Introduction

David Goldstein

3 May 2021
Copenhagen phylolinguistics workshop

Introduction

The overarching question

- How do we draw inferences about the unobservable linguistic past from observable linguistic (and non-linguistic) data?

My goals for this workshop

- Provide a conceptual introduction to Bayesian phylogenetic inference

My goals for this workshop

- Provide a conceptual introduction to Bayesian phylogenetic inference
- Introduce you to a .Rev script

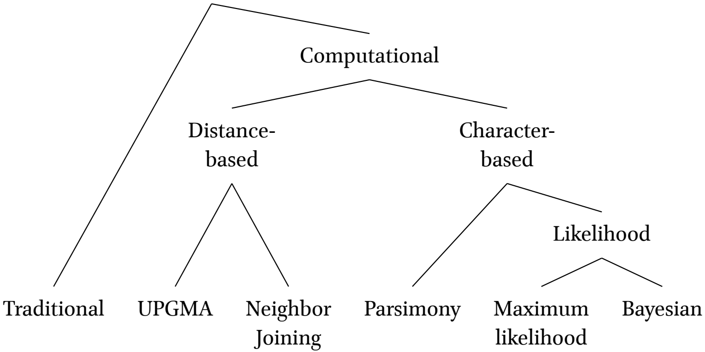
My goals for this workshop

- Provide a conceptual introduction to Bayesian phylogenetic inference
- Introduce you to a .Rev script
- Enable you to interpret some of the results of a Bayesian phylogenetic analysis

What are phylogenetic trees?

A phylogenetic tree is a hypothesis about the specific sequence of historical branching events leading from a common ancestor forwards in time to contemporary groupings, be they biological species or languages. (*Pagel 2017, p. 152*)

Overview of methods in linguistic phylogenetics



Comparing phylolinguistic methods

- Distance-based methods infer a phylogenetic tree on the basis of a distance measure among languages (e.g., the number of cognates/homologous characters) that they share.

Comparing phylolinguistic methods

- Distance-based methods infer a phylogenetic tree on the basis of a distance measure among languages (e.g., the number of cognates/homologous characters) that they share.
- Character-based methods infer a tree on the basis of linguistic characters. In principle, these can be continuous or discrete, but in practice they are overwhelmingly discrete.

Comparing phylolinguistic methods

- Distance-based methods infer a phylogenetic tree on the basis of a distance measure among languages (e.g., the number of cognates/homologous characters) that they share.
- Character-based methods infer a tree on the basis of linguistic characters. In principle, these can be continuous or discrete, but in practice they are overwhelmingly discrete.
- Likelihood and Bayesian methods are probabilistic and assume stochastic models of linguistic change.

Why Bayesian methods?

- They enable us to pursue questions that are otherwise intractable (e.g., ancestral state estimation, divergence-time estimation, diversification rates).

Why Bayesian methods?

- They enable us to pursue questions that are otherwise intractable (e.g., ancestral state estimation, divergence-time estimation, diversification rates).
- The results are straightforward to interpret.

Why Bayesian methods?

- They enable us to pursue questions that are otherwise intractable (e.g., ancestral state estimation, divergence-time estimation, diversification rates).
- The results are straightforward to interpret.
- Estimates of uncertainty (crucial in historical linguistics!) are built in to the model.

Why Bayesian methods?

- They enable us to pursue questions that are otherwise intractable (e.g., ancestral state estimation, divergence-time estimation, diversification rates).
- The results are straightforward to interpret.
- Estimates of uncertainty (crucial in historical linguistics!) are built in to the model.
- Flexibility—it's possible to create a wide array of models.

Why Bayesian methods?

- They enable us to pursue questions that are otherwise intractable (e.g., ancestral state estimation, divergence-time estimation, diversification rates).
- The results are straightforward to interpret.
- Estimates of uncertainty (crucial in historical linguistics!) are built in to the model.
- Flexibility—it's possible to create a wide array of models.
- Bayesian methods can build on traditional knowledge through the use of prior probability distributions.

Increasing prominence in the field

CLADE	LITERATURE
Iranian	Cathcart 2019, Cathcart 2020
Semitic	Kitchen et al. 2009
Dravidian	Kolipakam et al. 2018
Transeurasian	Robbeets and Bouckaert 2018
Pama-Nyungan	Bowern 2012, Bowern and Atkinson 2012, Bouckaert, Bowern, et al. 2018
Turkic	Savelyev and Robbeets 2020
Sino-Tibetan	Sagart et al. 2019, Zhang, Yan, et al. 2019, Zhang, Ji, et al. 2020
Japonic	Lee and Hasegawa 2011
Austronesian	Saunders 2005, Dunn et al. 2008, Gray, Drummond, et al. 2009, Greenhill and Gray 2009, Greenhill, Atkinson, et al. 2010
Indo-European	Gray and Atkinson 2003, Atkinson and Gray 2006, Bouckaert, Lemey, et al. 2012, Chang et al. 2015, Rama 2018
Slavic	Cathcart and Wandl 2020
Dene-Yeniseian	A. Sicoli and Holton 2014, Yanovich 2020
Bantu	Guillon and Mace 2016; Holden et al. 2005

Why RevBayes?

- Can’t be used as a black box—you have to write up the scripts for the analyses.

Why RevBayes?

- Can’t be used as a black box—you have to write up the scripts for the analyses.
- Large community of users (evolutionary biologists mainly)

Why RevBayes?

- Can't be used as a black box—you have to write up the scripts for the analyses.
- Large community of users (evolutionary biologists mainly)
- Tutorials (https://revbayes.github.io/tutorials/)

Why RevBayes?

- Can't be used as a black box—you have to write up the scripts for the analyses.
- Large community of users (evolutionary biologists mainly)
- Tutorials (https://revbayes.github.io/tutorials/)
- Google-user group (https://groups.google.com/g/revbayes-users)

The representation of the data

What kind of data can we use?

- Any heritable and potentially variable observable feature of a language. Characters states may be discrete or continuous.

What kind of data can we use?

- Any heritable and potentially variable observable feature of a language. Characters states may be discrete or continuous.
- In biology, a distinction is drawn between PHENOTYPIC and MOLECULAR data. In linguistics, all our data is phenotypic.

What kind of data can we use?

- Any heritable and potentially variable observable feature of a language. Characters states may be discrete or continuous.
- In biology, a distinction is drawn between PHENOTYPIC and MOLECULAR data. In linguistics, all our data is phenotypic.
- DISCRETE CHARACTER: Any homologous or cross-linguistically comparable trait or feature whose possible states are finite.

What kind of data can we use?

- Any heritable and potentially variable observable feature of a language. Characters states may be discrete or continuous.
- In biology, a distinction is drawn between PHENOTYPIC and MOLECULAR data. In linguistics, all our data is phenotypic.
- DISCRETE CHARACTER: Any homologous or cross-linguistically comparable trait or feature whose possible states are finite.
- It is possible to carry out phylogenetic inference on continuous characters, but I'm not aware of any attempts to do this in a linguistic context.

Multistate character (Lexical cognates)

Language	‘father’	Class
Old English	*fæder*	1
Gothic	*fadar, atta*	1, 2
Old Norse	*faðir*	1
Latin	*pater*	1
Rumanian	*tată*	3
Greek	πατήρ	1
Sanskrit	*pitṛ́*	1
Old Irish	*athir*	1
Armenian	*hayr*	1
Hittite	*attas*	2?
Tocharian A	*pācar*	1
Tocharian B	*pācer*	1
Lithuanian	*tėvas*	4
OCS	отьць	5
Russian	отец	5

Binary character (Lexical cognates)

Language	'father'	1	2	3	4	5
Old English	*fæder*	1	0	0	0	0
Gothic	*fadar*, *atta*	1	1	0	0	0
Old Norse	*faðir*	1	0	0	0	0
Latin	*pater*	1	0	0	0	0
Rumanian	*tată*	0	0	1	0	0
Greek	πατήρ	1	0	0	0	0
Sanskrit	*pitṛ́*	1	0	0	0	0
Old Irish	*athir*	1	0	0	0	0
Armenian	*hayr*	1	0	0	0	0
Hittite	*attas*	0	1?	0	0	0
Tocharian A	*pācar*	1	0	0	0	0
Tocharian B	*pācer*	1	0	0	0	0
Lithuanian	*tėvas*	0	0	0	1	0
OCS	отьць	0	0	0	0	1
Russian	отец	0	0	0	0	1

Partial cognates

- Very important, but for another day...

Partial cognates

- Very important, but for another day...
- See List 2016 if you're curious.

Let’s look at our dataset (the .nex file)

#NEXUS

BEGIN DATA;
DIMENSIONS NTAX=24 NCHAR=294;
FORMAT DATATYPE=STANDARD MISSING=? GAP=- INTERLEAVE=NO
symbols=“01”;
MATRIX
Hittite ?00000000000000000000001000000
Armenian 10001110000000000000000010000

“Observed data”

- We often talk about the “observed data” in a maximum likelihood or Bayesian context.

“Observed data”

- We often talk about the “observed data” in a maximum likelihood or Bayesian context.
- This term is slightly misleading, because it conceals the intervention of the researcher. The data have not simply been “observed,” they’ve been *selected*!

“Observed data”

- We often talk about the “observed data” in a maximum likelihood or Bayesian context.
- This term is slightly misleading, because it conceals the intervention of the researcher. The data have not simply been “observed,” they’ve been *selected*!
- Bayesian methods are amazing, but they are not pixie dust: posterior inferences can only be as good as the data on which they are based. There never has been and never will be an exception to this truth.

Let's look at our RevBayes script

- Open `copenhagen-ringe-mk-ard-cc-prior.Rev`

Bayesian inference

The fundamental idea

Informal statement

- What is the probability of something unobserved or unobservable (such as a phylogenetic tree) given data that we can observe?

Bayes' Theorem

Bayes' Theorem

$$p(\theta|y) = \frac{p(y|\theta) \cdot p(\theta)}{p(y)}$$

y	"Observed data"
θ	Unobserved parameter
$p(\theta\|y)$	Posterior probability
$p(y\|\theta)$	Likelihood
$p(\theta)$	Prior probability
$p(y)$	Marginal likelihood

Conditional probability

- $p(\theta|y)$ is a conditional probability.

Conditional probability

- $p(\theta|y)$ is a conditional probability.
- A conditional probability denotes the probability of an event or a parameter given the occurrence or presence of some value y.

Conditional probability

- $p(\theta|y)$ is a conditional probability.
- A conditional probability denotes the probability of an event or a parameter given the occurrence or presence of some value y.
- How probable is it that a player will draw a card from a particular suit? 13/52 = 0.25.

Conditional probability

- $p(\theta|y)$ is a conditional probability.
- A conditional probability denotes the probability of an event or a parameter given the occurrence or presence of some value y.
- How probable is it that a player will draw a card from a particular suit? 13/52 = 0.25.
- Now suppose that the player draws a spade. What is the probability that he will draw a second spade?

Conditional probability

- $p(\theta|y)$ is a conditional probability.
- A conditional probability denotes the probability of an event or a parameter given the occurrence or presence of some value y.
- How probable is it that a player will draw a card from a particular suit? $13/52 = 0.25$.
- Now suppose that the player draws a spade. What is the probability that he will draw a second spade?
- Since one spade has already been drawn, only 12 remain. Since one card has already been drawn, the total number of cards in the deck is 51. $P(\text{Draw spade}|\text{First draw a spade}) = 12/51 = 0.24$.

Bayes' Theorem restated

Bayes' Theorem without the normalizing constant (i.e., as an unnormalized density)

$$p(\theta|y) \propto p(y|\theta) \cdot p(\theta)$$

Bayes' Theorem in a phylogenetic context

Bayes' Theorem in phylogenetics

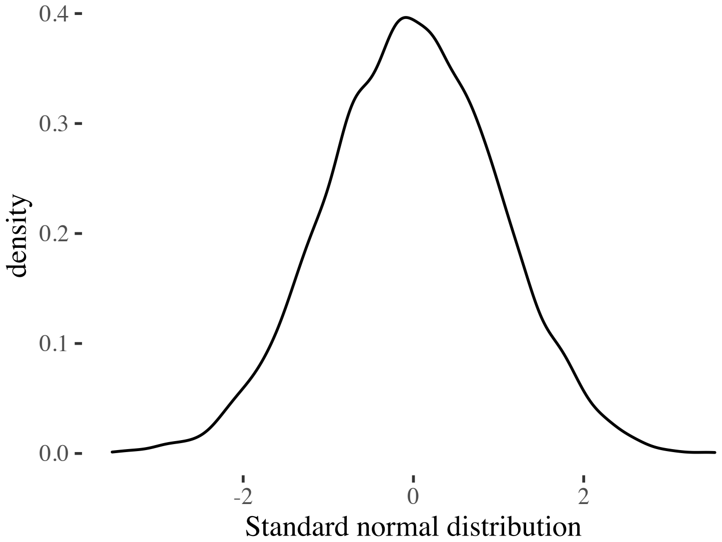
$$p(\Phi, \nu, \Phi | y) \propto p(y | \Psi, \nu, \Phi) \times p(\Psi, \nu, \Phi)$$

- Φ Tree topology
- ν Branch lengths
- Φ Parameters associated with transition model
- y Observed data

Excursus: Uniform distribution $\theta \sim \mathcal{U}(0, 4)$

Excursus: Standard normal distribution

$$\theta \sim \mathcal{N}(\mu = 0, \sigma = 1)$$



Bayesian analyses yield posterior distributions

- Phylogenetic trees

Bayesian analyses yield posterior distributions

- Phylogenetic trees
- Rates of change

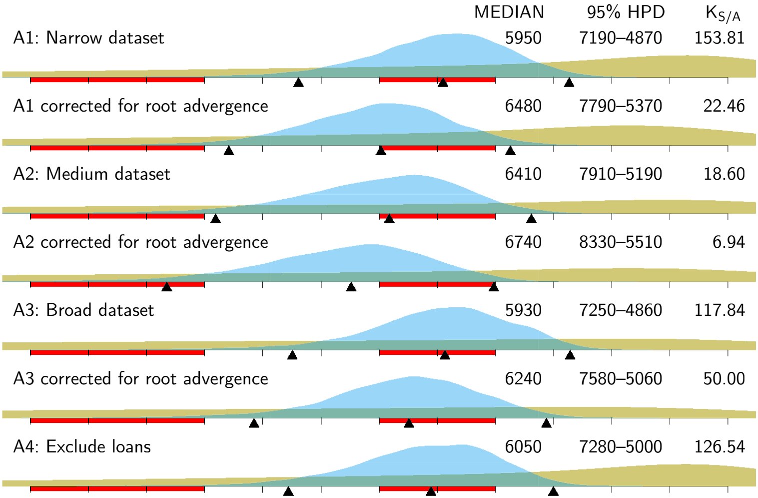
Bayesian analyses yield posterior distributions

- Phylogenetic trees
- Rates of change
- Branch lengths

Bayesian analyses yield posterior distributions

- Phylogenetic trees
- Rates of change
- Branch lengths
- Ancestral states

Posterior distributions for root age (Chang et al. 2015)



An illustration of Bayesian inference

Is the mint producing fair coins?

- Our mint is issuing coins and we're trying to establish whether or not they're fair.

Is the mint producing fair coins?

- Our mint is issuing coins and we're trying to establish whether or not they're fair.
- Fair coins are defined as having a probability heads or tails right about 50%.

Is the mint producing fair coins?

- Our mint is issuing coins and we're trying to establish whether or not they're fair.
- Fair coins are defined as having a probability heads or tails right about 50%.
- This value is a parameter—it controls the rate at which heads and tails show up.

Is the mint producing fair coins?

- To see if our mint is producing fair coins, we're going to take a sample of flips.

Is the mint producing fair coins?

- To see if our mint is producing fair coins, we're going to take a sample of flips.
- We're interested in $p(\theta|y)$.

Is the mint producing fair coins?

- To see if our mint is producing fair coins, we're going to take a sample of flips.
- We're interested in $p(\theta|y)$.
- To use Bayes' Theorem, we need to specify a prior probability distribution.

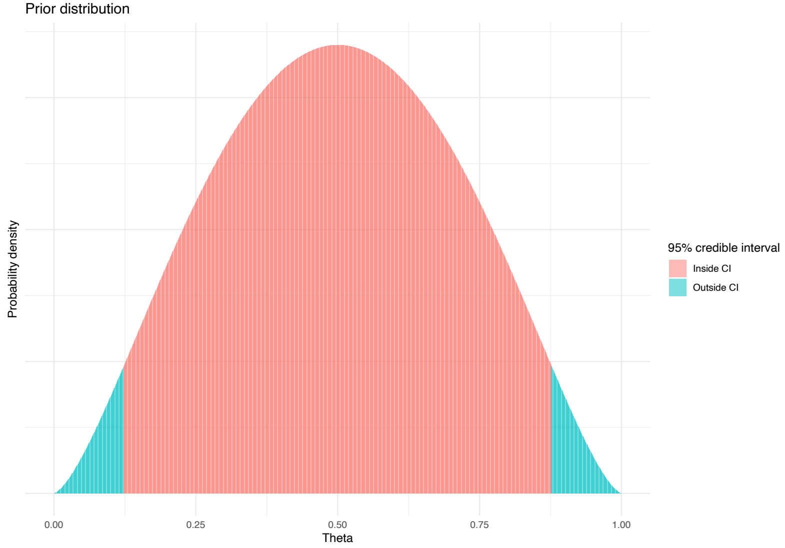
Is the mint producing fair coins?

- To see if our mint is producing fair coins, we're going to take a sample of flips.
- We're interested in $p(\theta|y)$.
- To use Bayes' Theorem, we need to specify a prior probability distribution.
- Let's say we really have no idea—the mint could produce anything.

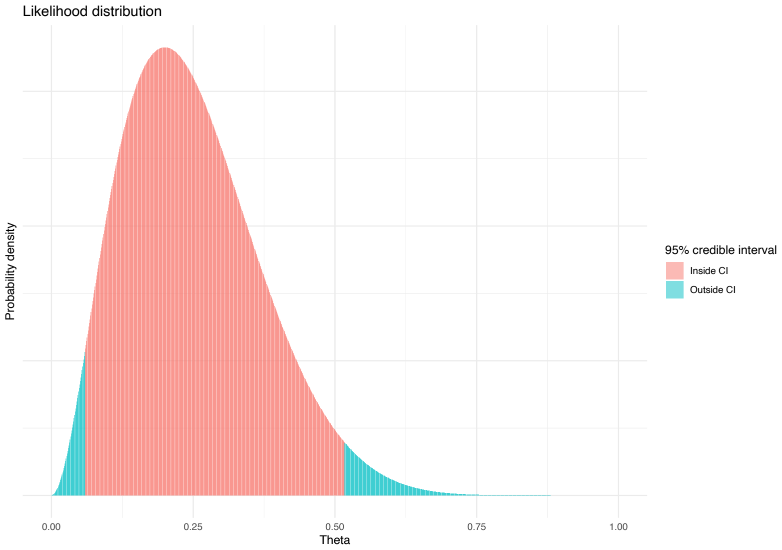
Is the mint producing fair coins?

- To see if our mint is producing fair coins, we're going to take a sample of flips.
- We're interested in $p(\theta|y)$.
- To use Bayes' Theorem, we need to specify a prior probability distribution.
- Let's say we really have no idea—the mint could produce anything.
- We flip the coin 10 times, and heads comes up twice.

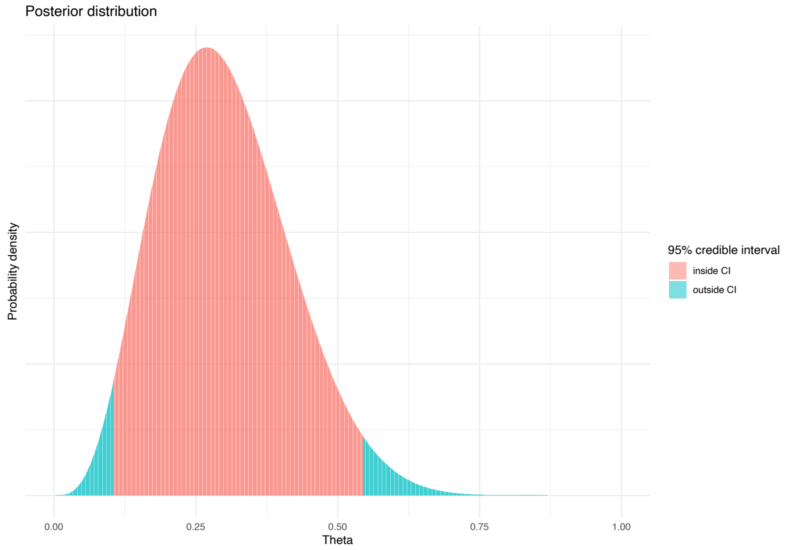
Prior probability distribution



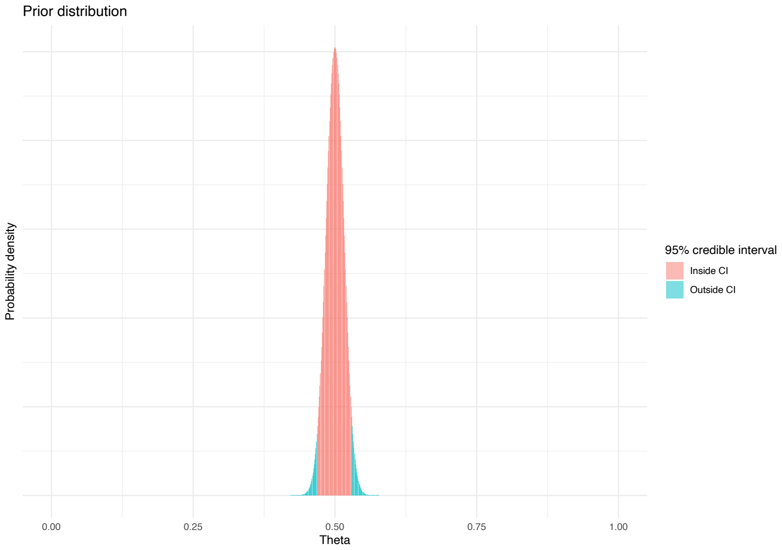
Likelihood $p(y|\theta)$



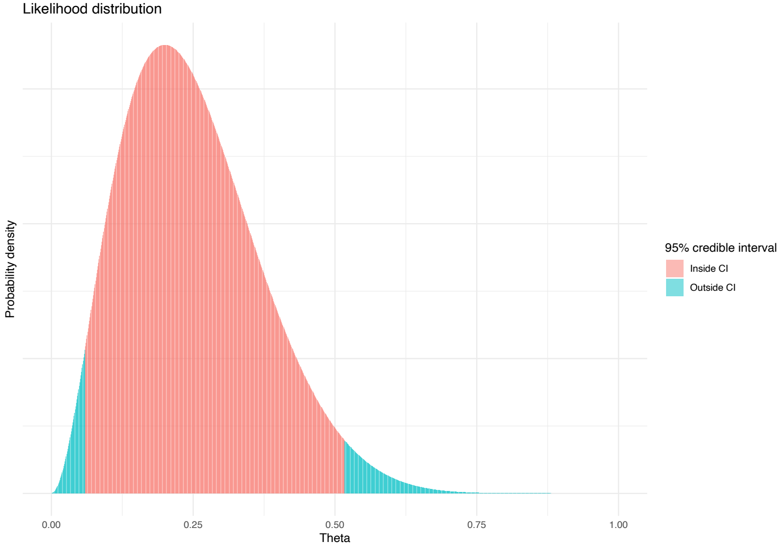
Posterior distribution



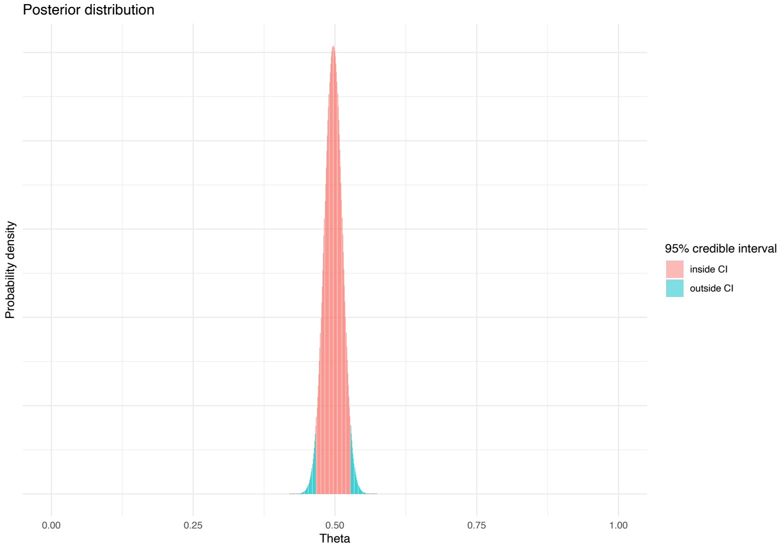
Now let’s change our prior beliefs



Likelihood $p(y|\theta)$ (no change)



Posterior is a compromise between the data and the priors



Rules of thumb

- Priors are a non-trivial aspect of Bayesian inference.

Rules of thumb

- Priors are a non-trivial aspect of Bayesian inference.
- The more observed data you have, the less they impact the posterior.

Rules of thumb

- Priors are a non-trivial aspect of Bayesian inference.
- The more observed data you have, the less they impact the posterior.
- The less observed data you have, the greater their impact on the posterior.

Rules of thumb

- Priors are a non-trivial aspect of Bayesian inference.
- The more observed data you have, the less they impact the posterior.
- The less observed data you have, the greater their impact on the posterior.
- Priors have been hugely controversial in the history of Bayesian inference.

Rules of thumb

- Priors are a non-trivial aspect of Bayesian inference.
- The more observed data you have, the less they impact the posterior.
- The less observed data you have, the greater their impact on the posterior.
- Priors have been hugely controversial in the history of Bayesian inference.
- Priors should be explicit and justifiable. If we can justify informative priors, then we should use them.

Coming back to phylogenetics

We need to able to calculate the likelihood of a sequence

- We ultimately want to calculate $p(\Psi|y)$, but to do that we need to be calculate the likelihood $p(y|\Psi)$.

We need to able to calculate the likelihood of a sequence

- We ultimately want to calculate $p(\Psi|y)$, but to do that we need to be calculate the likelihood $p(y|\Psi)$.
- That is, we need a way to calculate the likelihood of a particular sequence at the tips of the tree given a particular tree.

Calculating likelihood of a sequence

Two fundamental notions of probability theory

Independence

Independence

$P(X \cap Y) = P(X) \cdot P(Y)$ iff $P(Y|X) = P(Y)$

- Multiple flips of a coin are independent.

Independence

Independence

$P(X \cap Y) = P(X) \cdot P(Y)$ iff $P(Y|X) = P(Y)$

- Multiple flips of a coin are independent.
- The result of an earlier flip does not influence the probability of the current flip.

Mutually exclusive events

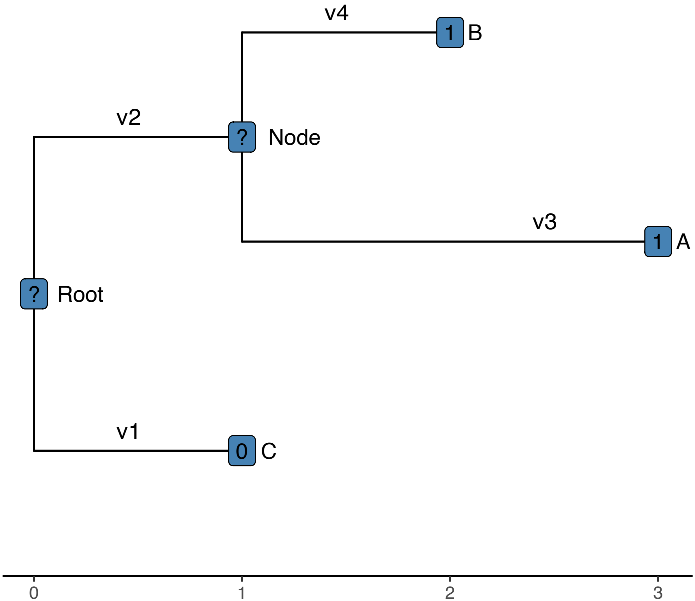
Probability of mutually exclusive (bzw. disjoint) events

When two events X and Y are mutually exclusive,

$$P(X \cup Y) = P(X) + P(Y)$$

We're going to use this concept to take into account uncertainty at interior nodes.

Hypothetical tree



What is the likelihood of our observed sequence?

- What is the likelihood of 110 at the tips given the above tree and its associated branch lengths?

$$p(y|\Psi,\nu) = \pi_0 \cdot p_{00}(\nu 1) \cdot p_{00}(\nu 2) \cdot p_{01}(\nu 3) \cdot p_{01}(\nu 4)$$

π_i: Stationary frequency
$p_{ij}(\nu_k) = p(i|j,\nu_k)$: Transition probability

What is the likelihood of our observed sequence?

- What is the likelihood of 110 at the tips given the above tree and its associated branch lengths?
- If we knew the states at the root and an interior node and if we could assume that all transitions were independent, we could simply do a lot of multiplication.

$$p(y|\Psi, \nu) = \pi_0 \cdot p_{00}(\nu 1) \cdot p_{00}(\nu 2) \cdot p_{01}(\nu 3) \cdot p_{01}(\nu 4)$$

π_i: Stationary frequency
$p_{ij}(\nu_k) = p(i|j, \nu_k)$: Transition probability

What is the likelihood of our observed sequence?

- What is the likelihood of 110 at the tips given the above tree and its associated branch lengths?
- If we knew the states at the root and an interior node and if we could assume that all transitions were independent, we could simply do a lot of multiplication.
- If the values at the root and interior node were both zero, for instance:

$$p(y|\Psi, \nu) = \pi_0 \cdot p_{00}(\nu 1) \cdot p_{00}(\nu 2) \cdot p_{01}(\nu 3) \cdot p_{01}(\nu 4)$$

π_i: Stationary frequency
$p_{ij}(\nu_k) = p(i|j, \nu_k)$: Transition probability

The unknown interior nodes

- Since we don't know the states at the root and the interior node, we have to take that uncertainty into account by repeating the calculation for each possible assignment of states.

The unknown interior nodes

- Since we don't know the states at the root and the interior node, we have to take that uncertainty into account by repeating the calculation for each possible assignment of states.

- $p(C=0, A=1, B=1|\Psi, \nu) =$

  $p(C=0, A=1, B=1|\Psi, \nu, \text{Root=0}, \text{Node=0}) =$
  $\pi_0 \cdot p_{00}(\nu 1) \cdot p_{00}(\nu 2) \cdot p_{01}(\nu 3) \cdot p_{01}(\nu 4) +$

  $p(C=0, A=1, B=1|\Psi, \nu, \text{Root=0}, \text{Node=1}) =$
  $\pi_0 \cdot p_{00}(\nu 1) \cdot p_{01}(\nu 2) \cdot p_{11}(\nu 3) \cdot p_{11}(\nu 4) +$

  $p(C=0, A=1, B=1|\Psi, \nu, \text{Root=1}, \text{Node=0}) =$
  $\pi_1 \cdot p_{10}(\nu 1) \cdot p_{10}(\nu 2) \cdot p_{01}(\nu 3) \cdot p_{01}(\nu 4) +$

  $p(C=0, A=1, B=1|\Psi, \nu, \text{Root=1}, \text{Node=1}) =$
  $\pi_1 \cdot p_{10}(\nu 1) \cdot p_{11}(\nu 2) \cdot p_{11}(\nu 3) \cdot p_{11}(\nu 4)$

The idea behind maximum likelihood

- Find the tree and branch lengths that yield the maximum likelihood for the observed data

The idea behind maximum likelihood

- Find the tree and branch lengths that yield the maximum likelihood for the observed data
- In other words, what values of Ψ and ν yield the highest value for $p(C = 0, A = 1, B = 1|\Psi, \nu)$?

The idea behind maximum likelihood

- Find the tree and branch lengths that yield the maximum likelihood for the observed data
- In other words, what values of Ψ and ν yield the highest value for $p(C = 0, A = 1, B = 1|\Psi, \nu)$?
- In a Bayesian context, we do more than search for the maximum-likelihood tree (and branch lengths) because the likelihood values are all multiplied by a prior distribution.

How do we calculate transition probabilities?

- We do this with a transition model

Transition models

Deterministic vs. stochastic processes

- DETERMINISTIC PROCESS The same output is always produced from a given input. (Dowbrow 2016, p. 1)

Lit.: Çınlar 1975, Liggett 2010, Dowbrow 2016, Durrett 2016.

Deterministic vs. stochastic processes

- DETERMINISTIC PROCESS The same output is always produced from a given input. (Dowbrow 2016, p. 1)
- STOCHASTIC PROCESS "A stochastic process is a system which evolves in time while undergoing chance fluctuations." (Coleman 1974, p. 1)

Lit.: Çınlar 1975, Liggett 2010, Dowbrow 2016, Durrett 2016.

Deterministic vs. stochastic processes

- DETERMINISTIC PROCESS The same output is always produced from a given input. (Dowbrow 2016, p. 1)
- STOCHASTIC PROCESS "A stochastic process is a system which evolves in time while undergoing chance fluctuations." (Coleman 1974, p. 1)
- What are some examples of stochastic processes?

Lit.: Çınlar 1975, Liggett 2010, Dowbrow 2016, Durrett 2016.

Deterministic vs. stochastic processes

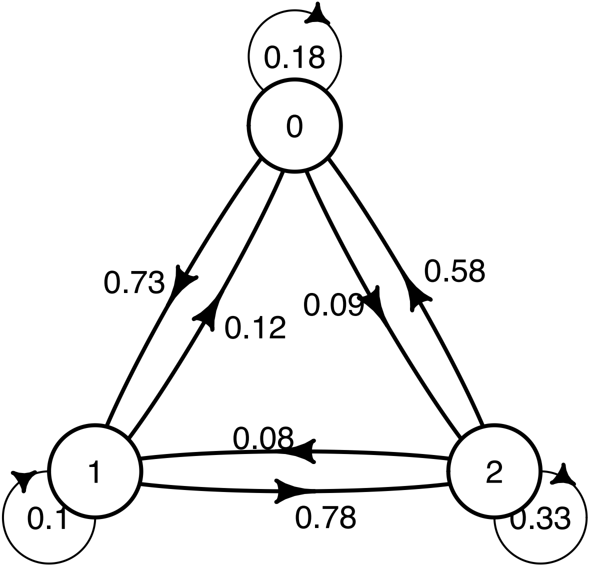
- DETERMINISTIC PROCESS The same output is always produced from a given input. (Dowbrow 2016, p. 1)
- STOCHASTIC PROCESS “A stochastic process is a system which evolves in time while undergoing chance fluctuations.” (Coleman 1974, p. 1)
- What are some examples of stochastic processes?
- Whether or not it will rain tomorrow, how many text messages you’ll receive in a given day, how long you’re going to have to wait in line...

Lit.: Çınlar 1975, Liggett 2010, Dowbrow 2016, Durrett 2016.

Markov chain

- A MARKOV CHAIN is a stochastic model describing a sequence of possible events in which the probability of each event depends only on the state attained in the previous event.

Markov chain as transition graph



Markov chain as matrix

$$
\begin{array}{cc}
 & \begin{array}{ccc} 0 & 1 & 2 \end{array} \\
\begin{array}{c} 0 \\ 1 \\ 2 \end{array} & \begin{bmatrix} 0.6 & 0.2 & 0.2 \\ 0.4 & 0.5 & 0.1 \\ 0.3 & 0.3 & 0.4 \end{bmatrix}
\end{array}
$$

- Each row in the above matrix sums to 1. Why?

Discrete- and continuous-time Markov chains

- DISCRETE-TIME MARKOV CHAIN Transitions can only happen at a discrete time value.

Discrete- and continuous-time Markov chains

- DISCRETE-TIME MARKOV CHAIN Transitions can only happen at a discrete time value.
- EXAMPLE: A board-game. A piece can only move when it's someone's turn.

Discrete- and continuous-time Markov chains

- DISCRETE-TIME MARKOV CHAIN Transitions can only happen at a discrete time value.
- EXAMPLE: A board-game. A piece can only move when it's someone's turn.
- CONTINUOUS-TIME MARKOV CHAIN Changes to the system can happen at any time along a continuous interval. A continuous-time Markov chain is known as a MARKOV PROCESS.

Discrete- and continuous-time Markov chains

- DISCRETE-TIME MARKOV CHAIN Transitions can only happen at a discrete time value.
- EXAMPLE: A board-game. A piece can only move when it's someone's turn.
- CONTINUOUS-TIME MARKOV CHAIN Changes to the system can happen at any time along a continuous interval. A continuous-time Markov chain is known as a MARKOV PROCESS.
- EXAMPLE: Whether or not it starts to rain. It can happen at any moment.

Discrete- and continuous-time Markov chains

- DISCRETE-TIME MARKOV CHAIN Transitions can only happen at a discrete time value.
- EXAMPLE: A board-game. A piece can only move when it's someone's turn.
- CONTINUOUS-TIME MARKOV CHAIN Changes to the system can happen at any time along a continuous interval. A continuous-time Markov chain is known as a MARKOV PROCESS.
- EXAMPLE: Whether or not it starts to rain. It can happen at any moment.
- Given a branch of a specific length, the CTMC provides us with probabilities for transitions among the possible states.

Continuous-time Markov Chain (CTMC)

- Likelihood and Bayesian methods model the history of characters on a phylogenetic tree as a MARKOV PROCESS.

Continuous-time Markov Chain (CTMC)

- Likelihood and Bayesian methods model the history of characters on a phylogenetic tree as a MARKOV PROCESS.
- CTMCs are the basis of most computational phylogenetic methods (Warnow 2018, p. 147).

Assumptions

- *Memoryless*
  The probability of a transition depends only on the current state.

Assumptions

- *Memoryless*
  The probability of a transition depends only on the current state.
- *Independent*
  Transitions are independent (what happens in one region of the tree is assumed to be independent of what happens elsewhere in the tree).

Assumptions

- *Memoryless*
  The probability of a transition depends only on the current state.
- *Independent*
  Transitions are independent (what happens in one region of the tree is assumed to be independent of what happens elsewhere in the tree).
- *Constant rate*
  For each rate parameter, a single rate is estimated for the entire tree, i.e., we're not going to allow rates to speed up or slow down along the tree.

Transition probability matrix $\mathbf{P}$

- The probability of a transition given a branch length.

Transition probability matrix $\mathbf{P}$

- The probability of a transition given a branch length.
- Each row vector will sum to 1 (Yang 2014, p. 4).

Transition probability matrix $\mathbf{P}$

- The probability of a transition given a branch length.
- Each row vector will sum to 1 (Yang 2014, p. 4).
- The transition probability matrix is derived from the instantaneous rate matrix.

Instantaneous rate matrix Q

- The transition probability matrix is derived from a instantaneous rate matrix Q, which represents the rate of change at an extremely small unit of time.

Instantaneous rate matrix **Q**

- The transition probability matrix is derived from a instantaneous rate matrix Q, which represents the rate of change at an extremely small unit of time.
- What we will work with in RevBayes is actually the instantaneous rate matrix Q and not the transition probability matrix P.

Instantaneous rate matrix **Q**

- The transition probability matrix is derived from a instantaneous rate matrix Q, which represents the rate of change at an extremely small unit of time.
- What we will work with in RevBayes is actually the instantaneous rate matrix Q and not the transition probability matrix P.
- There are many different varieties of rate matrices (or transition models).

Instantaneous rate matrix **Q**

- The transition probability matrix is derived from a instantaneous rate matrix Q, which represents the rate of change at an extremely small unit of time.
- What we will work with in RevBayes is actually the instantaneous rate matrix Q and not the transition probability matrix P.
- There are many different varieties of rate matrices (or transition models).
- Different transition models allow us to make different assumptions about linguistic change.

Equal-rate (ER) model

$$
\begin{array}{c c}
 & \begin{array}{cc} 0 & 1 \end{array} \\
\begin{array}{c} 0 \\ 1 \end{array} & \begin{bmatrix} -q & q \\ q & -q \end{bmatrix}
\end{array}
$$

- q_1 = rate at which character changes from 0 to 1 over a short interval dt

Equal-rate (ER) model

$$
\begin{array}{cc}
 & \begin{array}{cc} 0 & 1 \end{array} \\
\begin{array}{c} 0 \\ 1 \end{array} & \begin{bmatrix} -q & q \\ q & -q \end{bmatrix}
\end{array}
$$

- q_1 = rate at which character changes from 0 to 1 over a short interval dt
- q_2 = rate at which character changes from 1 to 0 over a short interval dt

Equal-rate (ER) model

$$
\begin{array}{cc}
 & \begin{array}{cc} 0 & 1 \end{array} \\
\begin{array}{c} 0 \\ 1 \end{array} & \begin{bmatrix} -q & q \\ q & -q \end{bmatrix}
\end{array}
$$

- q_1 = rate at which character changes from 0 to 1 over a short interval dt
- q_2 = rate at which character changes from 1 to 0 over a short interval dt
- The elements on the diagonal are determined by summing the non-diagonal elements in the row and negating them.

All-rates-different (ARD) model

$$
\begin{array}{cc}
 & \begin{array}{cc} 0 & \quad 1 \end{array} \\
\begin{array}{c} 0 \\ 1 \end{array} & \begin{bmatrix} -q_1 & q_1 \\ q_2 & -q_2 \end{bmatrix}
\end{array}
$$

The Mk model

- M stands for Markov

The Mk model

- M stands for Markov
- k is a variable over discrete character states

Other transition models

- F81

Other transition models

- F81
- HKY

Other transition models

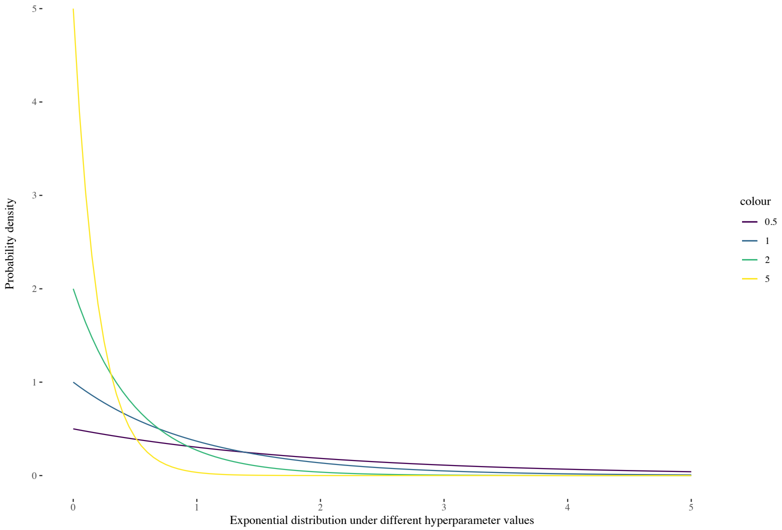
- F81
- HKY
- GTR

Other transition models

- F81
- HKY
- GTR
- Covarion

Prior distribution for the rates

Exponential distribution ($q_{ij} \sim \text{Exp}(\lambda)$)



Let's look at our .Rev script again

- Look at the code for the rate matrix Q

A. Sicoli, Mark and Gary Holton (Mar. 2014). "Linguistic phylogenies support back-migration from Beringia to Asia". In: *PLoS ONE* 9.3, e91722. DOI: 10.1371/journal.pone.0091722.

Atkinson, Quentin D. and Russell D. Gray (2006). "How old is the Indo-European language family? Illumination or more moths to the flame?" In: *Phylogenetic methods and the prehistory of languages*. Ed. by Peter Forster and Colin A. Renfrew. Cambridge: McDonald Institute for Archaeological Research, pp. 91–109.

Bouckaert, Remco R., Claire Bowern, and Quentin D. Atkinson (2018). "The origin and expansion of Pama-Nyungan languages across Australia". In: *Nature Ecology & Evolution*. DOI: 10.1038/s41559-018-0489-3.

Bouckaert, Remco R., Philippe Lemey, Michael Dunn, Simon J. Greenhill, Alexander V. Alekseyenko, Alexei J. Drummond, Russell D. Gray, Marc A. Suchard, and Quentin D. Atkinson (Aug. 2012). "Mapping the origins and expansion of the Indo-European language family". In: *Science* 337.6097, pp. 957–960. DOI: 10.1126/science.1219669.

Bowern, Claire (Nov. 2012). "The riddle of Tasmanian languages". In: *Proceedings of the Royal Society B: Biological Sciences* 279.1747, pp. 4590–4595. DOI: 10.1098/rspb.2012.1842.

Bowern, Claire and Quentin D. Atkinson (Dec. 2012). "Computational phylogenetics and the internal structure of Pama-Nyungan". In: *Language* 88.4, pp. 817–845. DOI: 10.1353/lan.2012.0081.

Cathcart, Chundra Aroor (2019). "Dialectal layers in West Iranian. A hierarchical Dirichlet process approach to linguistic relationships". In.

— (2020). "A probabilistic assessment of the Indo-Aryan Inner-Outer Hypothesis". In: *Journal of Historical Linguistics* 10.1, pp. 42–86.

Cathcart, Chundra Aroor and Florian Wandl (2020). "In search of isoglosses. Continuous and discrete language embeddings in Slavic historical phonology". In.

Chang, Will, Chundra Aroor Cathcart, David P. Hall, and Andrew J. Garrett (Mar. 2015). "Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis". In: *Language* 91.1, pp. 194–244. DOI: 10.1353/lan.2015.0005.

Çınlar, Erhan (1975). *Introduction to stochastic processes*. Englewood Cliffs, NJ: Prentice-Hall.

Coleman, Rodney (1974). *Stochastic processes*. Dordrecht: Springer. DOI: 10.1007/978-94-010-9796-3.

Dowbrow, Robert P. (2016). *Introduction to stochastic processes with R*. Hoboken: Wiley.

Dunn, Michael, Stephen C. Levinson, Eva Lindström, Ger Reesink, and Angela Terrill (2008). "Structural phylogeny in historical linguistics. Methodological explorations applied in Island Melanesia". In: *Language* 84, pp. 710–759.

Durrett, Richard (2016). *Essentials of stochastic processes*. New York: Springer. DOI: 10.1007/978-3-319-45614-0.

Gray, Russell D. and Quentin D. Atkinson (2003). "Language-tree divergence times support the Anatolian theory of Indo-European origin". In: *Nature* 426, pp. 435–439. DOI: 10.1038/nature02029.

Gray, Russell D., Alexei J. Drummond, and Simon J. Greenhill (2009). “Language phylogenies reveal expansion pulses and pauses in Pacific settlement”. In: *Science* 323.5913, pp. 479–483. DOI: 10.1126/science.1166858.

Greenhill, Simon J., Quentin D. Atkinson, Andrew Meade, and Russell D. Gray (Aug. 2010). “The shape and tempo of language evolution. Biological Sciences”. In: *Proceedings of the Royal Society B* 277, pp. 2443–2450. DOI: 10.1098/rspb.2010.0051.

Greenhill, Simon J. and Russell D. Gray (2009). “Austronesian language phylogenies. Myths and misconceptions about Bayesian computational methods”. In: *Austronesian historical linguistics and culture history. A festschrift for Robert Blust*. Ed. by Alexander Adelaar and Andrew Pawley. Canberra: Pacific Linguistics, pp. 375–397.

Guillon, Myrtille and Ruth Mace (Mar. 2016). “A phylogenetic comparative study of Bantu kinship terminology finds limited support for its co-evolution with social organisation”. In: *PLoS ONE* 11.5, e0155170. DOI: 10.1371/journal.pone.0147920.

Holden, Clare Janaki, Andrew Meade, and Mark Pagel (2005). “Comparison of maximum parsimony and Bayesian Bantu language trees. A phylogenetic approach”. In: *The evolution of cultural diversity*. Ed. by Ruth Mace, Clare Janaki Holden, and Stephen J. Shennan. Walnut Creek: Left Coast Press, pp. 53–66.

Kitchen, Andrew, Christopher Ehret, Shiferaw Assefa, and Connie J. Mulligan (Apr. 2009). “Bayesian phylogenetic analysis of Semitic languages identifies an Early Bronze Age origin of Semitic in the Near East”. In: *Proceedings of the Royal Society B: Biological Sciences* 276.1668. DOI: 10.1098/rspb.2009.0408.

Kolipakam, Vishnupriya, Fiona M. Jordan, Michael Dunn, Simon J. Greenhill, Remco R. Bouckaert, Russell D. Gray, and Annemarie Verkerk (Mar. 2018). “A Bayesian phylogenetic study of the Dravidian language family”. In: *Royal Society Open Science* 5. DOI: 10.1098/rsos.171504.

Lee, Sean and Toshikazu Hasegawa (2011). “Bayesian phylogenetic analysis supports an agricultural origin of Japonic languages. Biological Sciences”. In: *Proceedings of the Royal Society B: Biological Sciences* 278.1725, pp. 3662–3669. DOI: 10.1098/rspb.2011.0518.

Liggett, Thomas M. (2010). *Continuous time Markov processes*. Providence: American Mathematical Society.

List, Johann-Mattis (July 2016). “Beyond cognacy. Historical relations between words and their implication for phylogenetic reconstruction”. In: *Journal of Language Evolution* 1.2, pp. 119–136. DOI: 10.1093/jole/lzw006.

Pagel, Mark (Feb. 2017). “Darwinian perspectives on the evolution of human languages”. In: *Psychonomic Bulletin & Review* 24.1, pp. 151–157. DOI: 10.3758/s13423-016-1072-z.

Rama, Taraka (2018). “Three tree priors and five datasets. A study of the effect of tree priors in Indo-European phylogenetics”. In: *Language Dynamics and Change* 8.2, pp. 182–218. DOI: 10.1163/22105832-00802005.

Robbeets, Martine and Remco R. Bouckaert (July 2018). “Bayesian phylolinguistics reveals the internal structure of the Transeurasian family”. In: *Journal of Language Evolution* 3.2, pp. 145–162. DOI: 10.1093/jole/lzy007.

Sagart, Laurent, Guillaume Jacques, Yunfan Lai, Robin J. Ryder, Valentin Thouzeau, Simon J. Greenhill, and Johann-Mattis List (2019). "Dated language phylogenies shed light on the ancestry of Sino-Tibetan". In: *Proceedings of the National Academy of Sciences of the United States of America*. DOI: 10.1073/pnas.1817972116.

Saunders, Arpiar (2005). "Linguistic phylogenetics of the Austronesian family. A performance review of methods adapted from biology". B.A. thesis.

Savelyev, Alexander and Martine Robbeets (Jan. 2020). "Bayesian phylolinguistics infers the internal structure and the time-depth of the Turkic language family". In: *Journal of Language Evolution* 5.1, pp. 39–53. DOI: 10.1093/jole/lzz010.

Warnow, Tandy (2018). *Computational phylogenetics. An introduction to designing methods for phylogeny estimation*. Cambridge: Cambridge University Press.

Yang, Ziheng (2014). *Molecular evolution. A statistical approach*. Oxford: Oxford University Press.

Yanovich, Igor (Apr. 2020). "Phylogenetic linguistic evidence and the Dene-Yeniseian homeland". In: *Diachronica* 37.3, pp. 410–446.

Zhang, Hanzhi, Ting Ji, Mark Pagel, and Ruth Mace (Nov. 2020). "Dated phylogeny suggests early Neolithic origin of Sino-Tibetan languages". In: *Scientific Reports* 10.20792. DOI: 10.1038/s41598-020-77404-4.

Zhang, Menghan, Shi Yan, Wuyun Pan, and Li Jin (2019). "Phylogenetic evidence for Sino-Tibetan origin in northern China in the Late Neolithic". In: *Nature* 569, pp. 112–115. DOI: 10.1038/s41586-019-1153-z.