Differentiable Search of Evolutionary Trees

Ramith Hettiarachchi[1], Avi Swartz[2], Sergey Ovchinnikov[3]

[1]FAS Division of Science, Harvard University, [2]Molecular and Cellular Biology Program, University of Washington
[3]JHDSF Program, Harvard University

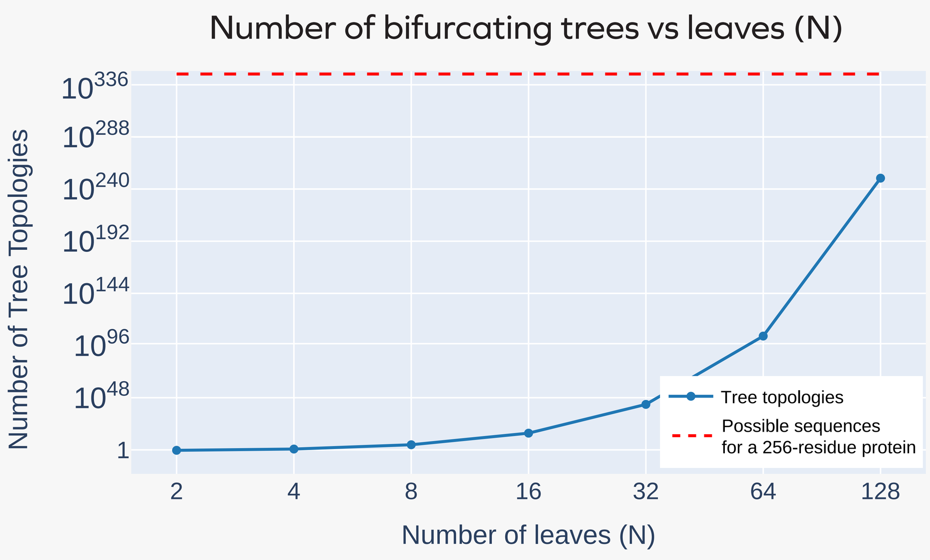
ICML International Conference On Machine Learning

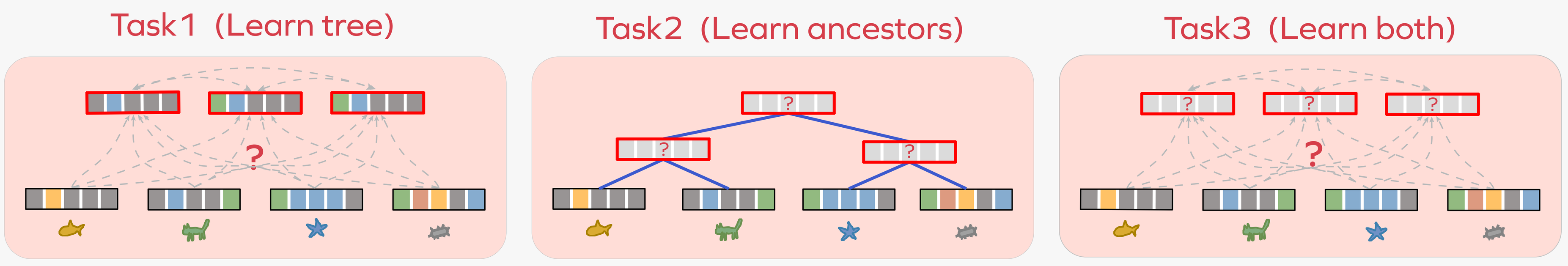
I'm looking for PhD opportunities for Fall 2024

TL;DR: We introduce a differentiable approach to search for phylogenetic trees.
We optimize the tree and ancestral sequences to reduce the total evolutionary steps (parsimony cost).

1 Introduction

- Evolutionary trees are used in various fields of science.
- Inferring the most parsimonious tree given leaves is a NP-hard problem.
- Due to this complexity, existing work consider heuristic search techniques.

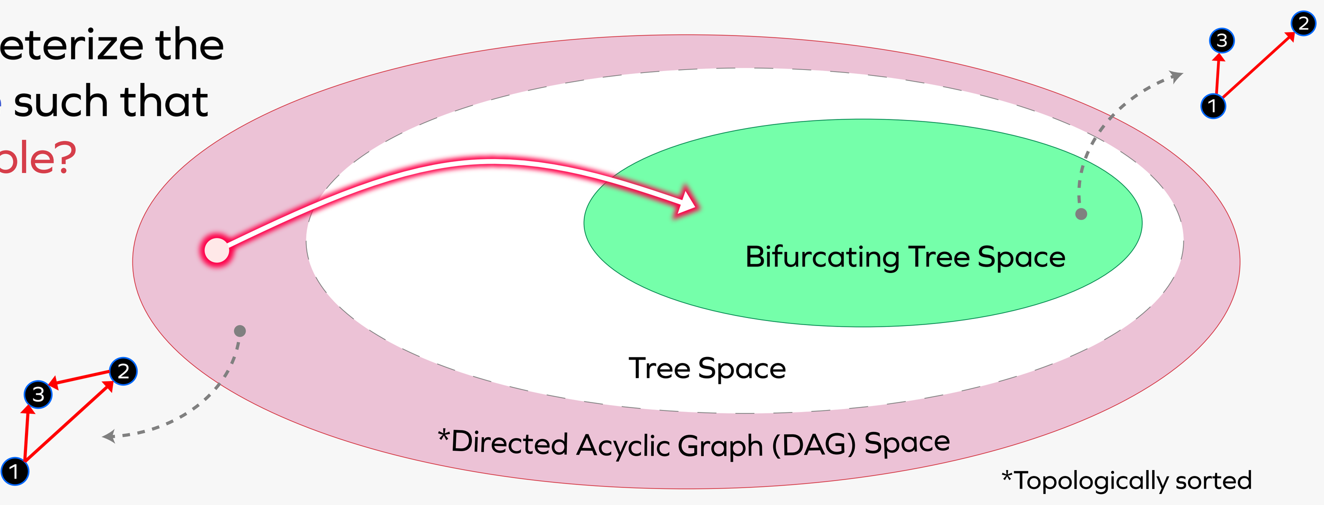




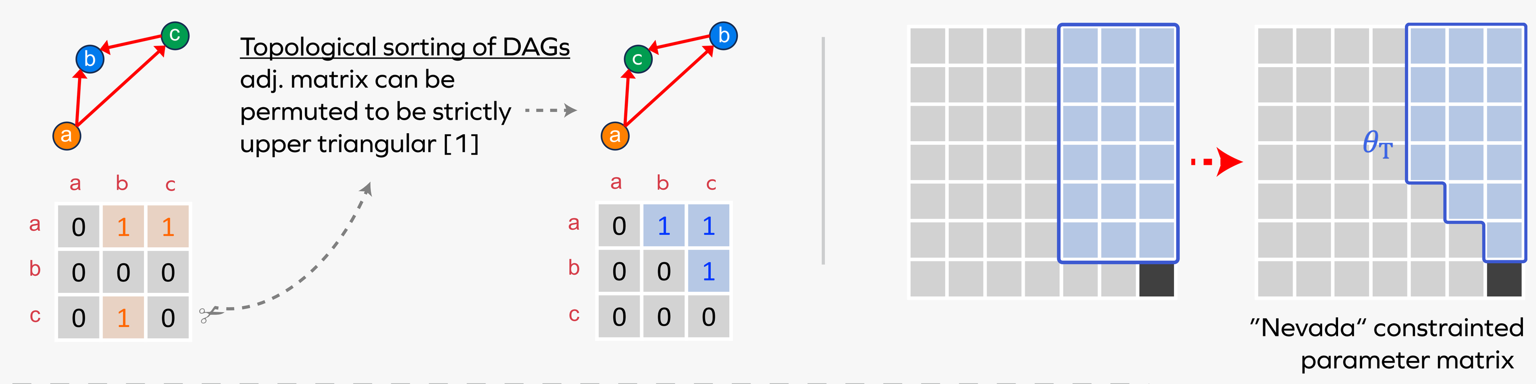
(small parsimony problem) (maximum parsimony problem)

2 Methodology

- Can we parameterize the tree (θ_T) space such that it is differentiable?



- How can we prevent cycles in our search space ? constraint to DAG space!



1) Obtain a probability distribution over the parents of each node.

$$A_{ij} = \frac{e^{\theta_{T_{ij}}/\tau_1}}{\sum_{k=1}^{N-1} e^{\theta_{T_{ik}}/\tau_1}}$$

2) Enforce bifurcating trees by regularizing the loss.

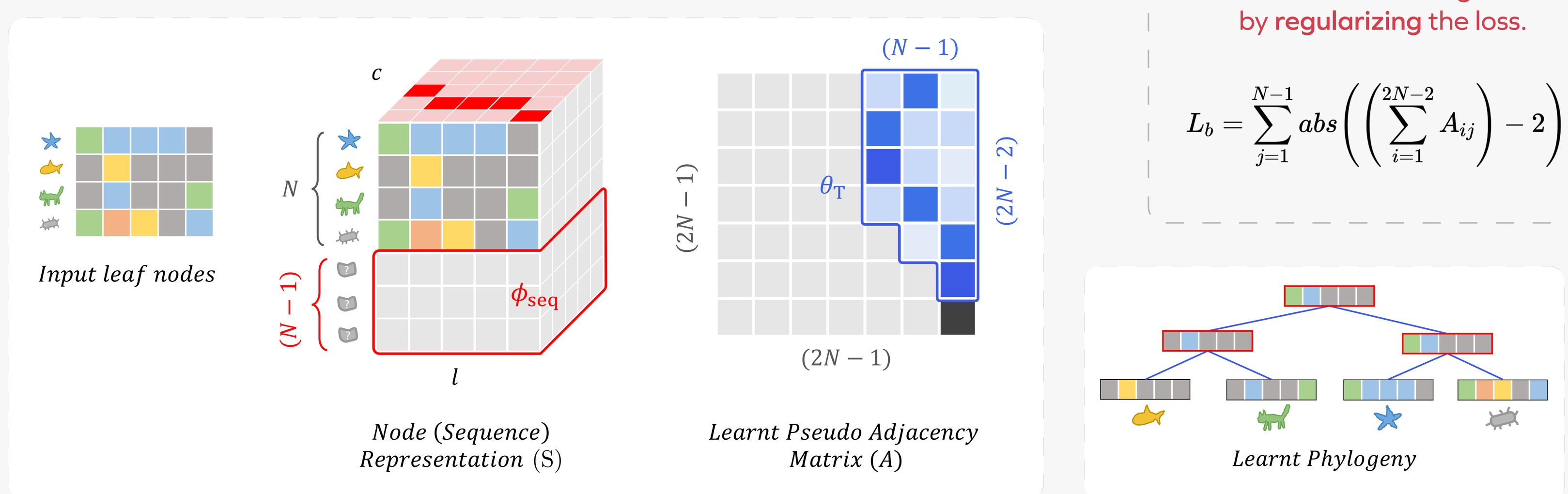
$$L_b = \sum_{j=1}^{N-1} abs\left(\left(\sum_{i=1}^{2N-2} A_{ij}\right) - 2\right)^2$$

- Making the sequence (ϕ_{seq}) space differentiable

Discrete nature of the categorical choices in the sequence representation is relaxed similarly by obtaining a probability distribution over the character space at each position.

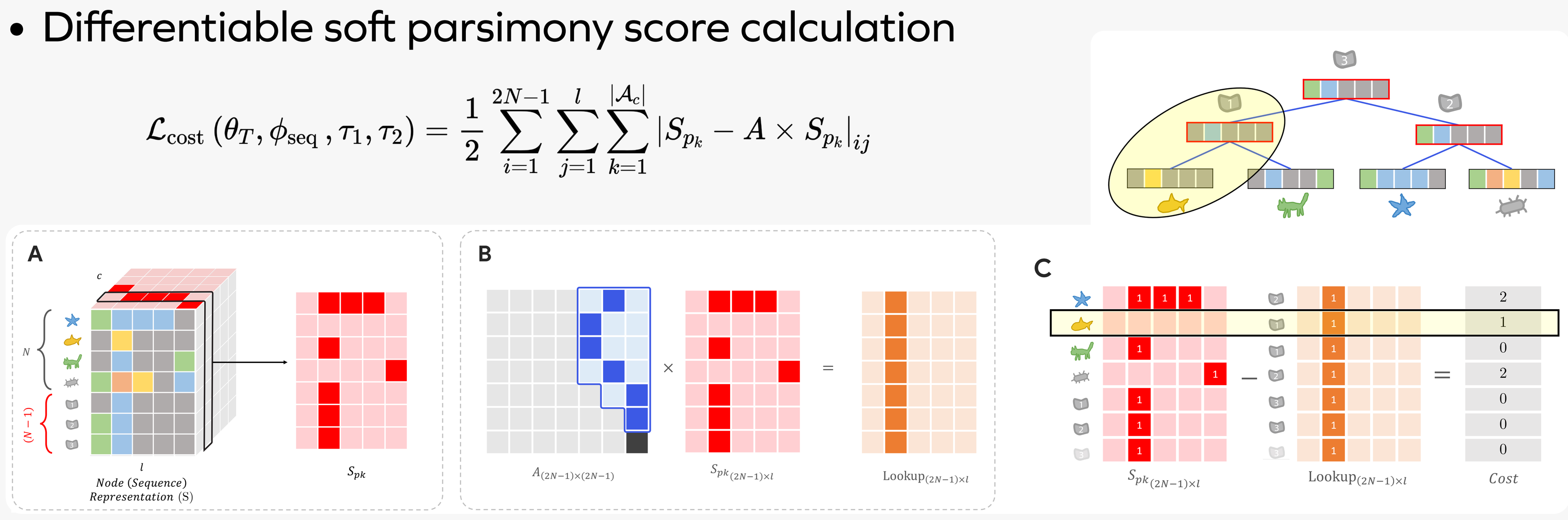
$$\hat{\phi}_{seq_{ijk}} = \frac{e^{\phi_{seq_{ijk}}/\tau_2}}{\sum_{m=1}^{c} e^{\phi_{seq_{ijm}}/\tau_2}}$$

- Therefore, we have the following two representations,

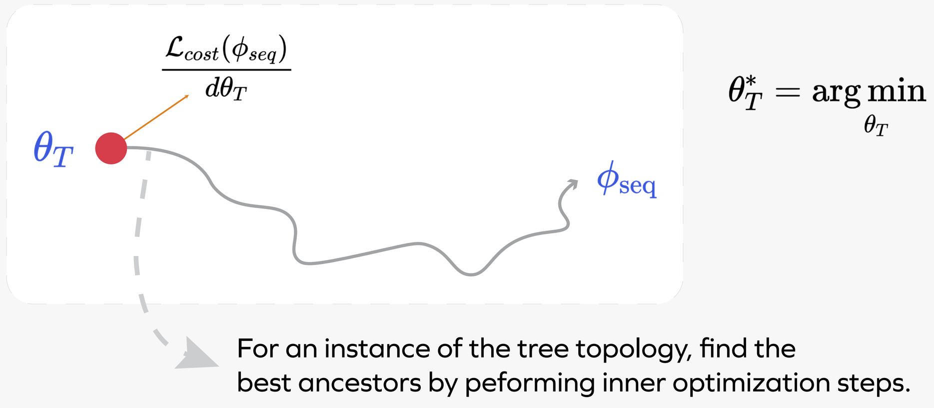


- Differentiable soft parsimony score calculation

$$\mathcal{L}_{cost}(\theta_T, \phi_{seq}, \tau_1, \tau_2) = \frac{1}{2}\sum_{i=1}^{2N-1}\sum_{j=1}^{l}\sum_{k=1}^{|\mathcal{A}_c|} |S_{p_k} - A \times S_{p_k}|_{ij}$$



- Bi-level optimization to find ancestors and tree



$$\theta_T^* = \arg\min_{\theta_T} \mathcal{L}_{cost}(\phi^*_{seq}(\theta_T), \theta_T) + \lambda L_b$$

$$\text{s.t. } \phi^*_{seq}(\theta_T) = \arg\min_{\phi_{seq}} \mathcal{L}_{cost}(\phi_{seq}, \theta_T) + \lambda L_b$$

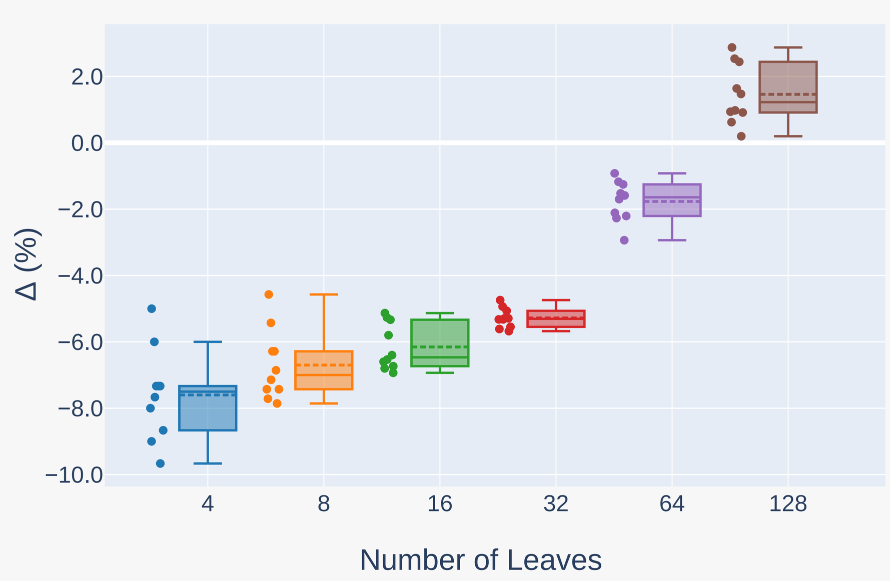
ϕ^*_{seq} is approximated by inner k gradient descent steps.

3 Results

- We compare the converged tree and ancestor solutions to the simulated solutions and the optimal solutions of tasks 1-3.

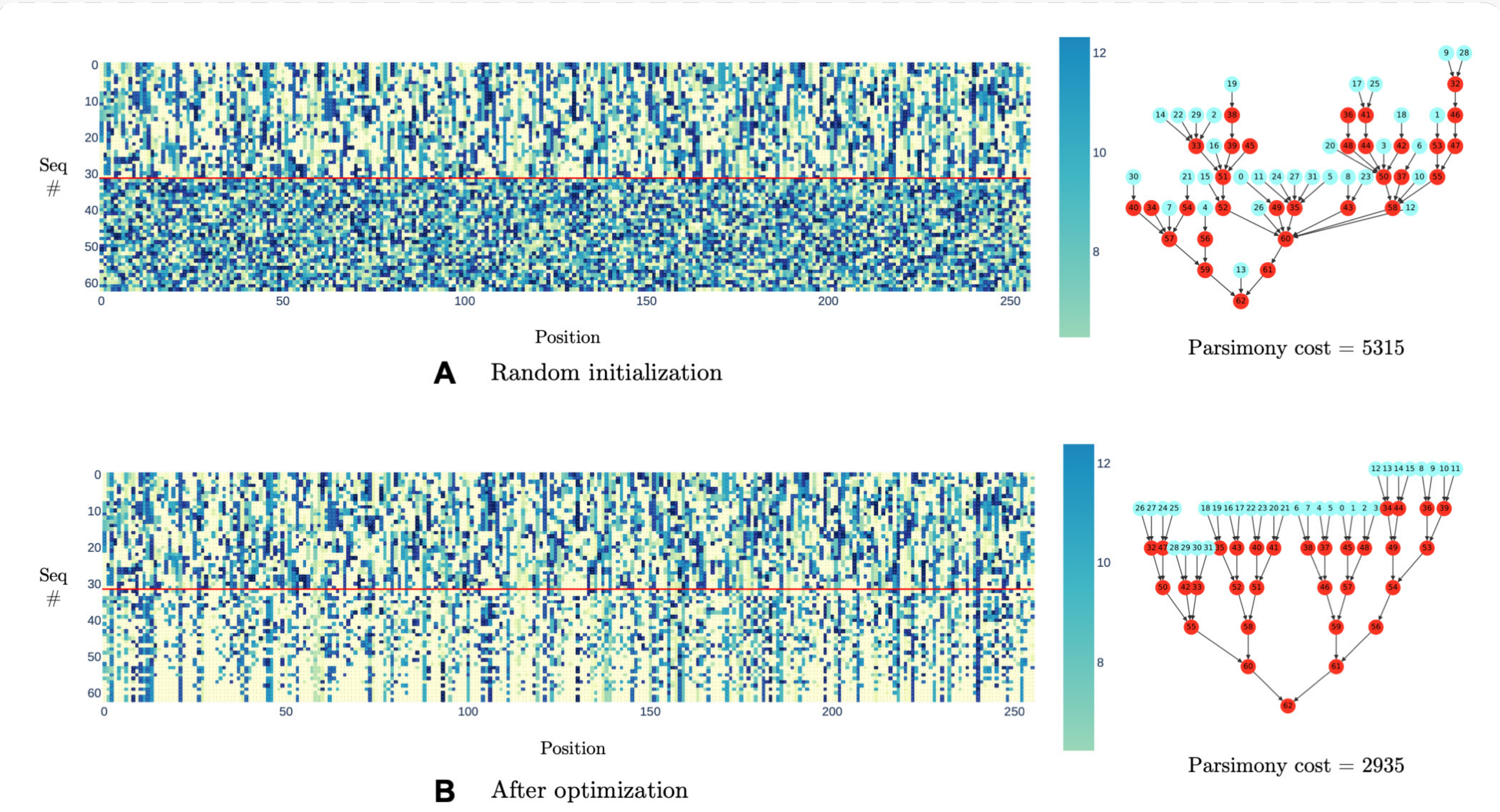
Table 1. Convergence analysis on the 3 tasks

Tree Complexity			Task 1	Task 2 (find seq given tree)			Task 3 (find both tree and seqs)		
N	Simulated solution	Mean optimal solution	Mean error	Mean solution	Mean error	Mean error as a % w.r.t optimal solution	Mean solution	Mean error	Mean error as a % w.r.t optimal solution
4	300	277.2	0.0	277.2	0.0 ± 0.0	0.000%	277.2	0.0 ± 0.0	0.000%
8	700	653.1	0.0	653.1	0.0 ± 0.0	0.000%	653.1	0.0 ± 0.0	0.000%
16	1500	1407.6	0.0	1407.6	0.0 ± 0.0	0.000%	1407.7	0.1 ± 0.3	0.007%
32	3100	2915.4	0.0	2915.4	0.0 ± 0.0	0.000%	2936.3	20.9 ± 7.4	0.717%
64	6300	5929.3	0.0	5929.3	0.0 ± 0.0	0.000%	6188.6	259.3 ± 27.4	4.373%
128	12700	11971.1	0.0	11971.3	0.2 ± 0.4	0.001%	12885.5	914.4 ± 99.6	7.638%



- Example experiment : (32 leaves, 256 length sequence)

(left : sequences, right : tree topology. optimal solution cost = 2913)



4 Conclusions & Future Work

- New approach for generating evolutionary trees by traversing a soft tree and sequence space.
- Even though task 2 can be solved with dynamic programming, it assumes site-wise independence. Yet, our method allows for lifting this restriction.
- This will allow the integration of distance calculations that model higher-order dependence, such as potts and protein language models.

Check out Our Github Page

References

[1] Nicholson, V. A. Matrices with permanent equal to one. Linear Algebra and Its Applications, 12(2), 1975.
[2] Felsenstein, J. Inferring phylogenies, volume 2. Sinauer associates Sunderland, MA, 2004.
[3] Sankoff, D. Minimal Mutation Trees of Sequences. SIAM Journal on Applied Mathematics, 28(1), 1975.
[4] Blondel, M., Berthet, Q., Cuturi, M., Frostig, R., Hoyer, S., Llinares-Lopez, F., Pedregosa, F., and Vert, J.-P. Efficient and modular implicit differentiation, 2021.