How far can we trust language phylogenies?

Emma Kopp, Robin J. Ryder
CEREMADE, CNRS, UMR 7534,
Université Paris-Dauphine, PSL University, 75016 Paris, France
Contact: emma.kopp@dauphine.psl.eu

Motivations

- **Framework:** observation of multiple modern languages.
- **Hypothesis** : we can represent the relationship between some languages as a phylogenic tree T.
- **Methods:** Bayesian algorithm to get a tree sample according to the posterior of the model.
- **Problem:** nodes estimation of the tree due to a lack of signal.
- **Languages datasets:** Sino-Tibetan (Sagart et al. [2019]), Bantu (Grollemund et al. [2015]) and Indo-European (Heggarty et al. [2023])

Phylogenetic inference

1. Observation of $\sigma_1, \dots, \sigma_K$ sequences for N languages.
2. Modelisation on the phylogeny according to a model $\mathcal{T}_\theta$.
3. Apply MCMC algorithm to sample posterior phylogenies $\left(\hat{T}_1, \dots, \hat{T}_M\right)$.

Question: How close to the truth is the sample $\left(\hat{T}_1, \dots, \hat{T}_M\right)$ obtained ?

Answer: We can't trust a phylogeny with a root older than ∼10 millenia.

Phylogeny Modelisation

A phylogenetic tree is written as $\mathcal{T} = (\mathcal{O}, \mathcal{D}, \sigma)$

- The **Tree model** ($\mathcal{O}$) is the topology of the tree. We put a Birth-Death (BD) prior on it.
- The **Clock model** ($\mathcal{D}$) estimates the ages of subgroups of T by attribuing a rate to each branch. Each site i in the sequence evolves at a rate λ_i. We put a strict clock prior on it ($\lambda_i = \lambda = 1$).
- The **Substitution model** (σ) describe substitution of each taxas. It is modelised by a transition matrix Q. We put a Binary Continuous Time Markov Chain (CTMC) prior on it ($\sigma_k^n \in \{0, 1\}$).

➞ We can express the likelyhood in closed form and apply an MCMC algorithm in order to sample $\left(\hat{T}_1, \dots, \hat{T}_M\right)$.

Mathematical evidence

- **Probability of exact topology reconstruction of T ($\mathcal{O}$)**

  Consider a phylogeny T. Let μ be the prior measure on a tree T. Then for any time $s > 0$ we can bound the probability of exact topology reconstruction $\Delta^T(s)$ by

$$\Delta^T(s) \leq \max_T \mu\left[T(s) = T\right] + k \sum_{v \in \partial T} M_{\mathcal{D}}\left(-q\left(t(v) - s\right)\right).$$

where $\begin{cases} M_{\mathcal{D}} & \text{is the moment generating function of } \mathcal{D} \\ \partial T & \text{are the leaves of } T \\ q = \sum_j \min_i Q_{i,j} \end{cases}$

💬 The definition of q doesn't allow the utilization of a binary covarion substitution model which stands as the most widely adopted model.

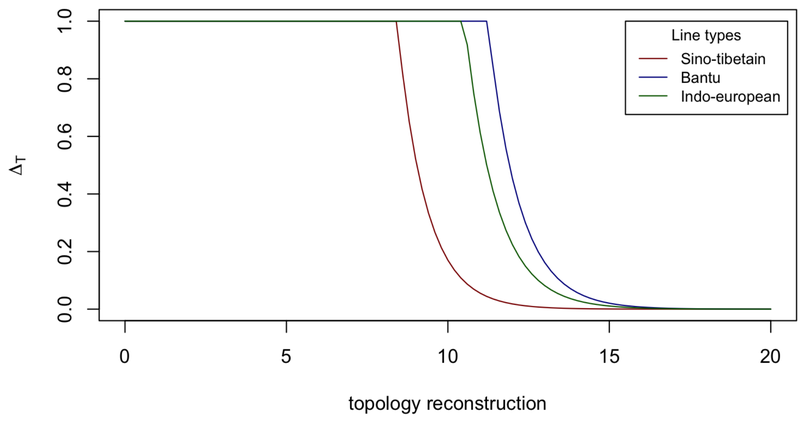
Elchanan Mossel. On the Impossibility of Reconstructing Ancestral Data and Phylogenies. Journal of Computational Biology, 2003.

Datasets

	Bantu	Sino-Tibetan	Indo-European
t	6.975	8.502	8.574
K	3859	3785	4990
N	422	50	161
Δ^T	513	0.765	287

Applications

We plot the upper bounds of the exact reconstruction for all fixed variables except the age of the tree.



➞ For a tree older than 10 millenia, it is impossible to recover the topology.

Simulated data

We simulate synthetic data on trees between 5 and 17 millennia old and check when the signal for the deep topology vanishes.

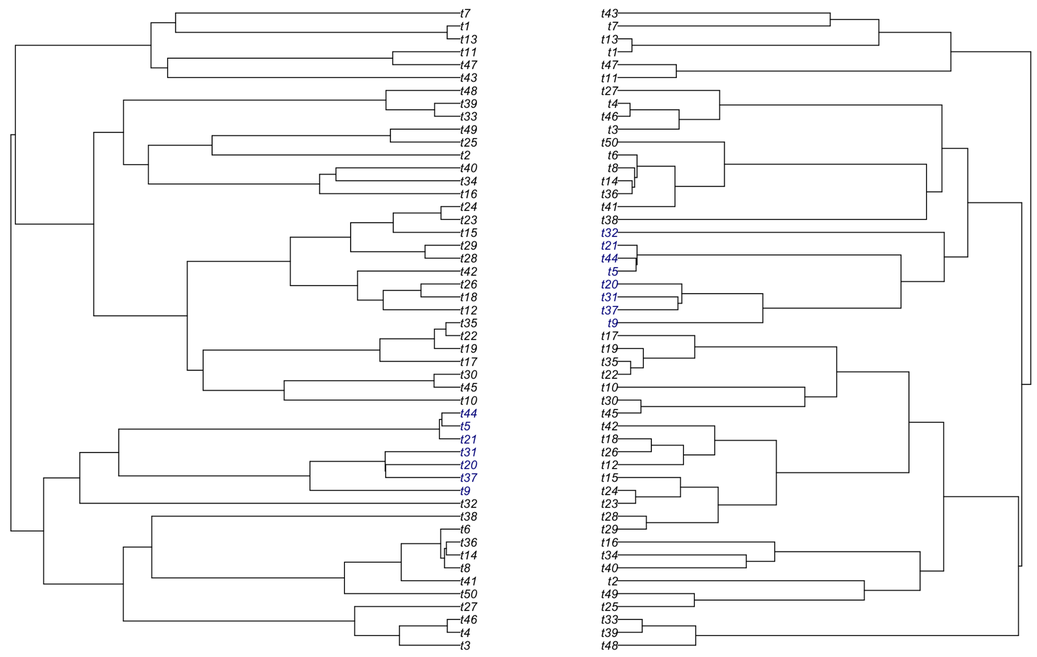
- **Topology step :** Simulate $T_1 \sim (\mathcal{O}, \mathcal{D})$.
- **Branch length step:** for $k = 2, \dots, 17$, set $T_k = (\mathcal{O}, k\mathcal{D})$. All phylogenies have the same topology $\mathcal{O}$. We rescaled $\mathcal{D}$ to obtain older trees.
- **Sequence step:** for $k = 1, \dots, 17$, $\quad T_k \sim (\mathcal{O}, k\mathcal{D}, \sigma_k)$
- **Inference:** sample $\left(\hat{T}_k^1, \dots, \hat{T}_k^M\right)$ via MCMC, with σ_k as observations.

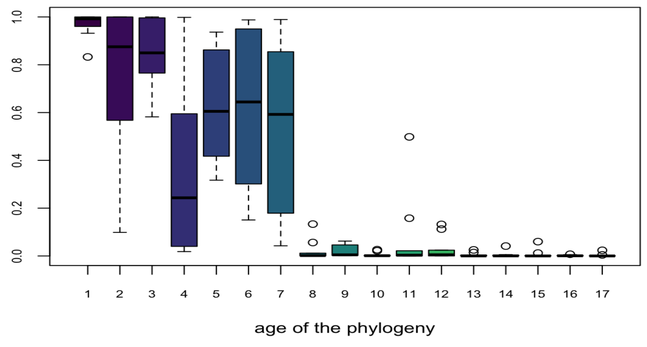
Question: How can we compare the true and the inferred trees?

- Compare T_k and $\left(\hat{T}_1, \dots, \hat{T}_M\right)$:

  Let $n_1, \dots, n_{10}$ be the ten deepest nodes of the tree. Let D_{n_i} be the set of descendants of n_i in $\mathcal{O}$. For each $k = 1, \dots, 17$

$$\hat{p}_{k,i} = \frac{1}{M} \sum_{l=1}^{M} \mathbb{I}_{\text{is.monophyletic }\left(\hat{T}_k^l, D_{n_i}\right)}.$$





➞ Trees older than 8 millenia are badly reconstructed.