BEAST2 Analysis of Algonquian Language Data

Benjamin Teo

University of Wisconsin-Madison

bteo@wisc.edu

September 11, 2020

Background

Tree of Languages

Figure: Illustration by cartoonist Minna Sundberg

Cognates: Words that have a common historical origin.

1. Night (English), Nacht (German), Nyx (Ancient Greek)
    1. Mozart's Eine kleine Nachtmusik ('A little night music')
    2. Nyx, the primordial deity of the night in Greek mythology

False Cognates: Possibly similar sounds/meaning, but different origin.

1. emoticon (English), emoji/絵文字 (Japanese)

Problem setup

Dataset: A 52-by-9 matrix representing cognate data from 52 words across 9 languages.

1. Each row corresponds to a different word, and uses non-negative integers to indicate the cognate classes for that word.
2. X's indicate cases where cognacy was undeterminable.

Pl Cree	SW Oj	Chey	Menom	Pot.	Meskw.	Shaw	Miami	Pass-M
0	0	2	1	0	0	0	0	0
2	0	X	0	0	3	0	1	4

Table: Rows 14-15 from the dataset. Row 14 indicates 3 cognate classes, while row 15 indicates 5 cognate classes and one language (Chey) for which cognacy was undeterminable.

Problem setup

Reformatted data: A 9-by-152 matrix of mostly 0's and 1's, with the exception of some ?'s.

1. Each row corresponds to a different language, and each column corresponds to a cognate class for some word.
2. 0, 1, and ? respectively indicate cognate absence, presence, and undeterminable cognacy.

Pl Cree	1	0	0	0	0	1	0	0
SW Oj	1	0	0	1	0	0	0	0
Chey	0	0	1	?	?	?	?	?
Menom	0	1	0	1	0	0	0	0
Pot.	1	0	0	1	0	0	0	0
Meskw.	1	0	0	0	0	0	1	0
Shaw	1	0	0	1	0	0	0	0
Miami	1	0	0	0	1	0	0	0
Pass-M	1	0	0	0	0	0	0	1

Table: Reformatted information from rows 14-15 of the original dataset. The light-gray columns correspond to row 14, and the gray columns correspond to row 15.

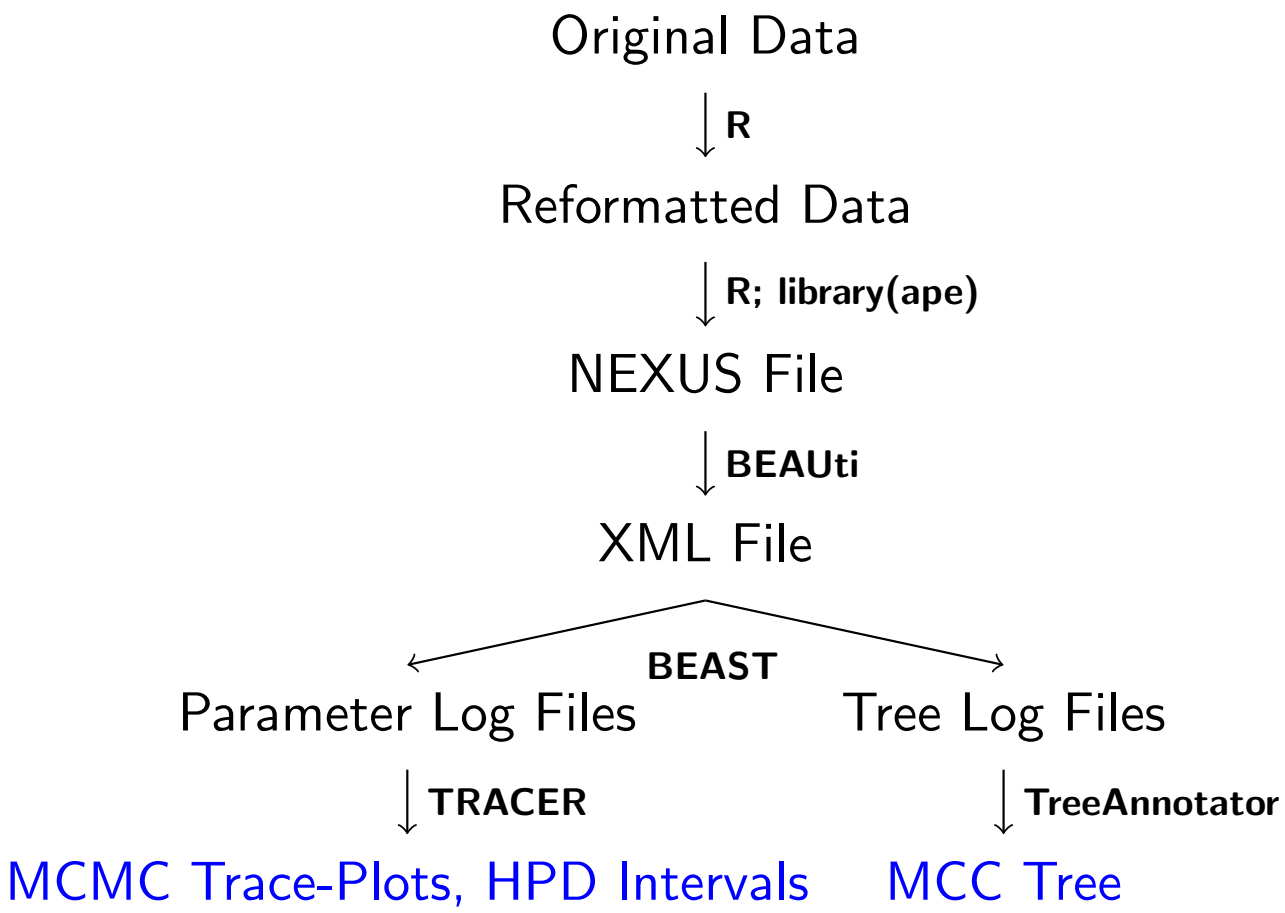
Problem setup

Objectives:

1. Compare how well 6 different models of cognate evolution explain the dataset.
    1. 3 possible site-state models: Binary CTMC, Binary Covarion, Stochastic Dollo.
    2. 2 possible clock models: Strict, Log-normal Relaxed.
2. Infer the phylogeny of these 9 languages, assuming this is a tree.

Why these models? An attempt to reproduce the methodology adopted in *Bowern, Claire and Atkinson, Quentin, 2012. Computational Phylogenetics and the Internal Structure of Pama-Nyungan, Language, Vol. 88, 817-845*

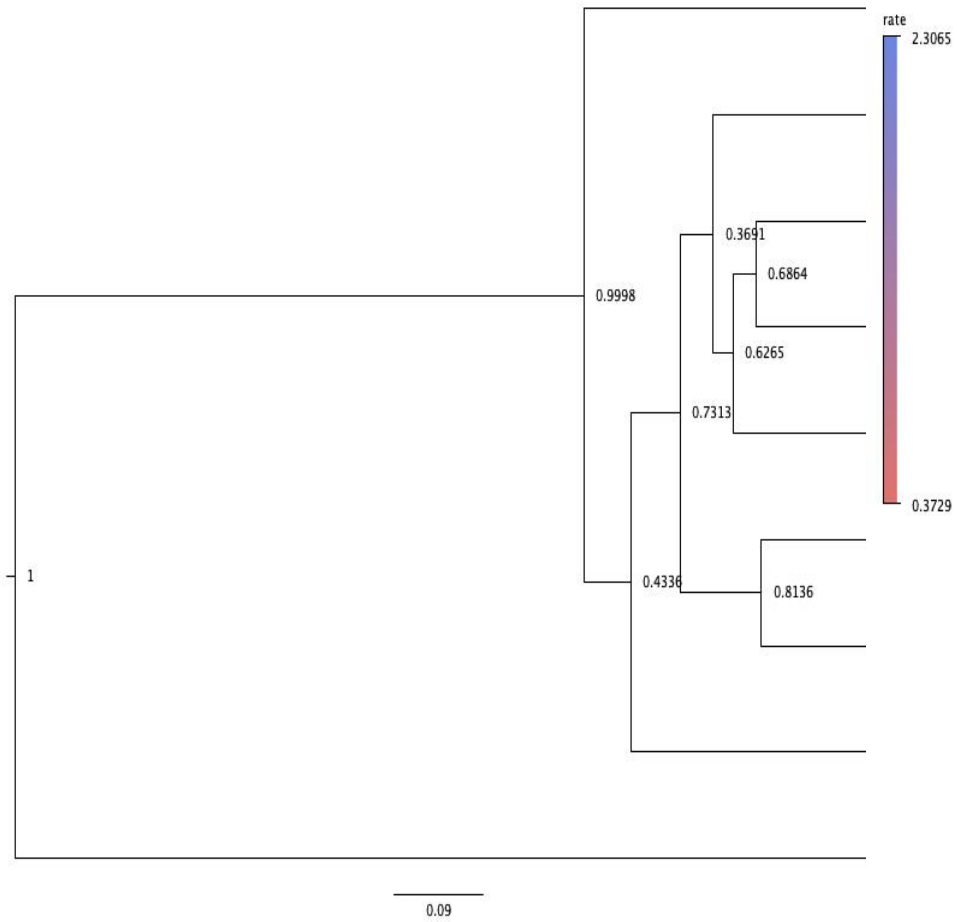
Pipeline



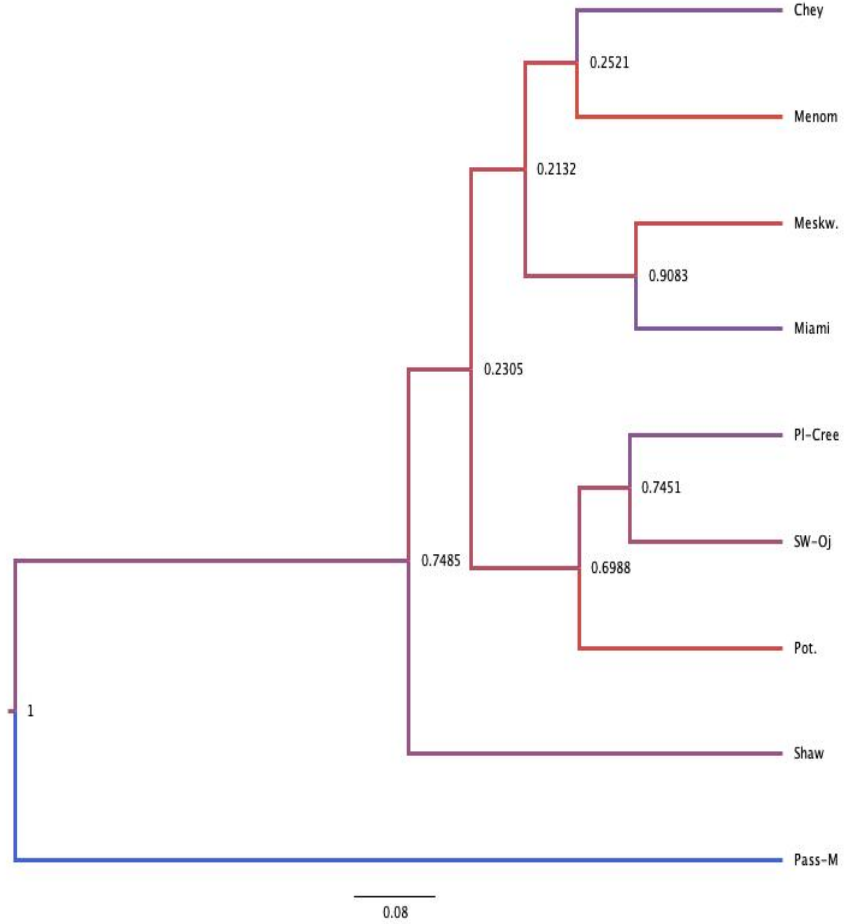
*HPD: Highest Posterior Density
*MCC: Maximum Clade Credibility

Results

MCC Trees: Visualized as ultrametric time-trees labeled with internal-edge posterior probabilities, and colored by rate.



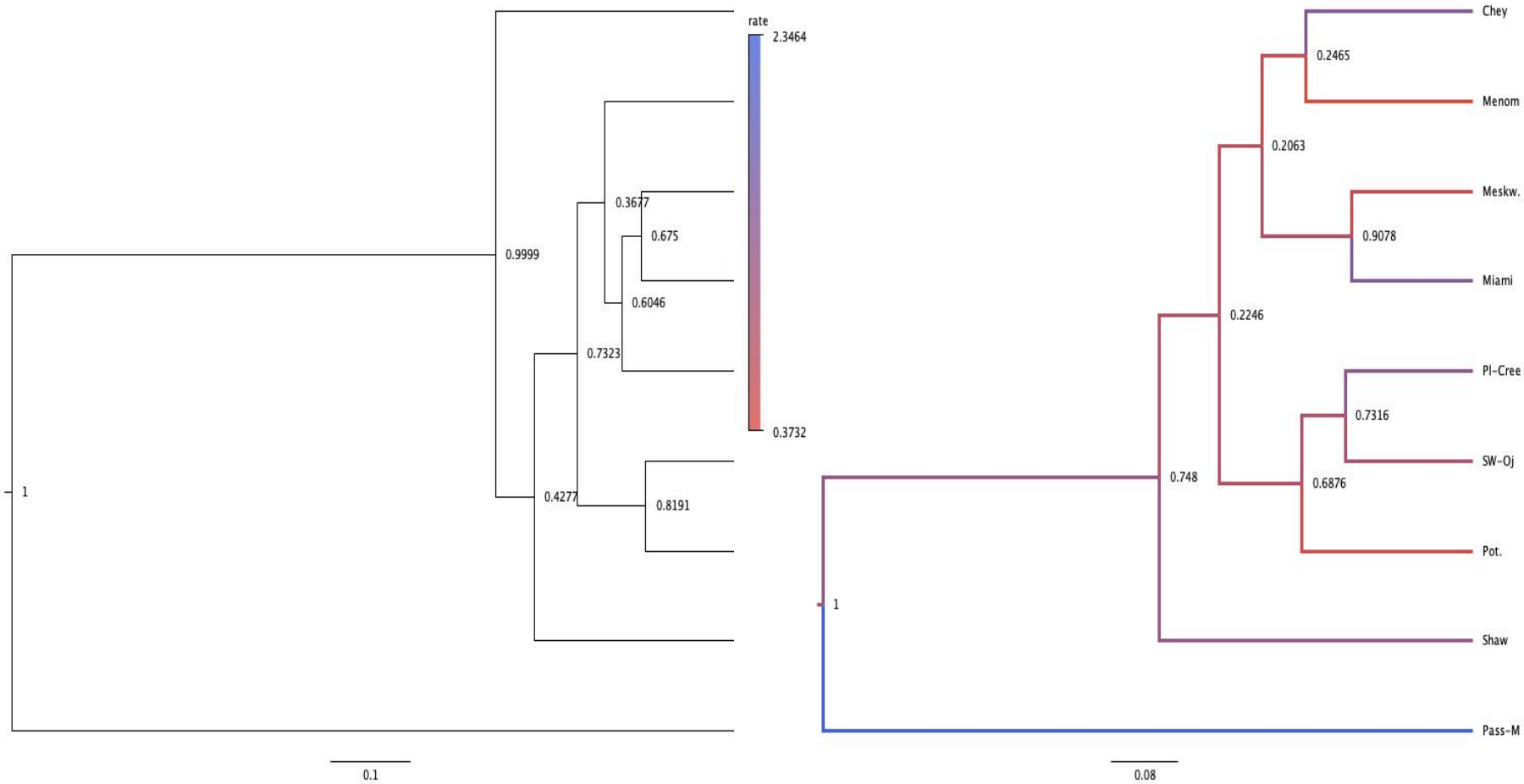
(a) Strict clock



(b) Relaxed clock

Figure: CTMC model

Results

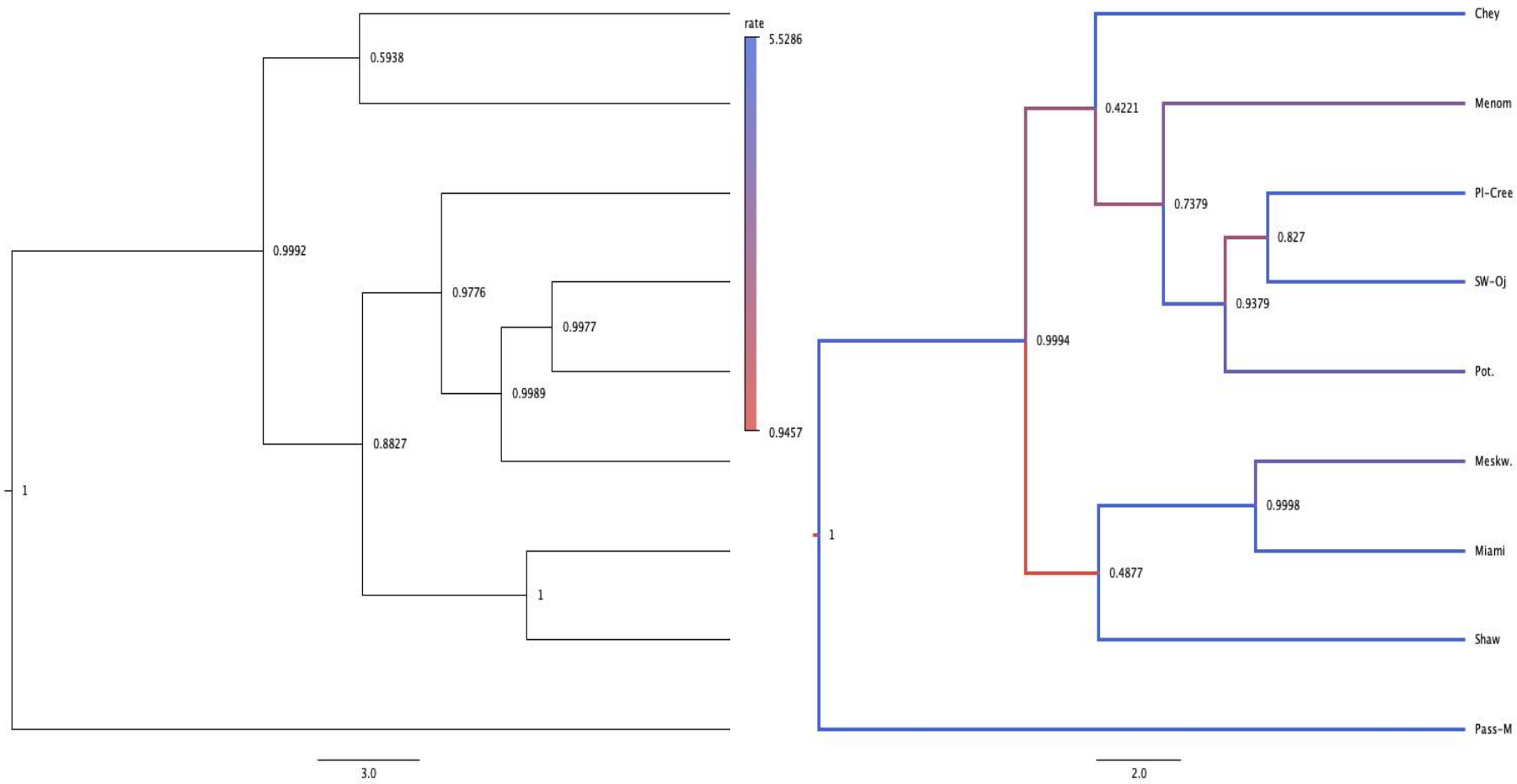


(a) Strict clock

(b) Relaxed clock

Figure: Covarion model

Results



(a) Strict clock **(b)** Relaxed clock

Figure: S. Dollo model

Results

Recall that $\text{BF}(\text{Model 1}, \text{Model 2}) = \frac{Pr(\text{Data}|\text{Model 1})}{Pr(\text{Data}|\text{Model 2})}$ so that:

$$\log BF(\text{M1}, \text{M2}) = \log Pr(\text{D} \mid \text{M1}) - \log Pr(\text{D} \mid \text{M2})$$

Bayes Factor Comparison:

Model	Marginal L estimate	CTMC		Covarion		S. Dollo	
		strict	relaxed	strict	relaxed	strict	relaxed
CTMC strict	-691	—					
CTMC relaxed	-646 (-687)	45 (4)	—				
Covarion strict	-693	-2	-47 (-6)	—			
Covarion relaxed	-647 (-689)	44 (2)	-1 (-2)	46 (4)	—		
S. Dollo strict	-873	-182	-227 (-186)	-180	-226 (-184)	—	
S. Dollo relaxed	-818	-127	-172 (-131)	-125	-171 (-129)	55	—

Table: Table of marginal likelihoods and their differences, all on the log scale. The row corresponding to the best performing model is highlighted in light-gray.

Improves model fit: Relaxed clock assumption
Does not improve model fit: Covarion assumptions
Worsens model fit: Stochastic Dollo assumptions

Results

(Binary CTMC, Relaxed clock) MCC Tree:

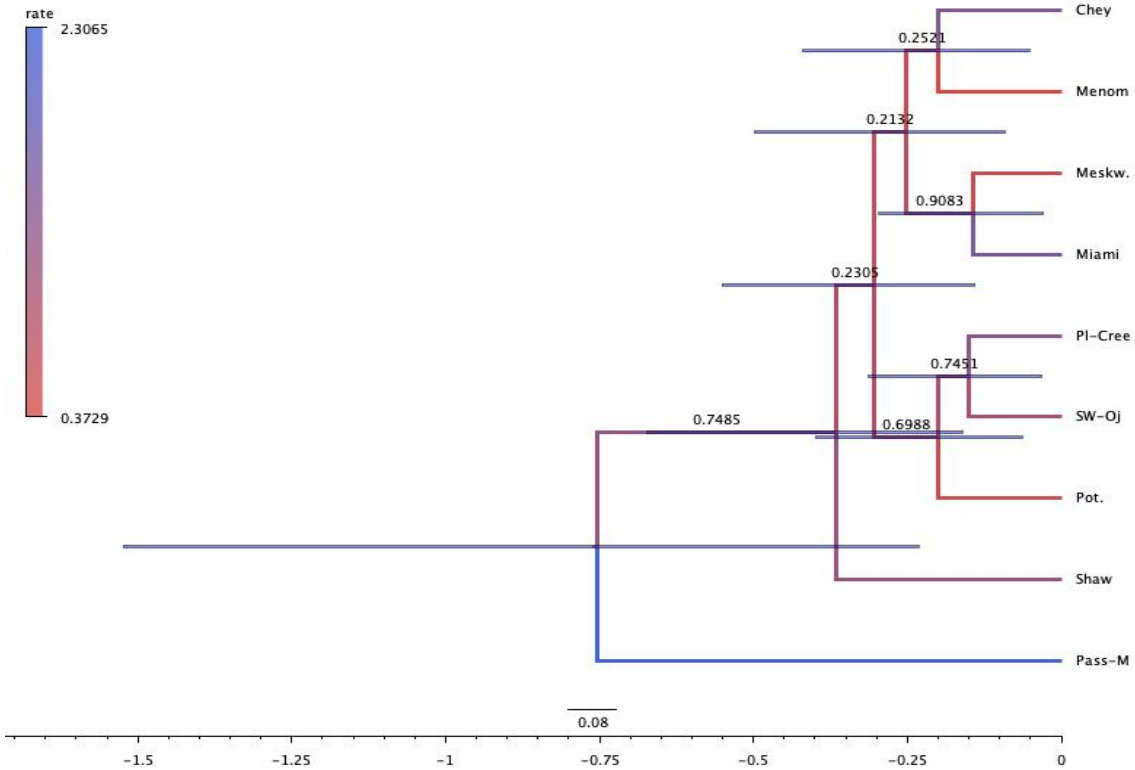


Figure: CTMC-relaxed model revisited. The 95% HPD intervals for node ages are shown as blue rectangles, whereas the internal edges are labeled by posterior support.

Story: The Miami-Meskw. clade is strongly supported. Other than that, many configurations of internal branching are plausible.

Models

Data-generating process

Generate time-tree: Yule process

↓

Adjust branch lengths to reflect evolutionary change: Clock model

↓

Propagate site-states down tree: Site-state model

Models

Yule process

1. Branching birth process with Exp(λ_y) waiting times.
    1. Time is proportional to evolutionary change.
    2. Elapsed time between parent and child is iid Exp(λ_y).
    3. $1/\lambda_y$ is the expected branch length or elapsed time between parent and child.
    4. $\uparrow \lambda_y \downarrow$ expected branch length, and $\downarrow \lambda_y \uparrow$ expected branch length.
    5. λ_y controls the expected total evolutionary change.

Clock model

1. Generate clock rates for each branch.
    1. Strict clock: the rates are all 1.
    2. Log-normal relaxed clock: the rates are iid Log-normal($-\sigma^2_{\text{clock}}/2, \sigma^2_{\text{clock}}$)
        1. $E[\text{Log-normal}(\mu, \sigma^2)] = \exp(\mu + \sigma^2/2)$
        2. $E[\text{Log-normal}(-\sigma^2_{\text{clock}}/2, \sigma^2_{\text{clock}})] = 1$
2. Scale each branch by its corresponding clock rate.
    1. The expected branch length is still λ_y.
    2. Instead of $1/\lambda_y^2$, the variance is now $(1/\lambda_y^2)(2\exp(\sigma^2_{\text{clock}}) - 1) \in (1/\lambda_y^2, \infty)$.
    3. σ^2_{clock} controls the variation in evolutionary change between parent and child.

Picture

Models

Binary CTMC

Equilibrium proportions, π and normalized rate matrix, Q

$$\pi = [f_0, (1 - f_0))]$$

$$Q = \frac{1}{2f_0(1 - f_0)} \begin{bmatrix} -(1 - f_0) & (1 - f_0) \\ f_0 & -f_0 \end{bmatrix}$$

1. 'Evolutionary change' quantified in terms of number of substitutions per site.
2. 'Normalized' means that $\pi_1|Q_{11}| + \pi_2|Q_{22}| = 1$ (i.e. the expected substitution rate is 1).
3. Expected number of substitutions per site between parent and child is λ_y.

Binary Covarion

Possible states = {slow-absent, slow-present, fast-absent, fast-present}
Switch rate, s. Scaling factor, α_{slow}. Normalizing constant, η.

$$\pi = [f_0/2, (1 - f_0)/2, f_0/2, (1 - f_0)/2]$$

$$Q = \frac{1}{\eta} \begin{bmatrix} -(\alpha_{\text{slow}}(1 - f_0) + s) & \alpha_{\text{slow}}(1 - f_0) & s & 0 \\ \alpha_{\text{slow}} f_0 & -(\alpha_{\text{slow}} f_0 + s) & 0 & s \\ s & 0 & -(1 - f_0 + s) & (1 - f_0) \\ 0 & s & f_0 & -(s + f_0) \end{bmatrix}$$

1. 'slow' and 'fast' are latent states like in a Hidden Markov Model. For the sake of time-reversibility, their equilibrium proportions are assumed to be 1/2.

Models

Stochastic Dollo

1. The waiting times for each cognate to be "born" on the tree are iid Exp(ν).
2. Once a cognate is born, it undergoes a branching death process with Exp(μ) waiting times.
    1. Problem: Can we ensure that the expected substitution rate is 1?
    2. If the expected substitution rate $\neq 1$, then branch length does not represent number of substitutions per site.
    3. If the expected substitution rate varies along the tree, then there is no simple interpretation for branch length.
        1. Furthermore, λ_y and μ cannot simultaneously be estimated.

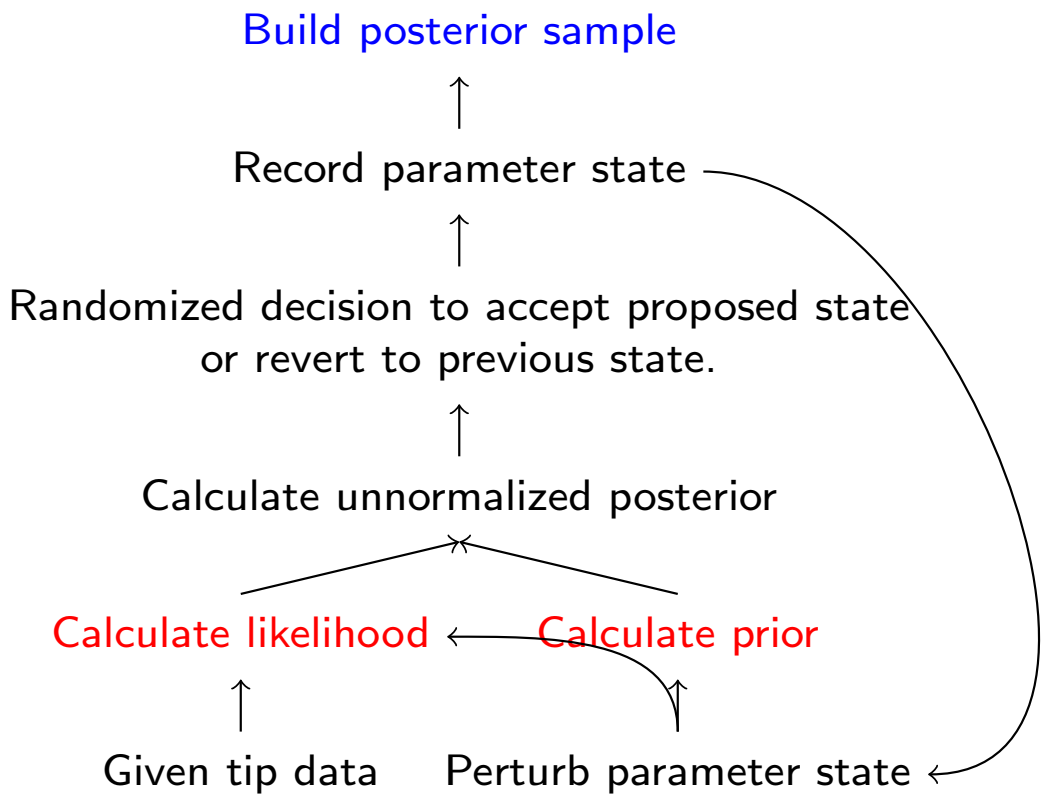
***Caveat: Site-heterogeneity**

1. Allow for variation across sites.
2. Site-specific rates, $r_i \overset{\text{iid}}{\sim}$ Discretized-Gamma($\alpha_{site}, 1/\alpha_{site}$), where the number of categories is 4.
    1. For the Binary CTMC/Covarion models, the site-specific transition probability matrix is given by $P_{l^*} = e^{r_i l^* Q}$, where l^* corresponds to branch length traversed.
        1. $E[\text{Gamma}(\alpha_{\text{site}}, 1/\alpha_{\text{site}})] = 1$ preserves the expected branch length.
    2. For the Stochastic Dollo model, the r_i's scale the birth and death rates (e.g. Exp($r_i\nu$) and Exp($r_i\mu$)).

Picture

Inference

Bayesian inference process



Inference

Estimated parameters

In practical terms, estimating a parameter means selecting a prior for it.

Parameter	CTMC		Covarion		S. Dollo	
	strict	relaxed	strict	relaxed	strict	relaxed
Yule birth rate, λ_y	✓	✓	✓	✓	✓	✓
Scale parameter (relaxed clock), σ_{clock}		✓		✓		✓
Scale parameter (site-heterogeneity), α_{site}	✓	✓	✓	✓	✓	✓
Equilibrium proportion (absent state), f_0	✓	✓	✓	✓		
Switch rate (between slow/fast modes), s			✓	✓		
Scale parameter (slow mode), α_{slow}			✓	✓		
Cognate birth rate, ν					✓	✓
Cognate death rate, μ					✓	✓

Table: Model parameters

Priors

Tree parameters:

1. $\lambda_y \sim \text{Unif}(0, \infty)$
2. $\sigma_{\text{clock}} \sim \text{Gamma}(\alpha_{\sigma_{\text{clock}}} = 0.5396, \beta_{\sigma_{\text{clock}}} = 0.3819)$

Site-state parameters:

1. $f_0 \sim \text{Unif}(0, 1)$, $s \sim \text{Gamma}(\alpha_s = 0.05, \beta_s = 10)$, $\alpha_{\text{slow}} \sim \text{Unif}(0, \infty)$
2. $\nu = 1$, $\mu \sim \text{Exp}(\text{mean} = 10^{-4})$, $\alpha_{\text{site}} \sim \text{Exp}(\text{mean} = 1)$

Model Selection

Marginal likelihood: $f(D \mid M) = \int_\Theta f(D \mid \theta, M)\pi(\theta \mid M)d\theta$

Importance sampling:

1. Regard the integral as an expectation, and estimate it as a sample mean.
2. Arithmetic Mean (AM) Method: Take $\pi(\theta \mid M)$ as the 'importance distribution'.

$$\int_\Theta f(D \mid \theta, M) \cdot \pi(\theta \mid M)d\theta = E_{\pi(\theta|M)}[f(D \mid \theta, M)]$$

$$\approx \frac{1}{n}\sum_{i=1}^{n} f(D \mid \theta_i, M), \ \theta_i \overset{\text{iid}}{\sim} \pi(\theta \mid M)$$

3. Disadvantage: 'If the likelihood is sharp compared with the prior, the AM estimate can have an unacceptably high variance' (Xie et al. 2011).

*Xie, Wangang & Lewis, Paul & Fan, Yu & Kuo, Lynn & Chen, Ming-Hui. (2011). Improving Marginal Likelihood Estimation for Bayesian Phylogenetic Model Selection. Systematic biology. 60. 150-60. 10.1093/sysbio/syq085.

Model selection

Stepping-stone sampling:

1. Regard the integral as a telescoping product, and estimate each term by importance sampling.
2. The importance distribution for each term is a different power-posterior distribution.

Letting $0 = \beta_0 < \beta_1 < ... < \beta_K = 1$, we have:

$$\begin{aligned}
\int_\Theta f(D \mid \theta, M)\pi(\theta \mid M)d\theta &= \prod_{k=1}^{K} \frac{\int_\Theta f(D \mid \theta, M)^{\beta_k}\pi(\theta \mid M)d\theta}{\int_\Theta f(D \mid \theta, M)^{\beta_{k-1}}\pi(\theta \mid M)d\theta}, \\
&= \prod_{k=1}^{K} \int_\Theta f(D \mid \theta, M)^{\beta_k - \beta_{k-1}} \cdot p_{\beta_{k-1}}(\theta \mid D, M)d\theta \\
&= \prod_{k=1}^{K} E_{p_{\beta_{k-1}}(\theta|D,M)}[f(D \mid \theta, M)^{\beta_k - \beta_{k-1}}] \\
&\approx \prod_{k=1}^{K} \frac{1}{n_k} \sum_{i=1}^{n_k} f(D \mid \theta_{(i,k-1)}, M)^{\beta_k - \beta_{k-1}}, \ \theta_{(i,k)} \overset{\text{iid}}{\sim} p_{\beta_k}, \ i \in \{1, ..., n_k\}
\end{aligned}$$

*$p_{\beta_k}(\theta \mid D, M) = \frac{f(D|\theta,M)^{\beta_k}\pi(\theta|M)}{\int_\Theta f(D|\theta,M)^{\beta_k}\pi(\theta|M)d\theta}$ is simulated through MCMC sampling.

*$\beta_k \overset{\text{iid}}{\sim} \text{Beta}(\alpha_{\text{step}} = 0.3, 1)$

Discussion

Extensions

1. Reduce computing time using high-performance library BEAGLE, which can make use of GPUs.
2. Node calibration.

Challenges

1. Computing time: To run an MCMC chain of length 100,000,000 takes about 2-2.5 hours.
2. Sensitivity analysis.

Thanks!

1. To the Linguisticians (or Linguists?) who shared their data and thoughts.
2. To Prof. Ané, for answering my numerous and repetitive questions.
3. To the audience, for listening.