Journal of Language Evolution, 2021, 1–17
doi: 10.1093/jole/lzab005
Tutorial

Bayesian phylogenetic analysis of linguistic data using BEAST

Konstantin Hoffmann [1], Remco Bouckaert[2], Simon J. Greenhill[3,4,*] and Denise Kühnert[1,*]

[1]Transmission, Infection, Diversification & Evolution Group, Max Planck Institute for the Science of Human History, Kahlaische Straße 10, Jena 07745, Germany, [2]Centre for Computational Evolution, University of Auckland, 3 Symonds St, Auckland Central, Auckland 1010, New Zealand, [3]Department of Linguistic and Cultural Evolution, Max Planck Institute for the Science of Human History, Kahlaische Straße 10, Jena 07745, Germany and [4]ARC Centre of Excellence for the Dynamics of Language, ANU College of Asia and the Pacific, The Australian National University, Coombs Building, Liversidge St, Acton ACT 2601, Australia

*Corresponding authors: kuehnert@shh.mpg.de; greenhill@shh.mpg.de

Abstract

Bayesian phylogenetic methods provide a set of tools to efficiently evaluate large linguistic datasets by reconstructing phylogenies—family trees—that represent the history of language families. These methods provide a powerful way to test hypotheses about prehistory, regarding the subgrouping, origins, expansion, and timing of the languages and their speakers. Through phylogenetics, we gain insights into the process of language evolution in general and into how fast individual features change in particular. This article introduces Bayesian phylogenetics as applied to languages. We describe substitution models for cognate evolution, molecular clock models for the evolutionary rate along the branches of a tree, and tree generating processes suitable for linguistic data. We explain how to find the best-suited model using path sampling or nested sampling. The theoretical background of these models is supplemented by a practical tutorial describing how to set up a Bayesian phylogenetic analysis using the software tool BEAST2.

Key words: language evolution; historical linguistics; Bayesian methods; phylogenetics

1. Introduction

Studying the evolution of language families through time allows us to reconstruct pieces of the human past, locally and globally. Bayesian phylogenetic methods enable the reconstruction of the evolutionary relationships among languages and the estimation of the time and place at which their most recent common ancestor (MRCA) existed. These methods are powerful tools for reconstructing evolutionary histories. This power comes from a robust statistical and inferential framework that can incorporate known information about processes and patterns. Once these relationships—phylogenies—are reconstructed there are an array of tools that can use these phylogenies to test evolutionary hypotheses.

Bayesian 'phylolinguistic' studies have become increasingly prevalent in linguistics over the last decade. The first application by Gray and Atkinson (2003) controversially inferred the date of origins of the

Indo-European language family to around 7,800–9,800 years ago in contrast to the previously proposed age of about 6,000 years. While the debate about Indo-European is ongoing (e.g. Bouckaert et al. 2012; Chang et al. 2015), Bayesian phylogenetic methods have become an integral part of the historical linguistics toolkit, exploring language families from Austronesian (Gray et al. 2009)—where the inferred dates match those predicted by historical linguistics (Greenhill et al. 2010b)—to Bantu (Grollemund et al. 2015), Chapacuran (Birchall et al. 2016), Dravidian (Kolipakam et al. 2018), Korean (Lee 2015), Japonic (Lee and Hasegawa 2011), Pama-Nyungan (Bouckaert et al. 2018), Semitic (Kitchen et al. 2009), Sino-Tibetan (Sagart et al. 2019; Zhang et al. 2019), Tupi-Guarani (Michael et al. 2015), and Uralic (Honkola et al. 2013; Lehtinen et al. 2014).

Why has there been such a rapid uptake of Bayesian phylogenetic methods in linguistics? Historical linguistics has dabbled with computational methods before—lexicostatistics and its offshoot glottochronology (Swadesh 1950; Lees 1953)—but despite some initial enthusiasm, these methods were quickly rejected.

The first major criticism of lexicostatistics was that it discounted the distinction between shared retentions and shared innovations. In historical linguistics, since Brugmann (1884), innovations (usually phonological innovations) have been used to identify language relationships, while traits that are just retentions from an earlier stage are not considered indicative of relationships (Blust 2000). Lexicostatistics builds a tree by summarizing all changes between pairs of languages as a single distance score, which collapses this fundamental distinction (Blust 2000). In contrast, phylogenetic methods model where traits originate and where they are retained,[1] a distinction just as fundamental to modern taxonomy as it is to historical linguistics (Hennig 1996).

The second major criticism of lexicostatistics concerned its fundamental assumption of a constant rate of change to infer tree topology and timing (Bergsland and Vogt 1962). Critics noted that language change proceeds at widely varying rates over time and lineages. One prominent critique used lexicostatistics to date the divergence of Icelandic and Old Norse to less than 200 years ago, when historically we know it diverged 1,000 years ago (Bergsland and Vogt 1962). In contrast, Bayesian phylogenetic methods implement a range of approaches to model and account for rate variation between parts of the data, between lineages, and over time[2].

The final major criticism of lexicostatistics is that it could not account for nontree like processes in language change, such as borrowing (Moore 1994). Indeed, this criticism has long been lurking in the background of all historical linguistic approaches (Heggarty et al. 2010) and identifying loan words and contact effects is always a priority. However, although we know that evolutionary processes hardly ever follow a pure binary branching process, this simplified process often approximates the truth well, such that important scientific questions can be tackled. Notably, Bayesian phylogenetic methods are robust to the effects of borrowing as they infer and quantify the uncertainty in their estimates of parameters and tree topologies. In addition to inferring language relationships, they provide probabilistic measures of support for each given subgrouping. Hence, borrowing is revealed through high levels of uncertainty in the affected groupings, while unaffected parts of the tree are reconstructed accurately (Greenhill et al. 2009).

Here we explain the fundamental concepts of Bayesian phylolinguistic analysis of lexical data. We concentrate on the concepts that are relevant for language evolution. Setting up a phylolinguistic analysis requires a sophisticated choice of models and their parameters. The presentation of the models comes with an explanation on how and why to use them and a detailed description of their parameters. We provide mathematical details to fully understand how these models interact and contribute to the Bayesian posterior distribution. The supplement contains a hands-on tutorial on how to set up and run a phylolinguistic analysis in BEAST2, which is also part of the community teaching material resource 'Taming the BEAST' (https://taming-the-beast.org, last accessed 03/08/2021) (Barido-Sottani et al. 2017).

2. Bayesian phylogenetics

The advantage of Bayesian methods is the use of probability distributions for model parameters. This use of distributions is in stark contrast to other phylogenetic approaches like maximum parsimony or maximum likelihood, where a single 'best' value for each parameter is estimated (Greenhill and Gray 2009). In Bayesian statistics, we aim to incorporate the full uncertainty around each estimate. This means that all possible values of the parameters, for example, the diversification rate, are allowed with a corresponding probability. The possible values are expressed as our prior belief in the form of a probability distribution. Priors can be 'neutral' (i.e. without a strong claim that the age of a language group is between 2,000 and 20,000 years) or highly informative (e.g. a specific group is between 1,000 and 1,200 years old with more weight toward 1,100 years).

The Bayesian approach is extremely powerful for linguistic analyses for several reasons. First, diverse

scenarios can be included in the calculations, weighted by how likely we believe they are. For example, we often have documented evidence from early historical sources that a group of people spoke a particular language, for example, the Chapacuran language Tor. Tor is first mentioned in a letter by a Jesuit priest in 1,714 (Menéndez 1981), so we know that the language existed around 300 years ago, but do not know when it originated. Birchall et al. (2016) used this information to reconstruct the Chapacuran language family tree with a prior probability distribution for Tor's emergence stretching from 300 to 780 years ago. Combining several such partial calibrations allowed them to infer the origin of the Chapacuran languages to an average of 1,039 years ago, with a 95% probability between 525 and 1,619 years. Their results suggest that many of the Chapacuran language splits occurred around the time the Spaniards entered lowland Bolivia.

Second, the output is not a single tree with the claim to be the true tree, but a sample of trees from the posterior. This *posterior probability distribution* of trees is important as it tells us which parts of the tree are well supported by the data, and which are more weakly attested. This enables us to make careful and nuanced statements of hypotheses that take into account uncertainty in the tree topology or parameters—and we can still make inferences even when there are moderate levels of uncertainty. For example, there is substantial debate about the first subgroup within Sino-Tibetan. One hypothesis proposes a primary split between Sinitic and Tibeto-Burman (Benedict 1972; Matisoff 2003), another places Sinitic and Tibetan in a lower-level subgroup (van Driem 2003; Blench and Post 2014), and a third hypothesis proposes that the deep structure of Sino-Tibetan is a rake with multiple branches (Peiros 1998). Sagart et al. (2019) aimed to identify the origins of Sino-Tibetan using Bayesian Phylogenetic methods. Rather than supporting a single hypothesis, they found that 33% of the trees placed the Sinitic languages as the first group, while 15% of the trees placed a West-Himalayish group first. Given this posterior distribution, Sagart et al. were able to rule out the 'rake' hypothesis and provide historical and archaeological evidence favoring the 'Sinitic first' hypothesis as twice as likely as the 'West-Himalayish' hypothesis.

2.1 Data structure

When speaking of language phylogenies, the complexity of a language is usually collapsed into a set of variables. Although the variables can theoretically be of any kind, they are in most cases binary and represent a feature that is present (and thus coded as 1) or absent (coded as 0) in the language. Throughout this article, a language is then always considered as a specific combination of these features. Variables could be grammatical features or lexical features in the form of cognates (Greenhill et al. 2017), or any other aspect of language that might show inheritance through descent with modification. In the lexical cognate case, the data are typically based on a list of basic vocabulary items (such as a *Swadesh list*) containing, for instance, simple verbs, kinship terms, or body parts. For each meaning in this list, lexemes from all languages are collected and classified into *cognate classes*. These are sets of homologous lexemes with similar meanings (Table 1). In the next step, every cognate class defines a feature. Either a language has or lacks a lexeme in that cognate class.

The whole dataset can be written as a matrix, where each row stands for a language and each column for a feature (Table 2). Following the terminology of *genetic* phylogenetics, the matrix is called an alignment and each column is a site.

Table 1. Example wordlist with two concepts ('to laugh' and 'wing') for five Oceanic languages. The lexemes are classified in cognate sets, where lexemes with shared (Oceanic) ancestry and meaning are summarized in the same class.

Language	'To laugh'		'Wing'	
	Lexeme	Cognate set	Lexeme	Cognate set
Fijian	dredre	tolaugh-A	taba-na	wing-A
Marquesan	kata	tolaugh-B	peheu	wing-B
Hawaiian	'aka	tolaugh-B	'heu	wing-B
Maori	kata	tolaugh-B	parirau	wing-C
Tahitian	'ata	tolaugh-B	pererau	wing-C

Table 2. The cognate class assignment from Table 1 is coded into a matrix of binary traits representing the presence or absence of a cognate class in the respective language.

Language	'To laugh'		'Wing'		
	A	B	A	B	C
Fijian	1	0	1	0	0
Marquesan	0	1	0	1	0
Hawaiian	0	1	0	1	0
Maori	0	1	0	0	1
Tahitian	0	1	0	0	1

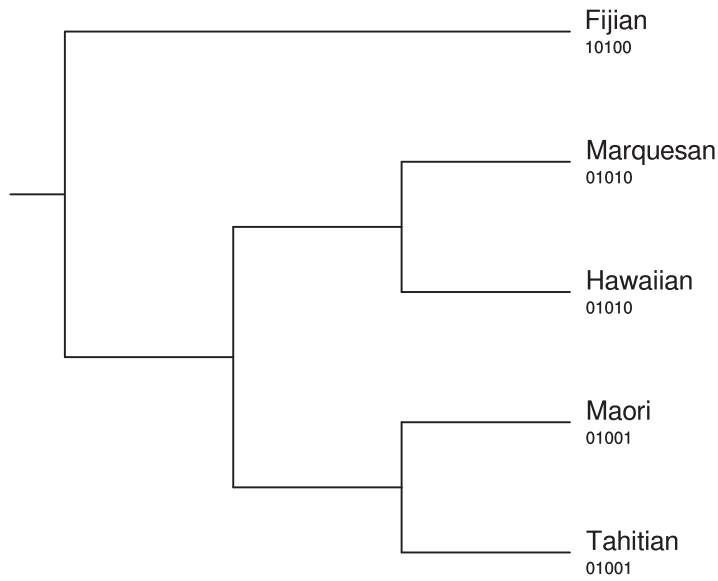

Figure 1. Phylogeny of the example data from Tables 1 and 2. Each tip relates to a language that is represented by a binary string of cognate codings. In this (most parsimonious) tree, the formation of the cognate classes tolaugh-A and wing-A only happened in the Fijian branch. The class tolaugh-B emerged in its sister branch that is ancestral to the four remaining languages. This branch splits up into two daughter lineages, where in the upper case (ancestral to Marquesan and Hawaiian) the feature tolaugh-B is gained and in the lower case (ancestral to Maori and Tahitian) the feature tolaugh-C is gained.

Note that it is not mandatory to have definite codings for all features and languages. Languages with missing data are omitted in the likelihood calculation for the sites without data. This is unproblematic as long as missing features are sparse.

A phylogeny is represented in a binary tree (Fig. 1). Its starting point is the root node corresponding to the MRCA of all the languages present. From there, lineages either branch into exactly two daughter lineages or end in a leaf node. The branch lengths correlate to the number of changes (i.e. loss or gain of a feature) from their starting point to their end (branching point or tip). Multifurcations into more than two lineages can be expressed by subsequent binary branching points with very short branches between them. The leaf nodes (or tips) represent the extinct or extant languages from the dataset. Internal branches are protolanguages of their children and a split is the point in time, where two languages start to accumulate changes independent of each other (Maurits et al. 2019).

2.2 Bayes' theorem

In the present context, Bayes' theorem states that the probability of a phylogenetic tree is based on an evolutionary model, the visible data and a set of prior beliefs about the unknown model parameters. This probability is called *posterior probability*. The prior beliefs are expressed as a set of probability distributions, which are called *priors*. If we knew the exact model of evolution together with its parameters, we could compute the *likelihood* that the data emerged from a given tree. In reality, however, the parameter values are unknown. The idea of Bayes' theorem is to multiply this likelihood with the prior probabilities of these parameters.

Explicitly for a tree T given the data D and an evolutionary model characterized by its model parameters θ, it is by Bayes' theorem that

$$P(T|D,\theta) \propto P(D|T,\theta)P(T,\theta), \quad (1)$$

where $P(D|T,\theta)$ is the likelihood of observing the data given the model and the tree and $P(T,\theta)$ is the prior probability of the tree T and the set of model parameters θ. The left-hand side of this relation is the posterior probability, which is *proportional to* $(\propto)$ the right-hand side. The absolute value of the posterior is often difficult to compute. But the relation implies that it is nevertheless possible to compare different outcomes of the parameters θ and trees T and by that produce samples according to the posterior probability.

2.3 Computing the likelihood

A phylogenetic tree is comprised of a set of leaf nodes connected to a series of internal nodes. Each language is associated with a leaf node. At the leaf nodes, we know

the exact configuration of traits (with the exception of ambiguous sites due to missing data), for the internal nodes it is unknown and we want to infer the most probable trait assignments. So, for every internal node configuration, a probability is calculated, which is the probability—given the evolutionary model—that this specific configuration leads to the state at the leaf nodes. When computing the likelihood, we sum over the probabilities of all possible internal node assignments. However, under the assumption that each trait of the data evolves independently, the likelihood can be calculated per site.

Let $x_1, ..., x_{2n-1}$ be the nodes of the tree with leaf nodes $x_1, \ldots, x_n$, internal nodes $x_{n+1}, \ldots, x_{2n-2}$, and root node x_{2n-1}. For $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, k\}$, let s_{ij} be the (known) data value of the leaf node x_i at the jth column in the alignment. For $i \in \{n+1, \ldots, 2n-1\}$, let s_{ij} be the (unknown) data value of the node x_i at position j. Assuming that the index of the parent node of x_i is m_i, the probability of observing s_{ij} for some j at node x_i is as follows:

$$\begin{aligned} P(x_i = s_{ij}) = P(x_i = s_{ij}|x_{m_i} = 0)P(x_{m_i} = 0) \\ +P(x_i = s_{ij}|x_{m_i} = 1)P(x_{m_i} = 1). \end{aligned} \tag{2}$$

Overall, the tree-likelihood at site j is computed as the sum over all possible states at the internal nodes:

$$\begin{aligned} P(D_j|T, \theta) = \sum_{s_{n+1}=0}^{1} \ldots \sum_{s_{2n-1}=0}^{1} \prod_{i=1}^{2n-2} P(x_i = s_{ij}|x_{m_i} = s_{m_i}, \theta) \\ \cdot P(x_{2n-1} = s_{2n-1}|\theta) \end{aligned} \tag{3}$$

and we get,

$$P(D|T, \theta) = \prod_{j=1}^{k} P(D_j|T, \theta). \tag{4}$$

This likelihood can be efficiently calculated using Felsenstein's pruning algorithm (Felsenstein 2004). The transition probabilities (2) are determined by the substitution model explained below.

It is worth noting that this method, at every step of the analysis, models where each trait originated and in which branches it was retained. The likelihood here is a direct analogue of the traditional distinction in historical linguistics between retentions and innovations. It is possible if cumbersome to log the inferred innovation point of a set of traits or to identify which particular traits define a subgroup of languages.

2.4 Ascertainment correction

In linguistic datasets, traits are typically included only if they are present in at least one of the languages. That is, linguists tend to exclude cognate sets that do not occur in the particular set of languages they are studying. This might seem like a strange thing to worry about, but if we think of our data as comprised of a set of cognate sets of varying size—from singletons containing one language, to maximal sets containing all languages—then it becomes clear that we have arbitrarily ignored one end of this distribution (i.e. cognate sets of size zero). To account for this bias, we replace $P(D_j|T, \theta)$ from Equation (3) with the corrected term:

$$P(D_j|D_j \neq 0, T, \theta) = \frac{P(D_j|T, \theta)}{1 - P(D_j = 0|T, \theta)} \tag{5}$$

where $D_j = 0$ means that all languages have a zero in the jth column of the alignment. That is, we ascertain at least one 1 in each cognate column (Felsenstein 2004). However, if the dataset is divided into partitions (e.g. one for every meaning class when considering cognate data), it is possible that a language has no data in a particular partition. In this case, the ascertainment correction described here is not sufficient (Chang et al. 2015), rather the correction should be applied on a per-word or meaning slot basis. Additionally, each zero in the 0-vector that corresponds to a language with missing data needs to be marked as missing. An example of a properly ascertained dataset can be found in the Supplementary tutorial in section 'The data-set'.

3. Models of evolution

The substitution model determines the probability of going from one state to another over a specific time period. The substitution process is modeled as a Markov process with an infinitesimal rate matrix $Q = (q_{ij})$, where q_{ij} is the exponential rate of going from state i to state j ($i \neq j$) and the diagonal elements $q_{ii} = -\sum_{j \neq i} q_{ij}$ are the outgoing rates. So, if state i was 0 and state j was 1, then q_{ij} is the rate of gaining this particular cognate while q_{ji} is the loss rate. The time-dependent transition probabilities between states are the entries of the matrix

$$P(t) = \exp(Qt). \tag{6}$$

To efficiently compute these probabilities, most of the models (and all of the models explained here with the exception of the pseudo-Dollo model) are based on the following assumptions: first that the process is stationary, which means that there are equilibrium frequencies $\pi = (\pi_0, \pi_1)$ such that

$$\pi Q = 0, \quad \pi P(t) = \pi, \quad \forall t. \tag{7}$$

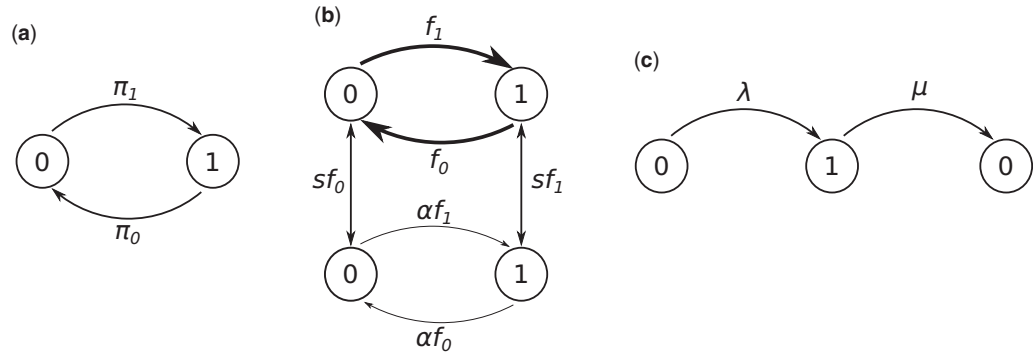

Figure 2. Overview of the binary substitution models. (a) The binary CTMC model, where mutations between 0 and 1 can happen in both directions. The binary covarion model (b) features slow and fast states. Switches between these states happen with rate s and mutations in the slow state happen with rate α. In the pseudo-Dollo model (c), a trait can be gained once with rate λ and permanently lost with rate μ.

This means that the evolutionary process has reached a state, where the overall amount of acquired traits is in balance. The second assumption is time-reversibility, that is

$$\pi_i P(t)_{ij} = \pi_j P(t)_{ji}, \quad \forall i, j, t. \tag{8}$$

This implies that evolution is not directional and that the probability of starting with 0 at one end of a branch and ending with 1 at the other is the same as the probability of starting with 1 and evolving to 0 (Felsenstein 2004). In the following, we describe the most commonly used models of linguistic evolution (Fig. 2).

3.1 Continuous time Markov chain model model

The continuous time Markov chain model (CTMC) (Gray and Atkinson 2003; Bouckaert et al. 2012) is the simplest substitution model presented here. It assumes that the data are generated by a time-reversible Markov process with two states (Fig. 2a) and that the distribution of zeroes and ones observed in the data follows the stationary distribution of this Markov chain. Each zero in the dataset can evolve into a one and vice versa. In terms of cognate evolution, this means that every language can gain or lose a cognate set with a state-specific rate and these rates are fixed according to the currently observed data. Given the stationary distribution $\pi = (\pi_0, \pi_1)$, the only possible infinitesimal rate matrix fulfilling these conditions is as follows:

$$Q = \beta \begin{pmatrix} - & \pi_0 \\ \pi_1 & - \end{pmatrix} \tag{9}$$

with normalizing constant $\beta = \frac{1}{\pi_0^2 + \pi_1^2}$.

Note that the off-diagonal entries are positive and represent rates of flow from i to j. The diagonal entries represent the total flow out of state i. The latter is the negative of the sum of off-diagonal other entries in each row and are left blank by convention.

3.2 Covarion model

The binary covarion model (Tuffley and Steel 1998; Penny et al. 2001) is widely used for cognate data (Gray et al. 2009; Bouckaert et al. 2012). The covarion model provides a powerful way of handling variation in rates of change. For example, many cognates are relatively stable over a long period of time but occasionally change in bursts. Bursts of change may be due to external events like language contact. The covarion model can account for that by letting states evolve at a slow 'background' rate during periods of stability and shifting into a faster rate category when bursts happen. In our experience, the covarion model is often the best performing model for lexical cognate data.

The covarion model contains two layers of states. The first and visible layer is the observable state of the site, 0, or 1. The second and hidden layer contains the additional information if the site is in a slow or fast state. In total, there are four hidden states, slow-0, fast-0, slow-1, and fast-1. Transitions can happen both between the fast and slow states and between the 0- and 1-states as it is shown in Fig. 2b. This is parameterized through several parameters: The stationary frequencies (f_0, f_1) of the observed states (0 resembles slow-0 and fast-0, and 1 resembles slow-1 and fast-1). The frequencies of the fast and slow states are set to $(0.5 \cdot f_0, 0.5 \cdot f_0, 0.5 \cdot f_1, 0.5 \cdot f_1)$ to ensure time-reversibility. The

switch rate s of changing between slow and fast states, and finally, the substitution rate α in the slow state. The substitution rate in the fast state is 1 making α practically to a relative value.

The rate matrix Q is then defined as follows:

$$\begin{matrix} & \\ \text{fast} & \\ \text{slow} & \end{matrix}\begin{matrix} 0: \\ 1: \\ 0: \\ 1: \end{matrix}\begin{pmatrix} - & f_1 & sf_0 & 0 \\ f_0 & - & 0 & sf_1 \\ sf_0 & 0 & - & \alpha f_1 \\ 0 & sf_1 & \alpha f_0 & - \end{pmatrix} = Q \tag{10}$$

3.3 Pseudo-Dollo model

The Dollo principle states that a feature can be gained only once, but can be lost several times (Dollo 1893). Once it is lost, it cannot be regained. This assumption appears suitable for language data as it is commonly believed that cognate sets are gained rarely but can be lost frequently (Nicholls and Gray 2008). However, in practice, this assumption is very restrictive as borrowings and semantic shifts may lead to multiple gains of a cognate. A variant accounting for this is the pseudo-Dollo model (Bouckaert and Robbeets 2017). It assumes that each language can gain a feature with an infinitesimal rate λ and lose it forever with rate μ (Fig. 2c). This model is neither time-reversible nor in an equilibrium state as the evolution is directional. There are three categories, a zero for not-yet-acquired, a one and a zero for feature-lost. The infinitesimal rate matrix is as follows:

$$Q = \begin{pmatrix} - & \lambda & 0 \\ 0 & - & \mu \\ 0 & 0 & 0 \end{pmatrix} \tag{11}$$

A covarion variant that can deal with rate variation through two slow states (similar to the binary covarion model) is available as well.

3.4 Markov model for multiple states

Nonbinary data with more than one state can be modeled using the M_k-model (Lewis 2001). This model is often useful for typological data or other characters with a small number of states that cannot be easily or sensibly converted to binary presence/absence coding. Under the M_k-model each trait can have one of the k states, and transitions are allowed between any pair of states at equal probabilities. The corresponding rate matrix is as follows:

$$Q = \begin{pmatrix} - & 1 & \dots & 1 \\ 1 & - & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & - \end{pmatrix} \tag{12}$$

and the equilibrium frequencies are $\left(\frac{1}{k}, \dots, \frac{1}{k}\right)$. Note that for $k = 2$, this is the CTMC-model with frequencies $\pi = (0.5, 0.5)$.

4. Rate variation and calibration

The rate matrix Q is normalized to yield on average one substitution per time unit of t. However, languages vary substantially in their rates of change across lineages, across different subsets of data, and over time (Bergsland and Vogt 1962; Greenhill et al. 2017). Bayesian phylogenetic methods provide several ways of varying the pace at which changes happen across sites and branches of the tree. This rate variation is usually accomplished by multiplying the matrix Q by several factors, which are summarized as the mutation rate μ.

4.1 Rate variation across sites

In order to allow rate variation across sites one must consider the characteristics of linguistic meaning classes. Empirically, sites in a meaning class with very few cognates evolve at a slower pace than sites in a meaning class with many members (Pagel and Meade 2006; Pagel et al. 2007). Hence, a simple model in which all meaning classes share a single mutation rate is unrealistic. Alternatively, one may allow one mutation rate for each meaning class. However, as there is often a large number of meaning classes in the data (i.e. 100 or 200 words or more) this bears a high risk of overfitting the data. An intermediate solution is to distribute the meaning-classes to several bins which share a mutation rate (e.g. one bin for meaning classes of size 1–10, 11–20, and so forth). This approach may capture the abovementioned variation without overfitting the model. The Supplementary tutorial explains how this is facilitated in BEAST2.[3]

Another possibility is to allow site variation governed by an approximated gamma distribution (Yang 1994). In contrast to the model above, where all sites in a meaning class share the same evolutionary rate, but rates may differ across meaning classes, in this approach all sites are allowed to vary in the same way. We assume that the distribution of mutation rates follows a gamma distribution with mean 1 and a shape parameter α. This distribution has the following properties: for $\alpha > 1$, it is bell-shaped centered around 1 with little variance, and for $\alpha \leq 1$, it is L-shaped, where values lower than 1 have a high density and high values have a low density (compare Fig. 3). So a dataset containing highly varying mutation rates, where most of the sites are fairly constant but some sites change at a fast pace, is best described by a gamma distribution with a low shape parameter. However, if all sites change at more or less

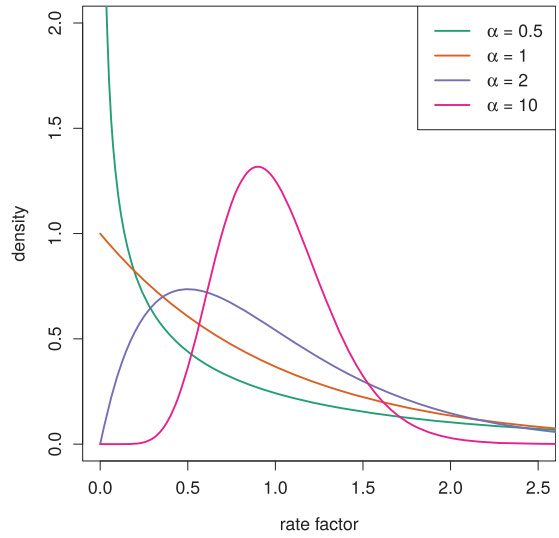

Figure 3. Densities of the Gamma function for varying shape parameters α. For higher α-values, most of its weight is centered around 1 meaning a low variation of rates. Low α-values lead to L-shaped distributions with higher probabilities of rate factors farther away from 1.

the same speed, we would expect a high α-value. Note that this method is recommended only if there is no other method governing mutation rates variation. For example, it is not advisable to combine a binary covarion substitution model with gamma rate heterogeneity as both include slow and fast rates.

Typically the continuous gamma distribution is approximated by a discrete version with a small number of rate categories. The range of rate multipliers from 0 to ∞ is split into segments of equal probability according to the gamma distribution. The weighted (by the gamma distribution) mean of each segment is its representative rate category. The likelihood is then calculated once for each category and the average is taken. (Yang, 1994). The amount of rate categories is a tradeoff between computational effort and the resolution of the captured variance. Harrison and Larsson (2014) showed that four rate categories are sufficient to capture most of the variance. In a Bayesian analysis, the gamma shape parameter can be estimated. As there is no prior information on which site belongs to which category the likelihood is calculated once for each category and the average is taken.

4.2 Clock models

To estimate the age of subgroups on a phylogeny, we need an evolutionary *clock* to convert the number of observed changes to time. The strict clock model assumes that substitutions happen at the same speed across the whole tree with a single parameter for the substitution rate. This 'clock rate' represents the average number of substitutions per site per time unit.

Some lineages in the tree might evolve faster than others, which can be accounted for through an uncorrelated relaxed clock model (Drummond et al. 2006; Douglas et al. 2021). Additionally to the clock rate, it samples a rate multiplier for each branch in the tree. The multipliers are drawn independently from a probability distribution, most commonly a log-normal distribution with mean 1 and standard deviation σ. The clock rate for a branch is then the product of the average clock rate c and the branch-specific multiplier. If the tree has n tips, this results in up to $2n$ clock model parameters: two from the distribution and up to one rate for each of the $2n-2$ branches. To avoid overfitting, this model should be used only when the data show at least a moderate temporal signal with reliable tip and/or internal node calibrations.

Further variations are possible, such as allowing a random local clock (Drummond and Suchard 2010) or a mixture of strict and relaxed clock models for different parts of the tree (Fourment and Darling 2018).

4.3 Calibrations

If we want to estimate divergence times, we need an informative prior distribution on the clock rate or reliable calibration points to estimate the ages of all nodes in the tree. One can either set a prior on the age of the MRCA of a language subgroup, on the parent of an MRCA, on the entire tree, or use the ages of ancient languages as

calibration points. If a set of languages belongs to a separated geographical unit (e.g. an island), an MRCA calibration could come from archaeological records. For example, the age of the first settlement of New Zealand can be securely dated to between 1230 and 1282 AD (Wilmshurst et al. 2010). Together with the assumption that after the initial colonization there was a fast spread across the islands and thus a geographical separation, this implies that the native languages spoken on these islands have a common ancestor, which must have existed around that time. Thanks to the Bayesian framework, such calibrations are not necessarily exact years, but can be included using prior distributions accounting for uncertainty and scholarly disagreement.

Although such calibrations can be incorporated for all clades and subclades in the tree, it is highly recommended to use few good calibrations and to refrain from nested calibrations as the combined prior distribution can induce unwanted side-effects (Heled and Drummond 2011). To check whether calibrations provide enough information to estimate the clock rate, one can perform a Bayesian evaluation of the temporal signal (Duchene et al. 2020). Date calibrations can also be used to validate the model and inference procedure, by reconstructing known dates (Ryder and Nicholls 2010). A useful discussion of how to best implement calibrations in a Bayesian analysis of languages can be found in (Maurits et al. 2019). A practical guide explaining how to set up calibrations can be found in the Supplementary tutorial (section 'The priors').

5. Tree priors

The last factor of Equation (1) is the prior distribution of the model parameters. A special role is played by the tree prior. The tree prior contains information about the process that gave rise to the phylogeny and itself introduces new parameters governing this tree-generating process. Consequently, the prior distribution expands to

$$P(T, \theta) = P(T|\theta)P(\theta), \tag{13}$$

where θ is the set of all model parameters, from the evolutionary model and the tree prior.

There is a range of tree priors available; however, not all are suitable for linguistic analyses. A very popular tree prior for biological applications is based on the coalescent process from population genetics. (Kingman 1982). The simple constant population size coalescent may not always fit language data well (Rama 2018; Ritchie and Ho 2019), but skyline variants (like the so-called Bayesian skyline (Drummond et al. 2005)) provides a flexible variant that requires conveniently little effort in specifying hyper priors. For languages, the most suitable models are the Yule and birth–death processes, but this is an ongoing area of research (Rama 2018; Ritchie and Ho 2019). It is best practice to test the robustness of phylogenetic results to different assumptions regarding the tree generating process.

5.1 The Yule model

The Yule model (Yule 1925) is a pure birth process with a single parameter λ, called the birth rate, determining the exponential rate of diversification in the tree. The probability that a branch splits into two at time t is given by the following equation:

$$p(t) = \lambda e^{-\lambda t}. \tag{14}$$

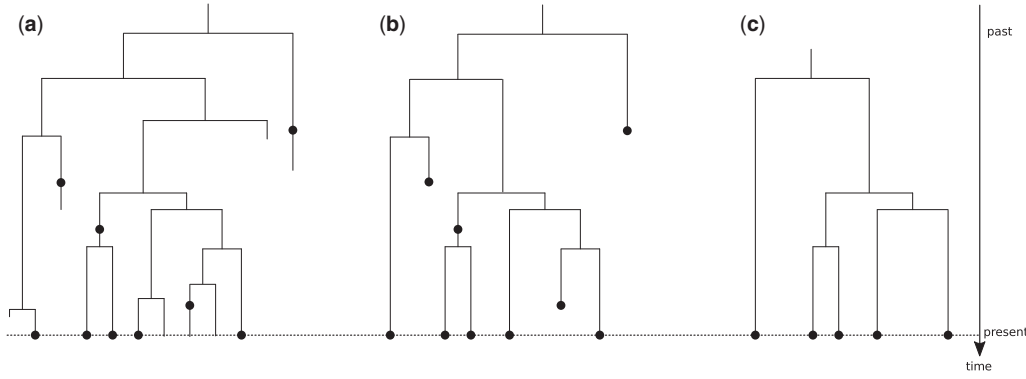

Figure 4. (a) Full tree, (b) reconstructed tree under a birth–death model with ψ-sampling through time, and (c) with contemporaneous ρ-sampling only. All dots refer to a sampling event. At present day, all extant lineages are sampled with probability ρ. All extinct and unsampled languages are pruned.

Condition on the tree having n tips and root age t_1, the prior distribution of a tree T depends on its branching times $t_2, \ldots, t_{n-1}$ (Gernhard 2008):

$$P(T, \lambda) = n! \lambda^{n-1} \prod_{i=1}^{n-1} e^{-\lambda t_i} P(\lambda) \quad (15)$$

Yule's simplicity is both its strength and its weakness. First, the Yule model assumes that the dataset includes *all* the languages in the family—that is, full sampling—which is unlikely. Second, the Yule model does not allow lineages to die. Therefore, while the Yule is the simplest tree prior model and commonly used in linguistic phylogenies, any dataset that includes doculects of extinct languages is invalid under the Yule model and may generate biased estimates of diversification rates and node ages.

5.2 Birth–death processes

Birth–death models are widely used in phylogenetics and come in different varieties (Gernhard 2008; Stadler et al. 2013; Gavryushkina et al. 2014). Like the Yule model, they all specify diversification events governed by a birth rate λ. In addition, each lineage can go extinct with an exponential rate μ, called the death rate. Furthermore, these models can account for an incomplete sampling of languages to give better estimates of the tree topology and timing.

In a perfect data set, we would have sampled all languages—both extinct and extant—of a language family. In this case, we would be able to reconstruct the full language family tree (Fig. 4a). Unfortunately, we often only observe extant languages such that the reconstructed language phylogeny represents a subset of the full language tree in which unobserved lineages are invisible (Fig. 4c). In this case, all languages in the dataset would be sampled at present, which can be modeled through a sampling probability ρ.

At least one of the three parameters—birth rate λ, death rate μ and the sampling probability ρ—needs strong prior information as they are highly correlated to each other (Stadler et al. 2013). A good candidate for a prior distribution on the sampling probability is the Beta distribution Beta(α, β) with the amount of sample languages α and the estimated amount of nonsampled languages β. The mean of this distribution is the proportion of sampled languages $\frac{\alpha}{\alpha+\beta}$. The attached tutorial contains a guide on how to specify the birth–death model (section 'The priors').

Datasets that include extinct languages need a sampling rate ψ through time. The sampling rate is an exponential rate at which lineages in the tree get sampled (Stadler et al. 2013). The reconstructed tree can also include ancient samples that may or may not have sampled descendants at present (Gavryushkina et al. 2014), which is illustrated as a black dot along the branch in Fig. 4b. A simple reparametrization of ψ as a sampling proportion $s = \frac{\psi}{\psi+\mu}$ allows the use of a Beta prior distribution.

Further extensions of the birth–death model are possible. For example, another way to include extinct languages is multi-rho sampling. A set of fixed times is specified, at each of which one or more samples are taken. For each of these times, a sampling probability parameter needs to be set as well, if suitable it can be a single ρ for all sampling times.

6. Choosing the best analysis

In this article, we have described a range of models for analyzing language data in a phylogenetic framework. The choice of model depends on a range of factors. In the following, we attempt to summarize them. One should try the models appropriate for the data at hand and evaluate which one performs best. If several models robustly lead to the same results, that is, the key aspects of the phylogeny are close, a model comparison procedure may not be required. If this is not the case, a model comparison should be performed and Bayes factors (BFs) computed.

6.1 Model preselection

The overall phylogenetic model consists of two major parts: the site model, consisting of a substitution model and a clock model, and the tree model. The choice of a site model depends on the data. If the data are nonbinary, that is, a site in the matrix can have more than two states, the only option currently implemented for linguistic data is an M_k-model (or, if appropriate, the data can be transformed into a binary form). For binary data (e.g. presence or absence of cognate sets or structural features), one should run a model comparison between the binary CTMC, pseudo-Dollo (covarion) and the binary covarion model.

If the rate of evolution varies amongst the sites in the data (i.e. some features evolve significantly quicker than others), one should additionally check, whether adding gamma rate categories improves the performance. However, this only makes sense for the CTMC, M_k, and pseudo-Dollo model as the covarion model naturally introduces two rate categories. Alternatively, the data can be partitioned into bins with a separate mutation rate for each bin, as described above.

Regarding the choice of clock model, a strict clock may lead to wrong time inferences if there is a significant variation in the branch rates. Perhaps unintuitively, a relaxed clock model performs well even if there is no variation at all (Drummond et al. 2006). However, an analysis with a relaxed clock model has a lot more parameters and thus requires informative data to converge. In an analysis with a relaxed clock model with log-normal distributed branch rates, the *coefficient of variation* can be computed. This is the estimated standard deviation of the branch rates divided by the mean clock rate and a measure of how clock-like the data are. A low coefficient of variation results from clock-like data, while higher values indicate higher variation in branch rates.

We suggest the following: first, run an analysis with a relaxed clock model. If the coefficient of variation is low (e.g. less than 0.1), there may not be any variation. In this case, the data may be better described by a strict clock model. Second, check if the clock rate converges with a strict clock model. If the parameters converge in both cases, do a model selection test between the two models and choose the better performing one.

For the tree prior, the birth–death model is most suitable if there are extinct languages in the data. Even if the data are contemporaneous a birth–death model is favored over the Yule model as the latter one assumes complete sampling and no extinction happening in the tree. As described above, there are several ways sampling can be modeled in a birth–death process. If all the languages are sampled from the present data (i.e. are contemporaneous) then the model requires a sampling probability (ρ) at the present day (e.g. if the analysis contains about 80% of the languages in the group, then ρ is 0.8). If, however, there are languages sampled at different times throughout (pre)history then this *sequentially sampled* data can be modeled using a sampling rate ψ through time (Stadler et al. 2013). Combinations of both contemporaneous and sequential sampling are possible, for example, when a larger number of sampled languages is acquired at a few different points in time ('multi-ρ sampling'), or when extinct (sequentially sampled) languages are combined with modern (contemporaneously sampled) languages (Stadler et al. 2013).

6.2 Model comparison

To compare two models M_1 and M_2, Bayesian model selection allows estimation of the so-called BF of M_1 with respect to M_2, which is calculated as follows:

$$BF_{1,2} = \frac{P(D|M_1)}{P(D|M_2)} \tag{16}$$

where $P(D|M_1)$ is the marginal likelihood (ML) for model M_1 and $P(D|M_2)$ the ML for M_2. If the BF is above 1, there is support for model M_1 and if it is below 1 there is support for M_2. The strength of support is reported using the following classification (Kass and Raftery 1995): a BF between 1 and 3 is low support, between 3 and 20 is moderate support, between 20 and 150 is strong support, and over 150 very strong support. Note that sometimes the log of the BF is reported:

$$\log BF_{1,2} = \log P(D|M_1) - \log P(D|M_2) \tag{17}$$

The calculation of the ML is computationally very expensive but essential. There are several ways to obtain estimates of the ML for a model, some of which are implemented as BEAST2 packages (Bouckaert et al. 2014), including path sampling/stepping stone (Baele et al. 2013) (model-selection package) and nested sampling (Maturana et al. 2018) (NS package). Nested sampling provides an estimate of the (log) ML together with its standard deviation, unlike most other methods. This additional information allows accounting for uncertainty in the ML estimates when comparing log BF estimates: if the log BF is larger than twice the sum of standard deviations, the difference is significant.

6.3 Model validation

The aim of model selection is to compare between two or more models. This comparison does not tell us if the models in the pool are acceptable or fail to describe key aspects of the phylogeny. Instead to evaluate if a given model is a good description of the underlying evolutionary process involved in generating the data, we can use an approach called *model validation*. In model validation, the aim is to check whether the chosen model is capable of producing the empirical data. The fit between the real empirical data and the data generated under the model is known as absolute model fit (in contrast to the relative model fit obtained by the selection methods). There are several ways to achieve this.

First, one can evaluate known facts that are not used as prior information. In an appropriate model, these known facts should be within the inferred 95% highest posterior density interval (i.e. the interval, in which 95% of the samples are located inside). For example, one could compare the obtained tree topology to see if the inferred clades make sense according to the historical linguistics literature (Greenhill et al. 2010b), or compare inferred dates to those known historically (Ryder and Nicholls 2010).

More sophisticated methods involve posterior predictive simulations (PPSs) (Gelman et al. 2013). This approach simulates a large number of synthetic datasets under the model of interest. These simulated datasets are then compared with the empirical dataset using various test statistics. In an adequate model, all the statistics of the real data should be within the respective distributions of the PPS statistics. There is a broad range of test statistics suitable for Bayesian phylogenetics assessing various components of the analysis such as clock model (Duchêne et al. 2015) or the tree prior (Duchene et al. 2018).

The choice of validation method is linked to the research question. The most significant aspects of the phylogeny for the topic of interest should be covered by a model adequacy test.

7. Exploring the space of trees using BEAST2

Sampling trees from the space of all possible trees is not trivial. State of art software packages use a Markov Chain Monte Carlo (MCMC) algorithm to explore the space of trees and parameters and return a sample of their posterior distribution. Starting with an initial random tree and a set of parameters, the Markov Chain iteratively proposes small changes to the tree and parameters in turn. The algorithm accepts any proposal that improves the posterior probability. Proposals that lead to a decrease in posterior likelihood are accepted with a probability proportional to the likelihood ratio between the proposed and current state. This ensures efficient exploration of the state space without the algorithm getting stuck in local maxima. In the initial 'burn-in' phase, the MCMC algorithm needs time to find the region of the highest posterior probability. After that, each step is a sample from the space of trees and parameters according to the posterior distribution. Since these are highly correlated, we only periodically log them to a trace file, and tree file, which can be considered a sample from the posterior.

Phylogenetic analyses of language evolution in BEAST2 (Bouckaert et al. 2014) are facilitated by the *Babel* package. Furthermore, the birth–death processes are part of the *bdsky* package, and the M_k-model is part of the *morph-models* package. A step-by-step tutorial on how to set up and run an analysis with BEAST2 is found in the supplement and on https://taming-the-beast.org/tutorials/LanguagePhylogenies/, last accessed 03/08/2021.

Further prominent examples of software packages for Bayesian phylogenetics are Mr Bayes (Huelsenbeck and Ronquist 2001), Bayes Phylogenies (Pagel and Meade 2004), TraitLab (G. K. Nicholls and Welch 2021), and RevBayes (Höhna et al. 2016).

7.1 Convergence

There are two parameters that can be computed from the posterior sample to give clues about the quality of the analysis. The first one is the auto-correlation time (ACT), which is the average number of steps in the Markov chain that two samples need to be apart to be uncorrelated to each other. The second is the effective sample size (ESS), which is the total amount of samples divided by the ACT. The ESS is the estimated number of truly independent samples from the posterior distribution and a good analysis should have—as a rule of thumb—an ESS of at least 200 for every estimated parameter.

A low ESS might have several reasons. First, the mixing may be poor as the MCMC method has not efficiently searched through the tree space. This can be improved by running the chain for longer, increasing the sampling frequency or optimizing the operator setup. The chain length needed to reach convergence can differ largely for different datasets. Second, the burn-in phase may be too short. This is a postprocessing issue and can easily be solved by increasing the fraction of initial samples being discarded (cf. Fig. 8 in the Supplementary tutorial).

The posterior probability is the product of the prior distributions and the likelihood. If this product is dominated by the prior distributions, the data may not contribute to our knowledge. Hence, one should test if the posterior distribution is significantly different to the prior distribution. It is best practice to run a prior-only analysis without data to investigate the (joint) prior distributions (see Supplementary tutorial section 'The MCMC tab'). If the posterior distribution is very close to the prior distribution, a sensitivity analysis is required, that is, the analysis needs to be rerun with different prior distributions. As an additional measure of convergence, one should always run two or more independent analyses (with different initial states) and check that the posterior distributions agree with each other.

7.2 Summarizing posterior distributions

The posterior distribution of trees obtained from an analysis can be visually displayed using DensiTree (Bouckaert 2010; Bouckaert and Heled 2014), which plots each tree from the posterior sample on top of one another (an example is given in Fig. 12 of the Supplementary tutorial). This is a good way to visualize uncertainty in a tree topology. Furthermore, the

posterior distribution can be summarized into a single consensus tree—of which a 'Maximum Clade Credibility Tree' (MCC tree) is the most common—by using the TreeAnnotator application distributed with BEAST2. For all trees in the posterior sample, each internal node (i.e. each clade) gets a posterior *credibility*. This is calculated as the amount of topologies in the posterior sample, where this specific clade is presently divided by the sample size. The topology of the MCC tree is the one that has the highest product of clade credibilities, while the node heights are either kept from the chosen tree or set to the mean or median of the corresponding node heights in the posterior sample (see also Fig. 11 of the Supplementary tutorial).

8. Hypothesis testing with trees

In linguistics, Bayesian phylogenies have been used to test a wide range of hypotheses. First, and most obviously, are hypotheses about language subgroups and their timing. For example, Gray et al. (2009) tested two different scenarios of the Austronesian expansion. Their tree topology strongly indicated that the root of the Austronesian language family existed in Taiwan around 5,200 years ago in striking concordance with findings from the comparative method (e.g. Blust 1999). They found no support for a deeper origin in Island South East Asia around 15,000 years ago despite this being a common assumption in genetic studies (e.g. Soares et al. 2011). Sicoli and Holton (2014) used Bayesian phylogenies to test the increasingly accepted Den-Yeniseian hypothesis that connects the Na-Den languages of North America to the Yeniseian languages in central Asia (Kari and Potter 2010). This hypothesis suggests a striking migration across the Bering land bridge. Sicoli and Holton (2014) employed phylogenetic methods to—guardedly—evaluate this hypothesis and propose that the linguistic data are more consistent with a spread of these languages from Beringia into both America and Asia (rather than a back migration from North America to Asia, or migration from Asia to North America). Robbeets and Bouckaert (2018) tested hypotheses of Trans-Eurasian families.

Second, a major application for Bayesian phylogenetic methods in linguistics has been to investigate rates and patterns of trait evolution. One strand of research has investigated the stability of grammatical features over time and space (e.g. Dediu and Levinson 2012; Greenhill et al. 2010a; Cathcart et al. 2018), or compared rates of change between different aspects of language (Greenhill et al. 2017). Another strand has investigated how rates of change are shaped by external factors such as their frequency of use in speech communities (Pagel et al. 2007), or the size of the speech communities themselves (e.g. Greenhill et al. 2018). Yet another strand has investigated the co-evolution of particular language subsystems from word-order (Dunn et al. 2011), to color naming in Pama-Nyungan (Haynie and Bowern 2016), to the evolution of higher numerals in Indo-European (Calude and Verkerk 2016), and noun-phrase recursion (Widmer et al. 2017). Other studies have used phylogenies as the backbone for making inferences about phonology, from reconstructing the proto-forms of proto-languages (Bouchard-Côté et al. 2013), modeling sound changes over time (Hruschka et al. 2015), and inferring an increase in labiodentals over time purportedly enabled by a shift in diet to agriculture (Blasi et al. 2019).

Third, a current area of major growth is *phylogeography*, which can be used to infer the geographical homelands of language groups (Lemey et al. 2009; Bouckaert 2016). Given the longstanding interest in linguistic homelands (e.g. Sapir 1916), it is unsurprising that these have a growing home in linguistics too. For example, (Bouckaert et al. 2012) used these tools to controversially infer the homeland of the Indo-European languages to Anatolia, while Walker and Ribeiro (2011) proposed a western Amazonian origin of Arawakan. Bouckaert et al. (2018) introduced a founder-dispersal model, which can take landscape heterogeneity in account, to test various hypotheses on Pama-Nyungan origins. Other studies have used phylogeographic methods to link language family expansions to climate change, such as Grollemund et al. (2015) proposing that the Bantu expansion was facilitated as savannah corridors opened up through the rainforest, or Lehtinen et al. (2014) suggesting that a warm period between 7,500 and 5,000 years ago facilitated the spread of the Uralic language family. Yet another research direction investigates the effect of geographical barriers on language spreads (Lee and Hasegawa 2014). Discrete phylogeographic models (De Maio et al. 2015; Kühnert et al. 2016; Müller et al. 2018) may also be useful if language groups are so isolated that an island model is more appropriate than a model of continuous geographic dispersal.

9. Conclusion

Due to applications in evolutionary biology, Bayesian phylogenetics has experienced enormous progress during the past decade. Sophisticated models are implemented in relatively easy-to-use software programs making them accessible to a wide audience of researchers in many fields. In this article, we are aiming to make these methods more accessible to scholars of language evolution. By

explaining the basic concepts of models relevant for linguistic evolution, we hope to enable scholars to understand what the components of a Bayesian analysis inferring a phylogeny of languages are, such that they can make informed decisions on which prior distributions to choose and how to interpret their analyses.

Supplementary data

Supplementary data is available at *JOLEV* online.

Data Availability

The data underlying this article are available in the GitHub repository: https://github.com/KonstantinHoffmann/LanguagePhylogenies, last accessed 03/08/2021. The datasets were derived from sources in the public domain: Austronesian Basic Vocabulary Database (https://abvd.shh.mpg.de/austronesian/, last accessed 03/08/2021). Additionally the data underlying this article are available in its online supplementary material.

Acknowledgements

The authors thank the anonymous reviewers for their valuable suggestions.

Funding

K.H., D.K., and S.J.G. are supported by the Max Planck Society. S.J.G. has also been supported by the Australian Research Council's Discovery Projects funding scheme (Grant CE140100041). R.B. has been supported by Marsden grant 18-UOA-096 from the Royal Society of New Zealand.

Conflict of interest statement

None declared.

Notes

1. See Section 2.3.
2. See Section 4.
3. See section 'Getting the data into BEAUti' and Fig. 4 in the Supplementary tutorial.

References

Baele, G. et al. (2013) 'Accurate Model Selection of Relaxed Molecular Clocks in Bayesian Phylogenetics', *Molecular Biology and Evolution*, 30/2: 239–43.

Barido-Sottani, J. et al. (2017) 'Taming the BEAST—A Community Teaching Material Resource for BEAST 2', *Systematic Biology*, 67/1: 170–4.

Benedict, P. K. (1972) *Sino-Tibetan: A Conspectus*. Cambridge: Cambridge University Press.

Bergsland, K., and Vogt, H. (1962) 'On the Validity of Glottochronology', *Current Anthropology*, 3/2: 115–53.

Birchall, J., Dunn, M., and Greenhill, S. J. (2016) 'A Combined Comparative and Phylogenetic Analysis of the Chapacuran Language Family', *International Journal of American Linguistics*, 82/3: 255–84.

Blasi, D. E. et al. (2019) 'Human Sound Systems Are Shaped by Post-Neolithic Changes in Bite Configuration', *Science*, 363/6432: eaav3218.

Blench, R., and Post, M. W. (2014) 'Rethinking Sino-Tibetan Phylogeny from the Perspective of North East Indian Languages'. In: Hill, T., Nathan W., and Owen-Smith (eds) *Trans-Himalayan Linguistics*, pp. 71–104. Berlin, Boston: Mouton de Gruyter.

Blust, R. (1999) 'Subgrouping, Circularity and Extinction: Some Issues in Austronesian Comparative Linguistics'. In: Elizabeth, Z., and Li, R. (eds) *Selected Papers from the Eighth International Conference on Austronesian Linguistics*, pp. 31–94. Taipei, Taiwan: Symposium Series of the Institute of Linguistics, Academia Sinica.

—— (2000) 'Why Lexicostatistics Doesn't Work: The 'Universal Constant' Hypothesis and the Austronesian Languages'. In: Renfrew, C., McMahon, A., and Trask, L. (eds) *Time Depth in Historical Linguistics*, pp. 311–331. Cambridge: McDonald Institute for Archaeological Research.

Bouchard-Côté, A., Hall, D., Griffiths, T. L., and Klein D. . (2013) 'Automated Reconstruction of Ancient Languages Using Probabilistic Models of Sound Change', *Proceedings of the National Academy of Sciences*, 110/11: 4224–9.

Bouckaert, R. R. (2016) 'Phylogeography by Diffusion on a Sphere: Whole World Phylogeography', *PeerJ*, 4: e2406.

—— R. and and Heled, J. (2014) 'DensiTree 2: Seeing Trees Through the Forest'. *bioRxiv*, 012401. doi: 10.1101/012401. (December 08, 2014).

—— R. and and Robbeets, M. (2017) 'Pseudo Dollo Models for the Evolution of Binary Characters Along a Tree'. *bioRxiv*, 207571. doi: 10.1101/207571. (October 23, 2017).

—— R. et al. (2012) 'Mapping the Origins and Expansion of the Indo-European Language Family', *Science*, 337/6097: 957–60.

Bouckaert, R. R. (2010) 'DensiTree: Making Sense of Sets of Phylogenetic Trees', *Bioinformatics*, 26/10: 1372–3.

—— et al. (2014) 'BEAST 2: A Software Platform for Bayesian Evolutionary Analysis', *PLoS Comput Biol*, 10/4: e1003537.

——, Bowern, C., and Atkinson, Q. D. (2018) 'The Origin and Expansion of Pama-Nyungan Languages across Australia', *Nature Ecology & Evolution*, 2/4: 741–9.

Brugmann, K. (1884) 'Zur Frage Nach Den Verwandtschaftsverhältnissen Der Indogermanischen Sprachen', *Internationale Zeitschrift Für Allgemeine Sprachewissenschaft*, 1: 226–56.

Calude, A. S., and Verkerk, A. (2016) 'The Typology and Diachrony of Higher Numerals in Indo-European: A Phylogenetic Comparative Study', *Journal of Language Evolution*, 1/2: 91–108.

Cathcart, C. et al. (2018) 'Areal Pressure in Grammatical Evolution', *Diachronica*, 35/1: 1–34.

Chang, W., Cathcart, C., Hall, D., and Garrett, A.. (2015) 'Ancestry-Constrained Phylogenetic Analysis Supports the Indo-European Steppe Hypothesis', *Language*, 91/1: 194–244.

Dediu, D., and Levinson, S. C. (2012) 'Abstract Profiles of Structural Stability Point to Universal Tendencies, Family-Specific Factors, and Ancient Connections between Languages', *PloS One*, 7/9: e45198.

Dollo, L. (1893) 'Les Lois de Lvolution', *Bulletin de la Socit Belge de Gologie*, 7: 164–6.

Douglas, J., Zhang, R., and Bouckaert, R. R. (2021) 'Adaptive Dating and Fast Proposals: Revisiting the Phylogenetic Relaxed Clock Model', *PLoS Computational Biology*, 17/2: e1008322.

Drummond, A. J., and Suchard, M. A. (2010) 'Bayesian Random Local Clocks, or One Rate to Rule Them All', *BMC Biology*, 8/1: 114.

Drummond, A. J., Rambaut, A., Shapiro, B., and Pybus, O. G. (.2005) 'Bayesian Coalescent Inference of Past Population Dynamics from Molecular Sequences', *Molecular Biology and Evolution*, 22/5: 1185–92.

Drummond, A. J., Ho, S. Y. W., Phillips, M. J., and Rambaut, A. (2006) 'Relaxed Phylogenetics and Dating with Confidence', *PLoS Biol*, 4/5: e88.

Duchêne, D. A., Duchêne, S., Holmes, E. C., and Ho, S. Y. W. . (2015) 'Evaluating the Adequacy of Molecular Clock Models Using Posterior Predictive Simulations', *Molecular Biology and Evolution*, 32/11: 2986–95.

Duchene, S. et al. (2018) 'Phylodynamic Model Adequacy Using Posterior Predictive Simulations', *Systematic Biology*, 68/2: 358–64.

—— et al. (2020) 'Bayesian Evaluation of Temporal Signal in Measurably Evolving Populations', *Molecular Biology and Evolution*, 37/11: 3363–79.

Dunn, M., Greenhill, S. J., Levinson, S. C., and Gray, R. D. . (2011) 'Evolved Structure of Language Shows Lineage-Specific Trends in Word-Order Universals', *Nature*, 473/7345: 79–82.

Felsenstein, J. (2004) *Inferring Phylogenies*, Vol. 2. Sunderland, MA: Sinauer associates.

Fourment, M., and Darling, A. E. (2018) 'Local and Relaxed Clocks: The Best of Both Worlds', *PeerJ*, 6: e5140.

Nicholls, G. K., Ryder, R. J., and Welch, D. (2021) 'Traitlab: a MATLAB Package for Fitting and Simulating Binary Trait-Like Data. Technical Report.' *Journal of Statistical Software*, VV/II: 79–82.

Gavryushkina, A., Welch, D., Stadler, T., and Drummond, A. J. (2014) 'Bayesian Inference of Sampled Ancestor Trees for Epidemiology and Fossil Calibration', *PLoS Computational Biology*, 10/12: e1003919.

Gelman, A., Carlin J. B. et al. (2013) *Bayesian Data Analysis*. New York: Chapman and Hall/CRC. doi: 10.1201/b16018.

Gernhard, T. (2008) 'The Conditioned Reconstructed Process', *Journal of Theoretical Biology*, 253/4: 769–78.

Gray, R. D., and Atkinson, Q. D. (2003) 'Language-Tree Divergence Times Support the Anatolian Theory of Indo-European Origin', *Nature*, 426/6965: 435–9.

——, Drummond, A. J., and Greenhill, S. J. (2009) 'Language Phylogenies Reveal Expansion Pulses and Pauses in Pacific Settlement', *Science*, 323/5913: 479–83.

Greenhill, S. J., and Gray, R. D. (2009) 'Austronesian Language Phylogenies: Myths and Misconceptions about Bayesian Computational Methods'. In: Adelaar, K. A., and Pawley, A. (eds) *Austronesian Historical Linguistics and Culture History: A Festschrift for Robert Blust*, pp. 375–97. Canberra: Pacific Linguistics.

——, Currie, T. E., and Gray, R. D. (2009) 'Does Horizontal Transmission Invalidate Cultural Phylogenies?', *Proceedings of the Royal Society, B. Biological Sciences*, 276/1665: 2299–306.

Greenhill, S. J., Atkinson, Q. D., Meade, A., and Gray, R. D. (.2010a) 'The Shape and Tempo of Language Evolution', *Proceedings of the Royal Society B: Biological Sciences*, 277/1693: 2443–50.

Greenhill, S. J., Drummond, A. J., and Gray, R. D. (2010b) 'How Accurate and Robust Are the Phylogenetic Estimates of Austronesian Language Relationships?', *PLoS One*, 5/3: e9573.

—— et al. (2017) 'Evolutionary Dynamics of Language Systems', *Proceedings of the National Academy of Sciences*, 114: 201700388.

—— et al. (2018) 'Population Size and the Rate of Language Evolution: A Test across Indo-European, Austronesian, and Bantu Languages', *Frontiers in Psychology*, 9: 1–18.

Grollemund, R. et al. (2015) 'Bantu Expansion Shows That Habitat Alters the Route and Pace of Human Dispersals', *Proceedings of the National Academy of Sciences*, 112/43: 13296–301.

Harrison, L. B., and Larsson, H. C. E. (2014) 'Among-Character Rate Variation Distributions in Phylogenetic Analysis of Discrete Morphological Characters', *Systematic Biology*, 64/2: 307–24.

Haynie, H. J., and Bowern, C. (2016) 'Phylogenetic Approach to the Evolution of Color Term Systems', *Proceedings of the National Academy of Sciences*, 113/48: 13666–71. 1613666113.

Heggarty, P., Maguire, W., and McMahon, A. (2010) 'Splits or Waves? Trees or Webs? How Divergence Measures and Network Analysis Can Unravel Language Histories', *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 365: 3829–43.

Heled, J., and Drummond, A. J. (2011) 'Calibrated Tree Priors for Relaxed Phylogenetics and Divergence Time Estimation', *Systematic Biology*, 61/1: 138–49.

Hennig, W. (1996) *Phylogenetic Systematics*. Champaign, IL: University of Illinois Press.

Höhna, S. et al. (2016) 'RevBayes: Bayesian Phylogenetic Inference Using Graphical Models and an Interactive Model-Specification Language', *Systematic Biology*, 65/4: 726–36.

Honkola, T. et al. (2013) 'Cultural and Climatic Changes Shape the Evolutionary History of the Uralic Languages', *Journal of Evolutionary Biology*, 26/6: 1244–53.

Hruschka, D. J. et al. (2015) 'Detecting Regular Sound Changes in Linguistics as Events of Concerted Evolution', *Current Biology*, 25/1: 1–9

Huelsenbeck, J. P. and Ronquist, F. (2001) 'MRBAYES: Bayesian Inference of Phylogenetic Trees', *Bioinformatics*, 17/8: 754–5.

Kari, J., and Potter, B. A. (2010) *The Dene-Yeniseian Connection. Anthropological papers of the University of Alaska*. Fairbanks, AK: University of Alaska Fairbanks.

Kass, R. E., and Raftery, A. E. (1995) 'Bayes Factors', *Journal of the American Statistical Association*, 90/430: 773–95.

Kingman, J. F. C. (1982) 'On the Genealogy of Large Populations', *Journal of Applied Probability*, 19/A: 27–43.

Kitchen, A., Ehret, C., Assefa, S., and Mulligan, C. J. (.2009) 'Bayesian Phylogenetic Analysis of Semitic Languages Identifies an Early Bronze Age Origin of Semitic in the near East', *Proceedings of the Royal Society B: Biological Sciences*, 270/1668: 2703–10.

Kolipakam, V. et al. (2018) 'A Bayesian Phylogenetic Study of the Dravidian Language Family', *Royal Society Open Science*, 5/3: 171504.

Kühnert, D., Stadler, T., Vaughan, T. G., and Drummond, A. J. . (2016) 'Phylodynamics with Migration: A Computational Framework to Quantify Population Structure from Genomic Data', *Molecular Biology and Evolution*, 33/8: 2102–16.

Lee, S. (2015) 'A Sketch of Language History in the Korean Peninsula', *Plos One*, 10/5: e0128448.

——, and Hasegawa, T. (2011) 'Bayesian Phylogenetic Analysis Supports an Agricultural Origin of Japonic Languages'. *Proceedings of the Royal Society B, Biological Sciences*, 278/1725: 3662–9.

—— and —— (2014) 'Oceanic Barriers Promote Language Diversification in the Japanese Islands', *Journal of Evolutionary Biology*, 27/9: 1905–12.

Lees, R. B. (1953) 'The Basis of Glottochronology', *Language*, 29/2: 113–27.

Lehtinen, J. et al. (2014) 'Behind Family Trees: Secondary Connections in Uralic Language Networks', *Language Dynamics and Change*, 4/2: 189–221.

Lemey, P., Rambaut, A., Drummond, A. J., and Suchard, M. A. (.2009) 'Bayesian Phylogeography Finds Its Roots', *PLoS Computational Biology*, 5/9: e1000520.

Lewis, P. O. (2001) 'A Likelihood Approach to Estimating Phylogeny from Discrete Morphological Character Data', *Systematic Biology*, 50/6: 913–25.

De Maio, N., Wu, C-H., O'Reilly, K. M., and Wilson, D. . (2015) 'New Routes to Phylogeography: A Bayesian Structured Coalescent Approximation', *PLOS Genetics*, 11/8: e1005421.

Matisoff, J. A. (2003) *Handbook of Proto-Tibeto-Burman, Volume 135 of University of California Publications in Linguistics*. Berkeley and Los Angeles: University of California press.

Maturana, P., Brewer, B. J., Klaere, S., and Bouckaert, R. R. (.2018) 'Model Selection and Parameter Inference in Phylogenetics Using Nested Sampling', *Systematic Biology*, 68: 219–33.

Maurits, L. et al. (2019) 'Best Practices in Justifying Calibrations for Dating Language Families', *Journal of Language Evolution*, 5/1: 17–38.

Menéndez, M. (1981) 'Uma Contribuição Para a Etno-História da Área Tapajós-Madeira', *Revista Do Museu Paulista*, 28: 289–388.

Michael, L. et al. (2015) 'A Bayesian Phylogenetic Classification of Tupi-Guarani', *LIAMES*, 15/2: 1–36.

Moore, J. H. (1994) 'Putting Anthropology Back Together Again: The Ethnogenetic Critique of Cladistic Theory', *American Anthropologist*, 96/4: 925–48.

Müller, N. F., Rasmussen, D., and Stadler, T. (2018) 'MASCOT: Parameter and State Inference under the Marginal Structured Coalescent Approximation', *Bioinformatics*, 34/22: 3843–8.

Nicholls, G. K., and Gray, R. D. (2008) 'Dated Ancestral Trees from Binary Trait Data and Their Application to the Diversification of Languages', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70/3: 545–66.

Pagel, M., and Meade, A. (2004) 'A Phylogenetic Mixture Model for Detecting Pattern-Heterogeneity in Gene Sequence or Character-State Data', *Systematic Biology*, 53/4: 571–81.

——, and —— (2006) 'Estimating Rates of Lexical Replacement on Phylogenetic Trees of Languages'. In: Forster P., and Renfrew, C. (eds) *Phylogenetic Methods and the Prehistory of Languages*, Vol. 1, pp. 173–182. Cambridge: McDonald Institute for Archaeological Research.

——, Atkinson, Q. D., and Meade, A. (2007) 'Frequency of Word-Use Predicts Rates of Lexical Evolution throughout Indo-European History', *Nature*, 449/7163: 717–20.

Peiros, I. (1998) *Comparative Linguistics in Southeast Asia*. Canberra: Australian National University.

Penny, D., McComish, B. J., Charleston, M. A., and Hendy, M. D. (.2001) 'Mathematical Elegance with Biochemical Realism: The Covarion Model of Molecular Evolution', *Journal of Molecular Evolution*, 53/6: 711–23.

Rama, T. (2018) 'Three Tree Priors and Five Datasets', *Language Dynamics and Change*, 8/2: 182–218.

Ritchie, A. M., and Ho, S. Y. W. (2019) 'Influence of the Tree Prior and Sampling Scale on Bayesian Phylogenetic Estimates of the Origin Times of Language Families', *Journal of Language Evolution*, 4/2: 108–23.

Robbeets, M. and Bouckaert, R. R. (2018) 'Bayesian Phylolinguistics Reveals the Internal Structure of the Transeurasian Family', *Journal of Language Evolution*, 3/2: 145–62.

Ryder, R. J., and Nicholls, G. K. (2010) 'Missing Data in a Stochastic Dollo Model for Binary Trait Data, and Its

Application to the Dating of Proto-Indo-European', *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 60/1: 71–92.

Sagart, L. et al. (2019) 'Dated Language Phylogenies Shed Light on the Ancestry of Sino-Tibetan', *Proceedings of the National Academy of Sciences*, 117/26: 14857–63. ISSN 0027–8424.

Sapir, E. (1916) *Time Perspective in Aboriginal American Culture: A Study in Method*. Ottawa: Government Printing Bureau.

Sicoli, M. A., and Holton, G. (2014) 'Linguistic Phylogenies Support Back-Migration from Beringia to Asia', *PloS One*, 9/3: e91722.

Soares, P. et al. (2011) 'Ancient Voyaging and Polynesian Origins', *American Journal of Human Genetics*, 88/2: 239–47.

Stadler, T., Kühnert, D., Bonhoeffer, S., and Drummond, A. J. (.2013) 'Birth–Death Skyline Plot Reveals Temporal Changes of Epidemic Spread in HIV and Hepatitis C Virus (HCV)', *Proceedings of the National Academy of Sciences*, 110/1: 228–33.

Swadesh, M. (1950) 'Salish Internal Relationships', *International Journal of American Linguistics*, 16/4: 157–67.

Tuffley, C., and Steel, M. (1998) 'Modeling the Covarion Hypothesis of Nucleotide Substitution', *Mathematical Biosciences*, 147/1: 63–91.

van Driem, G. (2003) 'Review of Thurgood and LaPolla 2003', *Bulletin of the School of Oriental and African Studies*, 66/2: 282–4.

Walker, R. S. and Ribeiro, L. A. (2011) 'Bayesian Phylogeography of the Arawak Expansion in Lowland South America'. *Proceedings of the Royal Society B, Biological Sciences*, 278/1718: 2562–7.

Widmer, M. et al. (2017) 'NP Recursion over Time: Evidence from Indo-European', *Language*, 93/4: 799–826.

Wilmshurst, J. M., Hunt, T. L., Lipo, C. P., and Anderson, A. J. (.2010) 'High-Precision Radiocarbon Dating Shows Recent and Rapid Initial Human Colonization of East Polynesia', *Proceedings of the National Academy of Sciences*, 108/5: 1815–20.

Yang, Z. (1994) 'Maximum Likelihood Phylogenetic Estimation from DNA Sequences with Variable Rates over Sites: Approximate Methods', *Journal of Molecular Evolution*, 39/3: 306–14.

Yule, G. U. (1925) 'A Mathematical Theory of Evolution, Based on the Conclusions of Dr. J. C. Willis, F.R.S', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 213/402–410: 21–87.

Zhang, M., Yan, S., Pan, W., and Jin, L. (.2019) 'Phylogenetic Evidence for Sino-Tibetan Origin in Northern china in the Late Neolithic', *Nature*, 569/7754: 112–5.