

An Empirical Analysis of MCMC Convergence in Phylogenetic Applications

Motivation

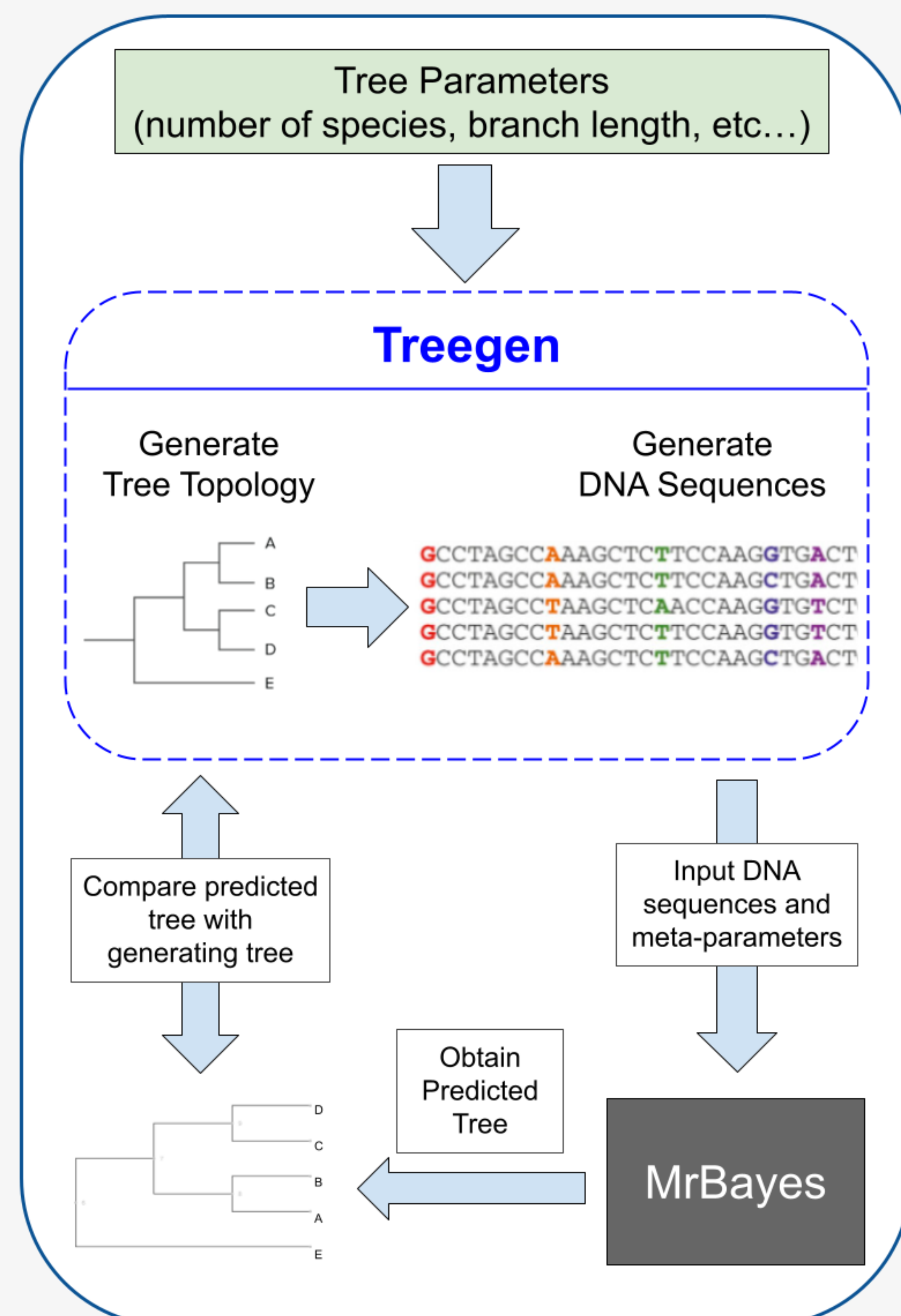
- Phylogenetic software is a vital tool that helps biologists analyze evolutionary history
- Underlying Markov Chain Monte Carlo (MCMC) process is treated like a black box
- Popular software like MrBayes are known to converge slowly and even inaccurately

Research Question

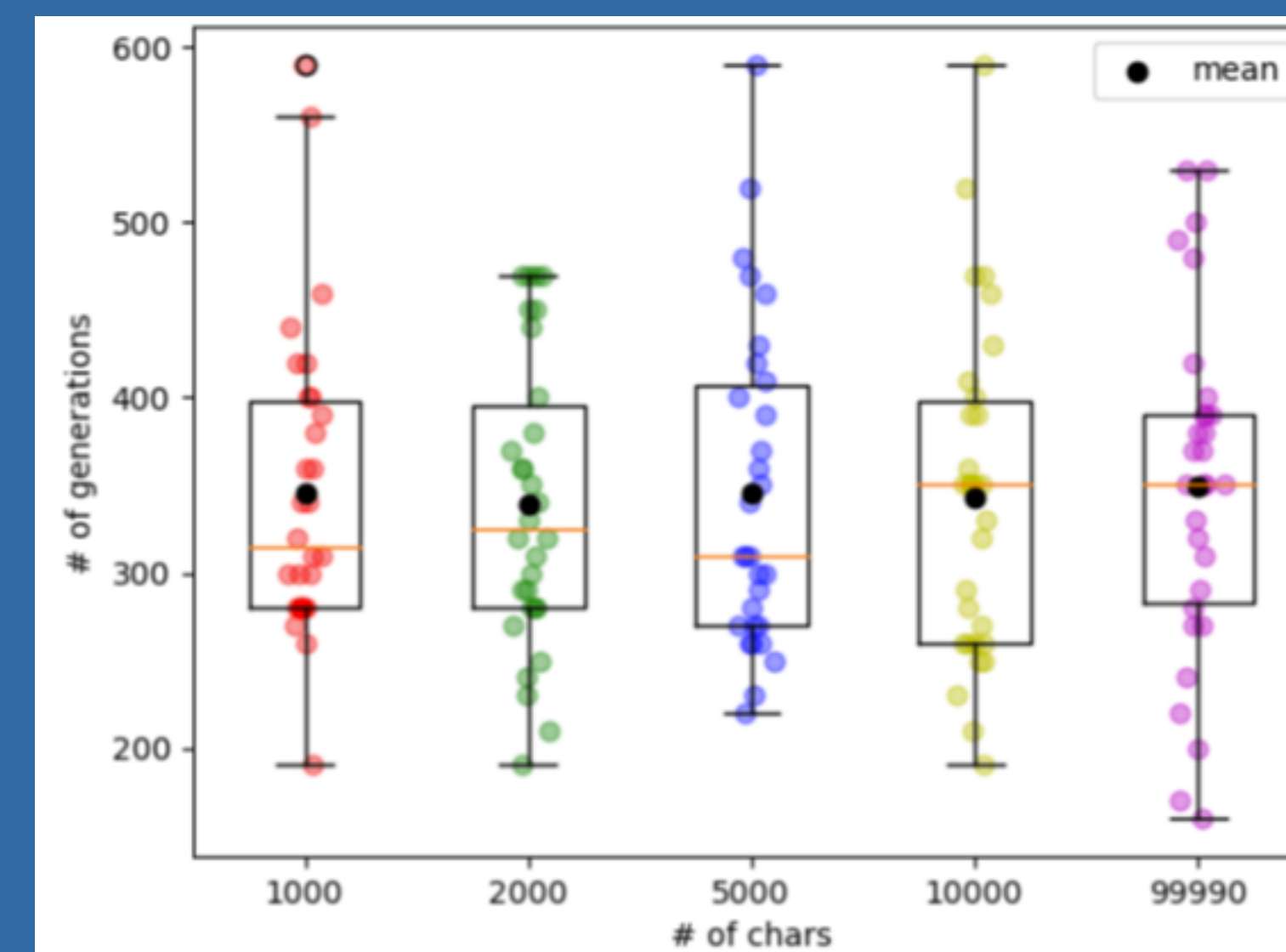
- How well does MrBayes perform under noisy input?
- What are the characteristics of tree topology and sequence generation that cause poor convergence?

Methodology

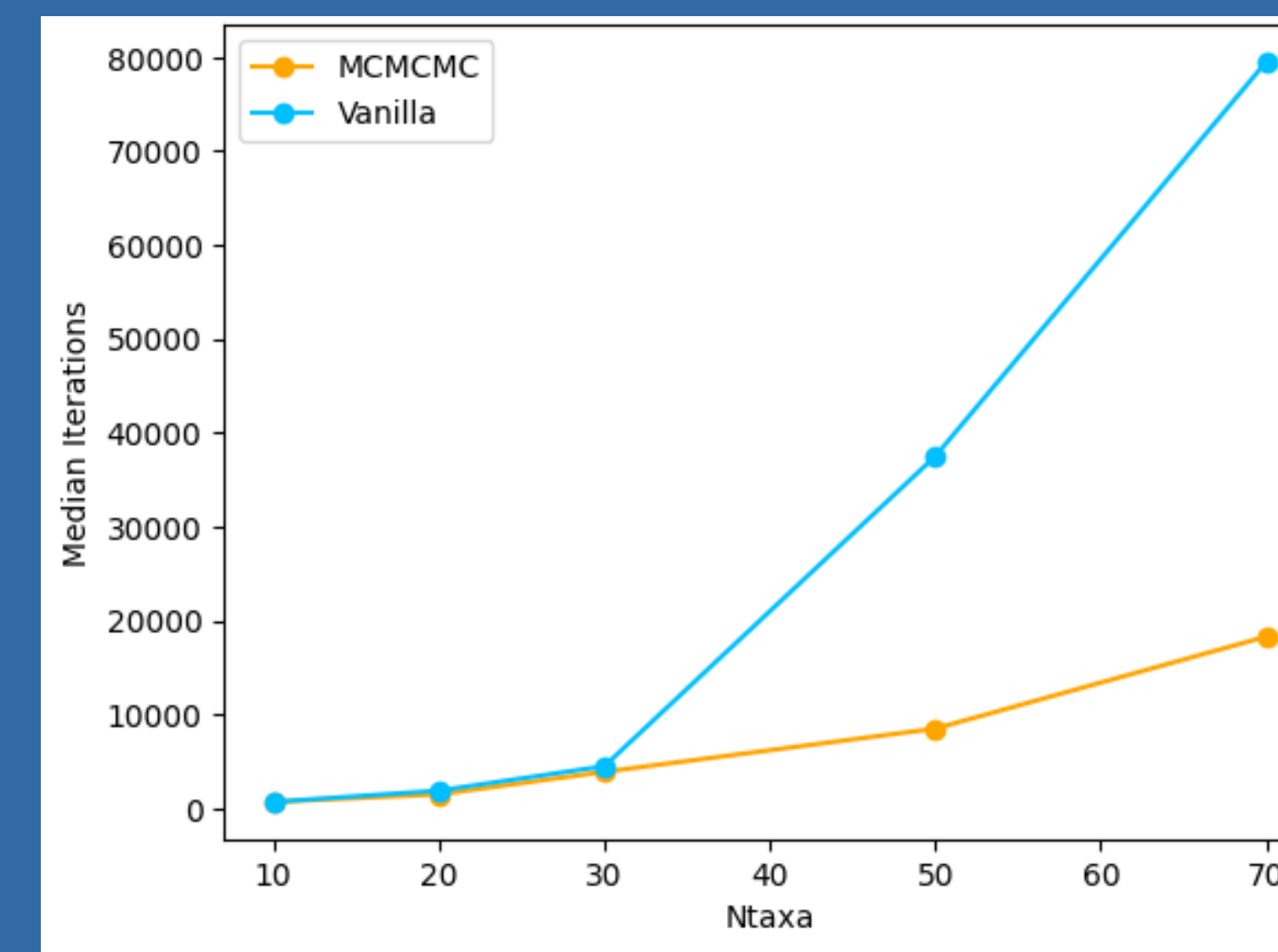
- Generate ground truth tree topology and DNA sequences with standard graph traversals
- Input leaf data into MrBayes and run until convergence
- Obtain the number of iterations to converge and the tree distance between the most probable output tree and the ground truth



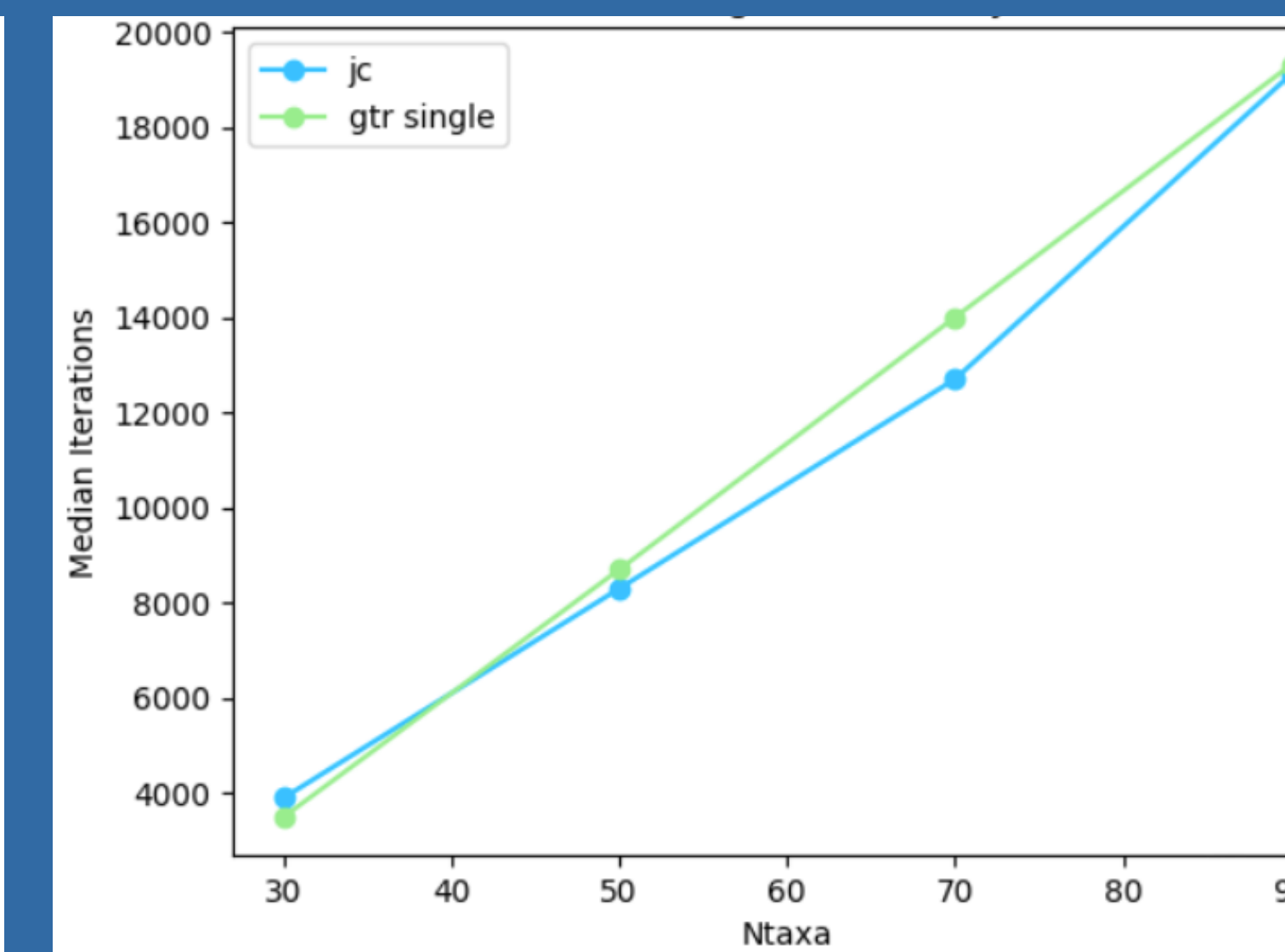
Results



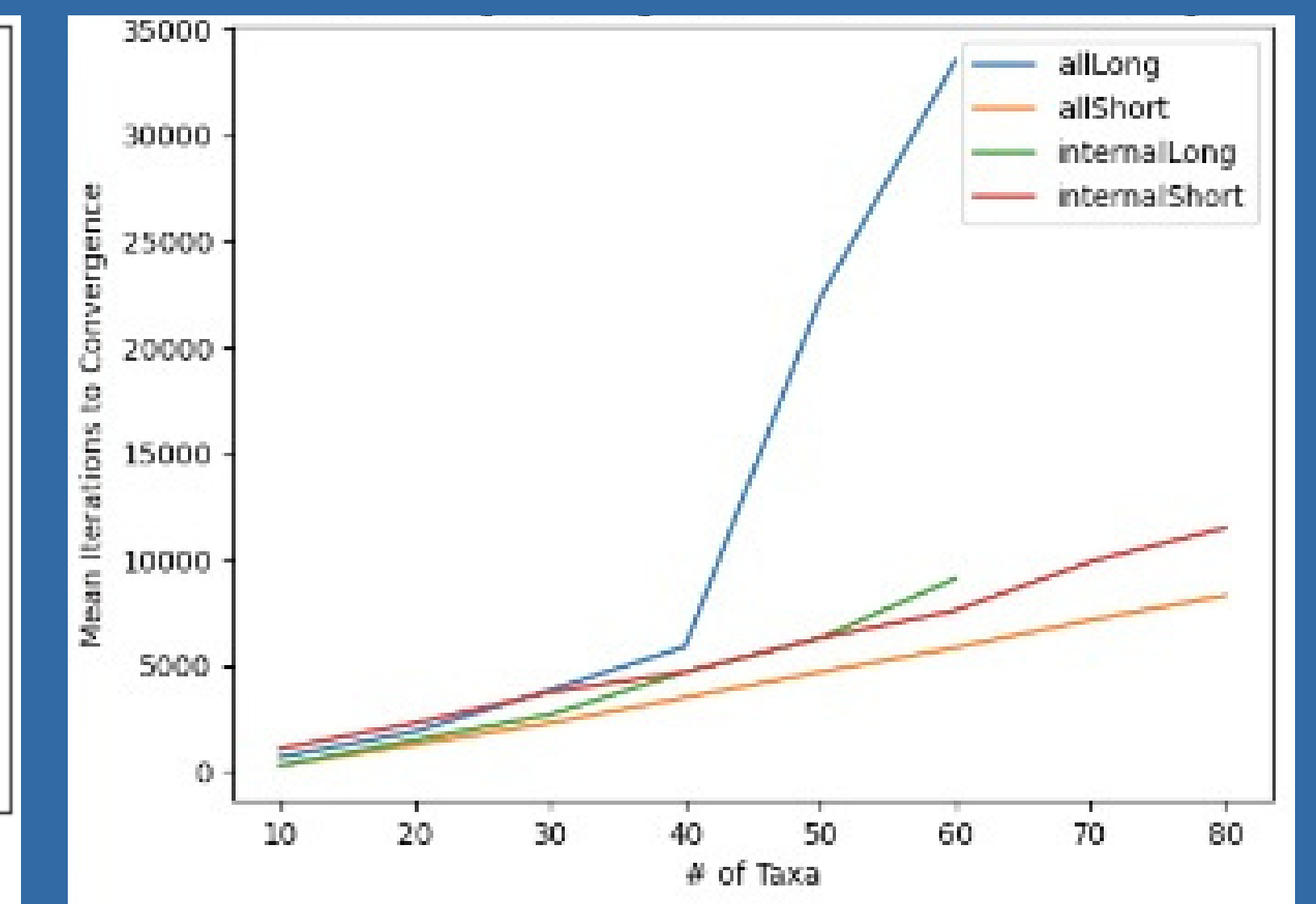
Mixing time as a function of sequence length. Shows mixing time is unaffected by sequence length



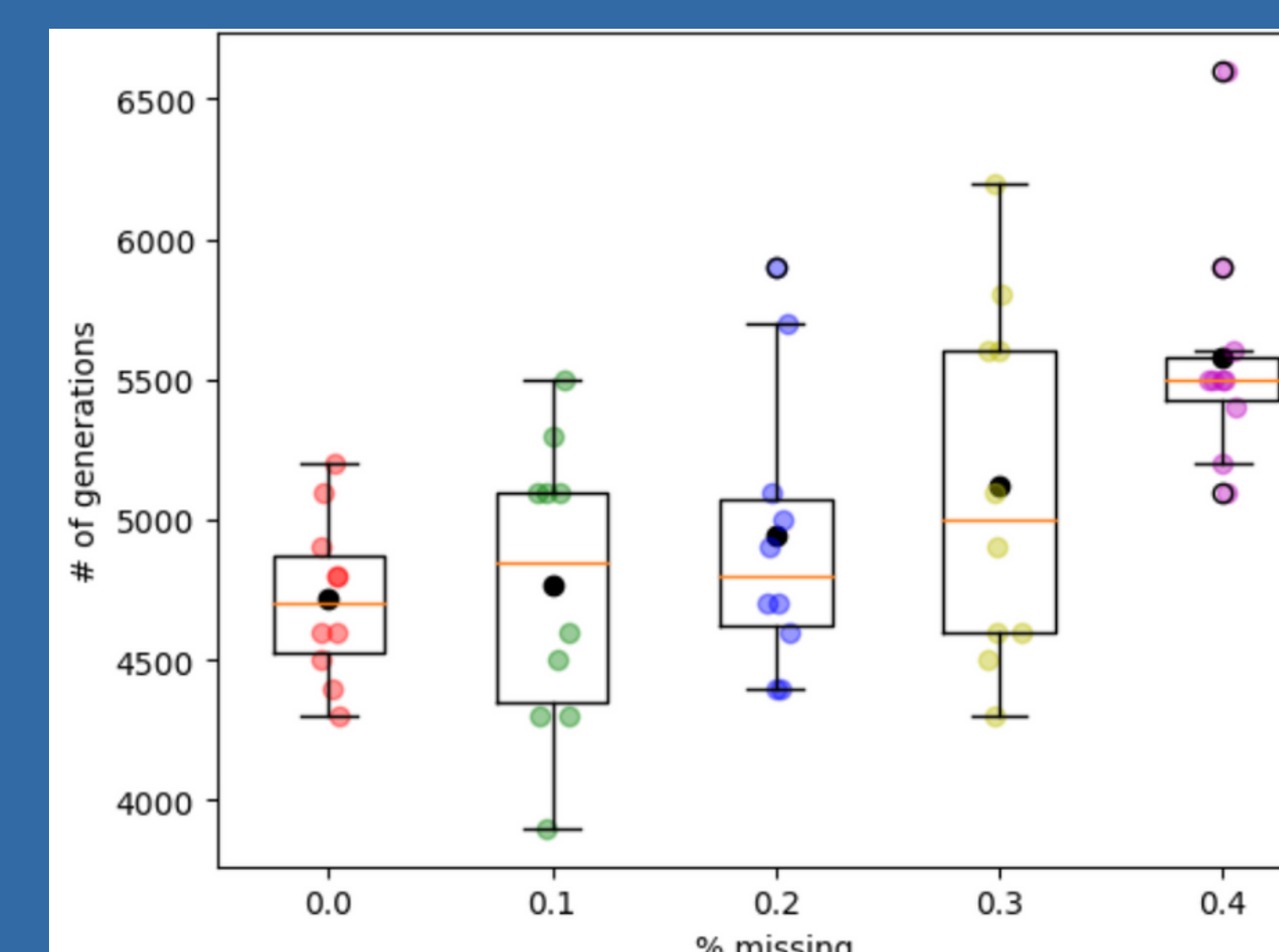
Mixing time of MC³ vs MCMC as a function of number of species; shows MC³ is faster than normal MCMC



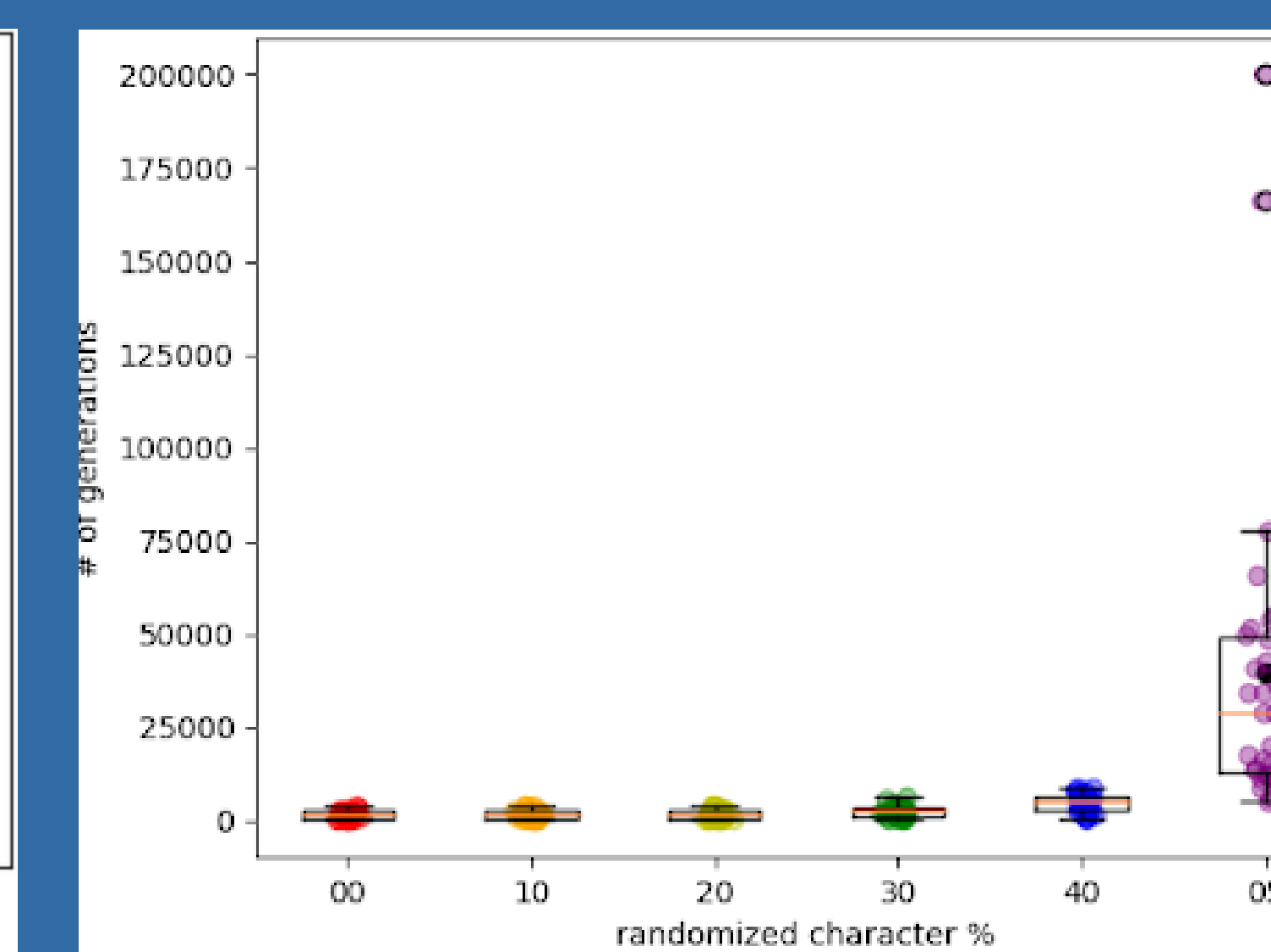
Generating data with GTR (6 parameter substitution model) vs. JC (1 parameter substitution model), while setting substitution model in MrBayes to JC.



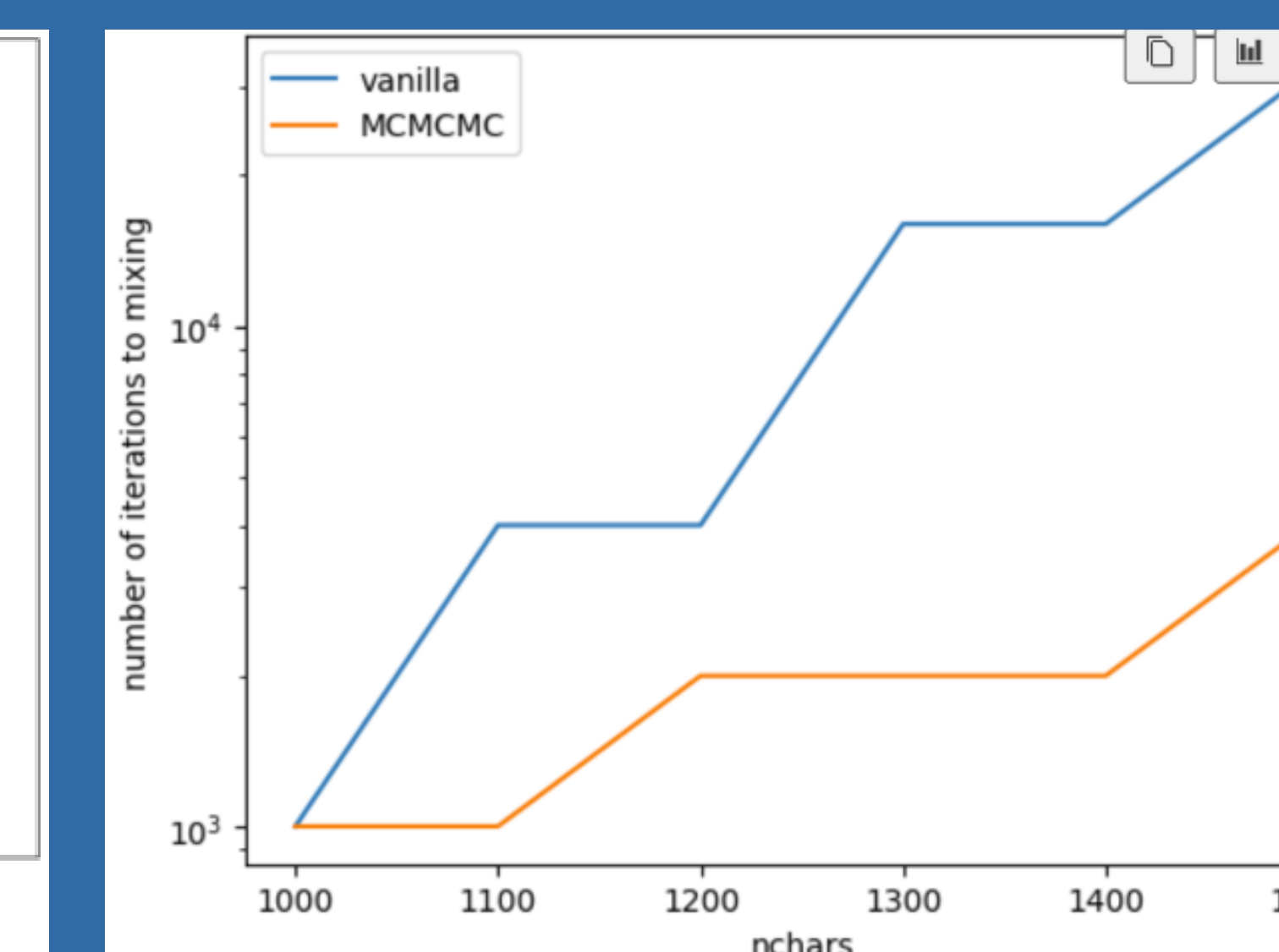
Convergence Time Scaling with # of Taxa vs Branch length Configuration (short = 0.1, long = 1.0). All long branches significantly increases convergence time.



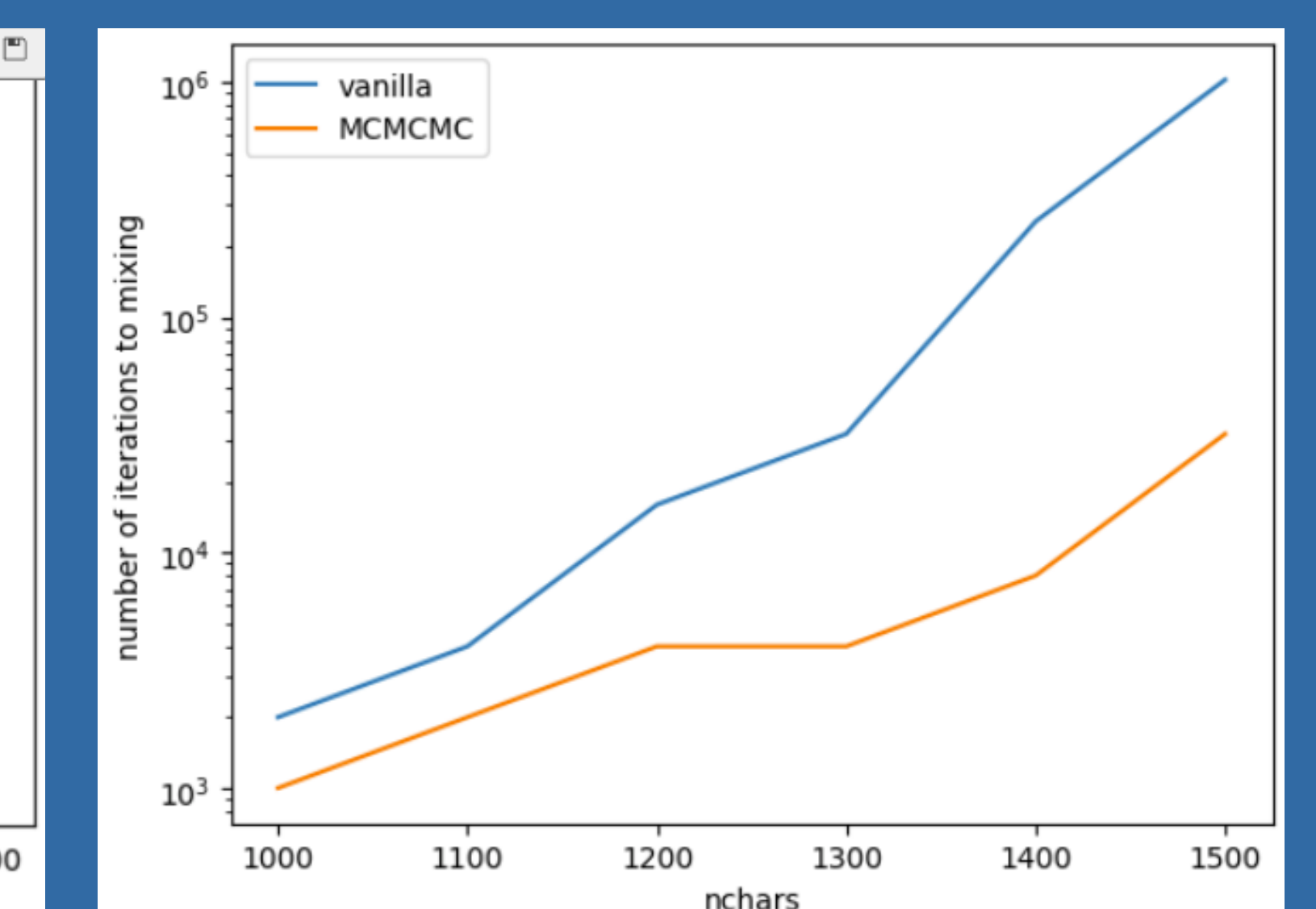
Effect of missing data (randomly dropping characters) on mixing time of MC³



Effect of randomized data (randomly changed characters) on mixing time of MC³; increased exponentially. With high enough sequence length, tree was correct up to 90% missing



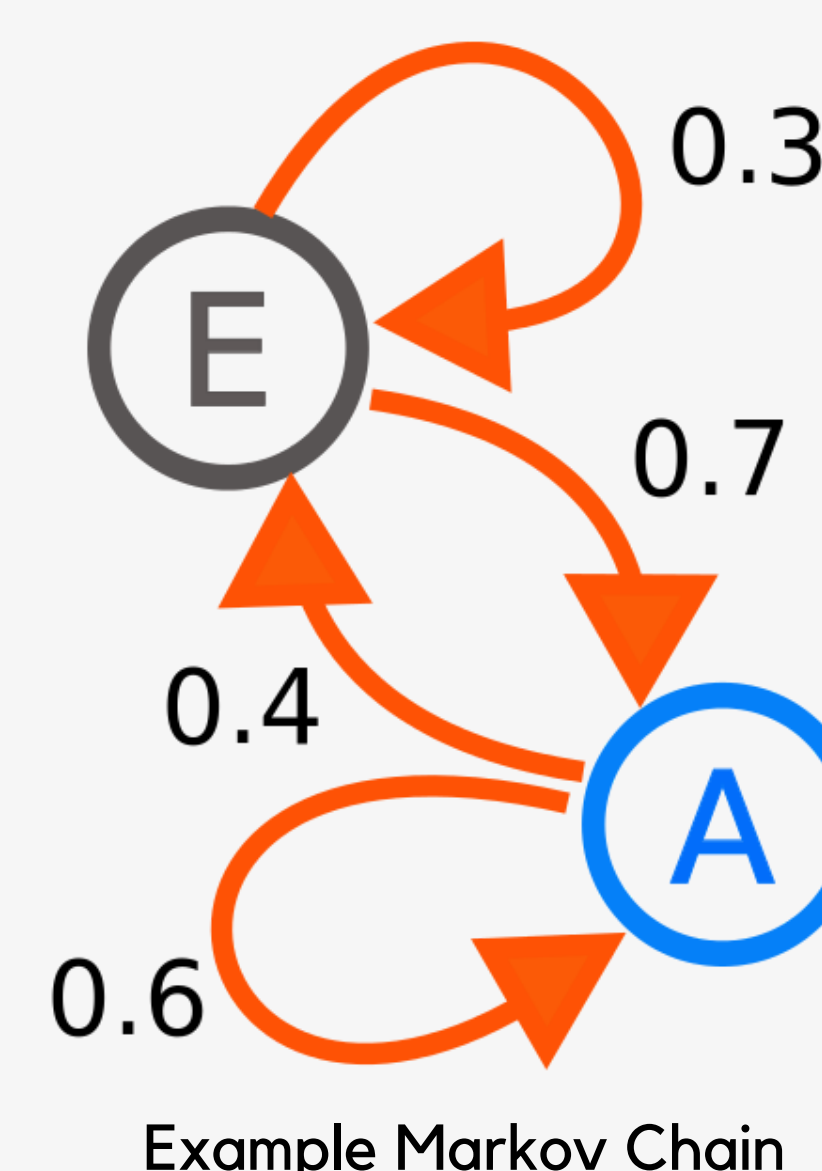
5 Leaf Mixture of Topologies from [1] Data generated from mixture distribution; exponential in sequence length (graph uses log scale)



Mixing time for 10 Leaf Mixture of Topologies; exponential in sequence length (graph uses log scale)

Background

- Phylogeny:** Studying evolutionary relationships of species
- Markov Chain:** Graph of states with probabilistic edges
- MCMC:** Performs a random walk until converging to the stationary distribution
- Stationary Distribution:** Huge space of all phylogenetic trees weighted by their likelihood
- Metropolis Coupled MCMC (MC³):** Optimization on the normal MCMC algorithm for faster convergence



MrBayes

- Input aligned DNA sequence for each species and outputs a sample from the posterior distribution (proportional to the likelihood)
- Default convergence heuristic is the average standard deviation of split frequency (ASDSF)
- Parameters include sequence length, number of species, branch lengths, and substitution model

Citations

[1]: Elchanan Mossel, Eric Vigoda. Phylogenetic MCMC Algorithms Are Misleading on Mixtures of Trees. Science, 2005.

Conclusion

- Converges accurately even with noisy data
- Exponential convergence on some mixed tree topologies, which real world data may resemble
- Exponential convergence requires specific conditions which are uncommon
- MC³ converges in fewer iterations and should generally be used over single chain MCMC

Acknowledgements

We would like to thank Professor Eric Vigoda, Professor Diba Mirza, Dr. Soojin Yi, Dr. Todd Oakley, and Chinmay Sonar for their support and guidance.