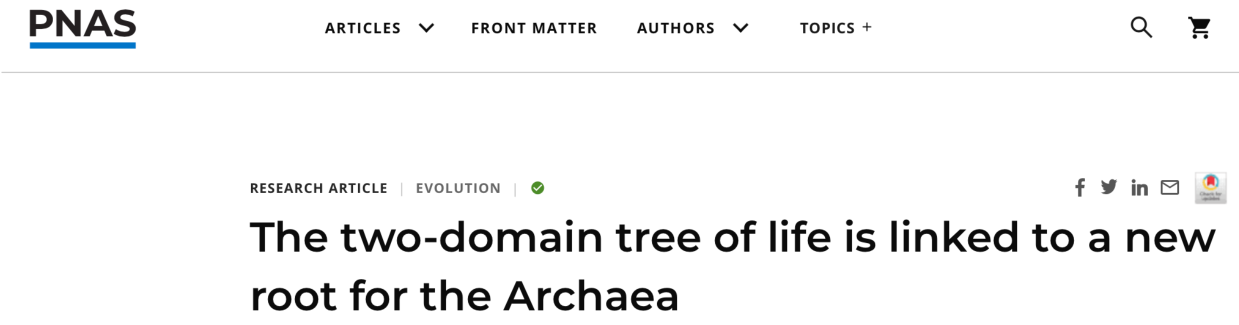
BioEssays

Problems and Paradigms | Full Access

Where is the root of the universal tree of life?

Patrick Forterre, Hervé Philippe

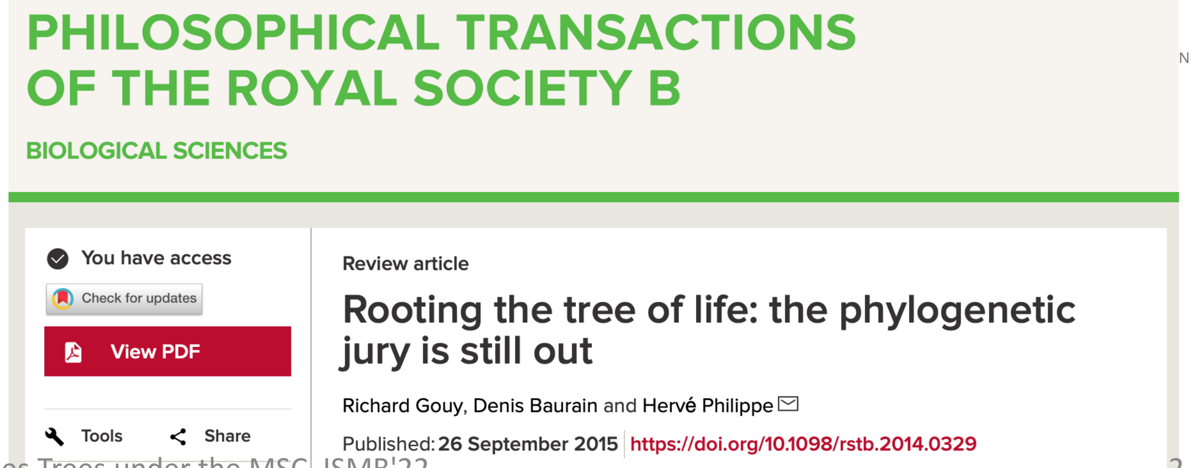
First published: 23 September 1999 | https://doi.org/10.1002/(SICI)1521-1878(199910)21:10<871::AID-BIES10>3.0.CO;2-Q | Citations: 151



RESEARCH ARTICLE | EVOLUTION |

The two-domain tree of life is linked to a new root for the Archaea

November 02, 2014)



PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

BIOLOGICAL SCIENCES

Review article

Rooting the tree of life: the phylogenetic jury is still out

Richard Gouy, Denis Baurain and Hervé Philippe

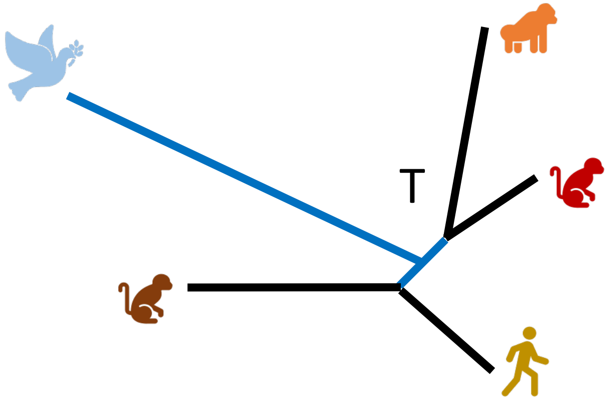
Published: 26 September 2015 | https://doi.org/10.1098/rstb.2014.0329

Current Approaches for Rooting Species Trees

Problem: Find the root position in a given unrooted species tree T.

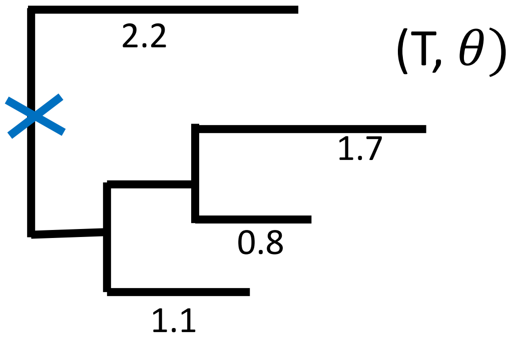
Outgroup Rooting

- Needs prior information about taxa
- Selecting a proper outgroup can be challenging



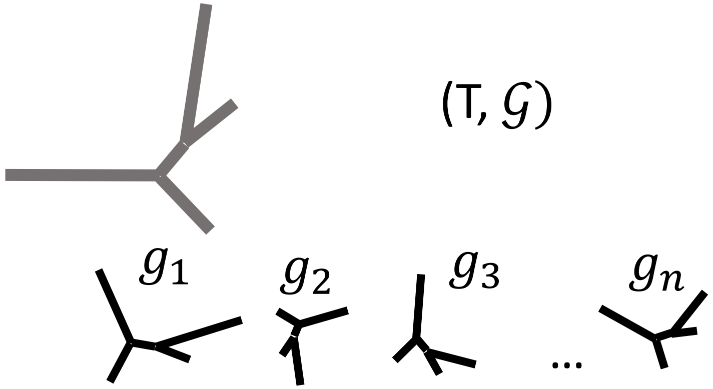
Distance-Based

- Species tree with branch lengths (e.g. Midpoint, MAD, MinVar, …)
- Most are sensitive to deviations from the molecular clock



Gene Based

- STRIDE (GDL-based): works with multi-copy gene trees
- Tian & Kubatko's site-based method [2017]: clock assumption



Most methods do not account for the biological processes that create discordance between gene trees and species trees.

Phylogenomics and Gene Tree Discordance

Causes of gene tree discordance:

- Incomplete Lineage Sorting (ILS)
- Gene Duplication and Loss (GDL)
- Horizontal Gene Transfer (HGT)
- …

Modeled by the Multi-Species Coalescent (MSC) model

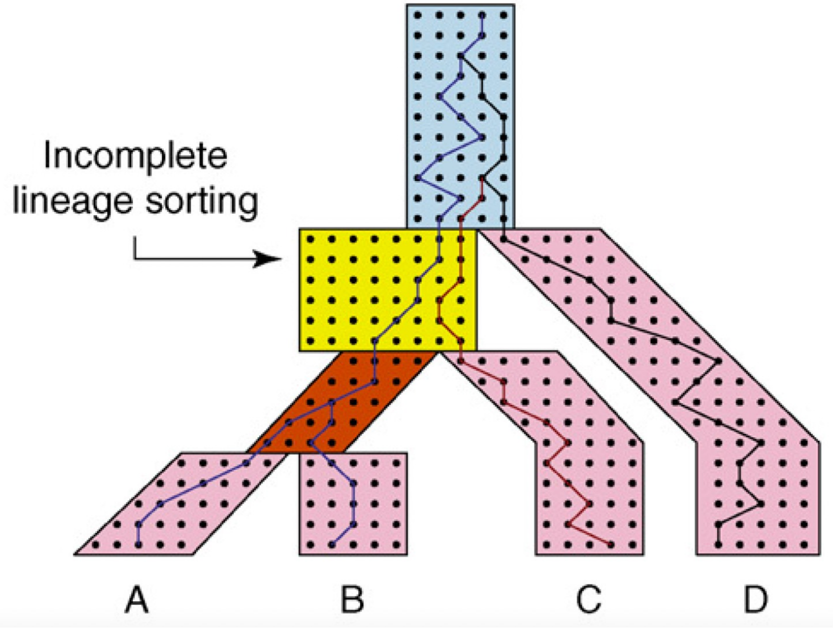


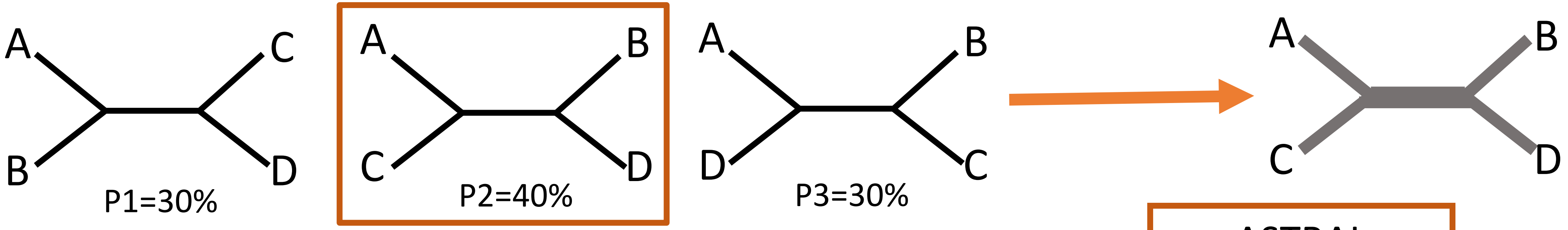
Image Credit: Degnan and Rosenberg, 2009, Trends in Ecology and Evolution

Identifiability of Unrooted Topology under MSC

Theorem: For 4 or more species, the unrooted topology of the species tree is identifiable from the probability distribution of the unrooted gene trees. [Allman, Degnan and Rhodes (ADR), J. Math. Biol, 2011]

Key property: For 4 species, the most probable unrooted gene tree has the same topology as the unrooted species tree

- Does not hold for more than 4 species



Quartet-Based species tree estimation methods →

ASTRAL
BUCKy-pop
wQFM
…

Identifiability of Rooted Topology under MSC

Theorem: For **5** or more species, the **rooted** topology of the species tree is identifiable from the probability distribution of the **unrooted** gene trees. [Allman, Degnan and Rhodes (ADR), J. Math. Biol, 2011]

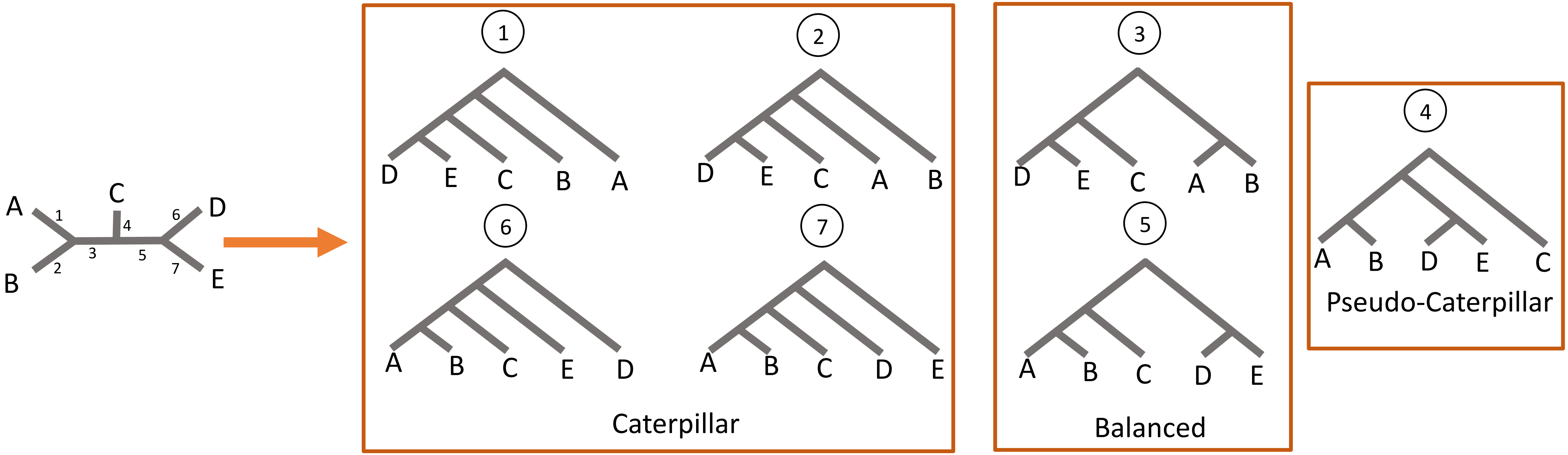
- ADR derive linear invariants and inequalities on the probability distribution of unrooted gene trees.
- They prove that these inequalities and invariants suffice to identify the rooted species tree

→ If the probability distribution of unrooted gene trees is exactly known, there will be exactly one rooted species tree topology satisfying all invariants and inequalities.

This result has not been used in any species tree estimation or rooting method before!

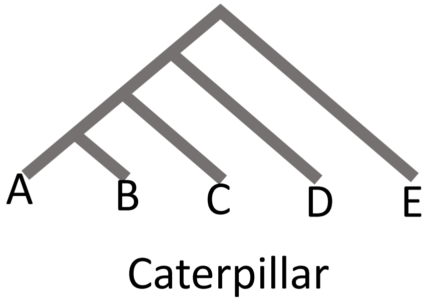
Properties of Quintet Trees

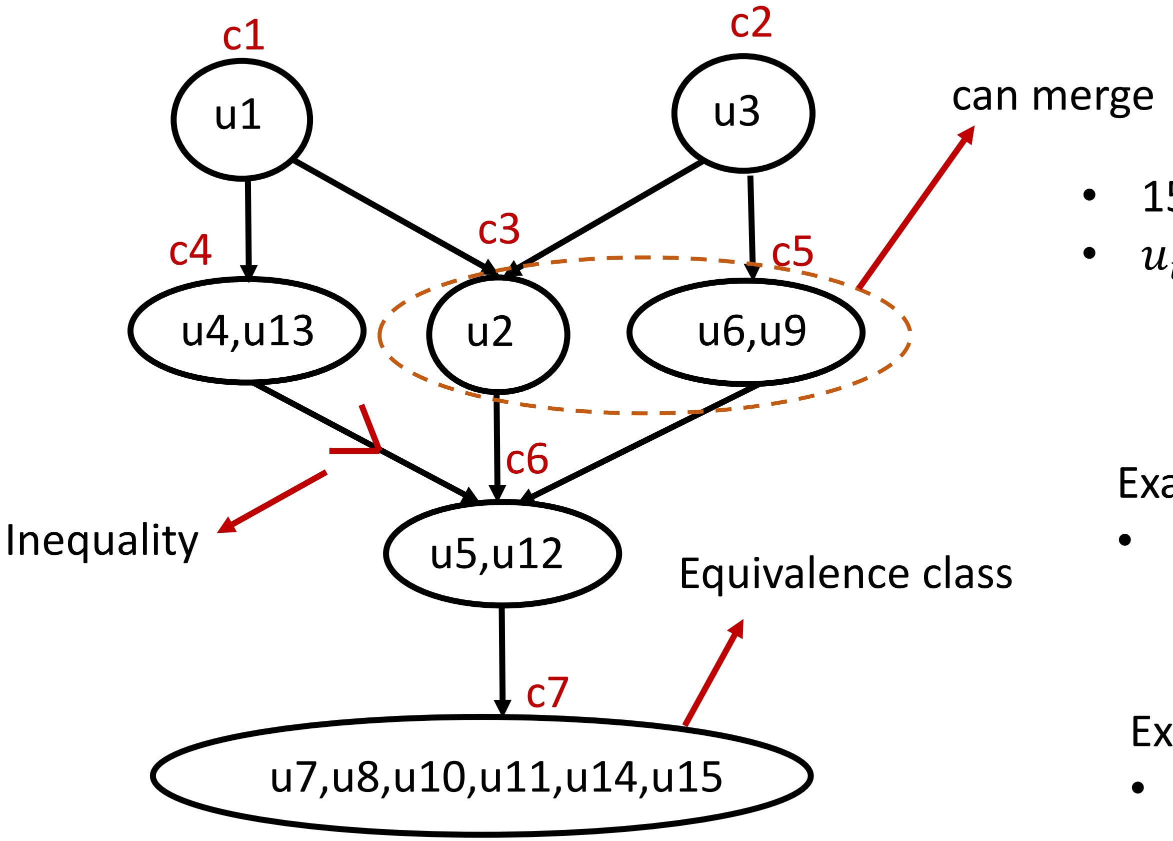
- There are **105** rooted binary trees and **15** unrooted binary trees on 5 taxa
- Each unrooted 5-taxon tree can be rooted on any of its **7** edges
- Rooted 5-taxon trees fall into **three** different shapes: caterpillar, balanced and pseudo-caterpillar [Rosenberg, 2007]



ADR Invariants & Inequalities

- ADR invariants and inequalities define a partial order on the distribution of unrooted gene trees $\vec{u}$
- The partial order for each tree shape can be shown with a Hasse diagram





- 15 5-taxon unrooted topologies $T_1, \ldots, T_{15}$
- $u_i = \mathbb{P}(T_i)$

Example of invariants:

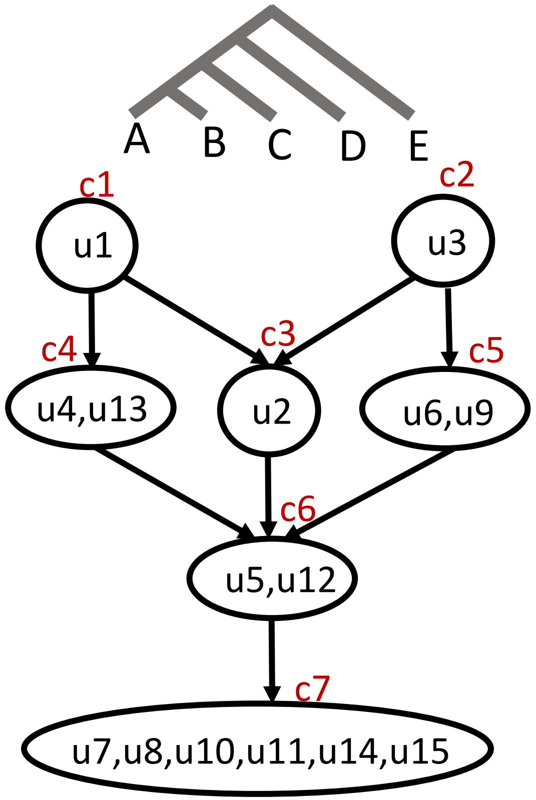
- $u_4 = u_{13}$

Example of inequalities:

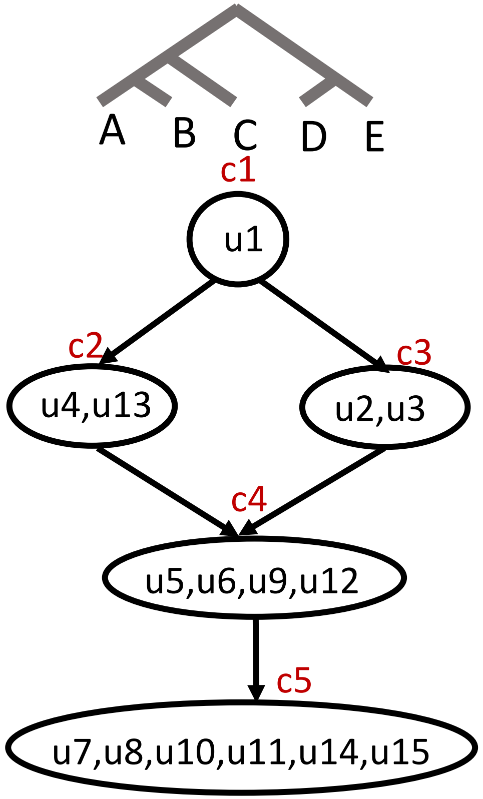
- $u_4 > u_5$

- Equivalence classes that are not related by inequalities can merge for some values of branch lengths

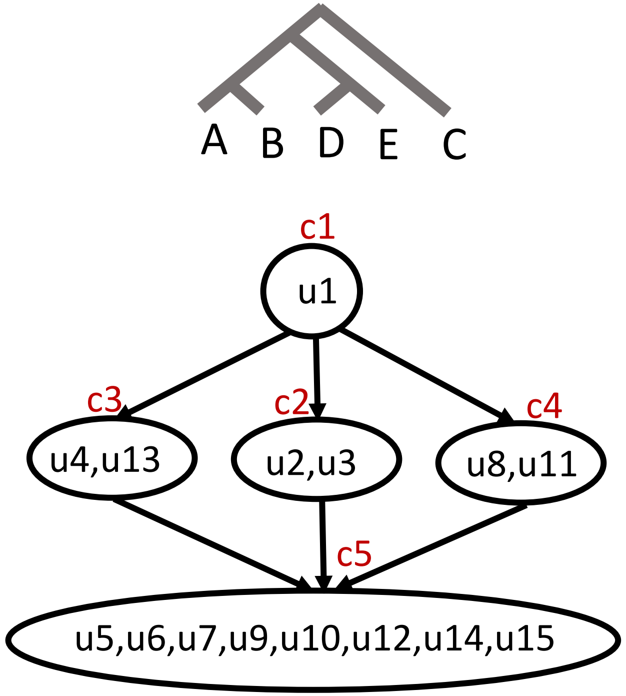
ADR Invariants & Inequalities



Caterpillar



Balanced



Pseudo-Caterpillar

- According to ADR theory, each 105 rooted binary tree corresponds to a unique Hasse diagram
- The shape of this diagram only depends on the topological shape of the tree
- The rooted tree is identifiable from the probability distribution $\vec{u}$ = $(u_1, u_2, \dots, u_{15})$

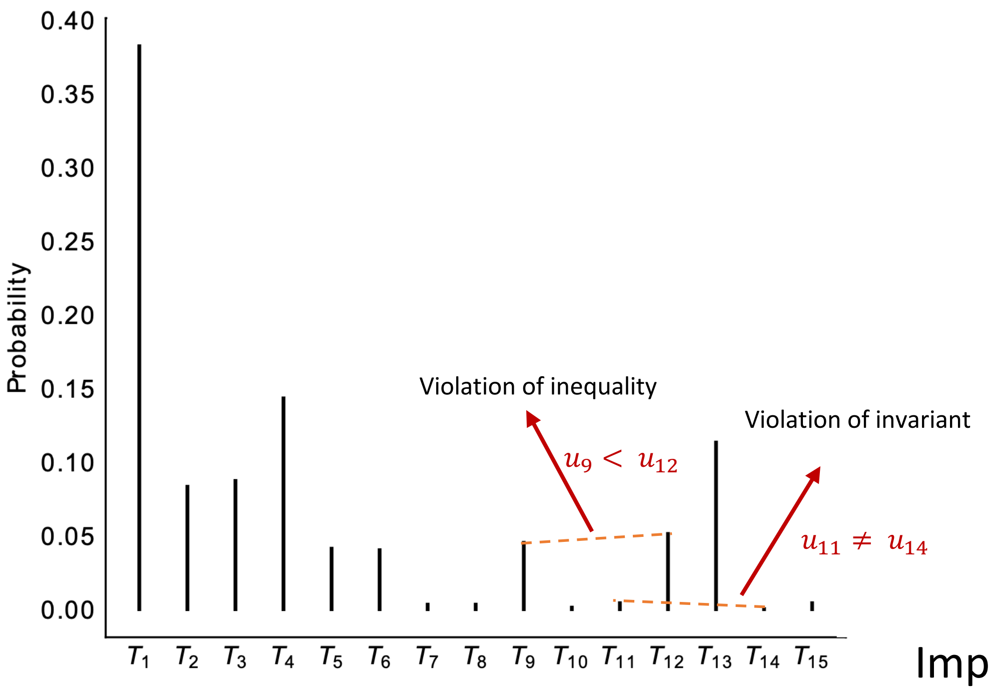
Our Question: How can we infer the model species tree when we’re given the estimated gene tree distribution?

- None of the invariants (i.e., equalities) exactly hold
- When probabilities are very close, even for good quality gene trees, the inequalities can be reversed
- In practice, gene tree estimation error add noise to the distribution

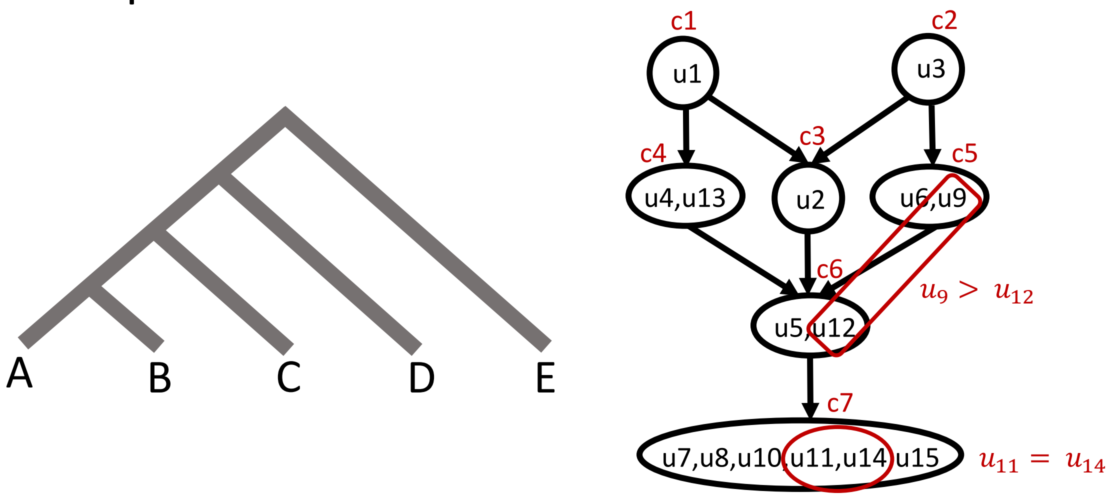
Question: Which of these partial orders better explain a given estimated distribution?

Fitness between a Tree and a Distribution

Idea: Penalizing each invariant and inequality in the partial order that is violated in the distribution



Estimated gene tree distribution $\vec{u}$



A model tree R and its partial order

Implied by the distribution:

- $u_{11} \neq u_{14}$
- $u_9 < u_{12}$

violations

Implied by the partial order:

- $u_{11} = u_{14}$
- $u_9 > u_{12}$

Cost Function $Cost(R, \vec{u})$**:**

- Measures the fitness between a distribution and a tree (i.e. its partial order)
- Linear combination of invariant and inequality penalty terms

Normalization: Correcting for Bias

- **Observation**: different tree shapes have different numbers of penalty terms
- The fitness cost might become biased towards one tree shape (Category Bias)

- **Idea**: Normalization by class sizes can help in practice!

$$Cost(R,\vec{\hat{u}}) = \underbrace{\sum_{c\in C_R} \frac{1}{|c|} \sum_{u_a,u_b\in c} |\hat{u}_a - \hat{u}_b|}_{\text{Invariants Penalty}} + \underbrace{\sum_{c>c'\in C_R} \frac{1}{|c'|} \sum_{u_a\in c,u_b\in c'} max(0,\hat{u}_b-\hat{u}_a)}_{\text{Inequalities Penalty}}$$

Normalization

Quintet Rooting: Rooting Species Trees under MSC

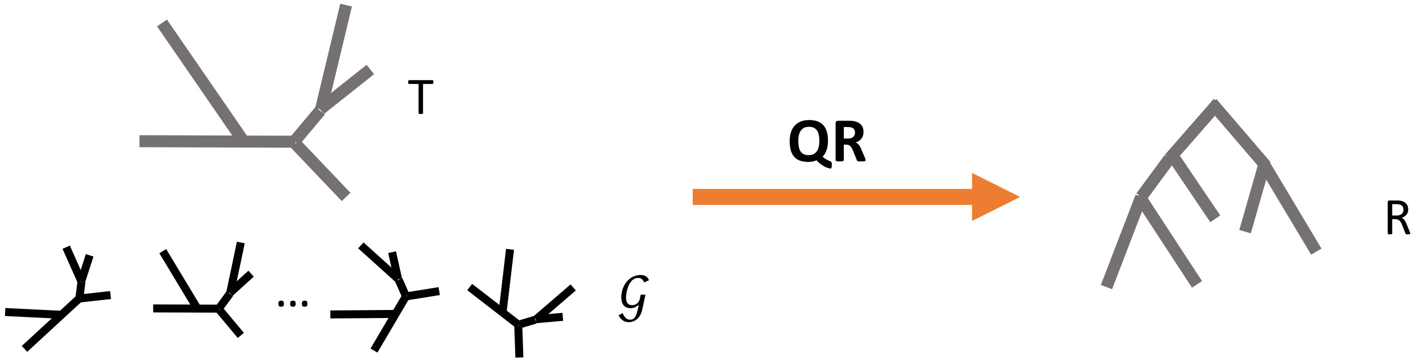
Input

- An unrooted species tree T.
- A set of k unrooted single-copy gene trees $\mathcal{G}$ on $\mathcal{L}(T)$.
- A cost function $Cost(R,\ \vec{u})$.

Output

- A rooted version of T that minimizes

  $Score(R,T) = \sum_{q \in Q^*} Cost(q, \overrightarrow{u_q})$



Quintet Rooting Algorithm

Rooting 5-taxon trees:

- Estimate the unrooted gene tree probability distribution $\vec{u}$ = $(u_1, u_2, \dots, u_{15})$
- For a given cost function $Cost(R,\ \vec{u})$, search all rooted versions of T to find $\hat{\mathrm{R}}$ such that

$$\hat{R}\ = argmin_R Cost(R, \vec{u})$$

Rooting Larger Trees:

- Select a set Q^* of quintets in the tree
  - at most $C(n, 5)$
- Preprocessing: Compute the cost for each quintet in set Q^*
- Score each of the $2n - 3$ rooted versions of T and return the one with the least score

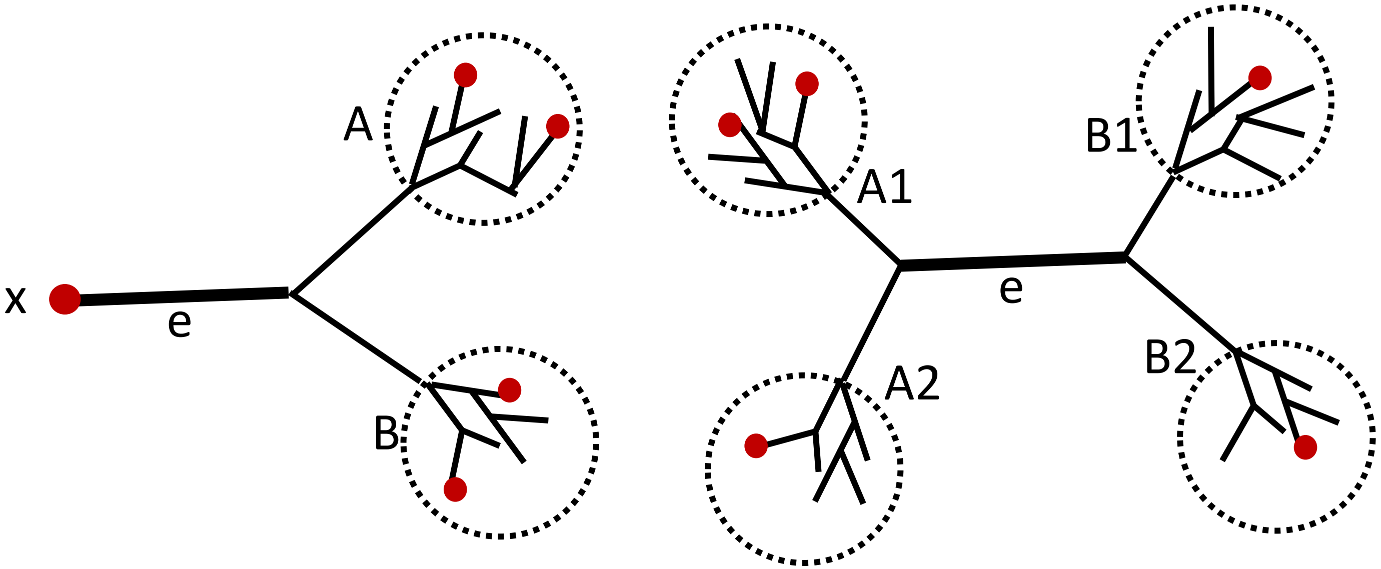
$$Score(R, T) = \sum_{q \in Q^*} Cost(q, \overrightarrow{u_q})$$

Runtime: $O(k|Q^*|) \xrightarrow{\text{using all quintets (default)}} O(n^5 k)$

Linear Encoding: Scaling QR to Larger Trees

- Reducing the size of set Q^*
  - runtime: $O(k|Q^*|)$
- Linear Encoding
  - Scoring a quintet for each edge in the unrooted topology T
- Reduces the runtime from $O(n^5k)$ to $O(nk)$

$|Q^*| = 2n - 3$



Lemma: For an MSC species tree R with n taxa, the root of R is identifiable from its unrooted topology T and the correct rootings of quintets in the set $Q_{LE}\ (T)$.

Experimental Study

Simulated Datasets:

- Avian (test) and Mammalian (training) simulated datasets [Mirarab et al., 2014]
- 800 to 1000 genes, Varying GTEE *, 5 to 30-leaf subsets, moderately high ILS

Rooting Methods:

- Midpoint
- MinVar [Mai et al., 2017]
- Minimum Ancestor Deviation (MAD) [Tria et al., 2017]
- RootDigger [Bettisworth and Stamatakis, 2021]
- Quintet Rooting

Biological Dataset:

- Avian biological dataset [Jarvis et al., 2014]
- 48 species, ~14500 genes, 5-leaf subsets

Pipeline:

- Branch length estimation: RAxML [Stamatakis, 2014] on concatenated sequence alignments
- Species tree estimation: ASTRAL [Zhang et al, 2018]

Evaluation Criteria:

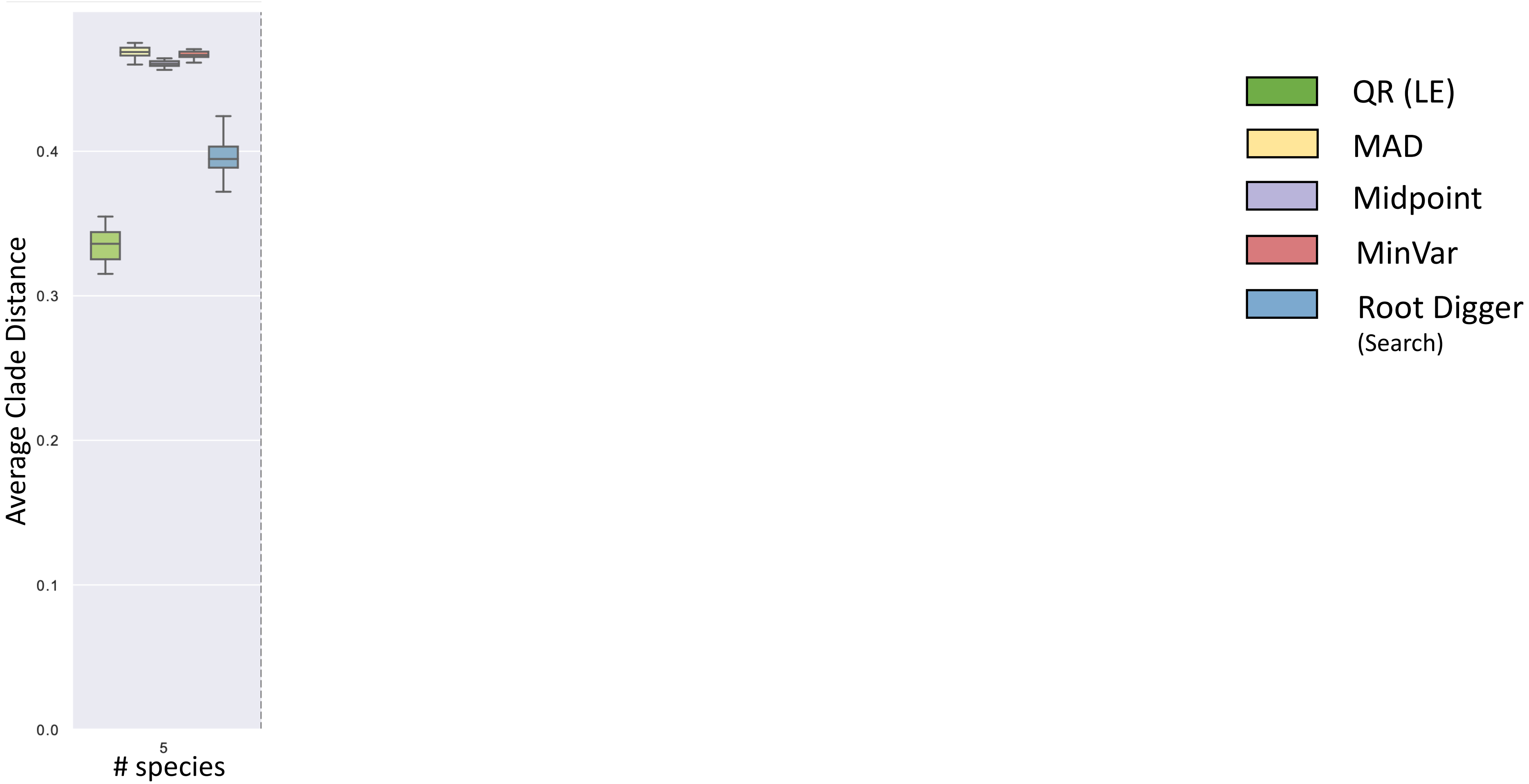
- Average normalized clade distance

* Gene tree estimation error

Comparison between Rooting Methods: Varying number of Species

- Avian simulated datasets, 5-30 species, 1000 genes, 1000bp sequences, 39% GTEE rate, 1X ILS, 200 samples, true species tree

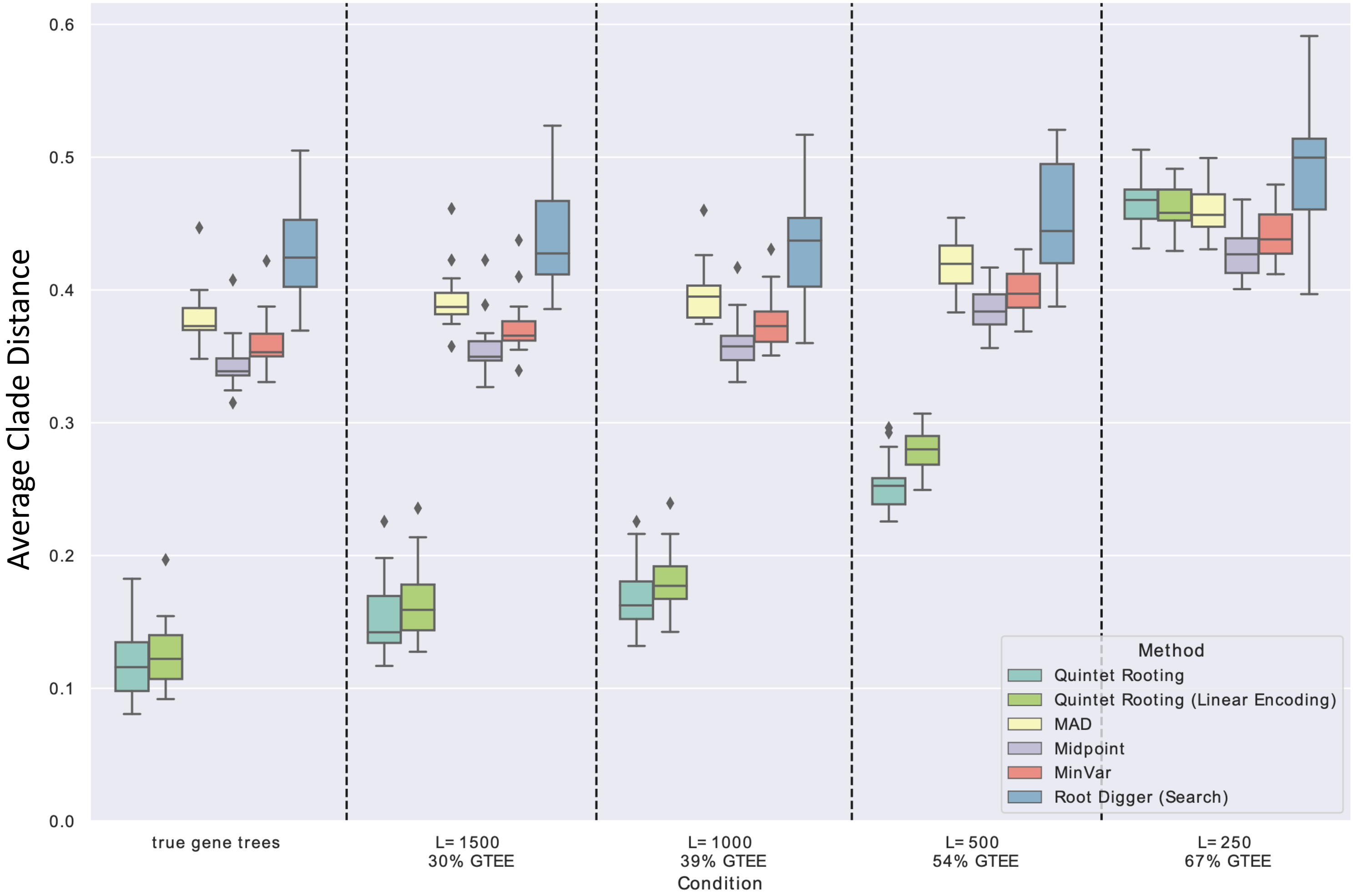
- Most methods except QR (LE) benefit from denser taxon sampling



Comparison between Rooting Methods: Varying rate of GTEE

- Avian simulated 10-taxon datasets, 1000 genes, Varying GTEE rates, 1X ILS, estimated (ASTRAL) species tree, 200 samples

- QR(LE) is slightly less accurate than QR, and both outperform other methods except under highest level of GTEE



% GTEE

🟩 QR

🟩 QR (LE)

🟨 MAD

🟪 Midpoint

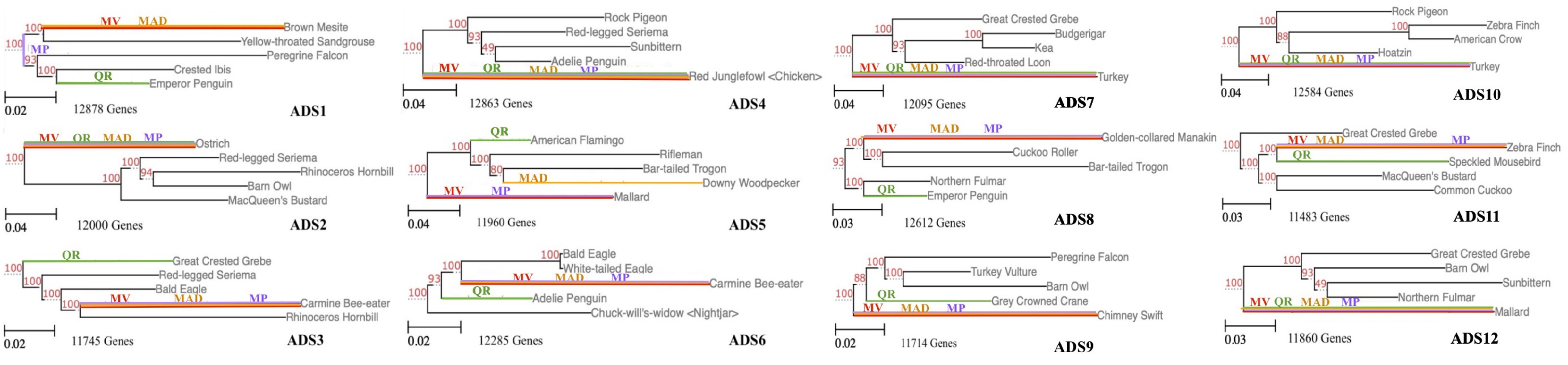
🟥 MinVar

🟦 Root Digger (Search)

Results on 5-leaf Subsets of the Avian Biological Dataset

- Analysis on 12 5-leaf subtrees of the avian biological dataset [Jarvis et al, 2014] TENT tree
- Outgroup rooting as the ground truth
- 11k-13k gene trees in each data subset
- Very high GTEE

Dataset	No. of genes	Quintet rooting	Midpoint	MinVar	MAD
Average	∼12 173	0.22	0.22	0.25	0.33



Summary & Future Directions

Summary

- Quintet Rooting is an ILS-aware method for rooting species trees
- It is designed based on a theoretical result of identifiability of rooted 5-taxon trees established by ADR
- In this study, it has shown better or comparable accuracy to existing rooting methods except under very high levels of GTEE

Future Directions

- Explore rooting methods under other model conditions, considering other causes of gene tree discordance (GDL, HGT, etc)
- Explore larger datasets (more taxa and more genes)
- Statistical Consistency
- Inferring rooted trees from unrooted gene trees

Acknowledgements

Thank you!

Tandy Warnow

Kowshika Sarker

Quintet Rooting (QR) is available on Github:
https://github.com/ytabatabaee/Quintet-Rooting

Paper is available at:
https://doi.org/10.1093/bioinformatics/btac224

Members of Warnow Lab

Funding:
CS@ILLINOIS
ISMB'22 Virtual Fellowship Award

COMPUTER SCIENCE

iSCB INTERNATIONAL SOCIETY FOR COMPUTATIONAL BIOLOGY

Computing Resources:
UIUC Campus Cluster

Backup Slides

Distance between two Rooted Trees

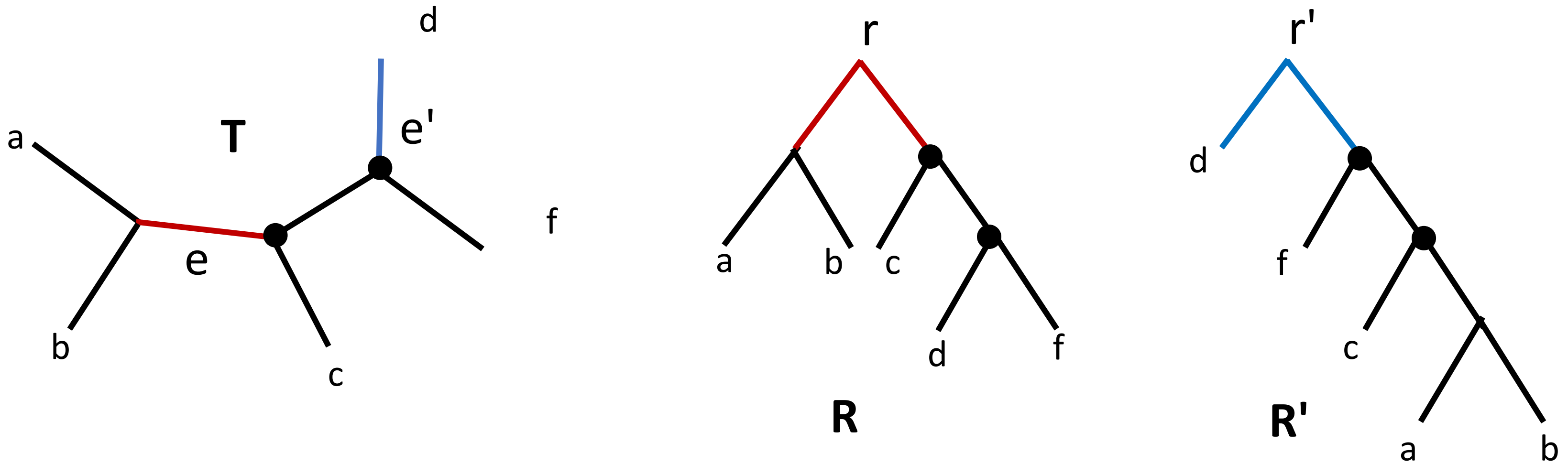
Root Distance:

Length of path between two root edges

Clade Distance:

|Clades(R)∆Clades(R’)| =
|Clades(R)\Clades(R’)| + |Clades(R’)\Clades(R)|

Lemma: For rooted binary trees R and R’ with unrooted topology T, we have CD(R,R’) = 2RD(R,R’).



RD(R,R') = 2

CD(R,R') = |{{a,b,c}, {a,b,c,f}, {d,f}, {c,d,f}}| = 4