Bayesian Inference of Phylogenetic Trees

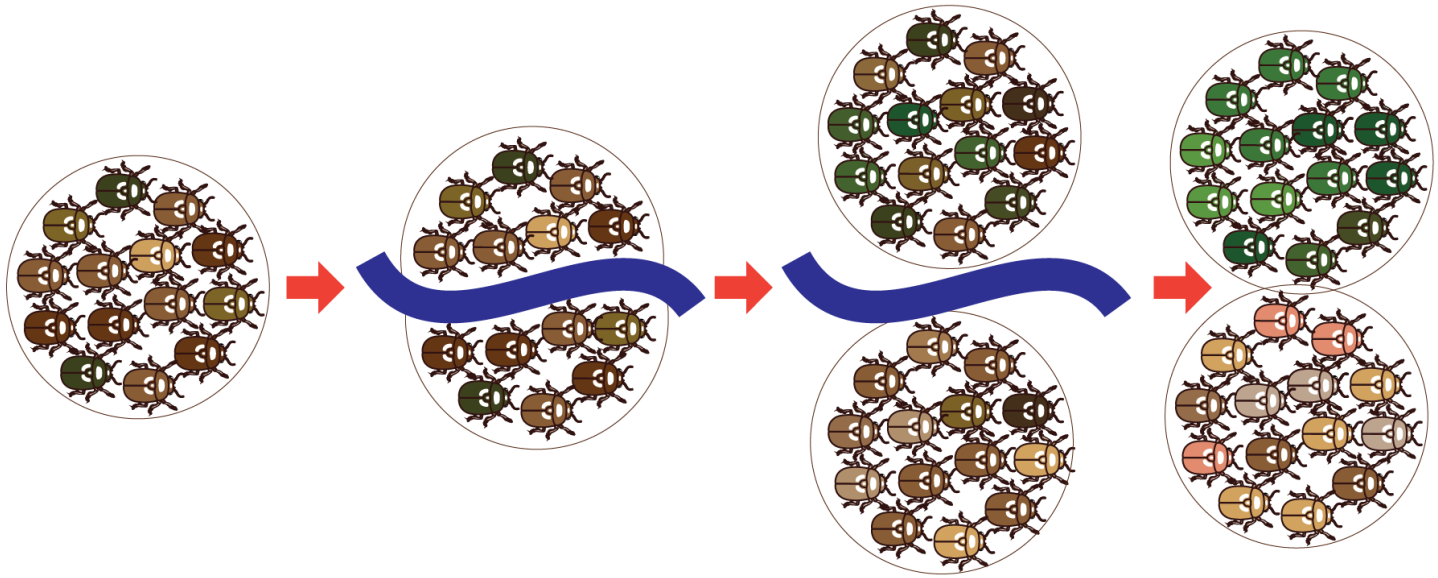
David Maxen, Georgios Aliatimis (Supervisor)

STOR-i Lancaster University

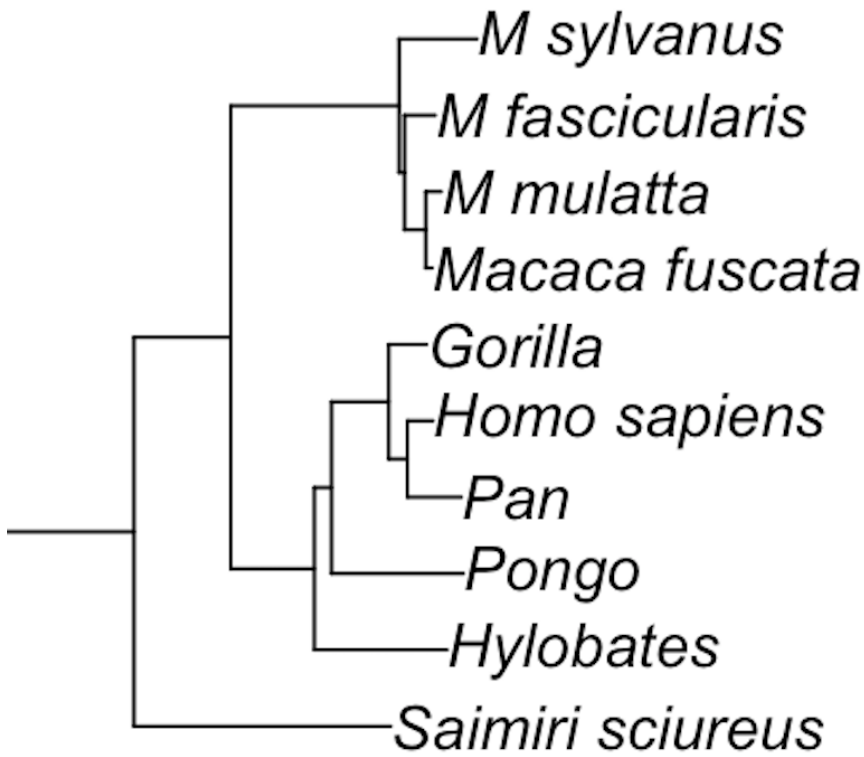
September 1, 2023

Table of Contents

Speciation

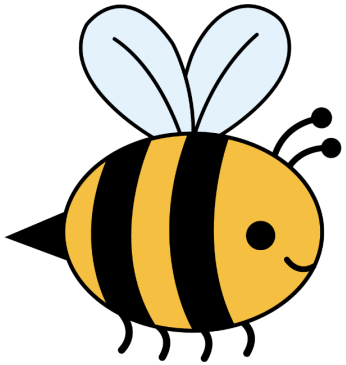


Phylogenetic Trees



Example Species

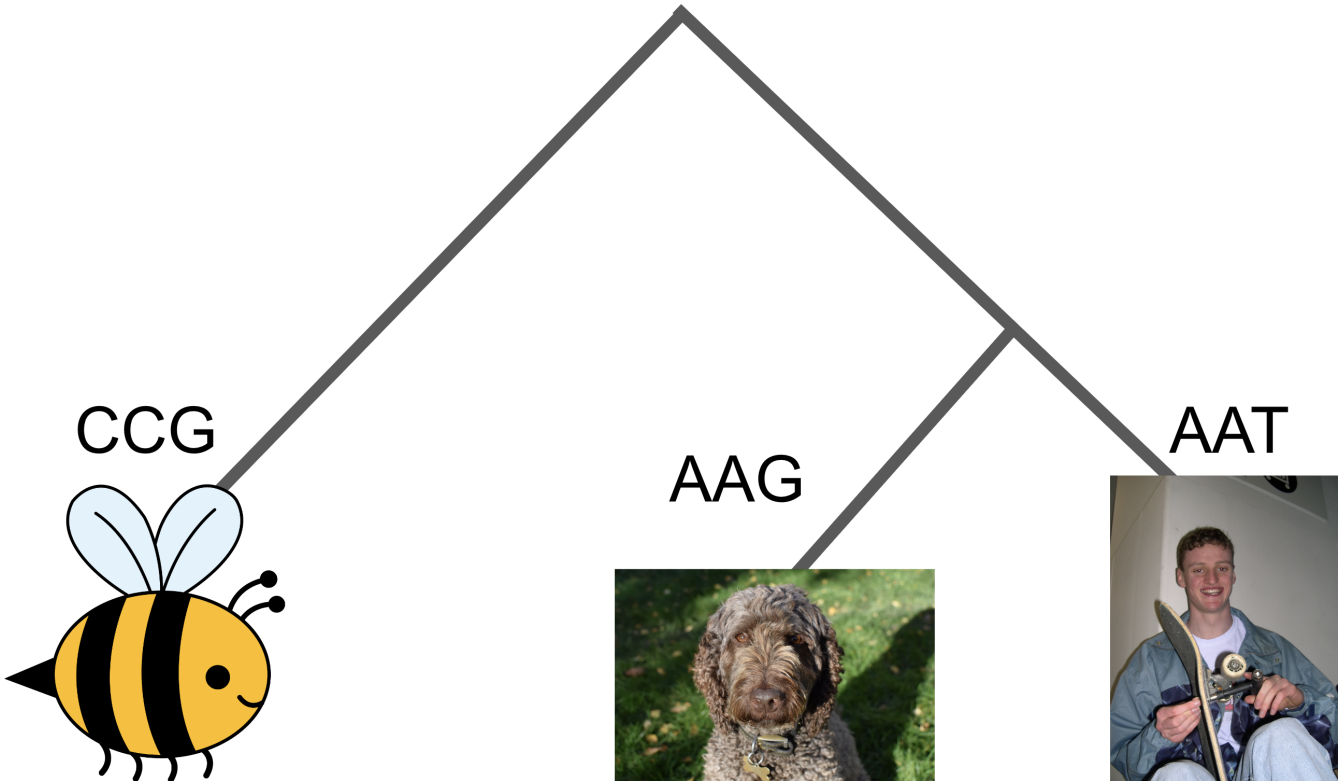
AAG



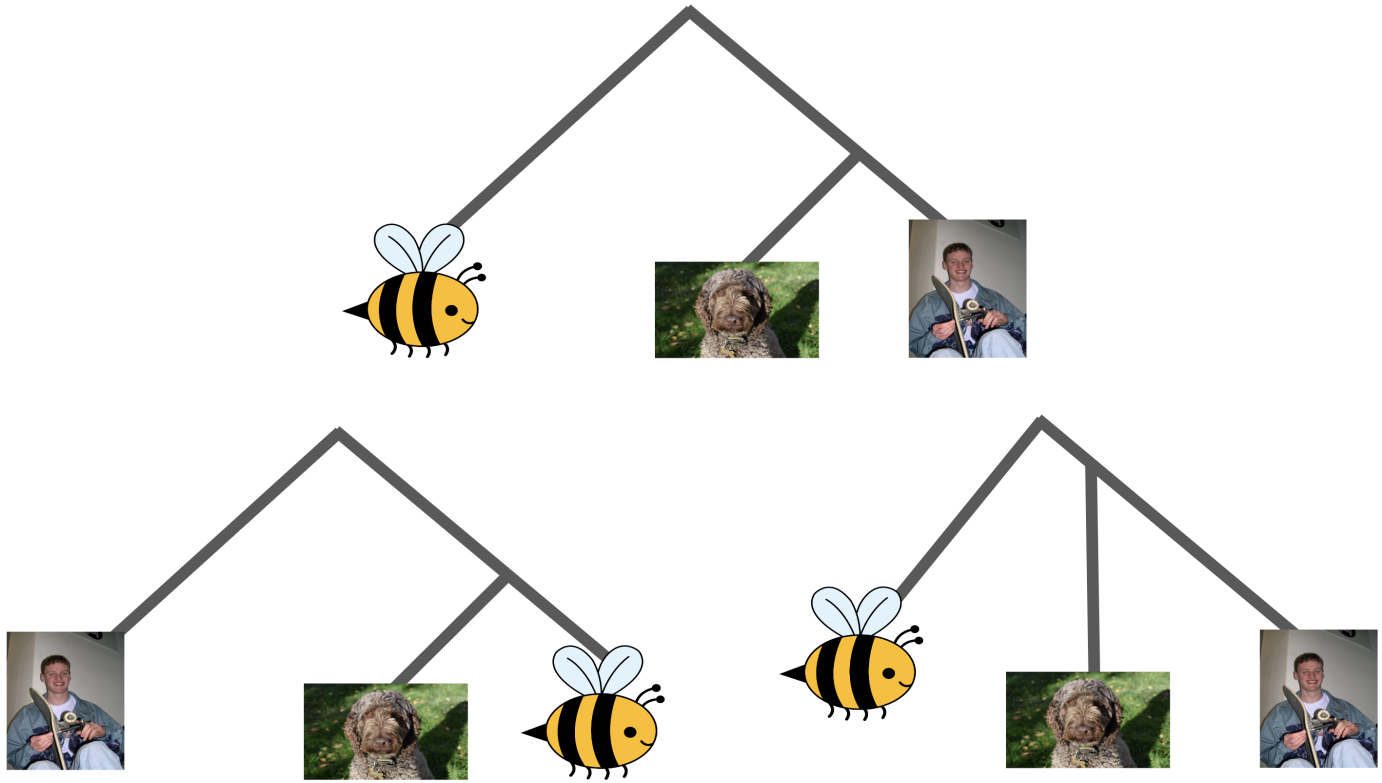
CCG

AAT

Example Phylogenetic Tree

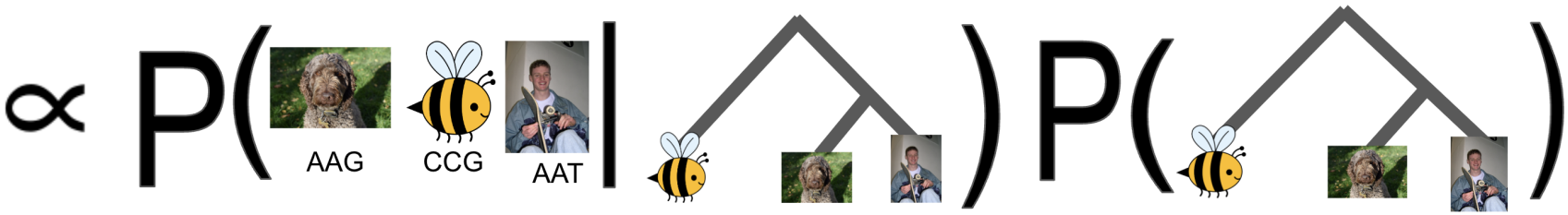


Example Phylogenetic Trees

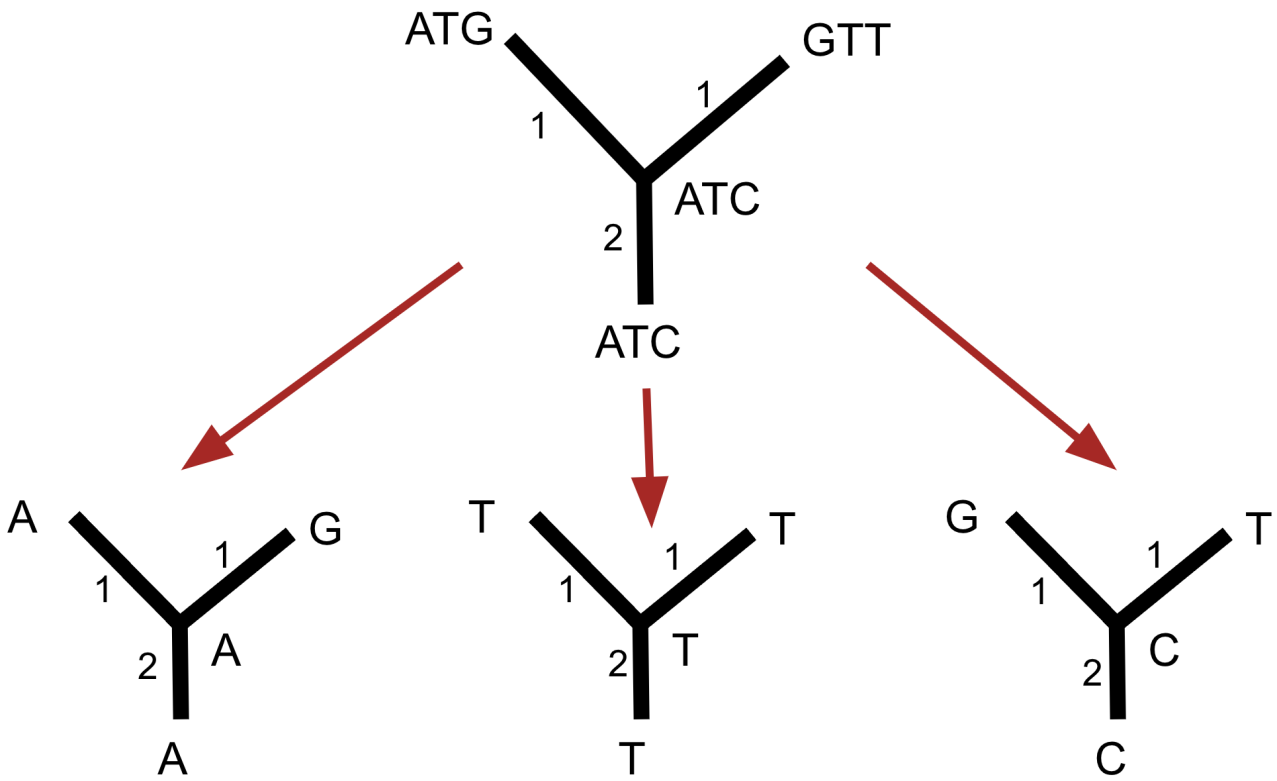


Target distribution

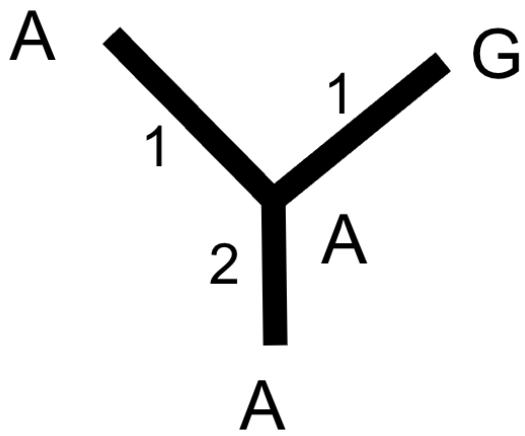
Target distribution



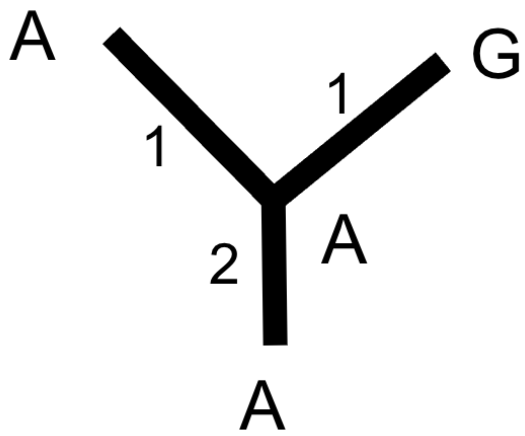
Independence Assumptions



Independence Assumptions



Independence Assumptions



$$\mathbf{P}_{AA}(1) \times \mathbf{P}_{AA}(2) \times \mathbf{P}_{AG}(1)$$

Transition Probabilities

$$P(t) = \begin{pmatrix} p_{\mathrm{AA}}(t) & p_{\mathrm{AG}}(t) & p_{\mathrm{AC}}(t) & p_{\mathrm{AT}}(t) \\ p_{\mathrm{GA}}(t) & p_{\mathrm{GG}}(t) & p_{\mathrm{GC}}(t) & p_{\mathrm{GT}}(t) \\ p_{\mathrm{CA}}(t) & p_{\mathrm{CG}}(t) & p_{\mathrm{CC}}(t) & p_{\mathrm{CT}}(t) \\ p_{\mathrm{TA}}(t) & p_{\mathrm{TG}}(t) & p_{\mathrm{TC}}(t) & p_{\mathrm{TT}}(t) \end{pmatrix}$$

Jukes and Cantor 1969

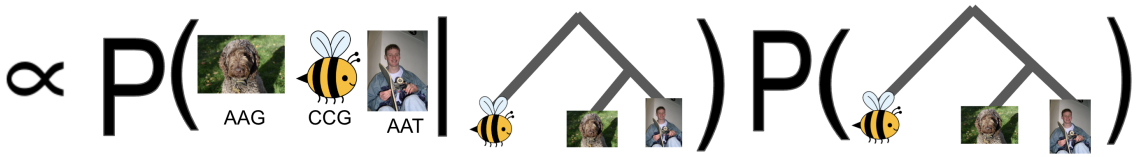
- Symmetrical rates
- Underlying base frequencies equal
- $P_{ii}(t) = \frac{1}{4} + \frac{3}{4}e^{-t}$
- $P_{ij}(t) = \frac{1}{4} - \frac{1}{4}e^{-t}, \quad i \neq j$

GTR Model, Tavaré 1986

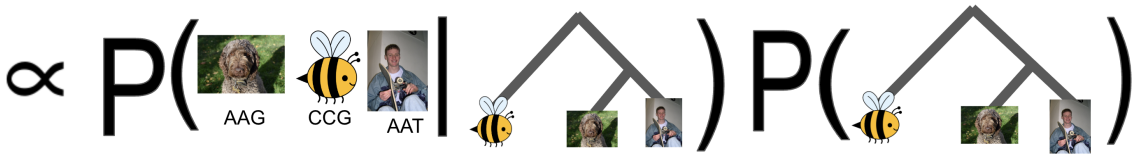
- Time-reversible
- Variable rate parameters $\alpha, \beta, \gamma, \delta, \epsilon, \eta$
- Underlying base frequencies π_i variable

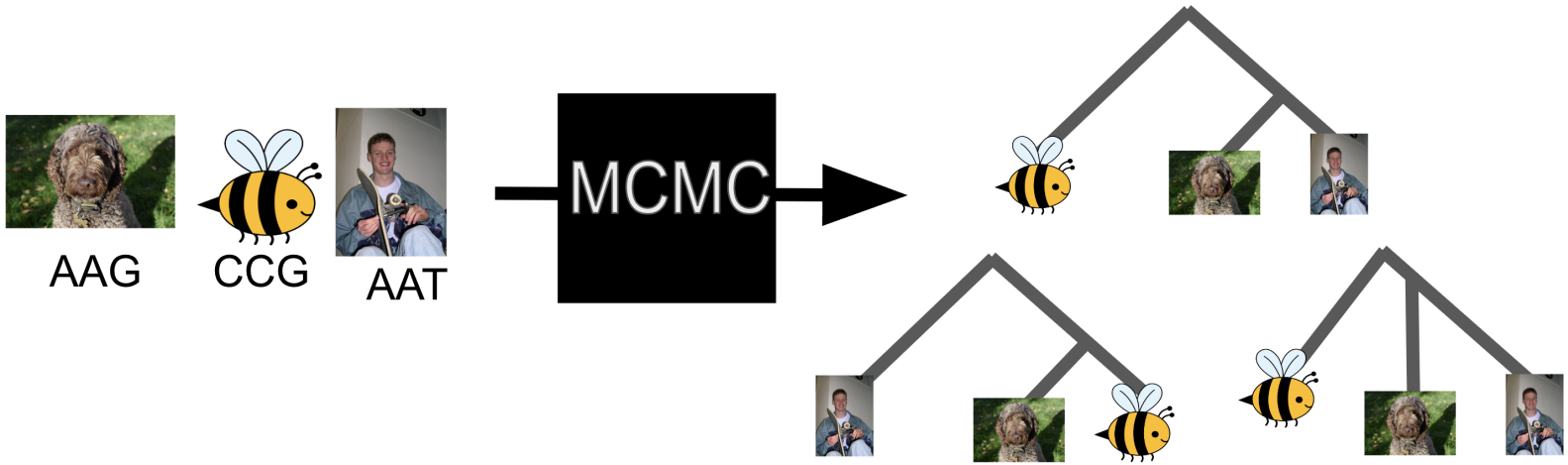
$$Q = \begin{pmatrix} * & \alpha\pi_G & \beta\pi_C & \gamma\pi_T \\ \alpha\pi_A & * & \delta\pi_C & \epsilon\pi_T \\ \beta\pi_A & \delta\pi_G & * & \eta\pi_T \\ \gamma\pi_A & \epsilon\pi_G & \eta\pi_C & * \end{pmatrix}$$

Markov chain Monte Carlo

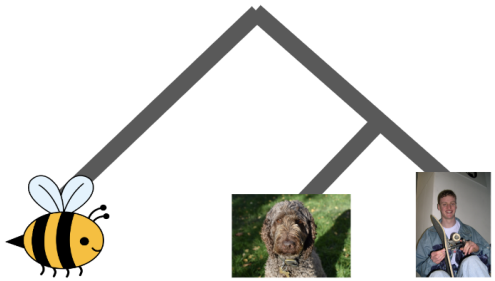
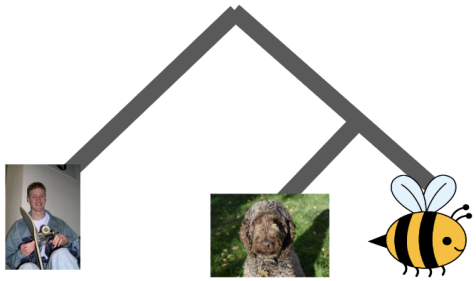


Markov chain Monte Carlo





Summary Statistics

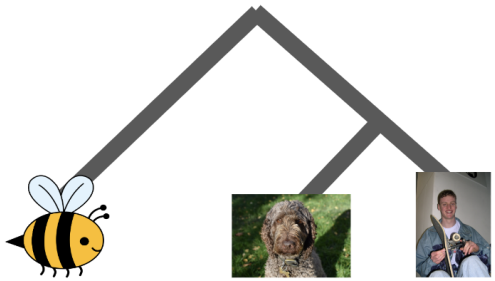
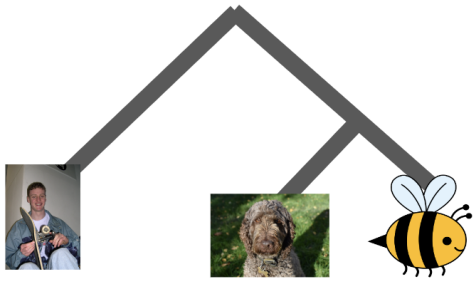
Topology	Proportion
	66.7%
	33.3%

Method

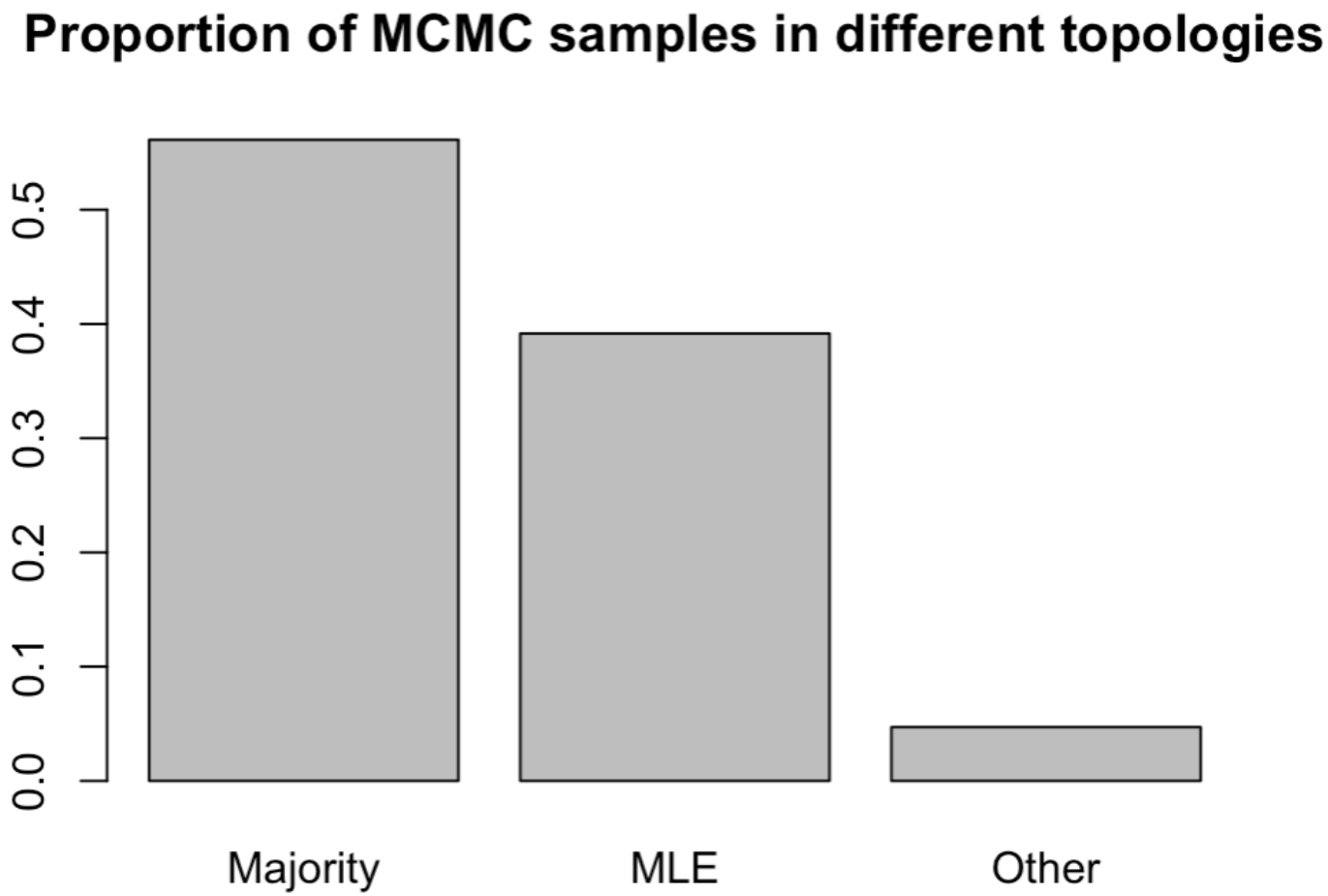
We analysed over 600 different genes. For each gene, we:

- Computed the tree which maximises the likelihood (the MLE) using IQ-TREE2 [1].
- Ran Markov chain Monte Carlo to sample thousands of trees from the posterior distribution using MrBayes 3 [2]. We then computed summary statistics of our samples.
- Explored relationship between MLE structure and our MCMC summary statistics.

Summary Statistics

Topology	Proportion
	66.7%
	33.3%

Gene 262 Topology Proportions



MLE not in Majority Topology

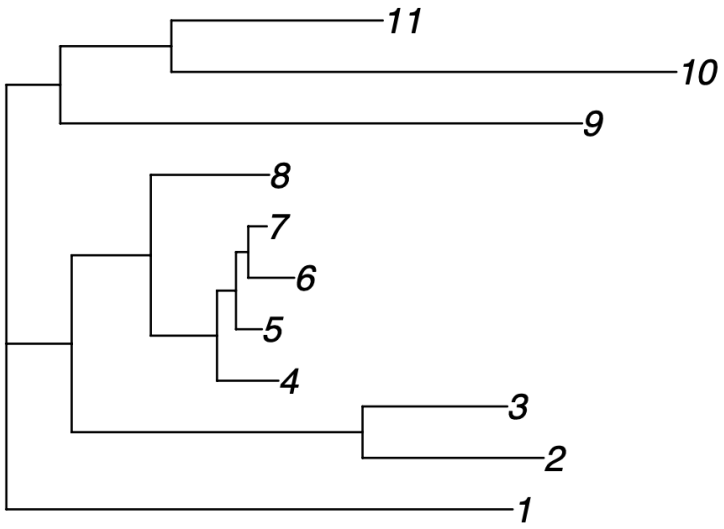


Figure: MLE Tree

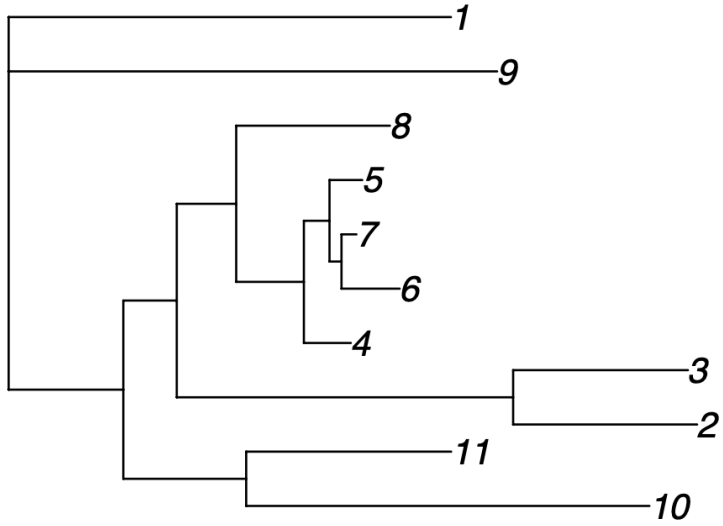
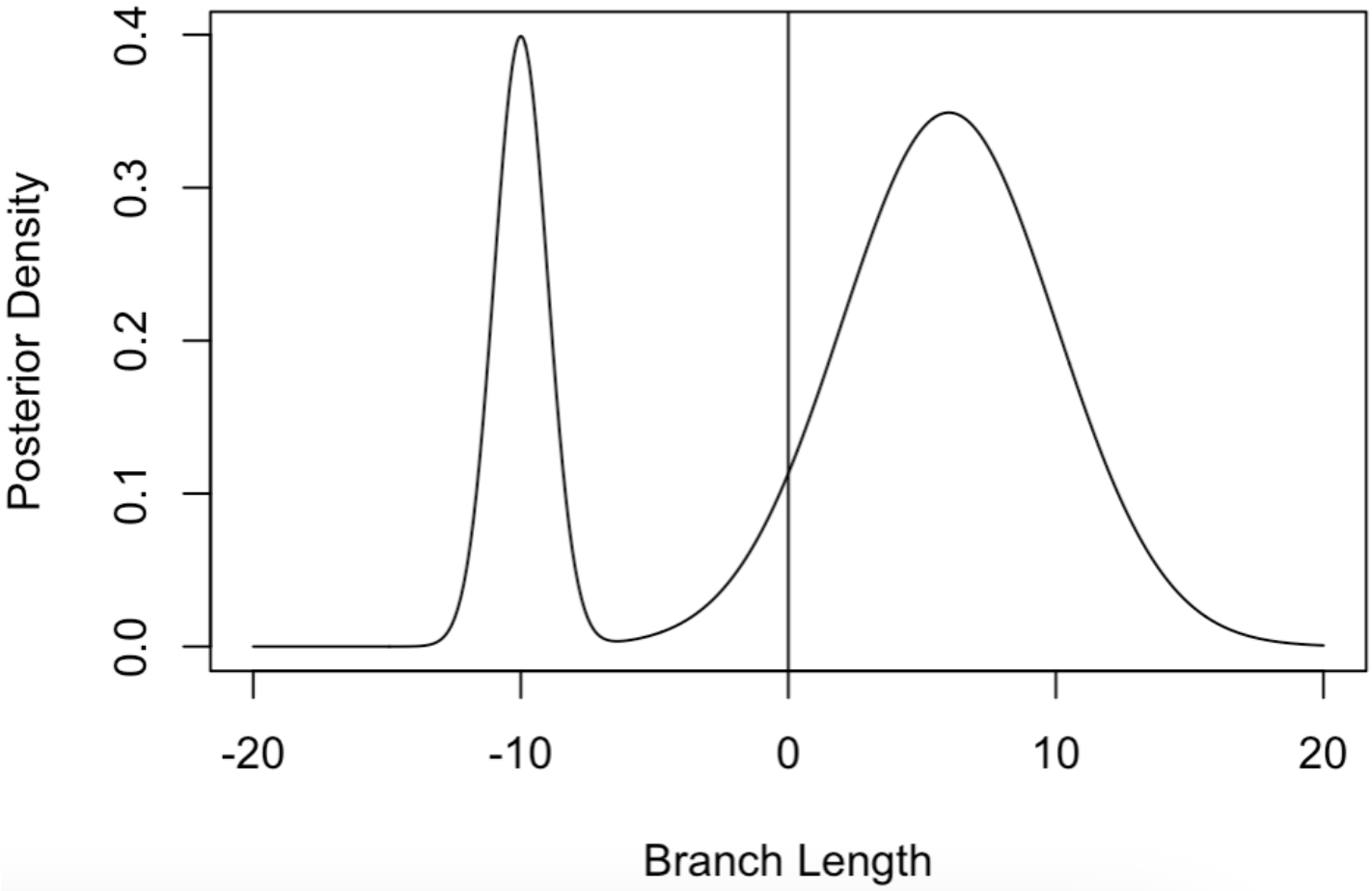
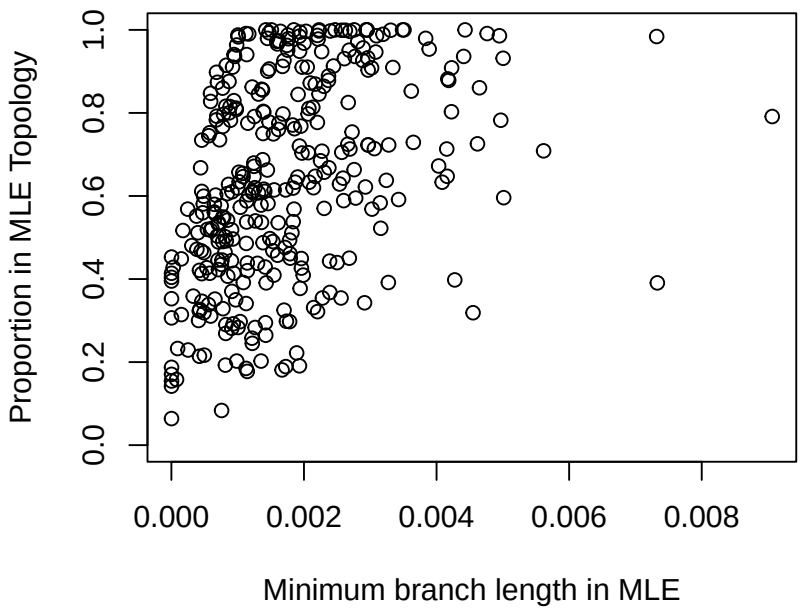


Figure: Tree in MCMC majority topology

Future Work



Correlation between MLE and samples



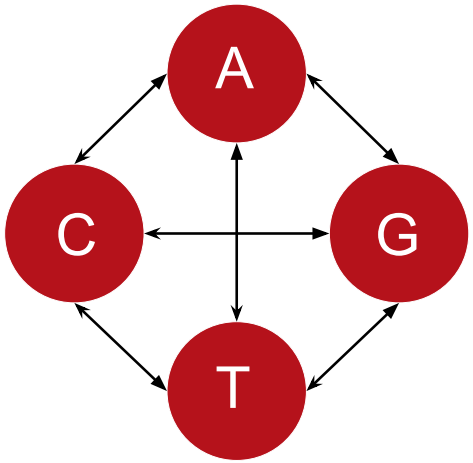
References

[1] Bui Quang Minh, Heiko A Schmidt, Olga Chernomor, Dominik Schrempf, Michael D Woodhams, Arndt von Haeseler, and Robert Lanfear. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. 2020.

[2] Fredrik Ronquist and John P. Huelsenbeck. MrBayes 3: Bayesian phylogenetic inference under mixed models. 2003.

Transition Rates

$$Q = \begin{pmatrix} -\mu_A & \mu_{AG} & \mu_{AC} & \mu_{AT} \\ \mu_{GA} & -\mu_G & \mu_{GC} & \mu_{GT} \\ \mu_{CA} & \mu_{CG} & -\mu_C & \mu_{CT} \\ \mu_{TA} & \mu_{TG} & \mu_{TC} & -\mu_T \end{pmatrix}$$



Continuous Time Markov Chains

$$P'(t) = P(t)Q$$

$$P(0) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Continuous Time Markov Chains

$$P'(t) = P(t)Q$$

$$P(0) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

$$P(t) = \exp(Qt) = \sum_{n=0}^{\infty} Q^n \frac{t^n}{n!}$$