Principal Component Analysis as a Sanity Check for Bayesian Phylolinguistic Reconstruction

Yugo Murawaki
Graduate School of Informatics, Kyoto University
Yoshida-honmachi, Sakyo-ku, Kyoto, 606-8501, Japan
murawaki@i.kyoto-u.ac.jp

Abstract

Bayesian approaches to reconstructing the evolutionary history of languages rely on the tree model, which assumes that these languages descended from a common ancestor and underwent modifications over time. However, this assumption can be violated to different extents due to contact and other factors. Understanding the degree to which this assumption is violated is crucial for validating the accuracy of phylolinguistic inference. In this paper, we propose a simple sanity check: projecting a reconstructed tree onto a space generated by principal component analysis. By using both synthetic and real data, we demonstrate that our method effectively visualizes anomalies, particularly in the form of *jogging*.

Keywords: Bayesian phylolinguistics, tree model, principal component analysis

1. Introduction

The tree model serves as the foundation for historical-comparative linguistics (Schleicher, 1853). Although manual inference has traditionally been dominant in the field (Campbell and Poser, 2008), the influence of evolutionary biology has led to a rapid rise to computation-heavy statistical analysis of linguistic data (Gray and Jordan, 2000; Gray and Atkinson, 2003; Bouckaert et al., 2012; Rama and Wichmann, 2018), spawning a multitude of papers built upon Bayesian phylolinguistic tools.

The tree model assumes that the evolutionary history of related languages can be represented as a tree. The root represents a single common ancestor and a number of branching events lead to the observed languages. Over time, modifications gradually accumulate along the branches, indicating that the distance between two languages on the tree approximately corresponds to the extent of divergence between them. Various methods have been proposed based on this intuition to address the inverse problem of reconstructing the tree from the observed languages (Felsenstein, 2004).

In reality, the tree model is violated to varying degrees. When languages come into contact, there is often a horizontal transmission of features between them, despite the assumption that they evolve independently. This horizontal transmission necessitates the addition of extra edges, resulting in a representation that is no longer a tree but a network (Nakhleh et al., 2005; Nelson-Sathi et al., 2011). Despite efforts to integrate horizontal transfer into statistical models (Kelly and Nicholls, 2017; Neureiter et al., 2022), achieving stable and scalable inference continues to pose a significant challenge. For this reason, the tree model retains its dominant position in phylolinguistics.

The modern proponents of the tree model are well aware of repeated criticisms that in fact date back centuries (Schmidt, 1872; Kalyan and François, 2018). Initially, they attempted to demonstrate the model's robustness against horizontal transmission by utilizing synthetic data (Greenhill et al., 2009; Barbançon et al., 2013). Subsequently, they focused their attention on examining the extent to which the tree model is applicable to real data (Gray et al., 2010; Auderset et al., 2023). Unfortunately, there is a disparity between the tree model and their analytical tools: NeighborNet (Bryant and Moulton, 2004), the δ score (Holland et al., 2002), and the Q-residual score (Gray et al., 2010). These tools are all based on distance-based approaches despite the use of Bayesian methods for phylolinguistic reconstruction.

Recent studies (Auderset et al., 2023) conduct additional analyses using Bayesian tree summarization tools such as DensiTree (Bouckaert, 2010), based on the speculation that a relative absence of disagreements within a summary tree may indicate endorsement of the tree model. However, uncertainty is an intrinsic characteristic of Bayesian inference that emerges regardless of whether the model's assumptions are valid. After all, the model itself lacks a direct means to assess the accuracy of its underlying assumptions. It indeed can be deceived by fundamentally non-tree-like generative processes (Murawaki, 2015).

In this paper, we present a simple and practical approach to directly analyzing Bayesian phylolinguistic reconstruction. We apply principal component analysis (PCA) to language states and project a reconstructed tree onto the PCA-

LREC-COLING 2024, pages 12999–13013
20-25 May, 2024.

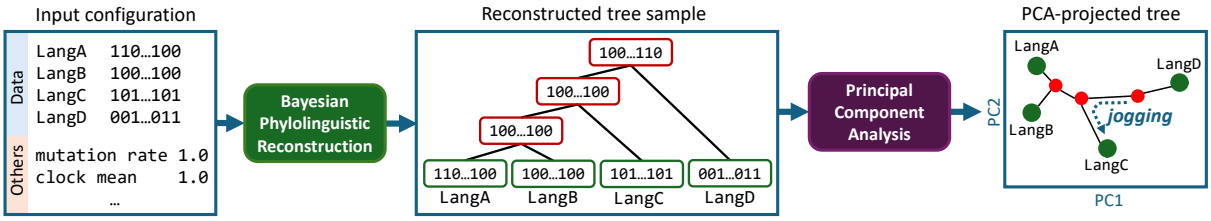


Figure 1: Overview of the proposed method. In this example, we reconstruct a phylogenetic tree for four modern languages, resulting in three ancestral nodes with explicitly represented states. The states of these seven languages are then subjected to principal component analysis (PCA), followed by projection onto a low-dimensional space. The downward path from the root to `LangC` exhibits *jogging*.

generated space (Figure 1). Our key idea is to leverage *continual diversification*, an aspect of tree-shaped evolution that usually falls outside the scope of the model's assumptions. We expect ancestor-descendant transitions to follow a unidirectional pattern along the first principal component axis. A gross violation of this unidirectionality, which we call *jogging*, can be seen as a deviation from the tree model, as is evident in Figure 4. To illustrate the usefulness of the proposed method, we provide demonstrations using both synthetic and real data, emphasizing its potential as a sanity check. The code is publicly available at https://github.com/murawaki/treepca.

2. Preliminaries

2.1. Binary Sequence Representations

In a typical Bayesian phylolinguistic reconstruction scenario, we are provided with a collection of observed languages, where each language is represented as a binary sequence. Most studies use binary-coded basic vocabulary data. These lexical data are originated from glottochronology (Swadesh, 1952), despite the decline of glottochronology itself due to substantial criticism (Bergsland and Vogt, 1962).

Basic vocabulary items such as `WATER` and `EAT` are assumed to be culture independent and resistant to change. The process of constructing lexical data involves two steps. Linguists begin by collecting words for these items in each language. Subsequently, they assess the cognacy (relatedness) of these words across languages. For example, English *water* and German *Wasser* share their etymological root whereas French *eau* and Italian *acqua* are cognates. By organizing these two cognate groups, we can represent English and German as `10` and French and Italian as `01`, where `1` and `0` indicate the presence and absence of a cognate group, respectively. Concatenating multiple basic vocabulary items, we typically obtain hundreds or thousands of binary features.

Although the evolutionary process of these binary features is assumed to follow a tree-like pattern, this assumption is not exempt from violations. One common type of deviation arises from loanwords. Since cognacy judgments rely on regular sound correspondences, loanwords can be identified by linguists and subsequently excluded from the dataset. However, older loanwords and borrowings between closely-related languages pose a higher risk of going undetected, thus potentially eluding removal from the analysis.

Thanks to the arbitrariness of meaning-symbol connection, it is generally assumed that a feature is gained only once throughout history. However, it is important to recognize that this assumption can be violated. One common cause of such deviations is semantic shift. For instance, the semantic shift from `PERSON` to `MAN` is universal and can happen in parallel, leading to multiple gains of the same word for `MAN` in a tree (Chang et al., 2015).

2.2. Bayesian Phylolinguistic Models

Bayesian phylolinguistic models encompass a range of advanced statistical techniques. For a comprehensive understanding of the details, we recommend referring to Drummond and Bouckaert (2015). Here, we will provide a high-level overview of the topic.

A phylolinguistic model assigns a probability to a generative process that begins with a common ancestor and extends to observed languages.[1] The probabilistic assessment can be subdivided into three primary components: a time-tree, state transitions, and rate variations. A time-tree represents a rooted tree where each node is associated with a calendar or relative date. The likelihood of a given time-tree is evaluated using a time-tree model.

Each node holds a binary sequence as its state, and the transition from a parent to a child involves gains ($0 \rightarrow 1$) and losses ($1 \rightarrow 0$). The probability of such transitions is assessed by a continuous-

[1]To be precise, coalescent variants of the time-tree model look backward in time.

time state transition model. For inference efficiency, the states of the unobserved languages are usually marginalized out, accounting for all possible combinations of the states (Felsenstein, 1981).

The state transition model is linked to a rate model. The strict clock model enforces a uniform rate of change in a tree, whereas various relaxed clock models investigate rate variations. By assigning different rates to different branches, we can analyze the potential alternation of rapid and slow phases of language change (Greenhill et al., 2017). Furthermore, assigning distinct rates to features or groups of features allows for the exploration of the hypothesis that certain vocabulary items display greater stability (Pagel et al., 2013).

With observed languages and optional hard constraints, the remaining portion of the generative process defines the search space. The conventional inference method is Markov Chain Monte Carlo (MCMC) sampling, which generates samples from the probability distribution. For our analysis, it is important to note that the sampler does not directly track the states of the unobserved languages because they are marginalized out. Nevertheless, it is easy to generate them using an algorithm analogous to forward filtering-backward sampling for sequence data (Scott, 2002).

MCMC sampling yields a vast number of time-trees, making it necessary to employ summarization techniques for human interpretation. One widely used approach is to construct a maximum clade credibility (MCC) tree by merging these samples. DensiTree (Bouckaert, 2010) offers another type of intuitive visualization that effectively highlights disagreements among the samples.

2.3. Principal Component Analysis (PCA)

Principal component analysis (PCA) linearly transforms high-dimensional data into a new coordinate system, where each principal component (PC) represents a new axis. Since the first few PCs usually capture key variance in the original data, PCA can be used for visualization.

While usually deemed irrelevant in phylolinguistics, PCA is ubiquitous in population genetics (Menozzi et al., 1978; Patterson et al., 2006). PCA itself is agnostic to the evolutionary process underlying genome data. In fact, whole-genome data do not follow a tree-like pattern either at a micro level due to recombination or at a macro level due to admixture (interbreeding of distinct populations). While population genetics gives weight to scalability (Galinsky et al., 2016), a naïve implementation suffices for small linguistic data.

Formally, let $\widetilde{X}$ be an $n \times p$ binary matrix, where n is the number of languages and p is the number of features. We first apply mean centering to $\widetilde{X}$:

$$X = \widetilde{X} - \boldsymbol{\mu},$$

where each element μ_i of the vector $\boldsymbol{\mu}$ represents the mean of the corresponding feature. We then apply singular value decomposition (SVD) to X:

$$X = U\Sigma V^{\mathsf{T}},$$

where U is an $n \times n$ orthogonal matrix containing the left singular vectors, Σ is an $n \times p$ diagonal matrix of singular values, and V^{T} is the transpose of an $p \times p$ orthogonal matrix containing the right singular vectors. Finally, we obtain the projection of X onto the i-th PC by

$$\hat{\boldsymbol{x}}_i = X\boldsymbol{u}_i,$$

where $\boldsymbol{u}_i$ is the i-th column of U.[2]

The proportion of variance explained by the i-th PC can be calculated as $\lambda_i / \sum_{i=1}^{k} \lambda_i$, where $\lambda_i = \sigma_i^2/(n-1)$. In this paper, we only use the first two PCs for visualization. In fact, the proportion of variance explained by the first two PCs for linguistic data (usually a few tens of percent) is much larger than that for genome data.

3. Proposed Method

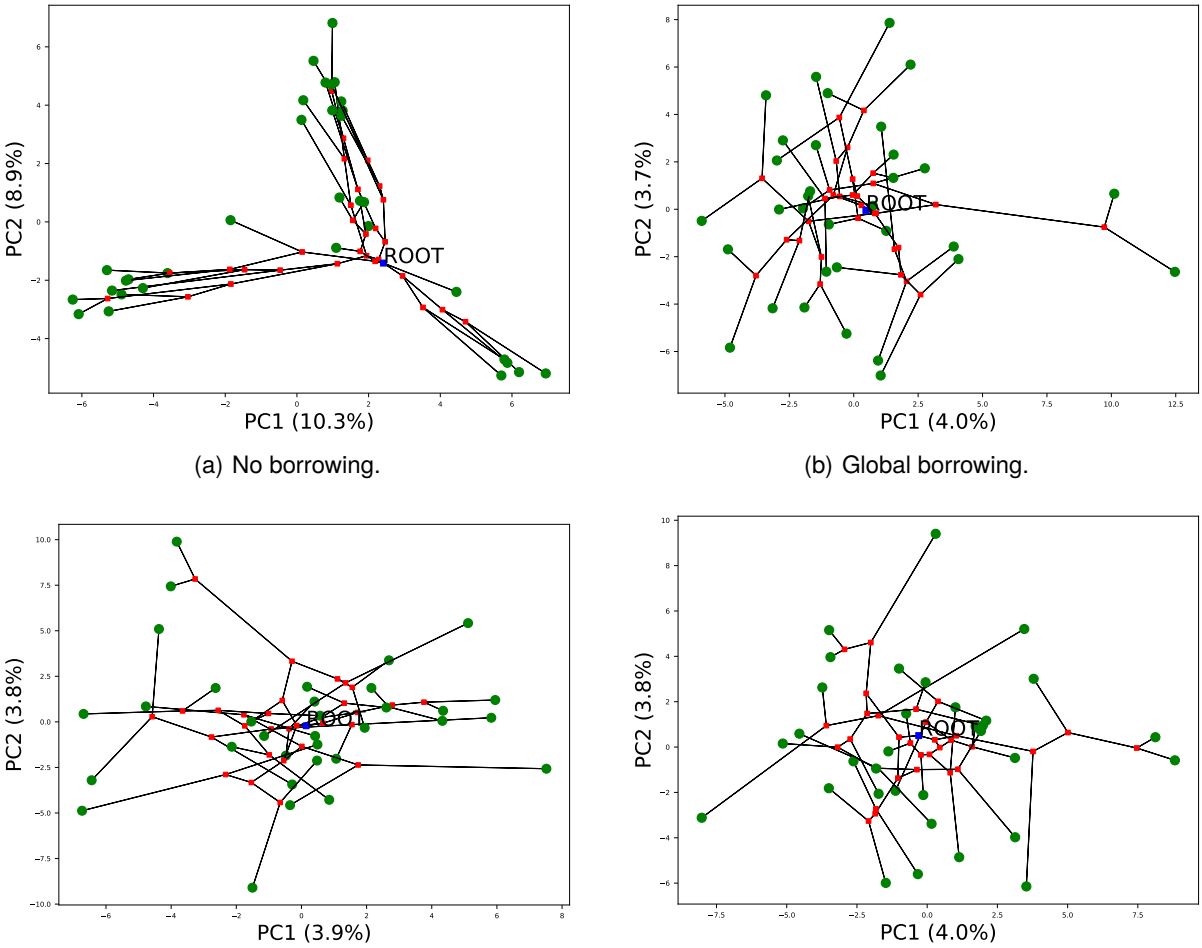
Our idea is fairy simple: project a reconstructed tree onto the two-dimensional space generated by PCA to check if it exhibits anomalies. To do this, we begin by applying PCA to $\widetilde{X}$, the states of the observed languages, to calculate $\hat{\boldsymbol{x}}_1$, $\hat{\boldsymbol{x}}_2$, $\boldsymbol{\mu}$, $\boldsymbol{u}_1$, and $\boldsymbol{u}_2$.[3] Next, We perform Bayesian phylolinguistic reconstruction and obtain a sample tree from the sampler. Let $\widetilde{Y}$ be an $(n-1) \times p$ matrix representing the states of the unobserved languages in the sample.[4] Using $\boldsymbol{\mu}$, $\boldsymbol{u}_1$, and $\boldsymbol{u}_2$, we map $\widetilde{Y}$ to $\hat{\boldsymbol{y}}_1$ and $\hat{\boldsymbol{y}}_2$. Finally, we draw a scatter plot of the entire set of languages, with additional straight lines connecting parents to children.

We anticipate that ancestor-descendant transitions will predominantly follow a unidirectional pattern along the first PC axis. This expectation aligns with the tree model's implication of continuous diversification. Note, however, that this is not a rigid

[2] Contrary to a belief mentioned in Elhaik (2022), PCA does not preserve distances between data points in the lower-dimensional space. For cases where distance preservation is crucial, which we suspect might not be prevalent, considering multidimensional scaling (MDS) may be more appropriate.

[3] Observed languages may contain missing features. In that case, we let the sampler impute these values.

[4] A bifurcating tree with n leaves has $n-1$ internal nodes including the root.



(a) No borrowing.

(b) Global borrowing.

(c) Local borrowing (1,000-year limit).

(d) Local borrowing (3,000-year limit).

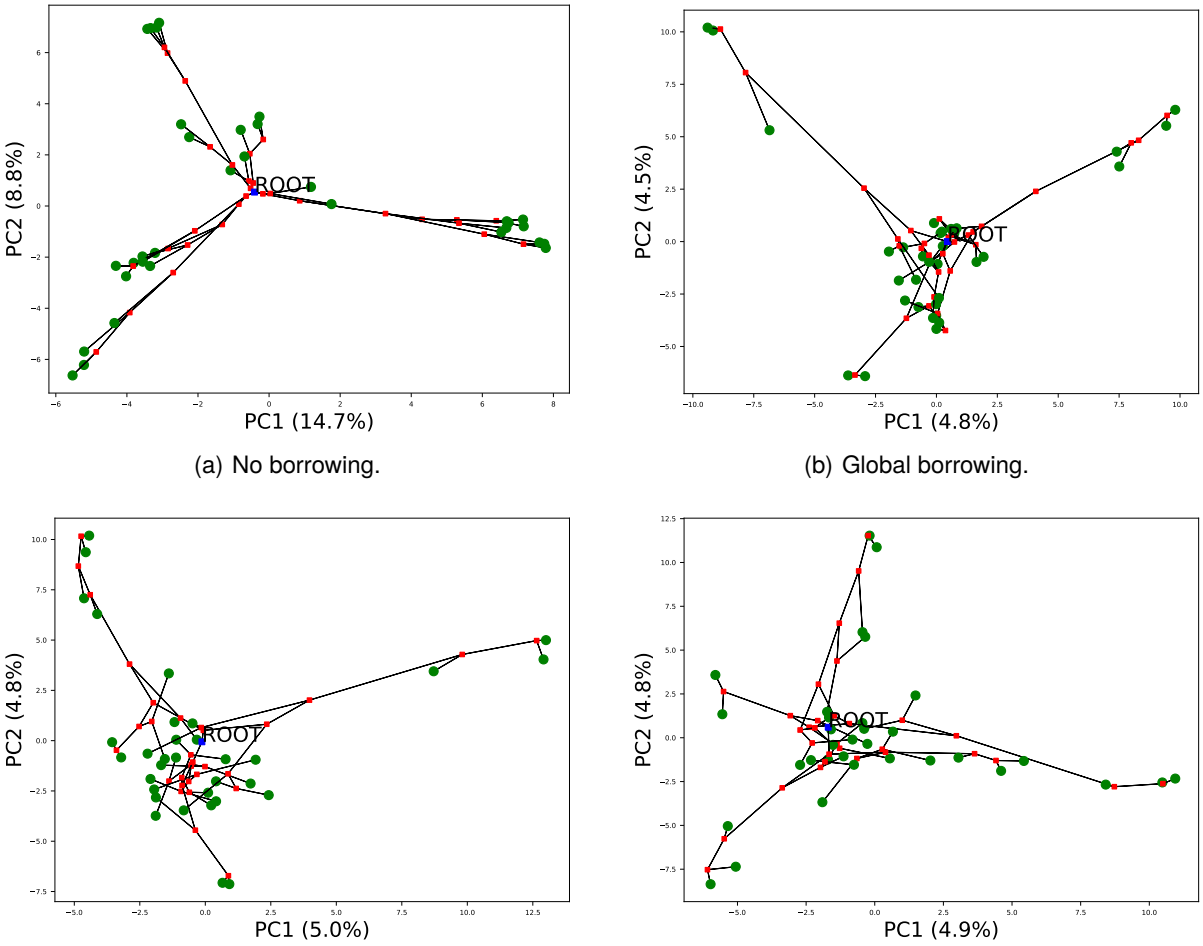
Figure 2: PCA of Bayesian phylolinguistic reconstruction for the skewed time-tree of data simulation, with four borrowing scenarios. We used the first two PCs, denoted as PC1 and PC2. A percentage indicates the amount of variance explained by the corresponding PC. Circles indicate observed leaf nodes while rectangles denote reconstructed internal nodes.

mathematical statement. To see why, let us consider three languages situated along a downward path in the tree. There are 2^3 possible combinations for each of the hundreds or thousands of binary features. The two changeless patterns ($0 \to 0 \to 0$ and $1 \to 1 \to 1$) can be ignored. The four single-change patterns ($0 \to 0 \to 1$, $0 \to 1 \to 1$, $1 \to 0 \to 0$, and $1 \to 1 \to 0$) together contribute to unidirectionality. Among the remaining two patterns, $0 \to 1 \to 0$ is a perfectly valid transition and yet goes against unidirectionality. The last pattern, $1 \to 0 \to 1$, is a violation of the assumption, with horizontal transmission being the main contributing factor, although sporadic parallel innovations cannot be entirely dismissed.

Recall that PCA is a linear transformation, and $\boldsymbol{u}_1$ acts as a weight vector for mean-shifted feature sequences. If the evolutionary process is indeed tree-like, we can ignore the last pattern and anticipate the dominance of the four progressive patterns over the first regressive pattern. The loss of a feature is expected to be largely compensated by the gain of another feature because every language is expected to have at least one word for a basic vocabulary item. To conclude, a gross violation of the unidirectionality, which we call *jogging*, can be seen as a deviation from the tree model.

Note that the absence of visible violations does not automatically imply the validity of the model for given data. Additionally, in the event that anomalies are detected, there is no feasible way to rescue the tree model. Therefore, the proposed method should primarily serve as a sanity check.

One obvious limitation of the proposed method is that it works on a single sample although Bayesian analysis conventionally draws conclusions by summarizing multiple samples. While applying PCA to multiple trees is possible, visualizing the outcome remains a challenge. If we focus on a specific clade, we can visualize a summary of



(a) No borrowing.

(b) Global borrowing.

(c) Local borrowing (1,000-year limit).

(d) Local borrowing (3,000-year limit).

Figure 3: PCA for the balanced time-tree of data simulation, with four borrowing scenarios.

multiple samples, as we see in Section 5.2.

The proposed method enables the verification of results from published papers. Note that slight modifications to the existing configuration files of Bayesian inference are required. This is necessary because, as mentioned in Section 2.2, the sampler does not track the states of unobserved languages by default. Technical details will be provided in Appendix A.

4. Simulation Experiments

4.1. Data Simulation

We evaluated the proposed method using synthetic data. To generate the data, we partly followed the procedure described by Greenhill et al. (2009). We obtained the same skewed and balanced time-trees (Supplementary Figure A.1) and used the software package TraitLab (Nicholls and Gray, 2006) to simulate evolutionary processes along the branches of each time-tree, with or without borrowing of features between branches.

TraitLab implemented the stochastic Dollo model, which assumes that a feature can only be gained once in history and that once lost in a branch, it is never regained by descendants. This assumption is suitable for simulation of lexical items although it is considered too stringent when fitting real data (Bouckaert and Robbeets, 2017).

TraitLab supported two borrowing scenarios for simulation. One was the global borrowing scenario, enabling borrowings among any contemporary languages, and the other is the local borrowing scenario which allowed borrowings only when the two languages shared a common ancestor within a specified time period. For each time-tree, we tested four scenarios: (1) no borrowing, (2) global borrowing, (3) local borrowing with the 1,000-year limit, and (4) local borrowing with the 3,000-year limit.

Regarding hyperparameters, we configured the loss rate to be 0.2 per 1,000 years and the mean number of traits (features) to be 200. For borrowing scenarios, we set the borrowing rate at 2.241, indicating that as many as 50% of features were

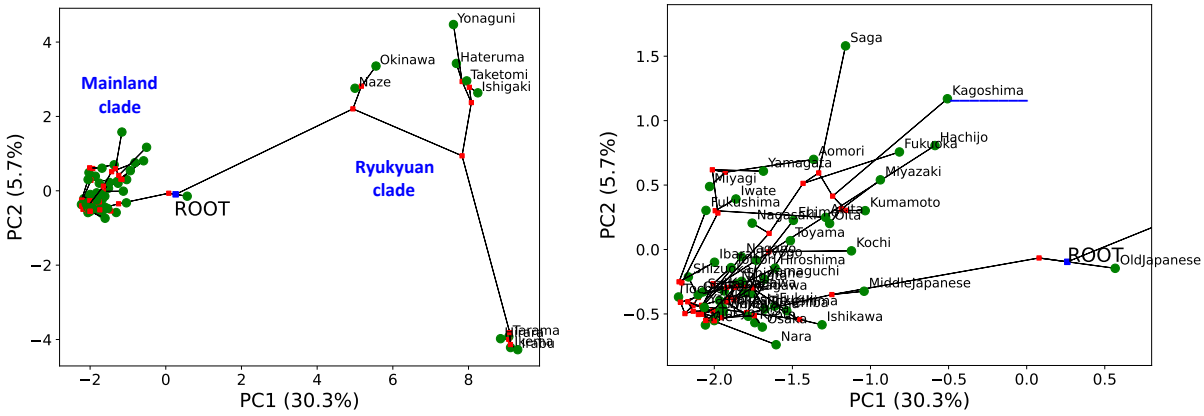


Figure 4: PCA for a Japonic sample tree. Left: The entire tree. Right: Zoomed-in view of the mainland portion. Kagoshima (underlined) is the closest modern mainland dialect to Old Japanese along PC1.

borrowed along an evolutionary path within a span of 1,000 years.

4.2. Phylolinguistic Reconstruction

We used the software package BEAST 2.7.5 (Drummond and Bouckaert, 2015) to reconstruct the evolutionary process from observed languages. For simplicity, we used a Yule tree prior as the time-tree model, a binary continuous-time Markov chain model as the state transition model, and a strict clock as the rate model. Since age calibration was not conducted, we only estimated relative dates. We manually modified auto-generated configuration files to output the node states. We perform MCMC with a total of 10 million steps and applied PCA to the final sample.

4.3. Results

Figures 2 and 3 show PCA projections of the tree samples. Our anticipation was validated by the synthetic data: In the absence of borrowing, the trees maintained near-perfect unidirectionality. In contrast, under the borrowing scenarios, all trees exhibited jogging.

The structural pattern observed under the no-borrowing scenario was better preserved in the balanced tree than in the skewed tree. This was likely due to the direct translation of high-level clades into the first two PCs.

5. Analyzing Real Data

5.1. Japonic

We reviewed an analysis of the Japonic languages by Lee and Hasegawa (2011). Using basic vocabulary data from 59 Japonic dialects, they conducted a phylolinguistic tree reconstruction, with a primary emphasis on determining the root age. They contended that the estimated root age aligned with the putative agricultural population expansion of Japonic speakers.

A peculiarity of their approach was that they analyzed closely-related dialects that were usually considered to be primarily characterized by horizontal transmission (Onishi, 2011). To our knowledge, no one had applied the comparative method of historical-comparative linguistics to analyze their primary source, a dialect dictionary (Hirayama, 1992–1994).[5] Although Lee and Hasegawa (2011) expressed some reservations about the non-tree-like nature of the data, they nonetheless persisted in utilizing the tree model.

Some effort was needed to replicate their analysis because no BEAST configuration file was published. We extracted binary-coded data from a supplementary PDF. We selected the model and hyperparameters based on the description of the paper although we replaced the relaxed clock model with a newer, more efficient one (Douglas et al., 2021). Although several errors had been identified in the data (Pellard, 2021), we only corrected language names. Our MCC tree (Supplementary Figure A.2) suggests that we replicated the original analysis to a large extent.

Figure 4 shows the PCA projection of the final tree sample. The first PC manifested a well-known division between the mainland and Ryukyuan, while also revealing considerable internal diversity within Ryukyuan. When examining the mainland, anomalies were evident. The extensive amount of jogging confirmed the inapplicability of the tree model to this dataset in an intuitive manner. The

[5] A recent phylolinguistic reconstruction of Japonic languages (Igarashi, 2021) is build on top of a careful manual selection of shared innovations, not a quantitative analysis of the entire lexical data.

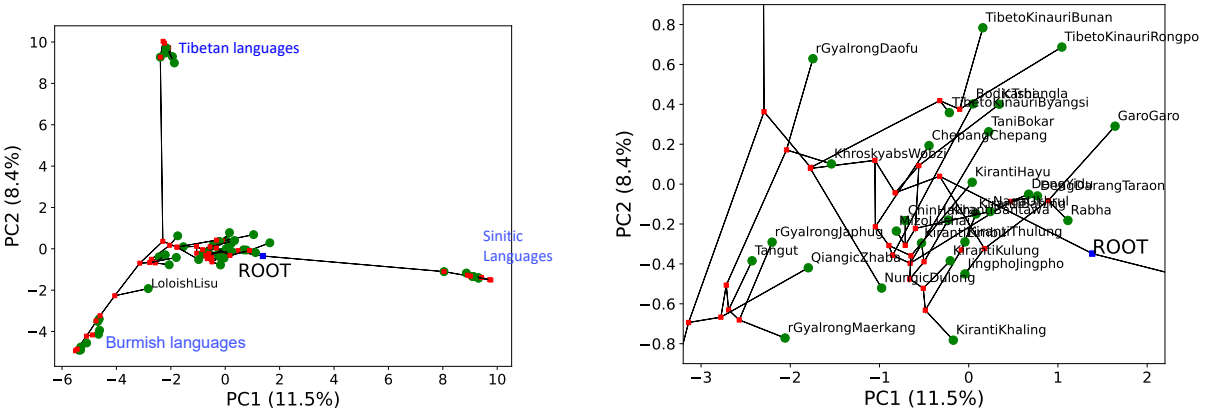


Figure 5: PCA for a Sino-Tibetan sample tree. Left: The entire tree. Right: Zoomed-in view of the central portion.

estimated root age is deemed unreliable because it was derived from the flawed trees.

Kagoshima, located at the southwestern tip of the mainland, exhibited the closest resemblance to Old Japanese along the first PC axis even though it ranked as the second least similar to Old Japanese among the mainland varieties if we switched to similarity based on binary sequences. A plausible explanation of this disparity is that the leftmost area of the figure was characterized by a multitude of overlapping diffusional patterns that covered vast areas but did not consistently reach their peripheries. In other words, Kagoshima underwent a relatively rapid change because it was less affected by dialect leveling, but the features it retained signaled archaism.

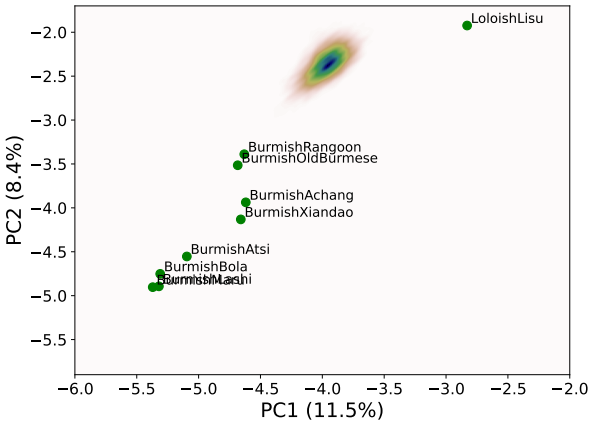


Figure 6: Kernel density estimation of the location of Proto-Lolo-Burmese.

5.2. Sino-Tibetan

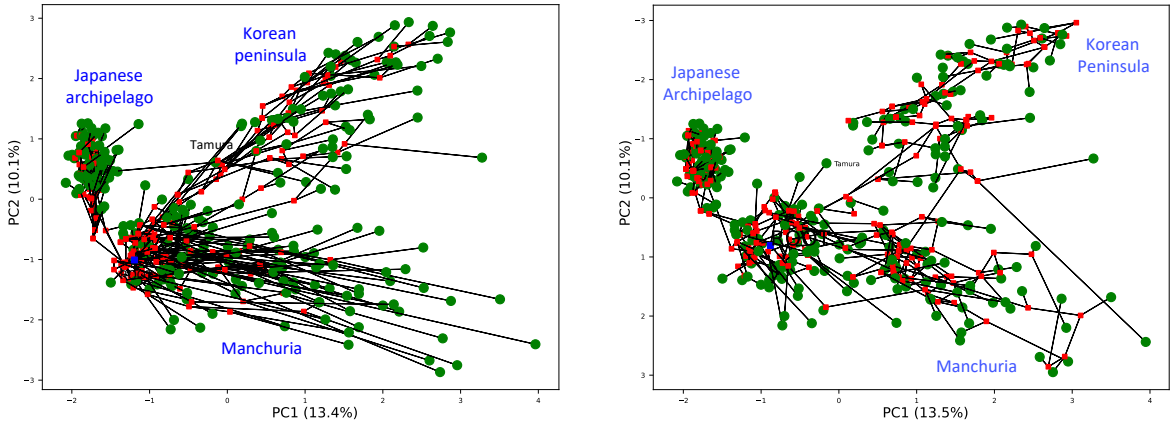
We turned our attention to Sagart et al. (2019), who investigated Sino-Tibetan phylogenies. Also known as Trans-Himalayan, the Sino-Tibetan language family encompasses not just Chinese, Burmese, and Tibetan but also numerous smaller languages found in the mountainous regions of Asia. The high-level structure of Sino-Tibetan, including whether Sinitic represents a primary branch, remains poorly understood. Recent studies have also explored a potential connection between the emergence of Sino-Tibetan branches and the early phases of agriculture in northern China.

Sino-Tibetan is renowned for posing significant challenges in historical-comparative linguistics, with its complex contact history being a key factor (DeLancey, 2021). With the world's leading Sino-Tibetan specialists on their team, Sagart et al. (2019) carefully compiled a lexical database themselves and excluded from their analysis languages known for intense contact such as Bai. Our interest lies in assessing whether their sophisticated methodology effectively addressed the problem of horizontal transmission.

We used a BEAST configuration file[6] published as part of the supplementary materials and slightly modified it to output node states. Our MCC tree (Supplementary Figure A.3) again indicates largely successful replication.

Figure 5 shows our PCA projection of the final tree sample. Within this subspace, three distinct clusters emerge at the extremities, namely Sinitic, Tibetan, and Burmish, all of which had long writing traditions and were oversampled in the dataset. The remaining languages form a clustered group near the root and exhibit noticeable levels of jogging. According to the model, the evolutionary paths from Proto-Sino-Tibetan (the root) toward these languages follow a trajectory that includes Burmish-like intermediate nodes before

[6] `sinotibetan-beast-covarion-relaxed-fbd.xml`



(a) Pseudo Dollo covarion model.

(b) Covarion model.

Figure 7: PCA for the Northeast Asian archaeological sites. (a) The model selected by Robbeets et al. (2021) used the pseudo Dollo covarion model for state transition. (b) The covarion model was used instead. The y-axis is inverted for the sake of facilitating comparison.

moving back in the direction of the root. While Sinitic occupies the opposite end of the axis, the relative positions of these languages do not seem to correlate with their similarity to Sinitic.

We conducted a further analysis of the Loloish language of Lisu, which was located slightly outside the cluttered group. With the posterior probability of nearly 100%, Lisu shared a direct common ancestor (Proto-Lolo-Burmese) with the Burmish languages. We collected the states of Proto-Lolo-Burmese from multiple samples and applied PCA projection. We then performed kernel density estimation to approximate the probability distribution of its location. The result is visualized in Figure 6. The phylolinguistic model demonstrated high confidence in determining the location of Proto-Lolo-Burmese, and thus in the presence of jogging in the evolutionary path to Lisu. Although we cannot conclude that the phylolinguistic reconstruction failed, the presence of anomalies necessitates further investigation.

5.3. Northeast Asian Archaeological Sites

Finally, we examined an analysis of archaeological sites of Northeast Asia by Robbeets et al. (2021), who advocated a version of the Altaic hypothesis under a new brand of Transeurasian. The highly controversial Altaic hypothesis posits a linguistic connection among Turkic, Mongolic, and Tungusic languages, and at times includes Koreanic and Japonic languages within this proposed single language family. It remains a minority view among historical linguists (Janhunen, 2023).

A striking characteristic of Robbeets et al. (2021) was their integration of archaeological, genetic, and linguistic evidence. However, all three types of evidence met biting criticism (Tian et al., 2022). In this paper, we focused on the archaeological data because the apparent lack of tree-like signal was the focal point of criticism (Tian et al., 2022).

We slightly edited a BEAST configuration file[7] published as part of the supplementary materials. It contained 171 binary-coded (presence/absence) typological features of archaeology, such as pottery, horse, and wheat. We replaced coupled MCMC (Müller and Bouckaert, 2019) with vanilla MCMC because the current implementation was incompatible with node state sampling. Comparing our MCC tree (Supplementary Figure A.4) with the published result, we can observe that the two agreed on low-level groupings. There were disagreements on high-level groupings, but they can be explained by their extremely low posterior probabilities. Even if the tree model was applicable, the phylolinguistic model was highly uncertain about the high-level structure of the data.

The PCA projection of the final sample is shown in Figure 7(a). The first PC featured a distinction between Japan (left) and the rest of Northeast Asia (right). Overall, the projected tree revealed a pattern of continual diversification. A notable exception was Tamura, a site on Japan, which was buried in the Asian continent in the subspace despite being clearly descended from a Japanese parent. This can be interpreted as hybridization, a violation of the tree model. The MCC tree alone shows no sign of such a deviation.

The scarcity of jogging raises suspicion, as the data was perceived as markedly non-tree-like (Tian et al., 2022). We argue that this stemmed from inappropriate model selection. The model selected by Robbeets et al. (2021) used a pseudo

[7] `pdcov-ucln-bsp-tips.xml`

Dollo model (Bouckaert and Robbeets, 2017) for state transition. This model loosely adheres to the Dollo principle, which suggests that a feature can be gained only once in a tree but lost multiple times. Because a naïve implementation of this principle is highly sensitive to borrowings, the pseudo Dollo model permits multiple gains of a feature in a tree while it still restricts languages from reacquiring a feature that their ancestor had lost. Robbeets et al. (2021) combined the pseudo Dollo model with the covarion model, which is widely used to capture fast and slow phases of evolution (Tuffley and Steel, 1998).

The pseudo Dollo covarion model yields the complete absence of the $1 \rightarrow 0 \rightarrow 1$ pattern, which strongly promotes unidirectionality. For comparison, we applied the PCA projection to the simple covarion model, based on the configuration file included in their supplementary materials.[8] As expected, this model choice resulted in a substantial quantity of jogging (Figure 7(b)).

While the the arbitrariness of meaning-symbol connection provides a rational basis for applying the Dollo principle to cognates, it is entirely plausible that a typological feature could potentially be reacquired. Although Robbeets et al. (2021) justified their model choice based on its superior fit to the data, our analysis suggests that the apparent lack of jogging was an artifact of the inappropriate model selection.

6. Conclusions

In this paper, we have introduced a method for projecting a tree sample using principal component analysis in order to identify anomalies in Bayesian phylolinguistic reconstruction. A departure from the tree model can be observed as a deviation along the first principal component axis, which we refer to as jogging.

The proposed method is strikingly simple and can be applied to a wide range of published data. Our primary focus is on binary-coded lexical data, as their meaning-symbol connection inherently enforces a unidirectional pattern under the tree model. Conducting a more comprehensive analysis of our approach's effectiveness on different data types would be a valuable avenue for further research.

7. Acknowledgments

We express our gratitude to Simon J. Greenhill for generously providing the code and data necessary for replicating the findings in Greenhill et al. (2009).

This work received partial support from JSPS KAKENHI Grant Numbers 21K12029 and 18KK0012.

[8] `cov-strict-bsp.xml`

8. Limitations

We investigated a critical assumption inherent in Bayesian phylolinguistic models, which is frequently violated in real-world scenarios. Our method aims to visualize deviations from the tree model, yet it is important to note that the absence of apparent violations does not guarantee the model's validity for the given data. Furthermore, in cases where anomalies are detected, there is no feasible way to rescue the tree model.

Principal component analysis is a parameter-free technique that operates under minimal assumptions. Nonetheless, it can be susceptible to bias when confronted with an overrepresentation of one or more clades within the dataset, potentially resulting in a skewed data representation.

9. Bibliographical References

Sandra Auderset, Simon J Greenhill, Christian T DiCanio, and Eric W Campbell. 2023. Subgrouping in a 'dialect continuum': A Bayesian phylogenetic analysis of the Mixtecan language family. *Journal of Language Evolution*, 8(1):33–63.

François Barbançon, Steven N. Evans, Luay Nakhleh, Don Ringe, and Tandy Warnow. 2013. An experimental study comparing linguistic phylogenetic reconstruction methods. *Diachronica*, 30(2):143–170.

Knut Bergsland and Hans Vogt. 1962. On the validity of glottochronology. *Current Anthropology*, 3(2):115–153.

Remco Bouckaert, Philippe Lemey, Michael Dunn, Simon J. Greenhill, Alexander V. Alekseyenko, Alexei J. Drummond, Russell D. Gray, Marc A. Suchard, and Quentin D. Atkinson. 2012. Mapping the origins and expansion of the Indo-European language family. *Science*, 337(6097):957–960.

Remco Bouckaert and Martine Robbeets. 2017. Pseudo Dollo models for the evolution of binary characters along a tree. *bioRxiv*.

David Bryant and Vincent Moulton. 2004. Neighbor-Net: An agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution*, 21(2):255–265.

Lyle Campbell and William J. Poser. 2008. *Language Classification*. Cambridge University Press.

Will Chang, Chundra Cathcart, David Hall, and Andrew Garrett. 2015. Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language*, 91(1):194–244.

Scott DeLancey. 2021. Classifying Trans-Himalayan (Sino-Tibetan) languages. In Paul Sidwell and Mathias Jenny, editors, *The Languages and Linguistics of Mainland Southeast Asia: A Comprehensive Guide*, pages 207–223. De Gruyter Mouton.

Jordan Douglas, Rong Zhang, and Remco Bouckaert. 2021. Adaptive dating and fast proposals: Revisiting the phylogenetic relaxed clock model. *PLOS Computational Biology*, 17(2):1–30.

Alexei J. Drummond and Remco R. Bouckaert. 2015. *Bayesian Evolutionary Analysis with BEAST*. Cambridge University Press.

Eran Elhaik. 2022. Principal component analyses (pca)-based findings in population genetic studies are highly biased and must be reevaluated. *Scientific Reports*, 12(1):14683.

Joseph Felsenstein. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376.

Joseph Felsenstein. 2004. *Inferring Phylogenies*. Sinauer Associates.

Kevin J. Galinsky, Gaurav Bhatia, Po-Ru Loh, Stoyan Georgiev, Sayan Mukherjee, Nick J. Patterson, and Alkes L. Price. 2016. Fast principal-component analysis reveals convergent evolution of ADH1B in Europe and East Asia. *The American Journal of Human Genetics*, 98(3):456–472.

Russell D. Gray and Quentin D. Atkinson. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, 426(6965):435–439.

Russell D. Gray, David Bryant, and Simon J. Greenhill. 2010. On the shape and fabric of human history. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1559):3923–3933.

Russell D. Gray and Fiona M. Jordan. 2000. Language trees support the express-train sequence of Austronesian expansion. *Nature*, 405(6790):1052–1055.

Simon J. Greenhill, Thomas E. Currie, and Russell D. Gray. 2009. Does horizontal transmission invalidate cultural phylogenies? *Proceedings of the Royal Society B: Biological Sciences*, 276(1665):2299–2306.

Simon J. Greenhill, Chieh-Hsi Wu, Xia Hua, Michael Dunn, Stephen C. Levinson, and Russell D. Gray. 2017. Evolutionary dynamics of language systems. *Proceedings of the National Academy of Sciences*, 114(42):E8822–E8829.

B. R. Holland, K. T. Huber, A. Dress, and V. Moulton. 2002. δ plots: A tool for analyzing phylogenetic distance data. *Molecular Biology and Evolution*, 19(12):2051–2059.

Yosuke Igarashi. 2021. Bunkigaku teki shuhō ni motozuita nichiryū shogo no keitō bunrui no kokoromi. In Yuka Hayashi, Tomohide Kinuhata, and Nobuko Kibe, editors, *Fīrudo to bunken kara miru Nichiryū shogo no keitō to rekishi*, pages 17–51. Kaitakusha. (in Japanese).

Juha A. Janhunen. 2023. The unity and diversity of Altaic. *Annual Review of Linguistics*, 9(1):135–154.

Siva Kalyan and Alexandre François. 2018. Freeing the comparative method from the tree model: A framework for historical glottometry. *Senri Ethnological Studies*, 98:59–89.

Luke J. Kelly and Geoff K. Nicholls. 2017. Lateral transfer in stochastic Dollo models. *The Annals of Applied Statistics*, 11(2):1146–1168.

Sean Lee and Toshikazu Hasegawa. 2011. Bayesian phylogenetic analysis supports an agricultural origin of Japonic languages. *Proceedings of the Royal Society B*, 278(1725):3662–3669.

Andrew Meade and Mark Pagel. 2022. Ancestral state reconstruction using BayesTraits. In Haiwei Luo, editor, *Environmental Microbial Evolution: Methods and Protocols*, pages 255–266. Springer US, New York, NY.

Paolo Menozzi, Alberto Piazza, and Luigi Cavalli-Sforza. 1978. Synthetic maps of human gene frequencies in Europeans. *Science*, 201(4358):786–792.

Nicola F. Müller and Remco Bouckaert. 2019. Coupled MCMC in BEAST 2. *bioRxiv*.

Yugo Murawaki. 2015. Spatial structure of evolutionary models of dialects in contact. *PLoS ONE*, 10(7):1–15.

Luay Nakhleh, Donald A. Ringe, and Tandy Warnow. 2005. Perfect phylogenetic networks: A new methodology for reconstructing the evolutionary history of natural languages. *Language*, 81(2):382–420.

Shijulal Nelson-Sathi, Johann-Mattis List, Hans Geisler, Heiner Fangerau, Russell D. Gray, William Martin, and Tal Dagan. 2011. Networks uncover hidden lexical borrowing in Indo-European language evolution. *Proceedings of the Royal Society B: Biological Sciences*, 278:1794–1803.

Nico Neureiter, Peter Ranacherand, Nour Efrat-Kowalsky, Gereon A. Kaiping, Robert Weibel, Paul Widmer, and Remco R. Bouckaert. 2022. Detecting contact in language trees: a Bayesian phylogenetic model with horizontal transfer. *Humanities and Social Sciences Communications*, 9(205).

Geoff K. Nicholls and Russell D. Gray. 2006. Quantifying uncertainty in a stochastic Dollo model of vocabulary evolution. In Peter Forster and Colin Renfrew, editors, *Phylogenetic Methods and the Prehistory of Languages*, pages 161–171. McDonald Institute for Archaeological Research.

Takuichiro Onishi. 2011. Analyzing dialectological distributions of Japanese. *Dialectologia: Revista Electrònica*, pages 123–135.

Mark Pagel, Quentin D. Atkinson, Andreea S. Calude, and Andrew Meade. 2013. Ultraconserved words point to deep language ancestry across Eurasia. *Proceedings of the National Academy of Sciences*, 110(21):8471–8476.

Nick Patterson, Alkes L. Price, and David Reich. 2006. Population structure and eigenanalysis. *PLoS Genetics*, 2(12):e190.

Thomas Pellard. 2021. Nichiryū shogo no keitō bunrui to bunki ni tsuite. In Yuka Hayashi, Tomohide Kinuhata, and Nobuko Kibe, editors, *Fīrudo to bunken kara miru Nichiryū shogo no keitō to rekishi*, pages 2–16. Kaitakusha. (in Japanese).

Taraka Rama and Søren Wichmann. 2018. Towards identifying the optimal datasize for lexically-based Bayesian inference of linguistic phylogenies. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1578–1590, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Martine Robbeets, Remco Bouckaert, Matthew Conte, Alexander Savelyev, Tao Li, Deog-Im An, Ken ichi Shinoda, Yinqiu Cui, Takamune Kawashima, Geonyoung Kim, , Junzo Uchiyama, Joanna Dolińska, Sofia Oskolskaya, Ken-Yōjiro Yamano, Noriko Seguchi, Hirotaka Tomita, Hiroto Takamiya, Hideaki Kanzawa-Kiriyama, Hiroki Oota, Hajime Ishida, Ryosuke Kimura, Takehiro Sato, Jae-Hyun Kim, Bingcong Deng, Rasmus Bjørn, Seongha Rhee, Kyou-Dong Ahn, Ilya Gruntov, Olga Mazo, John R. Bentley, Ricardo Fernandes, Patrick Roberts, Ilona R. Bausch, Linda Gilaizeau, Minoru Yoneda, Mitsugu Kugai, Raffaela A. Bianco, Fan Zhang, Marie Himmel, Mark J. Hudson, and Chao Ning. 2021. Triangulation supports agricultural spread of the Transeurasian languages. *Nature*, 599(7886):616–621.

Laurent Sagart, 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. *Proceedings of the National Academy of Sciences*, 116(21):10317–10322.

August Schleicher. 1853. Die ersten Spaltungen des indogermanischen Urvolkes. *Allgemeine Monatsschrift für Wissenschaft und Literatur*, 3:786–787. (in German).

Johannes Schmidt. 1872. *Die Verwandtschaftsverhältnisse der indogermanischen Sprachen*. Hermann Böhlau. (in German).

Steven L Scott. 2002. Bayesian methods for hidden markov models. *Journal of the American Statistical Association*, 97(457):337–351.

Morris Swadesh. 1952. Lexicostatistic dating of prehistoric ethnic contacts. *Proceedings of American Philosophical Society*, 96:452–463.

Zheng Tian, Yuxin Tao, Kongyang Zhu, Guillaume Jacques, Robin J. Ryder, José Andrés Alonso de la Fuente, Anton Antonov, Ziyang Xia, Yuxuan Zhang, Xiaoyan Ji, Xiaoying Ren, Guanglin He, Jianxin Guo, Rui Wang, Xiaomin Yang, Jing Zhao, Dan Xu, Russell D. Gray, Menghan Zhang, Shaoqing Wen, Chuan-Chao Wang, and Thomas Pellard. 2022. Triangulation fails when neither linguistic, genetic, nor archaeological data support the Transeurasian narrative. *bioRxiv*.

Chris Tuffley and Mike Steel. 1998. Modeling the covarion hypothesis of nucleotide substitution. *Mathematical Biosciences*, 147(1):63–91.

10. Language Resource References

Remco R. Bouckaert. 2010. DensiTree: making sense of sets of phylogenetic trees. *Bioinformatics*, 26(10):1372–1373.

Teruo Hirayama. 1992–1994. *Gendai Nihongo Hōgen Daijiten*. Meiji Shoin. (in Japanese).

A. Implementation Notes

To sample node states in the software package BEAST, we usually need to modify the existing configuration file. Specifically, we need to replace `TreeWithMetaDataLogger` with `AncestralSequenceLogger`. The "logger" does not just write logs but samples node states. The node states are output as node annotations in the NEXUS format. The logger requires the `tag` attribute specifying the key for NEXUS node annotations, the `data` attribute specifying the alignment data, the `siteModel` attribute specifying the site model, and the `branchRateModel` attribute specifying the branch rate model.

`AncestralSequenceLogger` is old and is included in the `beast-classic` package. It might not be compatible with newer modules.

Several recent studies define multiple site models to account for varying rates associated with basic vocabulary items. In such instances, a straightforward solution is to define a logger for each site model. Consequently, multiple copies of the same tree are generated, each providing distinct information about the node states. A postprocessing step is necessary to merge them into a single tree with complete node states.

Node state sampling can also be accomplished using the commonly utilized software BayesTraits (Meade and Pagel, 2022), which effectively models state transitions for a given tree or set of tree samples. While theoretically feasible to apply our method to analyses based on BayesTraits, it is important to acknowledge that our approach necessitates multiple features, whereas BayesTraits is often employed to analyze a singular feature.

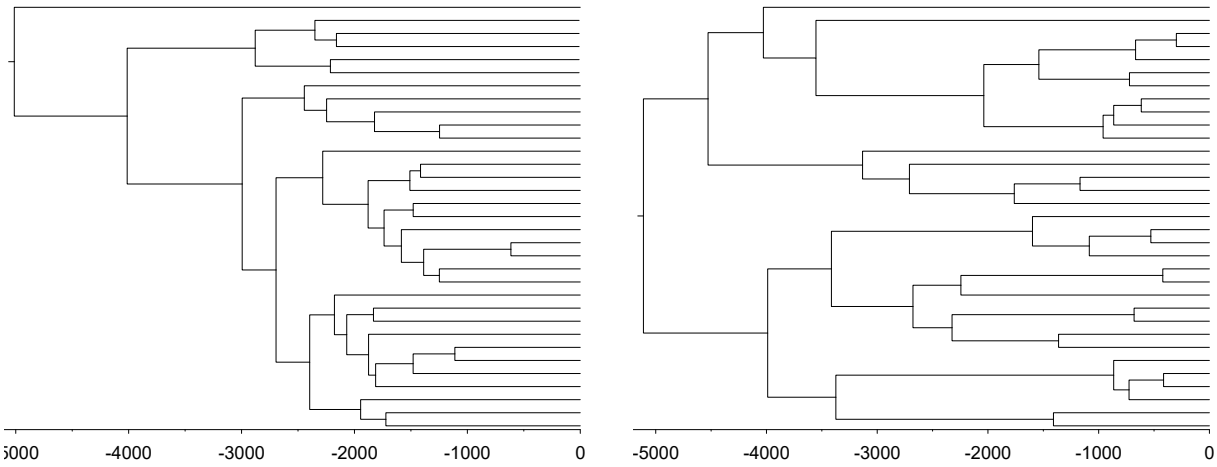


Figure A.1: Two time-trees used for data simulation by Greenhill et al. (2009). One is skewed while the other is balanced. The horizontal axis represents the passage of time, measured in years.

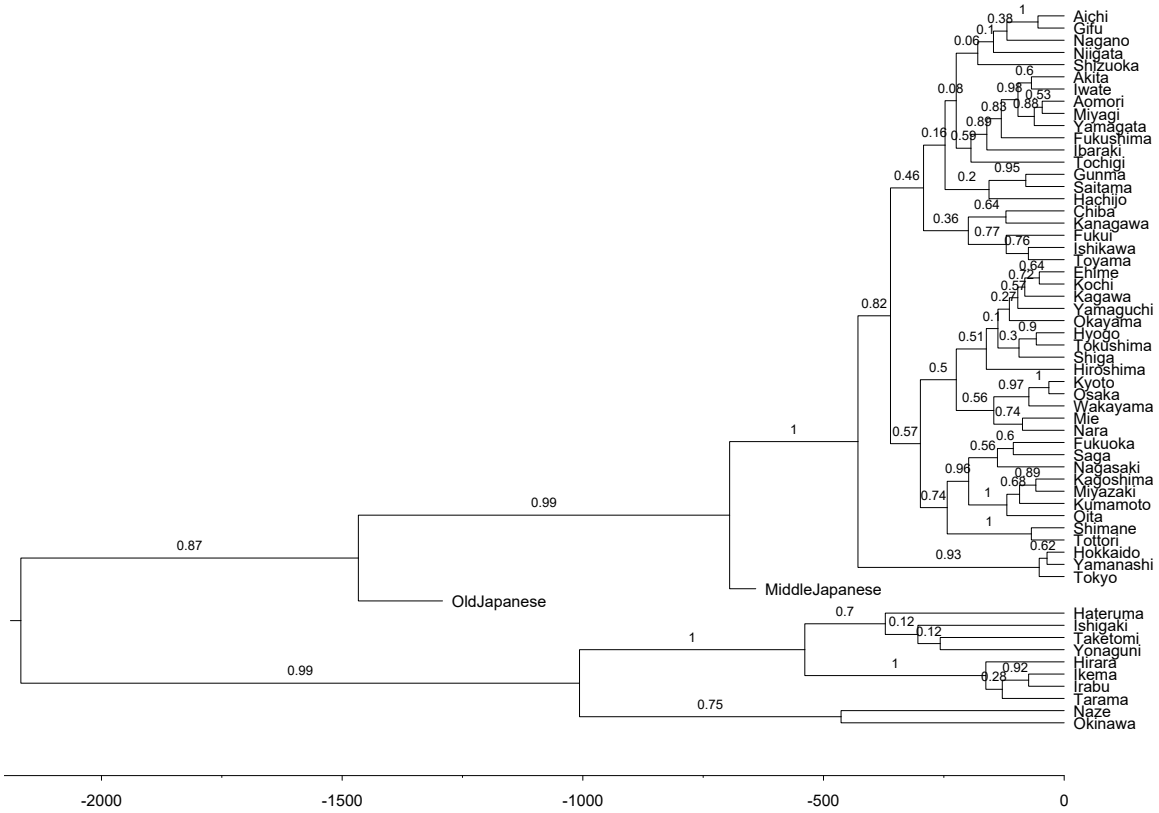


Figure A.2: The maximum clade credibility tree of the Japonic languages. A number positioned above a branch indicates the posterior probability of the corresponding clade.

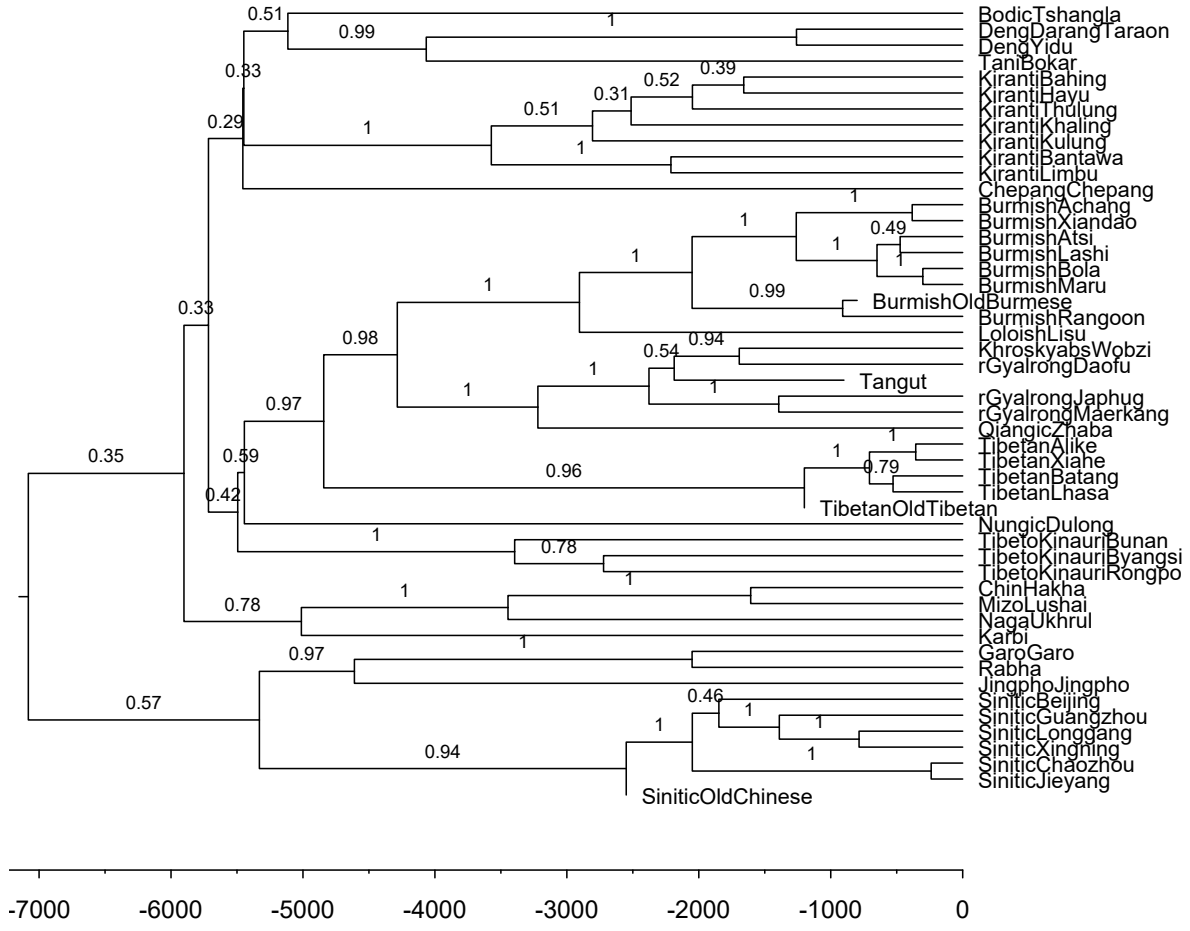


Figure A.3: The maximum clade credibility tree of the Sino-Tibetan languages. A number positioned above a branch indicates the posterior probability of the corresponding clade.

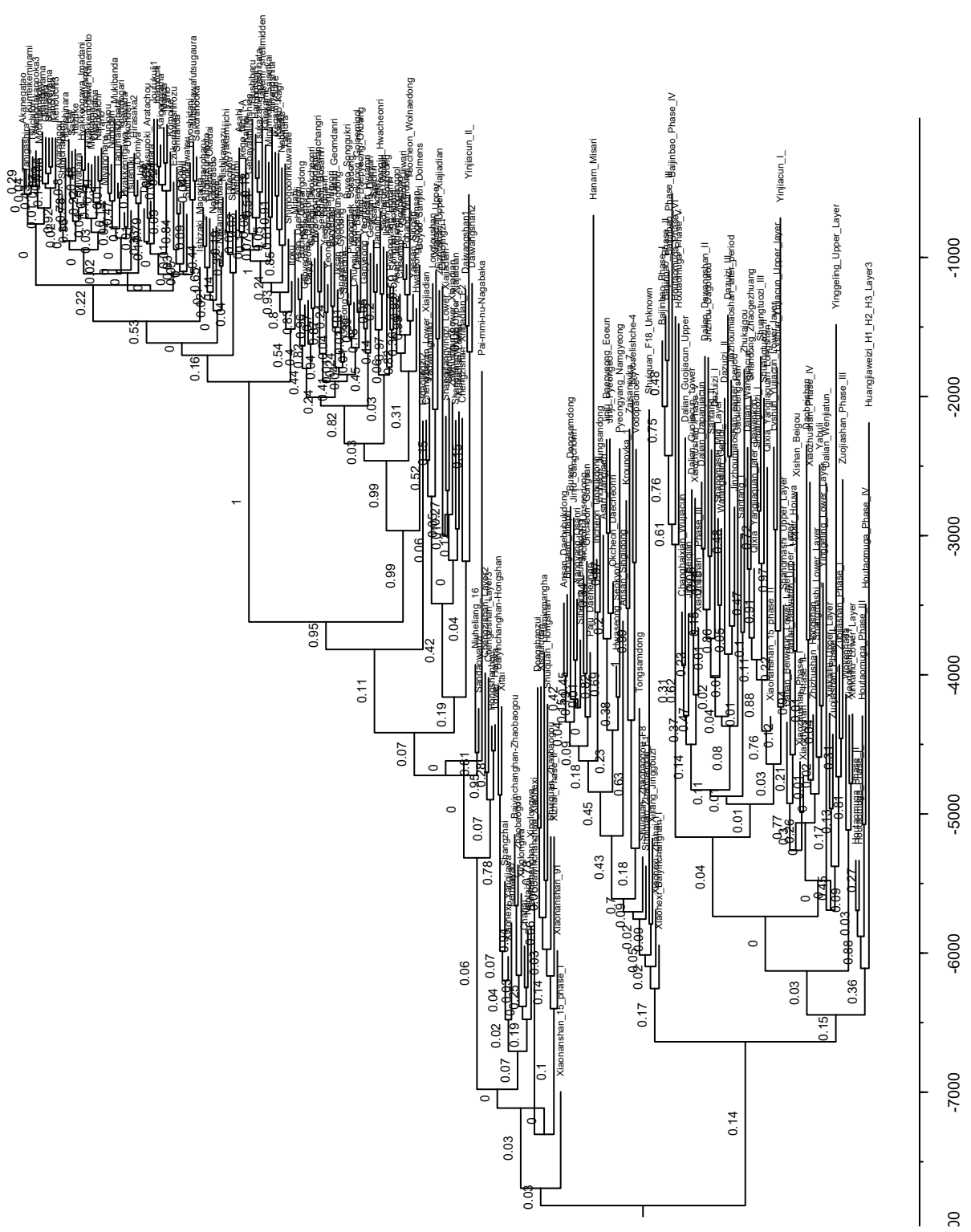


Figure A.4: The maximum clade credibility tree of the Northeast Asian archaeological sites. A number positioned above a branch indicates the posterior probability of the corresponding clade.