Lagged couplings for phylogenetic inference

Diagnosing convergence of MCMC on trees

Luke J. Kelly, Robin J. Ryder and Grégoire Clarté

Dauphine | PSL UNIVERSITÉ PARIS — CEREMADE UMR CNRS 7534

Summary

Existing methods do not diagnose MCMC convergence jointly across all components of a phylogenetic model

We couple Markov chains $(X_s)_s$ and $(Y_s)_s$ with stationary distribution π on trees, model parameters and latent variables

Chains coupled at lag l meet exactly at random finite time $\tau^{(l)}$ and remain together

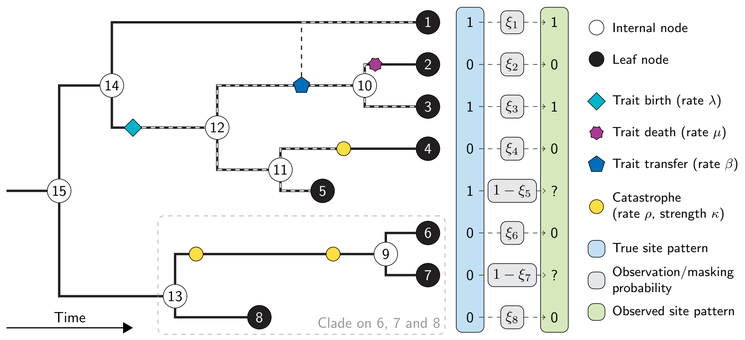
Estimate upper bound on convergence

$$d_{\mathrm{TV}}(\pi_s, \pi) \leq \mathbb{E}\left[0 \vee \left\lceil \frac{\tau^{(l)} - l - s}{l} \right\rceil\right]$$

Phylogenetics

Phylogenetics is the problem of inferring the ancestral history of a set of species

- Common traits indicate shared ancestry
- Phylogeny is typically a tree with data recorded at leaves



A branching process ○ on sets of traits is the phylogeny of the observed taxa ●, catastrophes ● represent bursts of evolutionary activity, trait sets diversify according to the Stochastic Dollo model ◆ ● ⬟

Inference problem

Difficult statistical problem

- Infer tree topology, branch lengths, rate parameters and catastrophe locations
- Intractable likelihood, often multimodal
- Many constraints and dependencies

MCMC for Bayesian phylogenetic inference

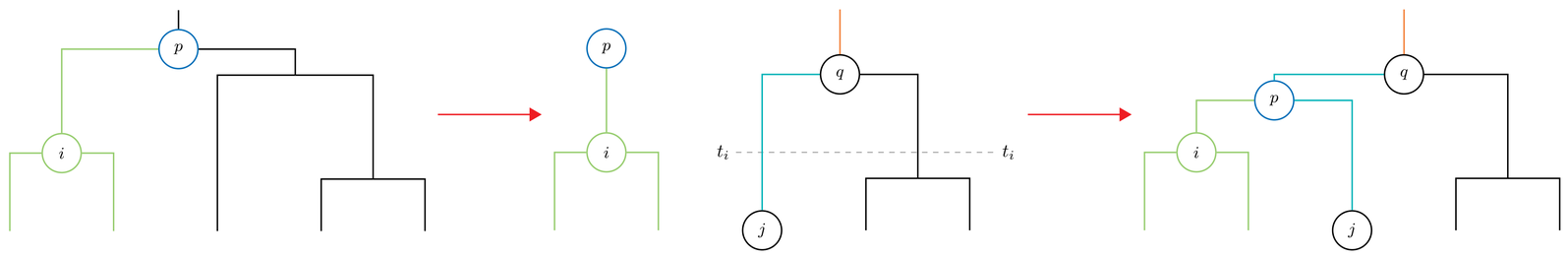
- Theoretical properties largely unknown
- Existing approaches compare many low-dimensional summaries so lack power to quantify MCMC behaviour in practice
- Difficult to separate modelling and fitting errors in inference

References

P.E. Jacob, J. O'Leary, and Y.F. Atchadé. *JRSS B*, 82(3):543–600, 2020.
N. Biswas, P.E. Jacob, and P. Vanetti. *NeurIPS*, pages 7389–7399, 2019.
L.J. Kelly, R.J. Ryder, and G. Clarté. *arXiv 2108.13328*, 2021.

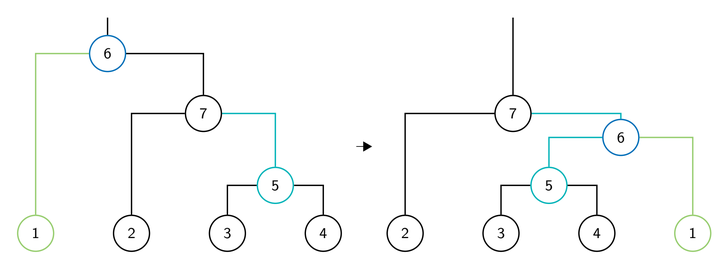
Coupling MCMC for phylogenetic models

In our mixture of proposal kernels, subtree prune-and-regraft (SPR) moves are the primary tool for exploring tree space

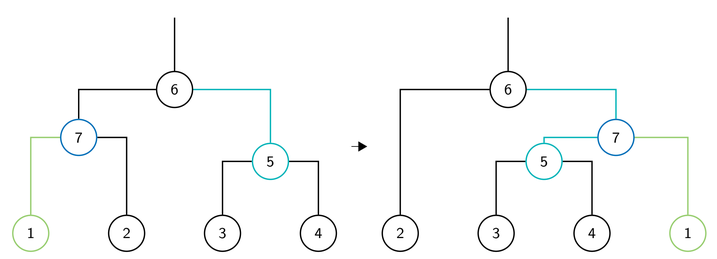


SPR proposal moves the subtree with root i to the branch leading into j while respecting ancestry constraints

Implicitly coupling moves: relabel nodes to identify similar tree components in current states (X, Y) and sample from a maximal coupling at each step of move to form proposals (X', Y')



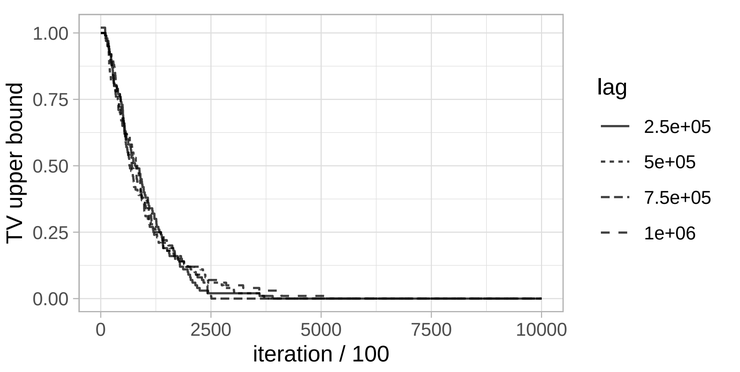
SPR $X \to X'$: move subtree $i^{(X)} = 1$ to branch $j^{(X)} = 5$



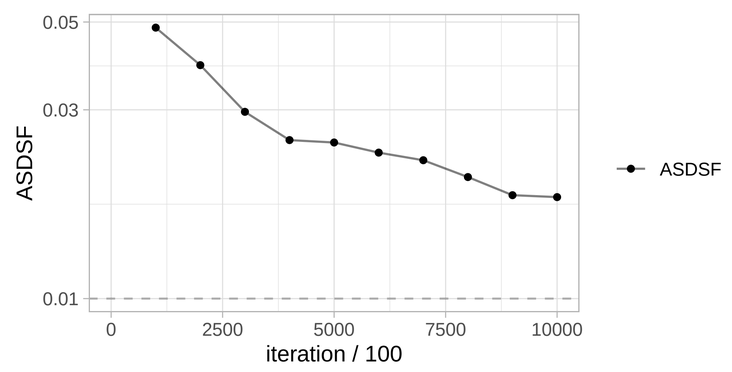
SPR $Y \to Y'$: move subtree $i^{(Y)} = 1$ to branch $j^{(Y)} = 5$

Eastern Polynesian languages

Fitting the Stochastic Dollo model to lexical trait data in 11 languages



Estimated TV bounds on tree and model parameters, 100 pairs of chains at each lag l

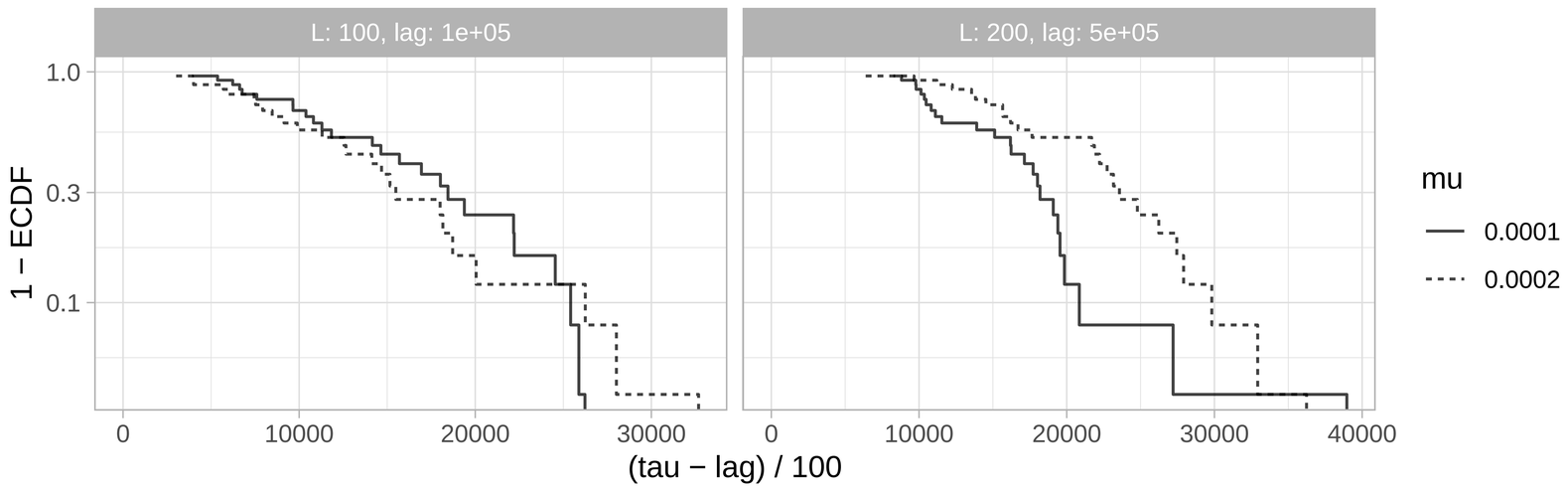


Average Standard Deviation of Split Frequencies (ASDSF) on sliding windows across 100 chains

Estimated TV bound diagnoses convergence across all components of the model and earlier than ASDSF or other methods

Larger trees

Coupling trees with 100 and 200 leaves, synthetic data with death rate μ



Future work

- Maximally coupling moves on trees
- Coupling trees with thousands of leaves
- Identify mixing or modelling issues from chains which fail to meet