PHYLOGENETICS · SPECTRAL METHODS · COMPUTATIONAL BIOLOGY

arXiv preprint
2603.10215

SDSR: A Spectral Divide-and-Conquer Approach for Species Tree Reconstruction

A scalable, statistically consistent method that leverages spectral graph theory to reconstruct species trees from multi-gene sequence data.

Ortal Reshef[1] **Ofer Glassman**[2] **Or Zuk**[1] **Yariv Aizenbud**[3] **Boaz Nadler**[2] **Ariel Jaffe**[1]

[1] Dept. of Statistics & Data Science, Hebrew University of Jerusalem [2] Dept. of CS & Applied Mathematics, Weizmann Institute of Science [3] Dept. of Applied Mathematics, Tel Aviv University

PRESENTER ortal.reshef@mail.huji.ac.il ADVISOR ariel.jaffe@mail.huji.ac.il

Motivation

- Species tree reconstruction from **multiple genetic markers** is fundamental to evolutionary biology.
- Modern datasets involve **thousands of species** and hundreds of genes, demanding scalable algorithms.

Key Challenges

Gene tree discordance: Due to Incomplete Lineage Sorting (ILS) and Horizontal Gene Transfer (HGT), individual gene trees differ from the true species tree, a single-gene approach is insufficient.

Scalability: Existing methods, such as ASTRAL and concatenation-based maximum likelihood analysis (CA-ML), can be computationally slow on large datasets.

Divide-and-conquer approaches: To handle datasets with thousands of species, divide-and-conquer methods break the problem into smaller tasks by dividing species into subsets, reconstructing a subtree for each, and merging results using a supertree algorithm.

Merging complexity: Prior divide-and-conquer approaches require solving **NP-hard** optimization during the merge step.

Our Approach: SDSR

SDSR is a **recursive spectral divide-and-conquer** method. It uses the **Fiedler eigenvector** of the averaged Graph Laplacian to bipartition species into groups that correspond to disjoint clans in the species tree.

NORMALIZED GRAPH LAPLACIAN PER GENE

$$L^g = I - (A^g)^{-0.5}\, S^g\, (A^g)^{-0.5}$$

where I is the identity matrix, A^g is the diagonal degree matrix, and S^g is the similarity matrix for gene g.

AVERAGED LAPLACIAN

$$\bar{L} = \frac{1}{K}\sum_{g=1}^{K} L^g$$

The average over all K gene trees yields a single Laplacian whose Fiedler eigenvector determines the bipartition.

Theoretical Guarantees

Our partitioning approach is **consistent under the MSC model**. When the number of gene alignments $K \to \infty$, species assigned to the same partition correspond to **clades in the species tree**.

Assuming that all observed species are equidistant to their root, we prove that the number of gene alignments K required to guarantee, with high probability, partitioning correctness scales as $O(m^2)$.

Main Advantages of SDSR

Unlike prior methods, SDSR produces **deterministic partitions**, enabling a provably correct and efficient merge step, **no NP-hard problem required**.

Any existing species tree method (CA-ML, ASTRAL, etc.) can serve as the **base subroutine** for reconstructing small subtrees.

SDSR is **trivially parallelizable**, independent subtree reconstructions can run on separate CPUs.

Complexity reduction: SDSR reduces the dominant ML search complexity from $O(m^2 n)$ to $O(\tau \cdot m \cdot n)$, where m is the number of taxa, n is the sequence length or number of sites, and τ is the recursion/partitioning threshold.

The SDSR Algorithm, Three-Step Pipeline

Given K gene alignments for m species, SDSR recursively applies the following steps until each subset is below a size threshold τ:

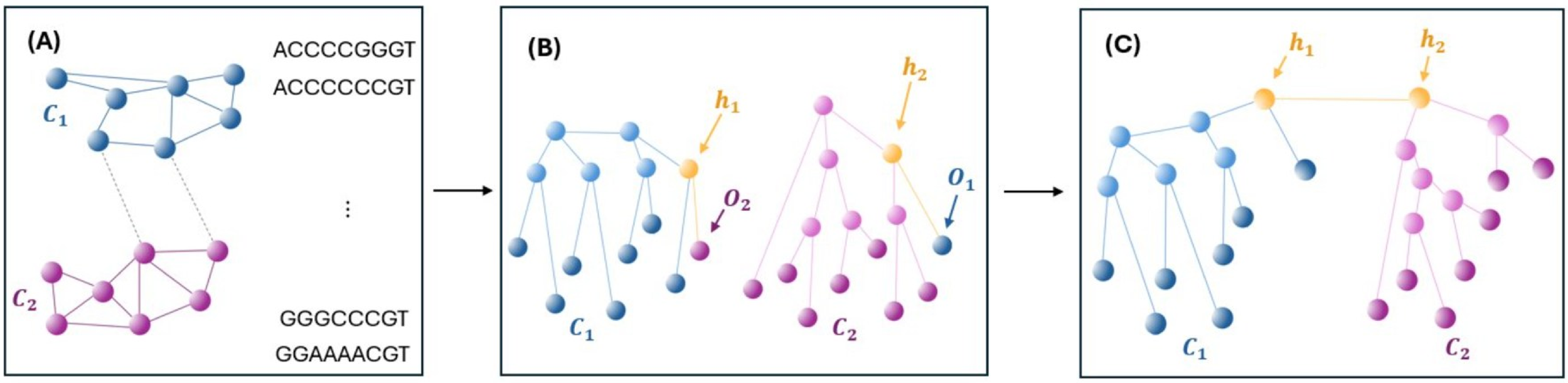
STEP 1 — **Partitioning & Add Outgroups**
Compute the averaged Graph Laplacian $\bar{L}$ across all genes. Use its **Fiedler vector** (2nd smallest eigenvector) to bipartition species into two groups. Add each partition one outgroup node from the other side.

→

STEP 2 — **Reconstruct**
Apply a user-chosen species tree method (e.g., CA-ML or ASTRAL) to reconstruct the subtree for each partition + outgroups.

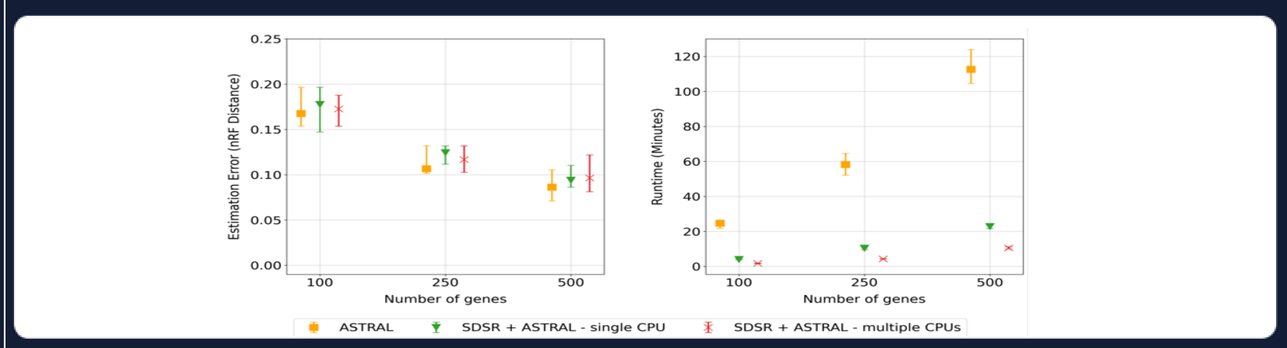
→

STEP 3 — **Reroot & Merge**
Reroot each subtree using the outgroup nodes to identify the correct root placement. Then prune outgroups and deterministically merge the two subtrees.

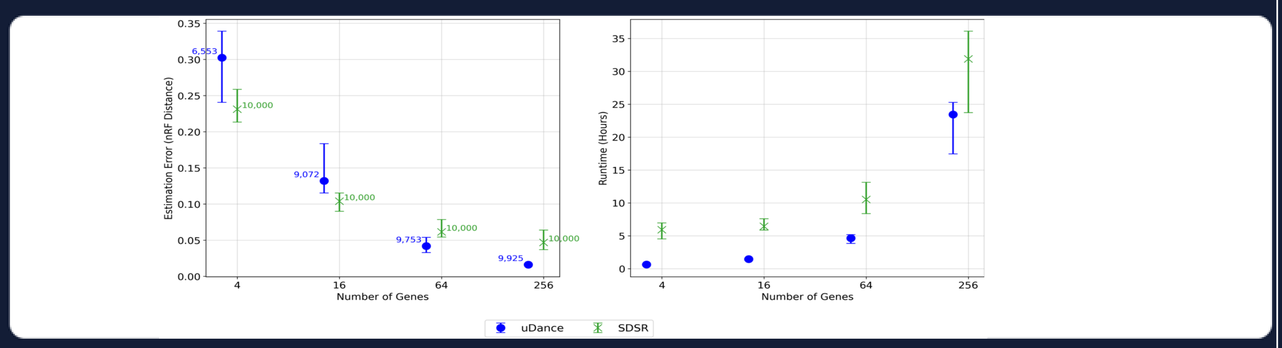


Results

Evaluated on simulative datasets with ILS and HGT



SDSR + ASTRAL vs. ASTRAL
200 species
(Molecular Biology and Evolution, 2022)



SDSR vs. uDance
10,000 species
(Nature biotechnology, 2024)

SDSR matches **ASTRAL** accuracy with reduced runtime, where ASTRAL recovers the full trees. **Outperforms uDance**, a divide-and-conquer approach, **using fewer genes to reconstruct** and **guarantees recovery of a tree spanning all species**.

Conclusions & Impact

- SDSR provides a **principled, spectral approach** to scalable species tree reconstruction that avoids NP-hard merging problems.
- Proven **statistical consistency** under the MSC model, with finite-sample guarantees for correct partitioning.
- Flexible design: works as a **meta-algorithm** that accelerates any existing species tree method (CA-ML, ASTRAL, etc.).
- Enables reconstruction of phylogenies for **thousands of species** on modest computational resources.
- Opens new directions for spectral methods in large-scale phylogenomics.

Key References

[1] Aizenbud et al. (2023), Spectral top-down recovery of latent tree models. *Information and Inference*.
[2] Zhang et al. (2018), ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*.
[3] Stamatakis (2014), RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*.
[4] Roch & Warnow (2015), On the robustness to gene tree estimation error of coalescent-based species tree methods. *Systematic Biology*.
[5] Molloy & Warnow (2019), TreeMerge: A new method for improving the scalability of species tree estimation methods. *Bioinformatics*.
[6] Balaban et al. (2024), Generation of accurate, expandable phylogenomic trees with uDance. *Nature biotechnology*.