Dating ancient humans splits by estimating Poisson rates from mitochondrial DNA parity samples

Keren Levinstein Hallak
Department of Statistics, University of Tel Aviv
and
Saharon Rosset
Department of Statistics, University of Tel Aviv

April 20, 2023

Abstract

We tackle the problem of estimating species divergence times, given a genome sequence from each species and a large known phylogenetic tree with a known structure (typically from one of the species). The number of transitions at each site from the first sequence to the other is assumed to be Poisson distributed, and only the parity of the number of transitions is observed. The detailed phylogenetic tree contains information about the transition rates in each site. We use this formulation to develop and analyze multiple estimators of the divergence between the species. To test our methods, we use mtDNA substitution statistics from the well-established Phylotree as a baseline for data simulation such that the substitution rate per site mimics the real-world observed rates. We evaluate our methods using simulated data and compare them to the Bayesian optimizing software BEAST2, showing that our proposed estimators are accurate for a wide range of divergence times and significantly outperform BEAST2. We then apply the proposed estimators on Neanderthal, Denisovan, and Chimpanzee mtDNA genomes to better estimate their TMRCA (Time to Most Recent Common Ancestor) with modern humans and find that their TMRCA is substantially later, compared to values cited recently in the literature.

1 Introduction

Dating species divergence has been studied extensively for the last few decades using approaches based on genetics, archaeological findings, and radiocarbon dating [8, 32]. Finding accurate timing is crucial in analyzing morphological and molecular changes in the DNA, in demographic research, and in dating key fossils. One approach for estimating the divergence times is based on the molecular clock hypothesis [38, 37] which states that the rate of evolutionary change of any specified protein is approximately constant over time and different lineages. Subsequently, statistical inference can be applied to a given phylogenetic tree to infer the dating of each node up to calibration.

Our work focuses on this estimation problem and proposes new statistical methods to date the TMRCA of two species given a detailed phylogenetic tree for one of the species with the same transition rates per site. We formulate the problem by modeling the number of transitions $(A \leftrightarrow G, C \leftrightarrow T)$ in each site using a Poisson process with a different rate per site; sites containing transversions are neglected due to their sparsity (indeed, we include sparse transversions in the simulations and show that our methods are robust to their occurrences). The phylogenetic tree is used for estimating the transition rates per site. Hence, our problem reduces to two binary sequences where the parity of the number of transitions of each site is the relevant statistic from which we can infer the time difference between them.

We can roughly divide the approaches to solve this problem into two. The **frequentist** approach seeks to maximize the likelihood of the observed data. Most notable is the PAML [36] package of programs for phylogenetic analyses of DNA and the MEGA software [18]. Alternatively, the **Bayesian** approach considers a prior of all the problem's parameters and maximizes the posterior distribution of the observations. Leading representatives of the Bayesian approach are BEAST2 [5] and MrBayes [27], which are publicly available programs for Bayesian inference and model choice across a wide range of phylogenetic and evolutionary models.

In this work, we developed several distinct estimators from frequentist and Bayesian approaches to find the divergence time directly. The proposed estimators differ in their assumptions on the generated data, the approximations they make, and their numerical stability. We explain each estimator in detail and discuss its properties.

A critical difference between our proposed solutions and existing methods is that we seek to estimate only one specific problem parameter. At the same time, software packages such as BEAST2 and PAML optimize over a broad set of unknown parameters averaging the error on all of them (the tree structure, the timing of every node, the per-site substitution rates, etc.). Subsequently, the resources they require for finding a locally optimal instantiation of the tree and dating all its nodes can be very high in terms of memory and computational complexity. Consequently, the amount of sequences they can consider simultaneously is highly limited. Thus, unlike previous solutions, we utilize transition statistics from all available sequences, in the form of a previously built phylogenetic tree.

We develop a novel approach to simulate realistic data to test our proposed solutions. To do so, we employ Phylotree [33] – a complete, highly detailed, constantly updated reconstruction of the human mitochondrial DNA phylogenetic tree. We sample transitions of similar statistics to Phylotree and use it to simulate a sequence at a predefined trajectory from Phylotree's root.

We then empirically test the different estimators on simulated data and compare our results to the BEAST2 software. Our proposed estimators are calculated substantially faster while utilizing the transitions statistics from all available sequences (Phylotree considers 24,275 sequences), unlike BEAST2 which can consider only dozens of sequences due to its complexity. Comparing with the ground truth, we show that BEAST2 overestimates the divergence time for low TMRCA values (e.g. human-Neanderthals and human-Denisovan), but performs an underestimation for larger divergence times (e.g. human-Chimpanzee), while our estimates provide more accurate results. Finally, we use our estimators to date the TMRCA (given in kya – kilo-years ago) of modern humans with Neanderthals, Denisovan and Chimpanzee based on their mtDNA. Surprisingly, the divergence times we find (human-Neanderthals ∼408 kya, human-Denisovans ∼824 kya, human-Chimpanzee ∼5,009 kya) – are considerably later than those accepted today.

2 Materials and Methods

2.1 Estimation methods

First, we describe an idealized reduced mathematical formulation for estimating divergence times and our proposed solutions. In Section 2.2, we describe the reduction process in greater detail.

Consider the following scenario: we have a set of n Poisson rates, denoted as $\{\lambda_i\}_{i=1}^{n}$ where $n \in \mathbb{N}$. Let $\vec{X}$ be a vector of length n such that each element X_i is independently distributed as $\mathrm{Pois}(\lambda_i)$. Similarly, let $\vec{Y}$ be a vector of length n such that each element Y_i is independently distributed as $\mathrm{Pois}(\lambda_i \cdot p)$ for a fixed unknown p. We denote $\vec{Z}$ as the coordinate-wise parity of $\vec{Y}$, meaning that $Z_i = 1$ if Y_i is odd and $Z_i = 0$ otherwise. **Our goal is to estimate p given $\vec{X}$ and $\vec{Z}$.**

Remark 1: Note that the number of *unknown* Poisson rate parameters n in the problem $\{\lambda_i\}_{i=1}^n$ grows with the number of observations $\{(X_i, Z_i)\}_{i=1}^n$. However, our focus is solely on estimating p, so additional observations do provide more information.

Remark 2: The larger the value of $p \cdot \lambda_i$, the less information on p is provided in Z_i as it approaches a Bernoulli distribution with a probability of 0.5. On the other hand, the smaller λ_i is, the harder it will be to infer λ_i from X_i. As a result, the problem of estimating p should be easier in settings where λ_i is high and p is low.

2.1.1 Preliminaries

First, we derive the distribution of Z_i; All proofs are provided in the Supplementary material (Section 1).

Lemma 1. *Let $Y \sim Pois(\Lambda)$ and Z be the parity of Y. Then $Z \sim Ber(\frac{1}{2}(1 - e^{-2\Lambda}))$.*

We use this result to calculate the likelihood and log-likelihood of p and $\vec{\lambda}$ given $\vec{X}$ and $\vec{Z}$. The likelihood is given by:

$$L\left(\vec{X}, \vec{Z}; p, \vec{\lambda}\right) = \prod_{i=1}^{n} e^{-\lambda_i} \frac{{\lambda_i}^{X_i}}{X_i!} \frac{1}{2} \left(1 + (-1)^{Z_i} e^{-2\lambda_i p}\right), \tag{1}$$

and the log-likelihood is:

$$l\left(\vec{X}, \vec{Z}; p, \vec{\lambda}\right) = \sum_{i=1}^{n} \left[-\lambda_i + X_i \log \lambda_i + \log\left(1 + (-1)^{Z_i} e^{-2\lambda_i p}\right)\right] + Const. \tag{2}$$

This result follows immediately from the independence of each coordinate.

2.1.2 Cramer-Rao bound

We begin our analysis by computing the Cramer-Rao bound (CRB; [7, 25]). In Section 3.1, we compare the CRB to the error of the estimators.

Theorem 1. *Denote the Fisher information matrix for the estimation problem above by $I \in \mathbb{R}^{(n+1,n+1)}$, where the first n indexes correspond to $\{\lambda_i\}_{i=1}^n$ and the last index $(n+1)$ corresponds to p. For clarity denote $I_{p,p} \doteq I_{n+1,n+1}, I_{i,p} \doteq I_{i,n+1}, I_{p,i} \doteq I_{n+1,i}$. Then:*

$$\forall i \neq j,\ 1 \leq i,j \leq n: \quad I_{i,j} = 0, \quad I_{i,i} = \frac{1}{\lambda_i} + \frac{4p^2}{e^{4\lambda_i p} - 1}, \quad I_{i,p} = I_{p,i} = \frac{4p\lambda_i}{e^{4\lambda_i p} - 1},$$
$$I_{p,p} = 4\sum_{i=1}^{n} \frac{\lambda_i^2}{e^{4\lambda_i p} - 1}. \tag{3}$$

Consequently, an unbiased estimator $\hat{p}$ holds:

$$\mathbb{E}\left[(p-\hat{p})^2\right] \geq \left[4\sum_{i=1}^{n} \frac{\lambda_i^2}{e^{4\lambda_i p} - 1 + 4p^2\lambda_i}\right]^{-1}. \tag{4}$$

If $\forall i = 1..n : \lambda_i = \lambda$, we can further simplify the expression:

$$\mathbb{E}\left[(p-\hat{p})^2\right] \geq \frac{e^{4\lambda p} - 1 + 4p^2\lambda}{4n\lambda^2}. \tag{5}$$

The CRB, despite its known looseness in many problems, provides insights into the sensitivity of the error to each parameter. This expression supports our previous observation that the error of an unbiased estimator increases exponentially with $\min_i\{\lambda_i \cdot p\}$. However, for constant $\lambda_i \cdot p$, the error improves for higher values of λ_i. We now proceed to describe and analyze several estimators for p.

2.1.3 Method 1 - Maximum Likelihood Estimator

Proposition 1. *Following equation 1, the maximum likelihood estimators $\hat{p}, \hat{\lambda}_i$ hold:*

$$\sum_{i=1}^{n} \hat{\lambda}_i = \sum_{i=1}^{n} X_i, \quad X_i = \hat{\lambda}_i + \frac{2\hat{p}\hat{\lambda}_i}{(-1)^{Z_i} e^{2\hat{\lambda}_i \hat{p}} + 1}, \quad \sum_{i=1}^{n} \frac{\hat{\lambda}_i}{(-1)^{Z_i} e^{2\hat{\lambda}_i \hat{p}} + 1} = 0. \tag{6}$$

Proposition 1 provides n separable equations for maximum likelihood estimation (MLE). Our first estimator sweeps over values of $\hat{p}$ (grid searching in a relevant area) and then for each $i = 1..n$ finds the optimal $\hat{\lambda}_i$ numerically. The solution is then selected by choosing the pair $(\hat{p}, \{\hat{\lambda}_i\}_{i=1}^{n})$ that maximizes the log-likelihood calculated using equation 2.

The obtained MLE equations are solvable, yet, finding the MLE still requires solving n numerical equations, which might be time-consuming. More importantly, MLE estimation is statistically problematic when the number of parameters is of the same order as the number of observations [6]. Subsequently, we propose alternative methods that may yield better practical results.

2.1.4 Method 2 - λ_i-conditional estimation

We propose a simple estimate of $\vec{\lambda}$ based solely on X_i, followed by an estimate of p as if $\vec{\lambda}$ is known, considering only $\vec{Z}$. This method is expected to perform well when λ_i values are large, as in these cases, X_i conveys more information about λ_i than Z_i. This approach enables us to avoid estimating both $\vec{\lambda}$ and p simultaneously, leading to a simpler numerical solution.

When $p \leq 1$, we can mimic Y_i's distribution as a sub-sample from X_i, i.e. we assume that $Y_i|X_i \sim Bin(n = X_i, p)$. Then, we find the maximum likelihood estimate of p:

Proposition 2. *If $Y_i|X_i \sim Bin(X_i, p)$, then:*

1. *$Y_i \sim Pois(\lambda_i \cdot p)$, which justifies this approach.*
2. *$Z_i|X_i \sim Ber\left(\frac{1}{2}\left(1 - (1-2p)^{X_i}\right)\right)$, so we can compute the likelihood of p without considering λ_i.*

3. The maximum likelihood estimate of p given $\sum\limits_{i=1}^{n} Z_i$ holds:

$$\sum_{i=1}^{n} (1-2\hat{p})^{X_i} = n - 2\sum_{i=1}^{n} Z_i \tag{7}$$

Remark: We use the maximum likelihood estimation of p given $\sum_{i=1}^{n} Z_i$ by applying Le-Cam's theorem [20]. This eliminates the need for a heuristic solution of the pathological case $X_i = 0, Z_i = 1$.

2.1.5 Method 3 - Gamma distributed Poisson rates

The Bayesian statistics approach incorporates prior assumptions about the parameters. A common prior for the rate parameters $\vec{\lambda}$ is the Gamma distribution, which is used in popular Bayesian divergence time estimation programs such as MCMCtree [36], BEAST2 [5], and MrBayes [27]. Specifically, we have $\lambda_i \sim \Gamma(\alpha, \beta)$, and for p, we use a uniform prior over the positive real line.

Proposition 3. *Let $\lambda_i \sim \Gamma(\alpha, \beta)$, then the maximum a posteriori estimator of p holds:*

$$\frac{\partial l}{\partial p} = \sum_{i=1}^{n} \frac{X_i + \alpha}{(-1)^{Z_i}\left(1 + \frac{2p}{\beta+1}\right)^{X_i+\alpha} + 1} = 0 \tag{8}$$

Subsequently, given estimated values for α and β, we can be find an estimator for p numerically to hold Equation 8. Unfortunately, the derivative with respect to α does not have a closed-form expression, nor is it possible to waive the dependence on $\vec{Z}, p$. Hence, we suggest using Negative-Binomial regression [12] to estimate α and β given $\vec{X}$.

2.2 Estimating ancient divergence times using a large modern phylogeny

In this section, we apply the methods described in Section 2.1 to estimate the non-calibrated divergence times between humans and their closest relatives by comparing mitochondrial DNA (mtDNA) sequences. Our approach assumes the following assumptions:

1. Molecular clock assumption - the rate of accumulation of transitions (base changes) over time and across different lineages is constant, as first proposed by Zuckerkandl and Pauling [38] and widely used since.

2. Poisson distribution - The number of transitions along the human and human's closest relatives mtDNA lineages follows a Poisson distribution with site-dependent rate parameter λ_i per time unit.

3. No transversions - We only consider sites with no transversions and assume a constant transition rate per site ($\lambda_{i,A\to G} = \lambda_{i,G\to A}$, or $\lambda_{i,T\to C} = \lambda_{i,C\to T}$).

4. Independence of sites - The number of transitions at each site is independent of those at other sites.

5. Phylogenetic tree - The phylogenetic tree presented in the Phylotree database includes all transitions and transversions that occurred along the described lineages.

As the Phylotree database is based on tens of thousands of sequences, the branches in the tree correspond to relatively short time intervals, making multiple mutations per site unlikely in each branch [29]. However, when considering the mtDNA sequence of other species, the branches in the tree correspond to much longer time intervals, meaning that many underlying transitions are unobserved. For instance, when comparing two human sequences that differ in a specific site, Phylotree can determine whether the trajectory between the sequences was $A \rightarrow G$, $A \rightarrow G \rightarrow A \rightarrow G$, or $A \rightarrow T \rightarrow G$. However, when comparing sequences of ancient species, an elaborate phylogenetic tree like Phylotree is not available, making it impossible to discriminate between these different trajectories.

We use the following notation:

1. Let $\vec{X}_{\text{mtDNA}}$ denote the number of transitions observed at each site along the human mtDNA phylogenetic tree as described by Phylotree. Each coordinate corresponds to a different site out of the 16,569 sites. The number of transitions at site i, $X_{\text{mtDNA,i}}$, follows a Poisson distribution with parameter λ_i.

2. Let $\vec{Y}$ denote the number of transitions between two examined sequences (e.g. modern human and Neanderthal). We normalize the length of the tree edges so that the sum of all Phylotree's edges is one. The estimated parameter p relates to the edge distance between the two examined sequences. Subsequently, Y_i follows a Poisson distribution with parameter $\lambda_i \cdot p$.

3. Let $\vec{Z}$ denote the parity of $\vec{Y}$.

Using $\vec{X}$ and $\vec{Z}$, we can estimate p using the methods in Section 2.1. The TMRCA is given by: $\frac{1}{2}(T_{\text{sequence 1}} + T_{\text{sequence 2}} + p)$ when $T_{\text{sequence 1,2}}$ are the estimated times of the examined sequences measured in (uncalibrated) units of phylotree's total tree length.

2.3 Calibration

Our methods output p, which is the ratio of two values:

1. The sum of the edges between the two examined sequences and their most recent common ancestor (MRCA).

2. The total sum of Phylotree's edges.

Similarly to BEAST2, to calibrate p to years, we use the per-site per-year substitution rate for the coding region given in [10] $\mu = 1.57$ x 10E-8. We then calculate the total sum of Phylotree's edges in years by dividing the average number of substitutions in the coding region per site (1.4) by μ.

2.4 Data Availability Statement

The code used in this work is available at: https://github.com/Kerenlh/DivergenceTimes. A full description of all simulations is available in the Materials and Methods section, pages 8-10, and in the Supplementary Material, pages 26-28.

3 Results

3.1 Comparative Study on Raw Simulations

To compare the performance of the three estimation methods described in 2.1, we conducted experiments using simulated data. The Poisson rates λ were generated to reflect the substitution rates observed in mtDNA data using either a Categorical or a Gamma distribution. The parameters for the Gamma distribution ($\alpha = 0.23, \beta = 0.164$) were estimated directly from the data, while the parameters for the Categorical distribution were chosen such that both distributions have the same mean and variance. One of the Categorical values ($\epsilon = 0.1$) corresponds to the rate of low activity sites in the mtDNA data. The other value ($a = 11.87$) and the probabilities $(0.11, 0.89)$ were chosen accordingly. The results of the comparison are shown in Figure 1 with the Cramer-Rao bound for reference. To provide a qualitative comparison, we performed a one-sided paired Wilcoxon signed rank test on every pair of models, correcting for multiple comparisons using the Bonferroni correction. Our results show that Method 2 has the lowest squared error while Method 1 has the highest squared error, for both distributions. It is noteworthy that although Method 3 assumes a Gamma distribution, it still performs well even when there is a model mismatch.

Figure 1: Estimation errors for different λ distributions.

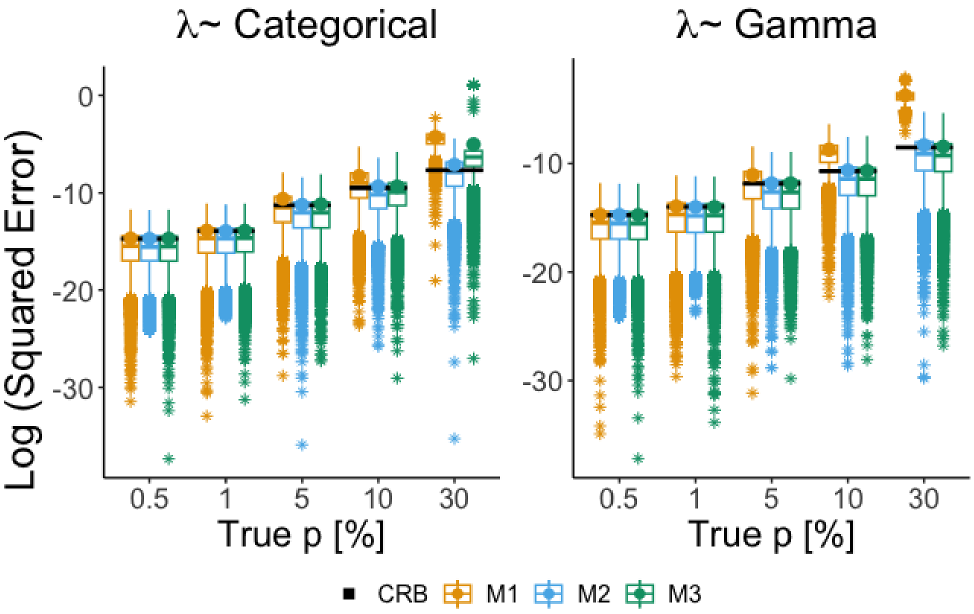


Figure 1. Box-plot of the log squared estimation errors of the three proposed methods for selected values of p, expressed as percentage of the total length of Phylotree's edges (outliers are marked with $*$). The simulations were run $10,000$ times for each value of p. The CRB is shown in black for reference and the circles represent the log of the mean values which are comparable to the CRB. The experiments were conducted for two different distributions of λ: (Left) Categorical distribution with two values: $\epsilon = 0.1$ with probability $\eta = 0.11$ and $a = 11.87$ with probability $1 - \eta$. (Right) Gamma distribution with parameters α and β.

3.2 Phylogenetic Tree Simulations

We validated our methods by testing their performance in a more realistic scenario of simulating a phylogenetic tree. Our methods take as input the observed transitions along Phylotree ($\vec{X}_{\mathrm{mtDNA}}$) and a binary vector $\vec{Z}$ denoting the differences between two sequences, which we aim to estimate the distance between. We compared our methods to the well-known BEAST2 software [5], which, similarly to other well-established methods (such as MCMCtree [36], MrBayes [27], etc.) considers sequences along with their phylogenetic tree to produce time estimations. The software BEAST2 performs Bayesian analysis using MCMC to average over the space of possible trees. However, it is limited in its computational capacity, so it cannot handle a large number of sequences like those in Phylotree. For this reason, we used a limited set of diverse sequences, including mtDNA genomes of 53 humans [13], the revised Cambridge Reference Sequence (rCRS) [2], the root of the human phylogenetic mtDNA tree, termed Reconstructed Sapiens Reference Sequence (RSRS) [4], and 10 ancient modern humans [10]. More details about the parameters used by BEAST2 are available in the Supplementary material, Section 2.2. To evaluate our methods, we added a simulated sequence with a predefined distance from the RSRS.

Our aim is to generate a vector $\vec{\lambda}$ that produces a vector $\vec{X}$ that has a similar distribution to $\vec{X}$mtDNA. The human mtDNA tree has 16,569 sites, of which 15,629 have no transversions. The MLE of λ_i at each site is the observed number of transitions, $X_{\mathrm{mtDNA,i}}$. However, simulating $\vec{\lambda}$ as $\vec{X}_{\mathrm{mtDNA}}$ leads to an undercount of transitions because 10,411 sites (67% of the total number of sites considered) had no transitions along the tree and their Poisson rate is taken to be zero. To mitigate this issue, the rates for these sites were chosen to be ϵ, the value that minimizes the Kolmogorov-Smirnov statistic [16, 28] (details are provided in the Supplementary material, Section 2.1).

The results are presented in Figure 2. BEAST2 overestimates the true p when p is smaller than approximately 2%, and underestimates it when p is higher. Additionally, BEAST2 has a much longer running time (roughly 3 hours) compared to our methods (less than a second). As shown in Figure 1, Method 1 has a larger error than Methods 2 and 3 for values of p within the simulated region, and the gap widens with increasing p. Methods 2 and 3 provide the best results for the entire range of p.

3.3 Real data results

As the final step of our experiments, we apply our methods on real-world data to determine the TMRCA of the modern human and Neanderthal, Denisovan, and chimpanzee mtDNA genomes. Table 1 displays the uncalibrated distances between modern human and each sequence, compared to the estimates from BEAST2. The presented TMRCA represents an average of the TMRCA obtained from 55 modern human mtDNA sequences of diverse origins [13]. Table 2 presents the TMRCA in kya (kilo-years ago) of the modern human and each sequence.

The estimates from real-world sequences presented in Table 1 are consistent with those obtained for the simulated dataset in Section 3.2. For low values of p, our three methods all produce similar estimates while BEAST2's has a slightly higher estimate. For the human-Chimpanzee uncalibrated distance, which is relatively high, Method 1 provides a higher estimate than that obtained by Methods 2 and 3, while BEAST2 provides a substantially lower estimate. The results in Table 2 show the TMRCA estimates, which are significantly smaller for our methods than those obtained from BEAST2 for human-Neanderthals and human-Denisovans. For example, BEAST2 estimated the human – Sima de los Huesos – Denisovans divergence time as ~ 934 kya, while our best-performing method (2) estimated it as ~ 824 kya. This divergence time is estimated as (540-1,410 kya) in [22]. Similarly,

Figure 2: Comparison of estimators applied on a simulated long branch.

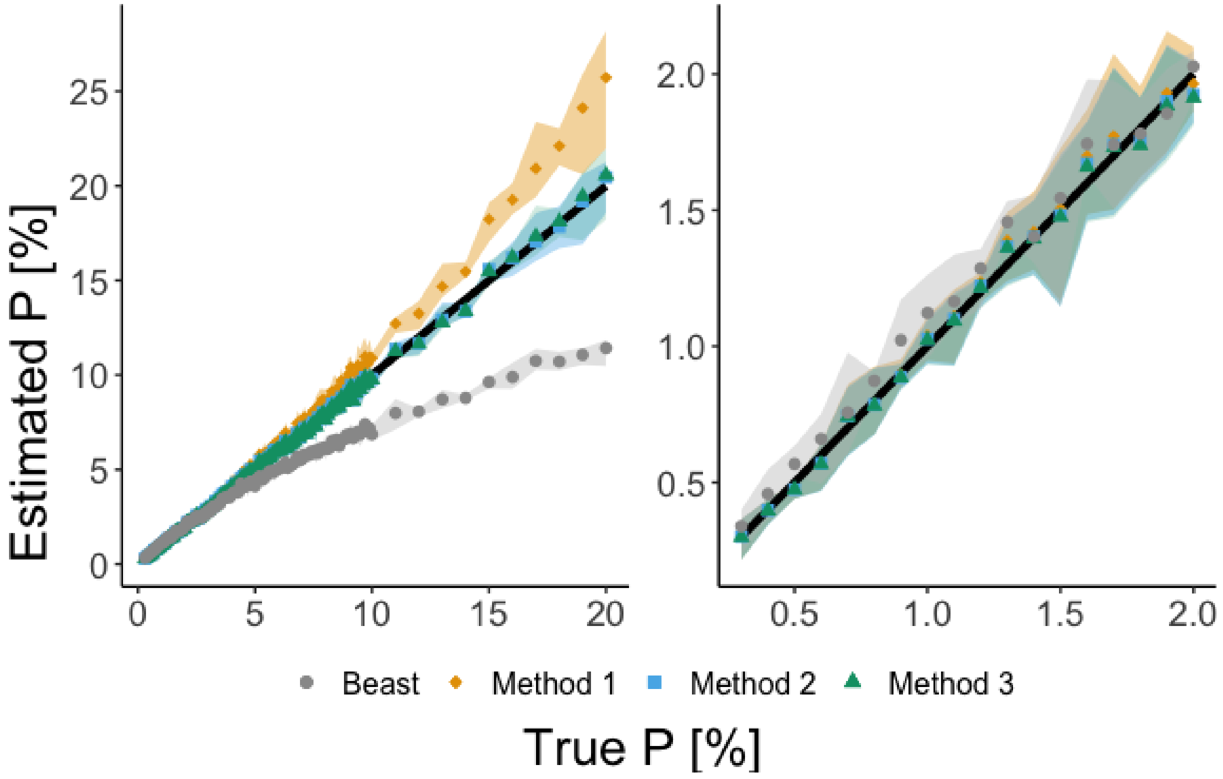


Figure 2. Comparison of our methods with BEAST2 estimator using simulated data. The right plot shows a zoom-in view of the left plot, focusing on values of p between 0 and 10%. Each point in the plot represents the average of 5 runs, while the shaded regions indicate the range of estimations obtained.

274 BEAST2 estimated the human – Neanderthal divergence time as ∼ 502 kya, while our methods estimated it as ∼ 408 kya. Preceding literature estimates this time closer to ours (∼400 kya [23, 9, 26]) while recent literature provides a much earlier estimate (∼800 kya [11]). Finally, BEAST2 estimates the human-Chimpanzee TMRCA as ∼3,712 kya whereas

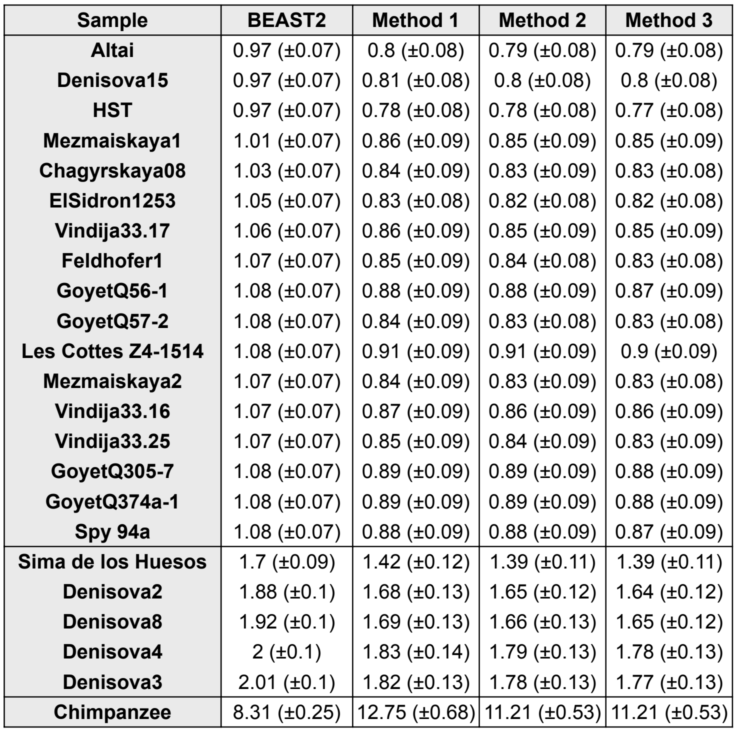
our estimate is ∼5,001 kya, much closer to the literature value of 5 – 8 million years ago [17, 19, 1, 30].

4 Conclusion

We investigated an estimation problem arising in statistical genetics when estimating divergence times between species. The problem's formulation, estimating Poisson rates from parity samples, leads to multiple estimators with varying assumptions. We calculated the CRB for this estimation problem and compared our methods against commonly used BEAST2 in different empirical settings, including a simple sampling scheme (Section 3.1), a more elaborate generative scheme based on real-world mtDNA data (Section 3.2), and the calculation of the TMRCA of modern humans and other hominins using their mtDNA genomes (Section 3.3).

Our results indicate that our proposed methods are significantly faster and more accurate than BEAST2, especially for earlier divergence times such as the human-Chimpanzee. Our methods utilize the transition statistics from the entire known human mtDNA phylogenetic tree (Phylotree) without the need for reconstructing a tree containing the sequences of interest. Our results show that the human – Neanderthal divergence time is $\sim 408,000$ years ago, considerably later than the values obtained by BEAST2 ($\sim 502,000$ years ago) and other values cited in the literature.

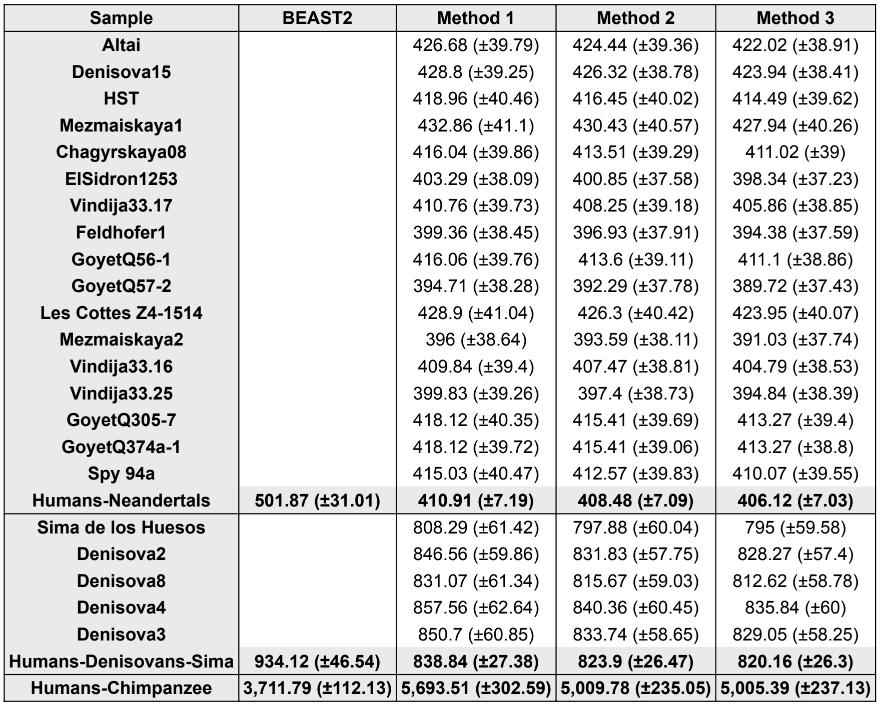
Table 1: Uncalibrated distances between modern humans and selected hominins.



Sample	BEAST2	Method 1	Method 2	Method 3
Altai	0.97 (±0.07)	0.8 (±0.08)	0.79 (±0.08)	0.79 (±0.08)
Denisova15	0.97 (±0.07)	0.81 (±0.08)	0.8 (±0.08)	0.8 (±0.08)
HST	0.97 (±0.07)	0.78 (±0.08)	0.78 (±0.08)	0.77 (±0.08)
Mezmaiskaya1	1.01 (±0.07)	0.86 (±0.09)	0.85 (±0.09)	0.85 (±0.09)
Chagyrskaya08	1.03 (±0.07)	0.84 (±0.09)	0.83 (±0.09)	0.83 (±0.08)
ElSidron1253	1.05 (±0.07)	0.83 (±0.08)	0.82 (±0.08)	0.82 (±0.08)
Vindija33.17	1.06 (±0.07)	0.86 (±0.09)	0.85 (±0.09)	0.85 (±0.09)
Feldhofer1	1.07 (±0.07)	0.85 (±0.09)	0.84 (±0.08)	0.83 (±0.08)
GoyetQ56-1	1.08 (±0.07)	0.88 (±0.09)	0.88 (±0.09)	0.87 (±0.09)
GoyetQ57-2	1.08 (±0.07)	0.84 (±0.09)	0.83 (±0.08)	0.83 (±0.08)
Les Cottes Z4-1514	1.08 (±0.07)	0.91 (±0.09)	0.91 (±0.09)	0.9 (±0.09)
Mezmaiskaya2	1.07 (±0.07)	0.84 (±0.09)	0.83 (±0.09)	0.83 (±0.08)
Vindija33.16	1.07 (±0.07)	0.87 (±0.09)	0.86 (±0.09)	0.86 (±0.09)
Vindija33.25	1.07 (±0.07)	0.85 (±0.09)	0.84 (±0.09)	0.83 (±0.09)
GoyetQ305-7	1.08 (±0.07)	0.89 (±0.09)	0.89 (±0.09)	0.88 (±0.09)
GoyetQ374a-1	1.08 (±0.07)	0.89 (±0.09)	0.89 (±0.09)	0.88 (±0.09)
Spy 94a	1.08 (±0.07)	0.88 (±0.09)	0.88 (±0.09)	0.87 (±0.09)
Sima de los Huesos	1.7 (±0.09)	1.42 (±0.12)	1.39 (±0.11)	1.39 (±0.11)
Denisova2	1.88 (±0.1)	1.68 (±0.13)	1.65 (±0.12)	1.64 (±0.12)
Denisova8	1.92 (±0.1)	1.69 (±0.13)	1.66 (±0.13)	1.65 (±0.12)
Denisova4	2 (±0.1)	1.83 (±0.14)	1.79 (±0.13)	1.78 (±0.13)
Denisova3	2.01 (±0.1)	1.82 (±0.13)	1.78 (±0.13)	1.77 (±0.13)
Chimpanzee	8.31 (±0.25)	12.75 (±0.68)	11.21 (±0.53)	11.21 (±0.53)

Table 1. Uncalibrated distances expressed as a percentage of the total length of Phylotree's edges, as determined by our methods compared with BEAST2. The values correspond to p, and indicate the estimation's location in Figure 2. In the parentheses we provide the standard deviation for each estimator, obtained from bootstrapping 100 site samples for every modern human – ancient sequence pair in the dataset. Note that the BEAST2 values presented here were de-calibrated as described in Section 2.3.

Table 2: Estimated divergence times between modern human and selected hominins.



Sample	BEAST2	Method 1	Method 2	Method 3
Altai		426.68 (±39.79)	424.44 (±39.36)	422.02 (±38.91)
Denisova15		428.8 (±39.25)	426.32 (±38.78)	423.94 (±38.41)
HST		418.96 (±40.46)	416.45 (±40.02)	414.49 (±39.62)
Mezmaiskaya1		432.86 (±41.1)	430.43 (±40.57)	427.94 (±40.26)
Chagyrskaya08		416.04 (±39.86)	413.51 (±39.29)	411.02 (±39)
ElSidron1253		403.29 (±38.09)	400.85 (±37.58)	398.34 (±37.23)
Vindija33.17		410.76 (±39.73)	408.25 (±39.18)	405.86 (±38.85)
Feldhofer1		399.36 (±38.45)	396.93 (±37.91)	394.38 (±37.59)
GoyetQ56-1		416.06 (±39.76)	413.6 (±39.11)	411.1 (±38.86)
GoyetQ57-2		394.71 (±38.28)	392.29 (±37.78)	389.72 (±37.43)
Les Cottes Z4-1514		428.9 (±41.04)	426.3 (±40.42)	423.95 (±40.07)
Mezmaiskaya2		396 (±38.64)	393.59 (±38.11)	391.03 (±37.74)
Vindija33.16		409.84 (±39.4)	407.47 (±38.81)	404.79 (±38.53)
Vindija33.25		399.83 (±39.26)	397.4 (±38.73)	394.84 (±38.39)
GoyetQ305-7		418.12 (±40.35)	415.41 (±39.69)	413.27 (±39.4)
GoyetQ374a-1		418.12 (±39.72)	415.41 (±39.06)	413.27 (±38.8)
Spy 94a		415.03 (±40.47)	412.57 (±39.83)	410.07 (±39.55)
Humans-Neandertals	**501.87 (±31.01)**	**410.91 (±7.19)**	**408.48 (±7.09)**	**406.12 (±7.03)**
Sima de los Huesos		808.29 (±61.42)	797.88 (±60.04)	795 (±59.58)
Denisova2		846.56 (±59.86)	831.83 (±57.75)	828.27 (±57.4)
Denisova8		831.07 (±61.34)	815.67 (±59.03)	812.62 (±58.78)
Denisova4		857.56 (±62.64)	840.36 (±60.45)	835.84 (±60)
Denisova3		850.7 (±60.85)	833.74 (±58.65)	829.05 (±58.25)
Humans-Denisovans-Sima	**934.12 (±46.54)**	**838.84 (±27.38)**	**823.9 (±26.47)**	**820.16 (±26.3)**
Humans-Chimpanzee	**3,711.79 (±112.13)**	**5,693.51 (±302.59)**	**5,009.78 (±235.05)**	**5,005.39 (±237.13)**

Table 2. The table displays the estimated divergence times (in kya) between modern humans and selected hominins, as determined by our methods and compared with BEAST2. The standard deviation, which arises from a combination of the standard deviation of our methods and the sample dating, is given in parentheses. It's important to note that BEAST2 calculates the TMRCA for all sequences in the same clade as a single estimate, while our methods estimate the TMRCA for each sample individually by taking the average of estimations derived from comparing the sample with every modern human sequence in the dataset.

Supplementary Material

1 Theoretical Details

1.1 Proof of Lemma 1

Let $Y \sim \text{Pois}(\lambda)$ and Z be the parity of Y. Then $Z \sim Ber(\frac{1}{2}(1-e^{-2\lambda}))$.

Proof.

$$P(Z_i = 1) = \sum_{n=0}^{\infty} P\left(Y_i = 2n+1\right) = \sum_{n=0}^{\infty} e^{-\lambda}\frac{\lambda^{2n+1}}{(2n+1)!} =$$

$$= e^{-\lambda}\frac{1}{2}\left(\sum_{n=0}^{\infty}\frac{\lambda^n}{n!} - \sum_{n=0}^{\infty}\frac{(-\lambda)^n}{n!}\right) = \frac{e^{-\lambda}}{2}\left(e^{\lambda}-e^{-\lambda}\right) = \frac{1}{2}\left(1-e^{-2\lambda}\right).$$

□

1.2 Proof of Theorem 1

Denote the Fisher information matrix for the estimation problem above by $I \in \mathbb{R}^{(n+1,n+1)}$, where the first n indexes correspond to $\{\lambda_i\}_{i=1}^n$ and the last index $(n+1)$ corresponds to p. For clarity denote $I_{p,p} \doteq I_{n+1,n+1}, I_{i,p} \doteq I_{i,n+1}, I_{p,i} \doteq I_{n+1,i}$. Then:

$$I_{i,j} = 0, \quad I_{i,i} = \frac{1}{\lambda_i} + \frac{4p^2}{e^{4\lambda_i p}-1}, \quad I_{i,p} = I_{p,i} = \frac{4p\lambda_i}{e^{4\lambda_i p}-1}, \quad I_{p,p} = 4\sum_{i=1}^{n}\frac{\lambda_i^2}{e^{4\lambda_i p}-1}. \tag{9}$$

Consequently, an unbiased estimator $\hat{p}$ holds:

$$\mathbb{E}\left[(p-\hat{p})^2\right] \geq \left[4\sum_{i=1}^{n} \frac{\lambda_i^2}{e^{4\lambda_i p} - 1 + 4p^2\lambda_i}\right]^{-1}. \tag{10}$$

If $\forall i = 1..n : \lambda_i = \lambda$, we can further simplify the expression:

$$\mathbb{E}\left[(p-\hat{p})^2\right] \geq \frac{e^{4\lambda p} - 1 + 4p^2\lambda}{4n\lambda^2}. \tag{11}$$

Proof. We calculate the second derivative of the log-likelihood. Denote:

$$\beta_i = -2\lambda_i p + j\pi Z_i, \quad \sigma(t) = \frac{e^t}{1+e^t},$$

then the first derivatives are given by:

$$\begin{aligned}\frac{\partial l}{\partial \lambda_i} &= -1 + \frac{X_i}{\lambda_i} + \frac{(-2p)\left(-1\right)^{Z_i}\exp\left(-2\lambda_i p\right)}{1+\left(-1\right)^{Z_i}\exp\left(-2\lambda_i p\right)} \\ &= -1 + \frac{X_i}{\lambda_i} - 2p\sigma\left(-2\lambda_i p + j\pi Z_i\right) \\ &= -1 + \frac{X_i}{\lambda_i} - 2p\sigma\left(\beta_i\right),\end{aligned} \tag{12}$$

and

$$\frac{\partial l}{\partial p} = \sum_{i=1}^{n} \frac{(-2\lambda_i)\left(-1\right)^{Z_i}\exp\left(-2\lambda_i p\right)}{1+\left(-1\right)^{Z_i}\exp\left(-2\lambda_i p\right)} = -2\sum_{i=1}^{n}\lambda_i\sigma\left(\beta_i\right). \tag{13}$$

The second derivatives are now given by:

$$\begin{aligned}\frac{\partial^2 l}{\partial \lambda_i \lambda_j} &= 0 \\ \frac{\partial^2 l}{\partial {\lambda_i}^2} &= -\frac{X_i}{\lambda_i^2} - 2p\left(-2p\right)\sigma\left(\beta_i\right)\left(1-\sigma\left(\beta_i\right)\right) = -\frac{X_i}{\lambda_i^2} + 4p^2\sigma\left(\beta_i\right)\left(1-\sigma\left(\beta_i\right)\right) \\ \frac{\partial^2 l}{\partial \lambda_i \partial p} &= -2p\left(-2\lambda_i\right)\sigma\left(\beta_i\right)\left(1-\sigma\left(\beta_i\right)\right) - 2\sigma\left(\beta_i\right) = 4p\lambda_i\sigma\left(\beta_i\right)\left(1-\sigma\left(\beta_i\right)\right) - 2\sigma\left(\beta_i\right) \\ \frac{\partial^2 l}{\partial p^2} &= \sum_{i=1}^{n} 4\lambda_i^2\sigma\left(\beta_i\right)\left(1-\sigma\left(\beta_i\right)\right)\end{aligned}$$

The expectation of these are given by:

$$\begin{aligned}
\mathbb{E}\left[\sigma\left(\beta_i\right)\right] &= \frac{1}{2}\left(1+\exp\left(-2\lambda_i p\right)\right)\frac{\exp\left(-2\lambda_i p\right)}{1+\exp\left(-2\lambda_i p\right)} + \frac{1}{2}\left(1-\exp\left(-2\lambda_i p\right)\right)\frac{(-1)\cdot\exp\left(-2\lambda_i p\right)}{1-\exp\left(-2\lambda_i p\right)} = 0 \\
\mathbb{E}\left[\sigma^2\left(\beta_i\right)\right] &= \frac{1}{2}\left(1+\exp\left(-2\lambda_i p\right)\right)\frac{\exp\left(-4\lambda_i p\right)}{\left(1+\exp\left(-2\lambda_i p\right)\right)^2} + \frac{1}{2}\left(1-\exp\left(-2\lambda_i p\right)\right)\frac{\exp\left(-4\lambda_i p\right)}{\left(1-\exp\left(-2\lambda_i p\right)\right)^2} = \\
&= \frac{1}{2}\exp\left(-4\lambda_i p\right)\left[\frac{1}{1+\exp\left(-2\lambda_i p\right)} + \frac{1}{1-\exp\left(-2\lambda_i p\right)}\right] = \frac{1}{e^{4\lambda_i p}-1} \\
\mathbb{E}\left[\frac{\partial^2 l}{\left(\partial\lambda_i\right)^2}\right] &= E\left[-\frac{X_i}{\lambda_i^2} + 4p^2\sigma\left(\beta_i\right)\left(1-\sigma\left(\beta_i\right)\right)\right] = -\frac{1}{\lambda_i} - \frac{4p^2}{e^{4\lambda_i p}-1} = -I_{i,i} \\
\mathbb{E}\left[\frac{\partial^2 l}{\partial\lambda_i\partial p}\right] &= E\left[4p\lambda_i\sigma\left(\beta_i\right)\left(1-\sigma\left(\beta_i\right)\right) - 2\sigma\left(\beta_i\right)\right] = -\frac{4p\lambda_i}{e^{4\lambda_i p}-1} = -I_{i,p} \\
\mathbb{E}\left[\frac{\partial^2 l}{\left(\partial p\right)^2}\right] &= E\left[\sum_{i=1}^{n} 4\lambda_i^2\sigma\left(\beta_i\right)\left(1-\sigma\left(\beta_i\right)\right)\right] = -\sum_{i=1}^{n}\frac{4\lambda_i^2}{e^{4\lambda_i p}-1} = -I_{p,p}.
\end{aligned}$$

By CRB, for an unbiased estimator:

$$\begin{aligned}
\mathbb{E}\left[\left(p-\hat{p}\right)^2\right] \geq [I^{-1}]_{p,p} &= \frac{1}{I_{p,p} - I_{p,i}I_{i,i}^{-1}I_{i,p}} \\
&= \left[\sum_{i=1}^{n}\frac{4\lambda_i^2}{e^{4\lambda_i p}-1} - \sum_{i=1}^{n}\frac{\frac{16p^2\lambda_i^2}{\left[e^{4\lambda_i p}-1\right]^2}}{\frac{1}{\lambda_i}+\frac{4p^2}{e^{4\lambda_i p}-1}}\right]^{-1} \\
&= \left[\sum_{i=1}^{n} 4\lambda_i^2\frac{\left(e^{4\lambda_i p}-1\right)\left(\frac{1}{\lambda_i}+\frac{4p^2}{e^{4\lambda_i p}-1}\right) - 4p^2}{\left[e^{4\lambda_i p}-1\right]^2\left(\frac{1}{\lambda_i}+\frac{4p^2}{e^{4\lambda_i p}-1}\right)}\right]^{-1} \\
&= \left[4\sum_{i=1}^{n}\frac{\lambda_i^2}{e^{4\lambda_i p}-1+4p^2\lambda_i}\right]^{-1}
\end{aligned}$$

□

1.3 Proof of Proposition 1

Proof. Following Equations 12, 13, we compare the first order derivatives to 0:

$$\frac{\partial l}{\partial \lambda_i} = -1 + \frac{X_i}{\lambda_i} - 2\hat{p}\frac{(-1)^{Z_i}e^{-2\lambda_i\hat{p}}}{\left(1+(-1)^{Z_i}e^{-2\lambda_i\hat{p}}\right)} = 0 \Rightarrow X_i = \hat{\lambda}_i + 2\hat{p}\hat{\lambda}_i\frac{(-1)^{Z_i}e^{-2\hat{\lambda}_i\hat{p}}}{\left(1+(-1)^{Z_i}e^{-2\hat{\lambda}_i\hat{p}}\right)}$$

$$\frac{\partial l}{\partial p} = -\sum_{i=1}^{n} 2\lambda_i\frac{(-1)^{Z_i}e^{-2\lambda_i\hat{p}}}{\left(1+(-1)^{Z_i}e^{-2\lambda_i\hat{p}}\right)} = -\sum_{i=1}^{n}\frac{\lambda_i}{\hat{p}}\left[-1+\frac{X_i}{\lambda_i}\right] =0 \Rightarrow \sum_{i=1}^{n}\hat{\lambda}_i = \sum_{i=1}^{n} X_i.$$

Summing the first equation for every i and substituting the second equation results in the last part in Equation 6. □

1.4 Proof of Proposition 2

If $Y_i|X_i \sim Bin(X_i, p)$, then:

1. $Y_i \sim \text{Pois}(\lambda_i \cdot p)$, which justifies this approach.

2. $Z_i|X_i \sim Ber\left(\frac{1}{2}\left(1-(1-2p)^{X_i}\right)\right)$, so we can compute the likelihood of p without considering λ_i.

3. The maximum likelihood estimate of p given Z_i holds:

$$\sum_{i=1}^{n}\frac{X_i}{1+(-1)^{Z_i}(1-2p)^{-X_i}} = 0 \tag{14}$$

and the maximum likelihood estimate of p given $\sum_{i=1}^{n} Z_i$ holds:

$$\sum_{i=1}^{n}(1-2\hat{p})^{X_i} = n - 2\sum_{i=1}^{n} Z_i \tag{15}$$

Proof. Denote $q \equiv 1 - p$. For item 1:

$$\begin{aligned}
\Pr(Y_i = k) &= \sum_{n=k}^{\infty} \Pr(X_i = n) \cdot \Pr(Bin(n,p) = k) \\
&= \sum_{n=k}^{\infty} \frac{\lambda_i^n \cdot e^{-\lambda_i}}{n!} \cdot \Pr(\binom{n}{k}) p^k q^{n-k} \\
&= \frac{(\lambda_i \cdot p)^k \cdot e^{-\lambda_i p}}{k!} \sum_{n=k}^{\infty} \frac{\lambda_i^{n-k} \cdot e^{-\lambda_i q}}{(n-k)!} \cdot q^{n-k} \\
&= \frac{(\lambda_i \cdot p)^k \cdot e^{-\lambda_i p}}{k!} \sum_{n=0}^{\infty} \frac{\lambda_i^{n} \cdot e^{-\lambda_i q}}{n!} \cdot q^{n} = \frac{(\lambda_i \cdot p)^k \cdot e^{-\lambda_i p}}{k!}.
\end{aligned}$$

Now moving on to item 2:

$$\begin{aligned}
\Pr(Z_i = 1 | X_i) &= \Pr(Y_i \text{ is odd} | X_i), \quad Y_i | X_i \sim Bin(n = X_i, p) \\
(q+p)^n &= \Sigma_{k=0}^{n} \binom{n}{k} p^k q^{(n-k)} = P(Y_i \text{ is even}) + P(Y_i \text{ is odd}) \\
(q-p)^n &= \Sigma_{k=0}^{n} \binom{n}{k} (-p)^k q^{(n-k)} = P(Y_i \text{ is even}) - P(Y_i \text{ is odd})
\end{aligned}$$

And summing up these two equations leads to:

$$P(Y_i \text{ is even}) = \frac{1}{2}\left((q+p)^n + (q-p)^n\right) = \frac{1}{2}\left(1 + (1-2p)^n\right).$$

Subsequently, the likelihood of Z_i is given by:

$$\begin{aligned}
l(\vec{Z}; p) &= \prod_{i=1}^{n} \frac{1}{2}\left(1 + (-1)^{Z_i}(1-2p)^{X_i}\right) \\
L(\vec{Z}; p) &= \sum_{i=1}^{n} \log\left(1 + (-1)^{Z_i}(1-2p)^{X_i}\right) + Const
\end{aligned}$$

322 Taking the derivative to 0:

$$\frac{\partial L}{\partial p} = \sum_{i=1}^{n} \frac{-2(-1)^{Z_i} X_i (1-2p)^{X_i - 1}}{(1 + (-1)^{Z_i}(1-2p)^{X_i})} = \sum_{i=1}^{n} \frac{-2X_i}{((-1)^{Z_i}(1-2p)^{1-X_i} + 1 - 2p)} = 0, \qquad (16)$$

and division by $\frac{-2}{1-2p}$ yields the solution.

Now, according to Le Cam's theorem[1] [20], $\sum\limits_{i=1}^{n} Z_i \sim \text{Pois}\left(\lambda = \sum\limits_{i=1}^{n} \frac{1}{2}\left(1-(1-2p)^{X_i}\right)\right)$, and the likelihood is therefore:

$$L\left(\sum_{i=1}^{n} Z_i = m|\vec{X};p\right) = \lambda^m \frac{e^{-\lambda}}{m!}.$$

Now we look at the log-likelihood and take the derivative with respect to p to zero:

$$\begin{aligned} l\left(\sum_{i=1}^{n} Z_i = m|\vec{X};p\right) &= m\log\lambda - \lambda + Const \\ &= m\log\left(\sum_{i=1}^{n}\frac{1}{2}\left(1-(1-2p)^{X_i}\right)\right) - \sum_{i=1}^{n}\frac{1}{2}\left(1-(1-2p)^{X_i}\right) + Const \\ \frac{\partial l}{\partial p} &= m\frac{\sum\limits_{i=1}^{n} X_i(1-2p)^{X_i-1}}{\sum\limits_{i=1}^{n}\frac{1}{2}\left(1-(1-2p)^{X_i}\right)} - \sum_{i=1}^{n} X_i(1-2p)^{X_i-1} \\ &= \left(\frac{m}{\sum\limits_{i=1}^{n}\frac{1}{2}\left(1-(1-2p)^{X_i}\right)} - 1\right)\sum_{i=1}^{n} X_i(1-2p)^{X_i-1} = 0 \end{aligned}$$

Leading to the solution:

$$\sum_{i=1}^{n}(1-2\hat{p})^{X_i} = n - 2m = n - 2\sum_{i=1}^{n} Z_i$$

□

[1] More precisely: $\sum\limits_{k=0}^{\infty} |P(\sum\limits_{i=1}^{n} Z_i = k) - \frac{1}{k!}(\sum\limits_{i=1}^{n}\frac{1}{2}(1-(1-2p)^{X_i}))^k e^{-\sum\limits_{i=1}^{n}\frac{1}{2}(1-(1-2p)^{X_i}}| < 2\sum\limits_{i=1}^{n}\left(\frac{1}{2}(1-(1-2p)^{X_i})\right)^2$.

1.5 Proof of Proposition 3

Let $\lambda_i \sim \Gamma(\alpha, \beta)$, then the maximum a posteriori estimator of p holds:

$$\frac{\partial l}{\partial p} = \sum_{i=1}^{n} \frac{X_i + \alpha}{(-1)^{Z_i}\left(1 + \frac{2p}{\beta+1}\right)^{X_i+\alpha} + 1} = 0 \tag{17}$$

Subsequently, estimated values for α, β can be substituted for a numerical estimator for p.

Proof. We first compute the probability for each observation:

$$\begin{aligned}
\Pr\left(X_i = k, Y_i \, is \, even\right) &= \int_0^\infty P\left(\lambda_i = \lambda\right) P\left(X_i = k | \lambda_i = \lambda\right) P\left(Y_i \, is \, even | \lambda_i = \lambda\right) d\lambda \\
&= \int_0^\infty \lambda^{\alpha-1} e^{-\lambda\beta} \frac{\beta^\alpha}{\Gamma\left(\alpha\right)} e^{-\lambda} \frac{\lambda^k}{k!} \frac{1}{2} \left(1 + e^{-2\lambda p}\right) d\lambda \\
&= \frac{\beta^\alpha}{2k!\Gamma\left(\alpha\right)} \left[\int_0^\infty \lambda^{\alpha-1+k} e^{-\lambda(\beta+1)} d\lambda + \int_0^\infty \lambda^{\alpha-1+k} e^{-\lambda(\beta+1+2p)} d\lambda\right] \\
&= \frac{\beta^\alpha}{2k!\Gamma\left(\alpha\right)} [\frac{\Gamma\left(\alpha+k\right)}{\left(\beta+1\right)^{\alpha+k}} \underbrace{\int_0^\infty \lambda^{\alpha-1+k} e^{-\lambda(\beta+1)} \frac{\left(\beta+1\right)^{\alpha+k}}{\Gamma\left(\alpha+k\right)} d\lambda}_{=1} + \\
&\qquad \frac{\Gamma\left(\alpha+k\right)}{\left(\beta+1+2p\right)^{\alpha+k}} \underbrace{\int_0^\infty \lambda^{\alpha-1+k} e^{-\lambda(\beta+1+2p)} \frac{\left(\beta+1+2p\right)^{\alpha+k}}{\Gamma\left(\alpha+k\right)} d\lambda]}_{=1} \\
&= \frac{\beta^\alpha \Gamma\left(\alpha+k\right)}{2k!\Gamma\left(\alpha\right)} \left[\frac{1}{\left(\beta+1\right)^{\alpha+k}} + \frac{1}{\left(\beta+1+2p\right)^{\alpha+k}}\right] \\
&= \frac{\Gamma\left(\alpha+k\right)}{2k!\Gamma\left(\alpha\right)} \left[\left(\frac{\beta}{\beta+1}\right)^\alpha \left(\frac{1}{\beta+1}\right)^k + \left(\frac{\beta}{\beta+1+2p}\right)^\alpha \left(\frac{1}{\beta+1+2p}\right)^k\right]
\end{aligned}$$

Hence, the likelihood is given by:

$$\begin{aligned}
&L\left(\vec{X}, \vec{Z}; p, \alpha, \beta\right) = \\
&= \prod_{i=1}^{n} \frac{\Gamma\left(\alpha+k\right)}{2k!\Gamma\left(\alpha\right)} \left[\left(\frac{\beta}{\beta+1}\right)^\alpha \left(\frac{1}{\beta+1}\right)^{X_i} + (-1)^{Z_i} \left(\frac{\beta}{\beta+1+2p}\right)^\alpha \left(\frac{1}{\beta+1+2p}\right)^{X_i}\right]
\end{aligned}$$

and the log-likelihood:

$$l\left(\vec{X},\vec{Z};p,\alpha,\beta\right)=\sum_{i=1}^{n}\log\frac{\Gamma\left(\alpha+X_i\right)}{X_i!\Gamma\left(\alpha\right)}+\alpha log\beta-\left(X_i+\alpha\right)\log(\beta+1)+\log\left[1+(-1)^{Z_i}\left(\frac{\beta+1}{\beta+1+2p}\right)^{X_i+\alpha}\right]$$

Now comparing the derivative with respect to p to zero:

$$\frac{\partial l}{\partial p}=\sum_{i=1}^{n}\frac{-\frac{2}{\beta+1}\left(-1\right)^{Z_i}\left(X_i+\alpha\right)\left(1+\frac{2p}{\beta+1}\right)^{-X_i-\alpha-1}}{1+\left(-1\right)^{Z_i}\left(1+\frac{2p}{\beta+1}\right)^{-X_i-\alpha}}=0$$

□

2 Simulation details

2.1 Phylogenetic tree simulations

The rate parameter for sites with no transitions along the tree is denoted as ϵ, and we estimate it using the following simulation-based method. To generate $\vec{\lambda}$, we use the following equation:

$$\min D=\sup_{x}|F(\vec{X}_{\text{mtDNA}})-F(\vec{X})|\quad s.t.\quad \lambda_i=\begin{cases}X_{\text{mtDNA},i} & X_{\text{mtDNA},i}\neq 0\\ \epsilon & X_{\text{mtDNA},i}=0\end{cases}\tag{18}$$

The value of ϵ is chosen to minimize the Kolmogorov–Smirnov statistic. Figure 3 shows a simulation of $D(\epsilon)$, with the mean of 1,000 runs for each ϵ value. The minimum value of D is obtained for $\epsilon=0.0913$ (marked in red).

To make the simulated data closer to the real data, we also model transversions. We estimate the transversion rate per site in the same manner as the transition rate, using the Kolmogorov–Smirnov statistic to account for sites with no transversions. This results in

Figure 3: Kolmogorov–Smirnov statistic as a function of ϵ.

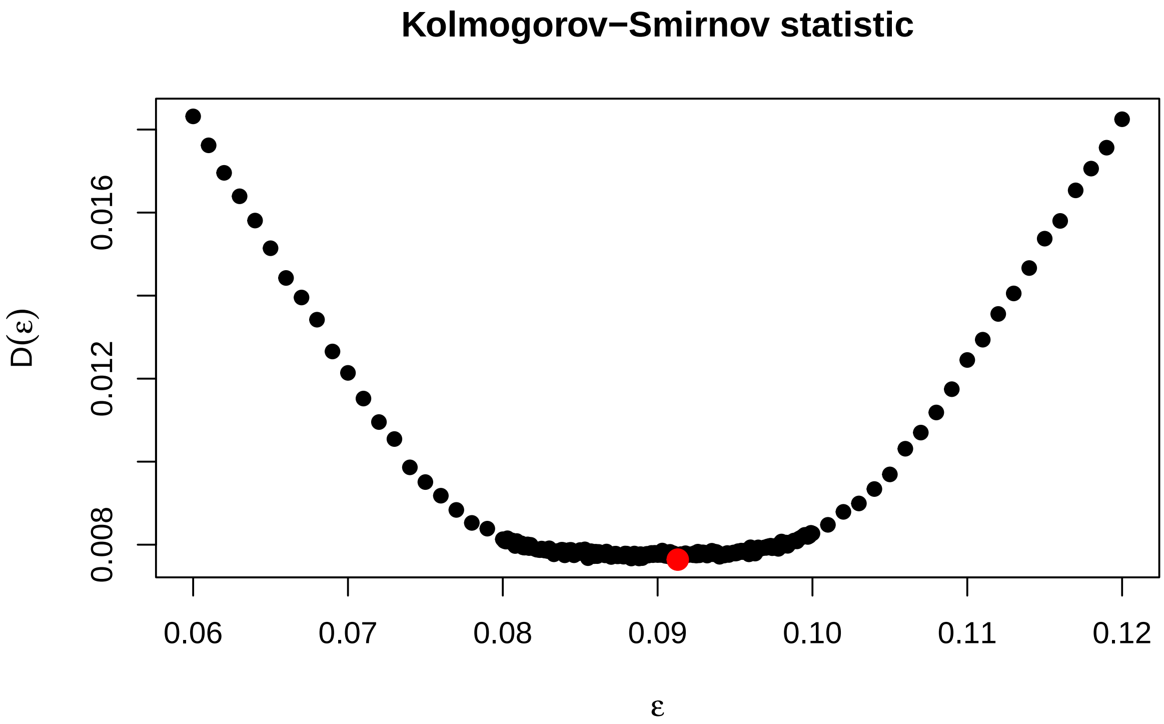


Figure 3. We performed 1,000 runs for each value of epsilon. The minimal $D(\epsilon)$ is marked red and equals $\epsilon = 0.0913$.

$\epsilon_{\text{transversion}} = 0.0149$. To determine the nucleotide at a given site, we sample whether an odd number of transversions have occurred. If so, a random nucleotide is sampled from the two available transversion options. The resulting sequence is then input into BEAST2, but our methods still use only the sites without observed transversions. Finally, the analysis is limited to the gene regions in the genome (11,341 sites).

2.2 BEAST2 run parameters

The sequences used in this work were aligned using mafft [15], and the 11.3 kb of protein-coding genes were extracted and used for the analysis. The analysis followed the approach described in [34], where the best fitting clock and tree model for the tree were identified using path sampling with the model selection package in BEAST2 [14, 3, 21]. Each model test was run with 40 path steps, a chain length of 25 million iterations, an alpha parameter of 0.3, a pre-burn-in of 75,000 iterations, and an 80% burn-in of the entire chain. The mutation rate was set to 1.57 x 10E-8 and a normal distribution (mean: mutation rate, sigma: 1.E-10) was used for a strict clock model [10]. The TN93 substitution model [31] was used for all models. The tree was calibrated with carbon dating data from ancient humans and Neanderthals, where available [24, 10, 35], and modern samples were set to a date of 0.

References

[1] Guy Amster and Guy Sella. Life history effects on the molecular clock of autosomes and sex chromosomes. *Proceedings of the National Academy of Sciences*, 113(6):1588–1593, 2016.

[2] Richard M Andrews, Iwona Kubacka, Patrick F Chinnery, Robert N Lightowlers, Douglass M Turnbull, and Neil Howell. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nature genetics*, 23(2):147–147, 1999.

[3] Guy Baele, Wai Lok Sibon Li, Alexei J Drummond, Marc A Suchard, and Philippe Lemey. Accurate model selection of relaxed molecular clocks in Bayesian phylogenetics. *Molecular biology and evolution*, 30(2):239–243, 2012.

[4] Doron M Behar, Mannis Van Oven, Saharon Rosset, Mait Metspalu, Eva-Liis Loogväli, Nuno M Silva, Toomas Kivisild, Antonio Torroni, and Richard Villems. A Copernican reassessment of the human mitochondrial DNA tree from its root. *The American Journal of Human Genetics*, 90(4):675–684, 2012.

[5] Remco Bouckaert, Joseph Heled, Denise Kühnert, Tim Vaughan, Chieh-Hsi Wu, Dong Xie, Marc A Suchard, Andrew Rambaut, and Alexei J Drummond. BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS computational biology*, 10(4):e1003537, 2014.

[6] David Roxbee Cox and Nancy Reid. Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 49(1):1–18, 1987.

[7] Harold Cramer. Mathematical methods of statistics, Princeton Univ. *Press, Princeton, NJ*, 1946.

[8] Mario Dos Reis, Philip CJ Donoghue, and Ziheng Yang. Bayesian molecular clock dating of species divergences in the genomics era. *Nature Reviews Genetics*, 17(2):71–80, 2016.

[9] Phillip Endicott, Simon YW Ho, and Chris Stringer. Using genetic evidence to evaluate four palaeoanthropological hypotheses for the timing of Neanderthal and modern human origins. *Journal of human evolution*, 59(1):87–95, 2010.

[10] Qiaomei Fu, Alissa Mittnik, Philip LF Johnson, Kirsten Bos, Martina Lari, Ruth Bollongino, Chengkai Sun, Liane Giemsch, Ralf Schmitz, Joachim Burger, et al. A revised timescale for human evolution based on ancient mitochondrial genomes. *Current biology*, 23(7):553–559, 2013.

[11] Aida Gómez-Robles. Dental evolutionary rates and its implications for the Neanderthal–modern human divergence. *Science advances*, 5(5):eaaw1268, 2019.

[12] Joseph M Hilbe. *Negative binomial regression*. Cambridge University Press, 2011.

[13] Max Ingman, Henrik Kaessmann, Svante Pääbo, and Ulf Gyllensten. Mitochondrial genome variation and the origin of modern humans. *Nature*, 408(6813):708–713, 2000.

[14] Robert E Kass and Adrian E Raftery. Bayes factor and model uncertainty. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

[15] Kazutaka Katoh and Hiroyuki Toh. Parallelization of the MAFFT multiple sequence alignment program. *Bioinformatics*, 26(15):1899–1900, 2010.

[16] Andrey Kolmogorov. Sulla determinazione empirica di una lgge di distribuzione. *Inst. Ital. Attuari, Giorn.*, 4:83–91, 1933.

[17] Sudhir Kumar, Alan Filipski, Vinod Swarna, Alan Walker, and S Blair Hedges. Placing confidence limits on the molecular age of the human–chimpanzee divergence. *Proceedings of the National Academy of Sciences*, 102(52):18842–18847, 2005.

[18] Sudhir Kumar, Koichiro Tamura, and Masatoshi Nei. MEGA: molecular evolutionary genetics analysis software for microcomputers. *Bioinformatics*, 10(2):189–191, 1994.

[19] Kevin E Langergraber, Kay Prüfer, Carolyn Rowney, Christophe Boesch, Catherine Crockford, Katie Fawcett, Eiji Inoue, Miho Inoue-Muruyama, John C Mitani, Martin N Muller, et al. Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. *Proceedings of the National Academy of Sciences*, 109(39):15716–15721, 2012.

[20] Lucien Le Cam. An approximation theorem for the Poisson binomial distribution. *Pacific Journal of Mathematics*, 10(4):1181–1197, 1960.

[21] Adam D Leaché, Matthew K Fujita, Vladimir N Minin, and Remco R Bouckaert. Species delimitation using genome-wide SNP data. *Systematic biology*, 63(4):534–542, 2014.

[22] Matthias Meyer, Qiaomei Fu, Ayinuer Aximu-Petri, Isabelle Glocke, Birgit Nickel, Juan-Luis Arsuaga, Ignacio Martínez, Ana Gracia, José María Bermúdez de Castro, Eudald Carbonell, et al. A mitochondrial genome sequence of a hominin from Sima de los Huesos. *Nature*, 505(7483):403–406, 2014.

[23] James P Noonan, Graham Coop, Sridhar Kudaravalli, Doug Smith, Johannes Krause, Joe Alessi, Feng Chen, Darren Platt, Svante Paabo, Jonathan K Pritchard, et al. Sequencing and analysis of Neanderthal genomic DNA. *Science*, 314(5802):1113–1118, 2006.

[24] Cosimo Posth, Christoph Wißing, Keiko Kitagawa, Luca Pagani, Laura van Holstein, Fernando Racimo, Kurt Wehrberger, Nicholas J Conard, Claus Joachim Kind, Hervé Bocherens, et al. Deeply divergent archaic mitochondrial genome provides lower time boundary for African gene flow into Neanderthals. *Nature communications*, 8(1):1–9, 2017.

[25] C Radhakrishna Rao. Information and the accuracy attainable in the estimation of statistical parameters. *Reson. J. Sci. Educ*, 20:78–90, 1945.

[26] Adrien Rieux, Anders Eriksson, Mingkun Li, Benjamin Sobkowiak, Lucy A Weinert, Vera Warmuth, Andres Ruiz-Linares, Andrea Manica, and François Balloux. Improved calibration of the human mitochondrial clock using ancient genomes. *Molecular biology and evolution*, 31(10):2780–2792, 2014.

[27] Fredrik Ronquist, Maxim Teslenko, Paul Van Der Mark, Daniel L Ayres, Aaron Darling, Sebastian Höhna, Bret Larget, Liang Liu, Marc A Suchard, and John P Huelsenbeck. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic biology*, 61(3):539–542, 2012.

[28] Nickolay Smirnov. Table for estimating the goodness of fit of empirical distributions. *The annals of mathematical statistics*, 19(2):279–281, 1948.

[29] Pedro Soares, Luca Ermini, Noel Thomson, Maru Mormina, Teresa Rito, Arne Röhl, Antonio Salas, Stephen Oppenheimer, Vincent Macaulay, and Martin B Richards.

Correcting for purifying selection: an improved human mitochondrial molecular clock. *The American Journal of Human Genetics*, 84(6):740–759, 2009.

[30] Anne C Stone, Fabia U Battistuzzi, Laura S Kubatko, George H Perry Jr, Evan Trudeau, Hsiuman Lin, and Sudhir Kumar. More reliable estimates of divergence times in Pan using complete mtDNA sequences and accounting for population structure. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1556):3277–3288, 2010.

[31] Koichiro Tamura and Masatoshi Nei. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular biology and evolution*, 10(3):512–526, 1993.

[32] Royal Ervin Taylor and Ofer Bar-Yosef. *Radiocarbon dating: an archaeological perspective*. Routledge, 2016.

[33] Mannis Van Oven and Manfred Kayser. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Human mutation*, 30(2):E386–E394, 2009.

[34] Benjamin Vernot, Elena I Zavala, Asier Gómez-Olivencia, Zenobia Jacobs, Viviane Slon, Fabrizio Mafessoni, Frédéric Romagné, Alice Pearson, Martin Petr, Nohemi Sala, et al. Unearthing Neanderthal population history using nuclear and mitochondrial DNA from cave sediments. *Science*, 372(6542):eabf1667, 2021.

[35] Rachel Elizabeth Wood, Thomas FG Higham, Trinidad De Torres, Nadine Tisnérat-Laborde, Hélène Valladas, José E Ortiz, Carles Lalueza-Fox, Sergio Sánchez-Moral, Juan Carlos Cañaveras, Antonio Rosas, et al. A new date for the Neanderthals from El Sidrón cave (Asturias, northern Spain). *Archaeometry*, 55(1):148–158, 2013.

[36] Ziheng Yang. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution*, 24(8):1586–1591, 2007.

[37] E Zuckerkandl and L Pauling. In evolving genes and proteins, ed. by V. Bryson & HJ Vogel, 1965.

[38] E Zuckerkandl, L Pauling, M Kasha, and B Pullman. Horizons in biochemistry. *Horizons in Biochemistry*, pages 97–166, 1962.