PhyloGen: Language Model-Enhanced Phylogenetic Inference via Graph Structure Generation

ChenRui Duan,
Zelin Zang, Siyuan Li, Yongjie Xu
Stan Z. Li

Zhejiang University, College of Computer Science and Technology;

Westlake University

NEURAL INFORMATION PROCESSING SYSTEMS

Background

(a) The inputs are aligned sequences, and topologies are learned from existing tree structures using methods like SBNs, which rely on MCMC-based methods for pre-generated candidate trees without considering branch lengths directly.
(b) The inputs are aligned sequences, and then tree structures and branch lengths are directly inferred by variational inference and biological modules. These methods optimize tree topology and branch lengths separately.
(c) The inputs are raw sequences processed by a pre-trained language model to generate species representations. Then, an initial topology is generated through a tree construction module, and the topology and branch lengths are co-optimized by the tree structure modeling module.

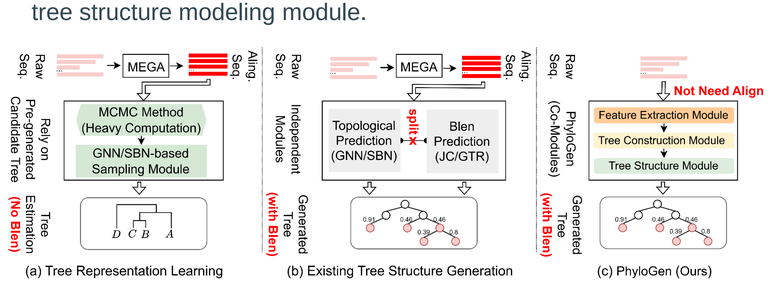


Figure1: Background: Comparison of PhyloTree Tree Inference Methods .

Our Method

Reformulate as graph generation problem.

As depicted in Fig.1 (c), we propose a novel approach based on a pre-trained genome language model. Our model does not rely on evolutionary models or the requirement to align input sequences to equal lengths and fully exploits the prior knowledge embedded in biological sequences. PhyloGen models phylogenetic tree inference as a conditional-constrained tree structure generation problem, aiming to generate and optimize the tree topology and branch lengths jointly. We map species sequences into a continuous geometric space and perform end-to-end variational inference without restricting topological candidates. To ensure the topology-invariance of phylogenetic trees, we incorporate distance constraints in the latent space to maintain translational rotation invariance. Our approach demonstrates effectiveness and efficiency on the eight real-world benchmark datasets and verifies its robustness through data augmentation and noise addition. In addition, we propose a new scoring function to guide the model towards a more stable and faster gradient descent.

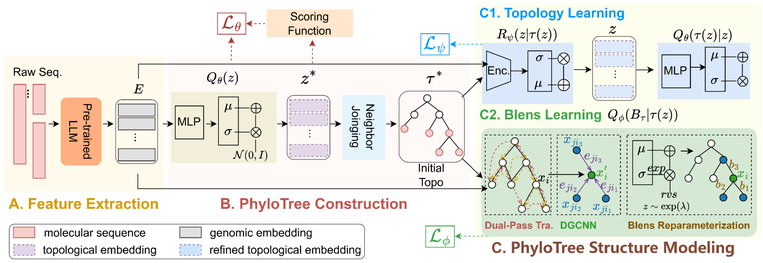
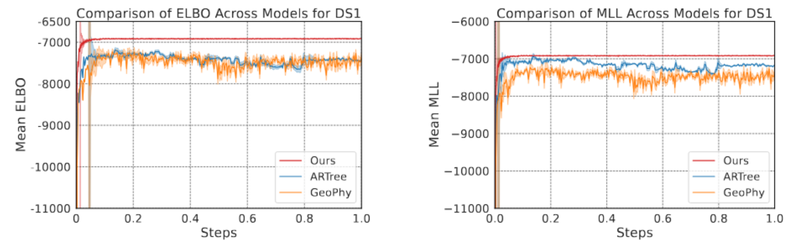


Figure2: Our Method: Framework of PhyloGen

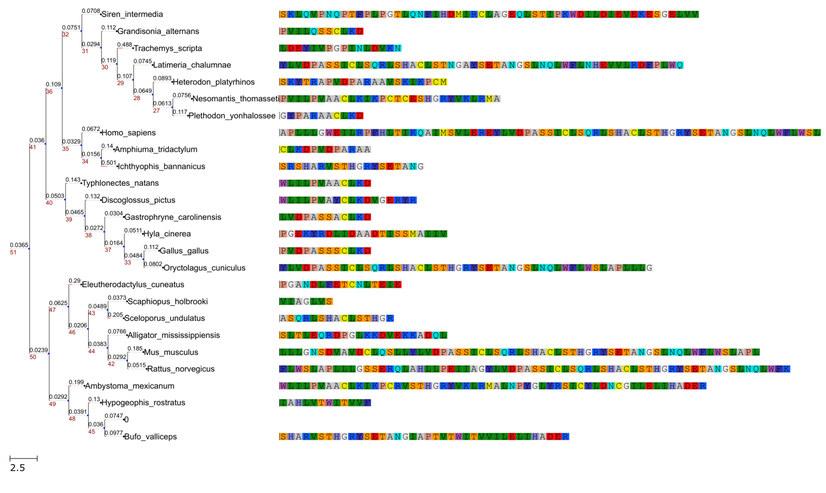
A. Feature Extraction module extracts genome embeddings E from raw sequences Y using a pre-trained language model.
B. PhyloTree Construction module uses E to compute topological parameters, which generate an initial tree structure τ^* via the Neighbor-Joining algorithm.
C. PhyloTree Structure Modeling module jointly model τ and B_τ through the topology learning component (TreeEncoder R and TreeDecoder Q) and the branch length~(Blens) learning component (dual-pass traversal, DGCNN network, Blens reparameterization).



	Metric	PhyloGFN	GeoPhy	GeoPhy LOO(3)+	Ours w/o KL	Ours w/o S	Ours
setting1	**ELBO (Δ)**	NA	-7721.82 (-100)	-7729.28 (-107)	-6725.49 (+12)	-6713.01(+15)	**-6711.47(+14)**
	MLL (Δ)	-6705.55 (-93)	-7440.38 (-198)	-7599.85 (-357)	-6564.51 (+18)	-6547.32 (+20)	**-6542.75 (+22)**
	Time	18h28min	6h16min	15h23min	6h43min	7h42min	**6h32min**
setting2	**ELBO (Δ)**	NA	-11802.07 (-98)	-11676.29 (-76)	-10678.24(+3)	-10655.02 (+2)	**-10674.28(+4)**
	MLL (Δ)	-12565.76 (-233)	-11763.40 (-131)	-11630.16 (-98)	-10422.12 (+6)	-10654.53 (+12)	**-10432.71(+5)**
	Time	24h35min	18h13min	12h24min	7h54min	8h6min	**6h37min**

Table 1: Comparison of the MLL (↑) with different approaches in eight benchmark datasets. VBPI and VBPI-GNN use pre-generated tree topologies in training and thus **are not directly comparable**. **Boldface** for the highest result, underline for the second highest from traditional methods, and underline for the second highest from tree structure generation methods.

Methods	Dataset #Taxa (N)	DS1 27	DS2 29	DS3 36	DS4 41	DS5 50	DS6 50	DS7 59	DS8 64
MCMC-based	MrBayes	-7108.42 (0.18)	-26367.57 (0.48)	-33735.44 (0.50)	-13330.44 (0.54)	-8214.51 (0.28)	-6724.07 (0.86)	-37332.76 (2.42)	-8649.88 (1.75)
	SBN	-7108.41 (0.15)	-26367.71 (0.08)	-33735.09 (0.09)	-13329.94 (0.20)	-8214.62 (0.40)	-6724.37 (0.43)	-37331.97 (0.28)	-8650.64 (0.50)
Structure Representation	VBPI	-7108.42 (0.10)	-26367.72 (0.12)	-33735.10 (0.11)	-13329.94 (0.31)	-8214.61 (0.67)	-6724.34 (0.68)	-37332.03 (0.43)	-8650.63 (0.55)
	VBPI-GNN	-7108.41 (0.14)	-26367.73 (0.07)	-33735.12 (0.09)	-13329.94 (0.19)	-8214.64 (0.38)	-6724.37 (0.40)	-37332.04 (0.12)	-8650.65 (0.45)
Structure Generation	ARTree	-7108.41 (0.19)	-26367.71 (0.07)	-33735.09 (0.09)	-13329.94 (0.17)	-8214.59 (0.34)	-6724.37 (0.46)	-37331.95 (0.27)	-8650.61 (0.48)
	phi-CSMC	-7290.36 (7.23)	-30568.49 (31.34)	-33798.06 (6.62)	-13582.24 (35.08)	-8367.51 (8.87)	-7013.83 (16.99)	NA	-9209.18 (18.03)
	GeoPhy	-7111.55 (0.07)	-26379.48 (11.60)	-33757.79 (8.07)	-133342.71 (1.61)	-8240.87 (9.80)	-6735.14 (2.64)	-37377.86 (29.48)	-8663.51 (6.85)
	GeoPhy LOO(3)+	-7116.09 (10.67)	-26368.54 (0.12)	-33735.85 (0.12)	-13337.42 (1.32)	-8233.89 (6.63)	-6735.9 (1.13)	-37358.96 (13.06)	-8660.48 (0.78)
	PhyloGFN	-7108.95 (0.06)	-26368.9 (0.28)	-33735.6 (0.35)	-13331.83 (0.19)	-8215.15 (0.20)	-6730.68 (0.54)	-37359.96 (1.14)	-8654.76 (0.19)
	Ours	**-6910.02** (0.07)	**-26257.09** (0.06)	**-33481.57** (0.10)	**-13063.15** (1.34)	**-7928.4** (0.23)	**-6330.21** (0.31)	**-36838.42** (12.03)	**-8171.04** (0.96)



Bipartition Frequency Distribution (RQ3)

The closer the two curves are, the better, which suggests that PhyloGen is highly consistent with the gold standard MrBayes approach.

Fig.5 shows the bipartition frequency distributions of trees inferred by PhyloGenfor datasets DS1, DS2, and DS3.The horizontal axis indicates the ranking of the bipartitions in the tree topology, and the vertical axis indicates the normalized frequency of occurrence of the corresponding bipartitions. The **similarity** of our method's **curves** to those of MrBayes underscores its accuracy, demonstrating that PhyloGen consistently captures evolutionary patterns with reliability comparable to the gold standard.

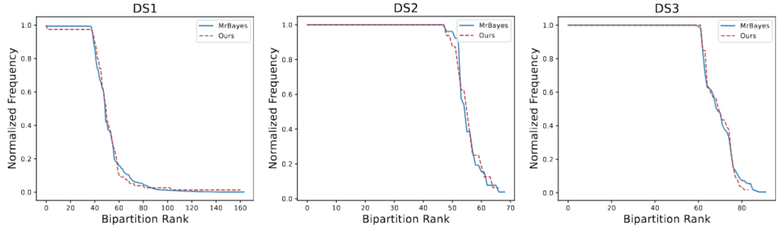


Figure 5: Comparative Bipartition Frequency Distribution in Tree Topologies for DS1, DS2, andDS3 datasets.